19e Colloque International de l'ADMEE-Europe

L'évaluation au 21e siècle : Vers de nouvelles formes, modélisations et pratiques de l'évaluation ?

11-13 septembre 2006, Luxembourg

Résumés des symposiums

	Titre de la communication	Auteur(s)
Symposium 1 : La modélisation de la mesure au 21e siècle : Nouveaux défis pour de vieux problèmes.	LA MODÉLISATION DE LA MESURE AU 21E SIÈCLE : NOUVEAUX DÉFIS POUR DE VIEUX PROBLÈMES	Michel Rousseau
	DU NOUVEAU DANS LE MONDE DE LA MESURE	Nathalie Loye
	UTILISATION DE L'ÉQUILIBRAGE DES ÉCHELLES DANS LE PROGRAMME D'ÉVALUATION DES APPRENTISSAGES DU NOUVEAUBRUNSWICK	Michel Rousseau, Richard Bertrand, Christian Arsenault & Robert Laurie
	LES IMPACTS DES ERREURS DE SPECIFICATION DE LA MATRICE Q SUR LES ESTIMES DE PARAMETRES ET LE TAUX D'ERREURS DE CLASSIFICATION DU MODELE DINA	André A. Rupp & Jon Templin
	LE MODÈLE DE RASCH : AVANTAGES, LIMITES, APPLICATIONS	Jean-Guy Blais, Gilles Raîche & Nathalie Loye
	COMPARAISON EMPIRIQUE DES PARAMÈTRES D'ITEMS/SUJETS DE LA THÉORIE DES RÉPONSES AUX ITEMS ET DE LA THÉORIE CLASSIQUE DES TESTS	Eric Frenette, Richard Bertrand, Pierre Valois, Marc Dussault & Marie-Hélène Hébert
	L'ÉLABORATION DE MODÈLES DE L'APPRENANT À PARTIR D'UNE APPROCHE BAYÉSIENNES	Michel C. Desmarais & Jean-Guy Blais
	INTEGRATION DE LA THEORIE DES REPONSES AUX ITEMS A L'EVALUATION DE LA COMPETENCE A RESOUDRE DES PROBLEMES MATHEMATIQUES	Nancy Boiteau, Dominic Voyer & Michel Rousseau
Symposium 2 : Etat des lieux et spécificité de l'évaluation en éducation artistique, physique et sportive au 21e siècle.	ETAT DES LIEUX ET SPECIFICITE DE L'EVALUATION ARTISTIQUE ET SPORTIVE AU 21ÈME SIECLE. QUELLES SONT LES NOUVELLES FORMES, MODELISATIONS ET PRATIQUES QUE LES DIDACTIQUES ARTISTIQUES ET SPORTIVES APPORTENT ?	Maugars Cédricia
	ETAT DES LIEUX ET SPECIFICITE DE L'EVALUATION MUSICALE	Cedricia Maugars
	QUELLE MODELE D'EVALUATION « EN LIGNE » AU SERVICE DE L'AMELIORATION DE L'EXPRESSION DE L'APPRENANT ?	Jacques Audran
	L'ANALYSE DES PRATIQUES ET LE PILOTAGE DES CONDUITES D'ENSEIGNEMENT : UNE ETUDE COMPARATIVE DES DISPOSITIFS D'EVALUATION EN RECHERCHE ET EN FORMATION EN DIDACTIQUES DE L'EDUCATION PHYSIQUE ET DES ARTS PLASTIQUES	Adrián Cordoba & René Rickenmann
	EVALUATION EN EPS : L'APPORT DE LA VIDEO TEL QUE PERCU PAR DES ETUDIANTS DE LA HEP FRIBOURG	Benoît Gisler & Pierre-François Coen

	PASSER D'UNE EVALUATION NORMATIVE A UNE EVALUATION COMPREHENSIVE EN FORMATION INSTRUMENTALE.	Pierre-François Coen
Symposium 3 : L'évaluation au cœur des activités de validation des acquis de l'expérience.	L'EVALUATION AU COEUR DES ACTIVITES DE VALIDATION DES ACQUIS DE L'EXPERIENCE.	Emmanuel Triby, Marie-Christine Presse, Isabelle Cherqui-Hout, Brigitte Pagnani, Pascal Tisserant & Martine Morisse
	L'ACTIVITE D'ACCOMPAGNEMENT EN VAE SAISIE PAR L'EVALUATION	Patricia Champy-Remoussenard & Emmanuel Triby
	LA MAÎTRISE DE LA LANGUE : OBJET CENTRAL DE L'EVALUATION DANS LA PROCEDURE VAE ?	Marie-Christine Presse
	LES PRATIQUES EVALUATIVES DES JURYS VAE DANS LES DISPOSITIFS DE L'ENSEIGNEMENT SUPERIEUR : QUELS PROCESSUS DE DECISION ? QUELS REFERENTS ?	Isabelle Cherqui-Houot & Paul Nkeng
	REGARDS CROISES SUR L'EVALUATION DANS LE PROCESSUS VAE : POINTS DE VUE DE CANDIDATS ET DE CONSEILLERS VAE A L'UNIVERSITE	Brigitte Pagnani & Frédérique Febvre-Buisson
	LA QUESTION DE L'EVALUATION DANS UN DISPOSITIF UNIVERSITAIRE DE TRAITEMENT DES DEMANDES DE VAE.	Pascal Tisserant & Cynthia Sellic
	ANALYSE DE L'ACTIVITE EVALUATIVE DES MEMBRES DE JURY DE VAE.	Matine Morisse
Symposium 4 : Problèmes métrologiques et validation d'outils d'évaluation.	PROBLÈMES MÉTROLOGIQUES ET VALIDATION D'OUTILS D'ÉVALUATION	Stéphane Germain, Pierre Valois, Marie-Hélène Hébert & Eric Frenette
	ESTIMATIONS DES COURBES CARACTÉRISTIQUES D'ITEMS PAR MAXIMUM DE VRAISEMBLANCE MARGINALE PÉNALISÉE	Stéphane Germain, Pierre Valois & Belkacem Abdous
	EFFET DE LA PROCÉDURE D'ESTIMATION DE CCI SUR LES TAUX DE DÉTECTION DE L'INDICE D'AJUSTEMENT DES SCORES INDIVIDUELS, Z l : UNE ÉTUDE MONTE CARLO.	Christina St-Onge, Pierre Valois, Belkacem Abdous & Stéphane Germain
	DÉVELOPPEMENT D'UN MODULE EXCEL D'ANALYSE D'ITEMS	Pierre Valois, Christina St-Onge & Stéphane Germain
	FIABILITE D'OUTILS D'EVALUATION DES APPRENTISSAGES DANS UNE APPROCHE PAR COMPETENCES : L'APPORT DE LA THEORIE DE LA GENERALISABILITE	Marie-Hélène Hébert & Virginie Duclos
	ADAPTATION OU TRADUCTION DE TEST : ÉTUDE DES DIFFÉRENTES FORMES DE BIAIS.	Eric Frenette & Richard Bertrand

Symposium 5 : Bilans d'innovations en matière d'évaluation : analyse critique et perspectives.	BILANS D'INNOVATIONS EN MATIERE D'EVALUATION : ANALYSE CRITIQUE ET PERSPECTIVES	Dominique Lafontaine, Alex Blanchet, Louise Bélair, Bruno Suchaut, Anne Soussi, Martine Wirthner, Ladislas Ntamakiliro, Patricia Gilliéron Giroud & Walter Tessaro
	LE DOSSIER DE L'ÉLÈVE (PORTFOLIO) DANS L'APPROCHE PAR COMPÉTENCE AU PRIMAIRE AU QUÉBEC ; QUELS APPORTS ? QUELLES PERSPECTIVES ? QUELLES LIMITES ?	Louise M. Bélair
	LE DOSSIER D'EVALUATION OU PORTFOLIO EN SUISSE ROMANDE : VISEES INSTITUTIONNELLES, MISES EN OEUVRE ET PERSPECTIVES D'AVENIR	Patricia Gilliéron Giroud
	PRATIQUES D'EVALUATION EXTERNE EN SUISSE ROMANDE : APPORTS ET LIMITES	Ladislas Ntamakiliro & Walther Tessaro
	LES ÉVALUATIONS NATIONALES EN FRANCE : OUTILS PÉDAGOGIQUES ET OUTILS DE PILOTAGE, INTERÊTS, LIMITES ET NOUVELLES PERSPECTIVES.	Bruno Suchaut
	NOUVELLES PRATIQUES D'EVALUATION : LA SITUATION DANS LE CANTON DE GENEVE	Anne Soussi
	BILAN DES REFORMES DE L'EVALUATION DU TRAVAIL DES ELEVES EN SUISSE ROMANDE	Martine Wirthner
	COMMENT CONCILIER LA NOUVELLE CULTURE DE L'EVALUATION ET LES CONDITIONS DE CERTIFICATION ? EVOLUTION DES PRATIQUES ET MODÈLES DE CONNAISSANCE	Alex Blanchet
	D'UN DISPOSITIF D'ÉVALUATIONS EXTERNES CENTRÉ SUR L'ENSEIGNANT Á UN DISPOSITIF CENTRÉ SUR L'ÉTABLISSEMENT ET LE SYSTÈME LE CAS DE LA COMMUNAUTÉ FRANÇAISE DE BELGIQUE	Dominique Lafontaine
Symposium 6 : Les dispositifs innovants d'évaluation des stages et leur prise en compte par les superviseurs et les formateurs de terrain	LES DISPOSITIFS INNOVANTS D'ÉVALUATION DES STAGES ET LEUR PRISE EN COMPTE PAR LES SUPERVISEURS ET LES FORMATEURS DE TERRAIN DANS UN CONTEXTE DE PROFESSIONALISATION : DIVERSES PERSPECTIVES	Louise M. Bélair
	ENTRETIENS DE COEVALUATION INSTRUMENTES PAR UN PORTFOLIO : POUR QUELLE RE-INTERPRETATION DE LA PROGRESSION DES COMPETENCES PROFESSIONNELLES ?	Lucie Mottier Lopez & Walther Tessaro
	TRANSFORMATION D'UN DISPOSITIF D'EVALUATION DES STAGES	Françoise Campanale & Hélène Gondrand

	PROFESSIONNELS : INCLURE LE CERTIFICATIF DANS UN PROCESSUS FORMATIF	
	RECHERCHE SUR LA CONSTRUCTION DES COMPÉTENCES : DÉFIS DANS LA RÉVISION DES DISPOSITIFS D'ÉVALUATION DES STAGES	Colette Gervais & Michel Lepage
	UN DISPOSITIF FORMATIF D'ACCOMPAGNEMENT DES STAGES DES FUTURS INSTITUTEURS : QUELS APPORTS AU NIVEAU REFLEXIF ?	Catherine Van Nieuwenhoven
	L'IMPLANTATION GRADUELLE D'UN DISPOSITIF D'ÉVALUATION DES COMPÉTENCES EN STAGE À L'UQTR : ENJEUX ET LIMITES	Christine Lebel & Louise M. Bélair
Symposium 7 : Évaluations des acquis en langues et dans d'autres domaines	EVALUATION DES ACQUIS EN LANGUES ET DANS D'AUTRES DOMAINES : QUELQUES PROBLEMES POSES PAR LEURS COMPOSANTES LINGUISTIQUES ET QUELQUES SOLUTIONS POUR LES RESOUDRE	Monique Reichert, Christian Monseur, Ariane Baye, Claude Houssemand, Franck Desroches & Marc Demeuse
	LE C-TEST : UN NOUVEAU CHEMIN IÉAL POUR UNE ÉVALUATION VALIDE DE COMPÉTENCES LINGUISTIQUES ?	Monique Reichert, Ulrich Keller & Romain Martin
	PROCÉDURES DESTINÉES À MINIMISER, À PRIORI, LES BIAS SUSCEPTIBLES D'ALTÉRER L'UNIDIMENSIONALITÉ INTERNATIONALE DES ÉCHELLES DE RENDEMENT.	Christian Monseur & Béatrice Halleux
	L'IMPACT DE TESTS TRADUITS SUR LES PERFORMANCES DES ÉLÈVES	Ariane Baye & Dominique Lafontaine
	ANALYSE SECONDAIRE ET QUALITATIVE DE L'ÉTUDE INTERNATIONNALE PISA	Claude Houssemand
	LES BIAIS CULTURELS DANS LES TESTS INTERNATIONAUX DE FRANÇAIS LANGUE ÉTRANGÈRE : DIAGNOSTIC ET ANALYSE	Desroches Franck
	EVALUATION ET AUTOÉVALUATION EN FRANÇAIS LANGUE ÉTRANGÈRE : APPROCHE EMPIRIQUE DU CADRE EUROPÉEN COMMUN DE RÉFÉRENCE POUR LES LANGUES (CECR) DU CONSEIL DE L'EUROPE À TRAVERS LE TEF ET L'ÉCHELLE D'AUTOÉVALUATION DIALANG	Demeuse Marc, Artus Frédérique, Casanova Dominique, Crendal Alexandra, Desroches Franck & Renaud François
Symposium 8 : L'évaluation de programme : notions, modèles et méthodologies.	L'EVALUATION DE PROGRAMME : NOTIONS, MODELES ET METHODOLOGIES	Gérard Figari, Marthe Hurteau, Nathalie Younes, Maria Palmira Alves & Marie-Françoise Fave-Bonnet
	MODÉLISER L'ÉVALUATION DE PROGRAMME, ENTRE THÉORIE ET PRATIQUE - VERS UNE COMPARAISON DES APPROCHES AU QUÉBEC ET EN France	Gerard Figari & Claire Tourmen

	L’EVALUATION DE PROGRAMME : NOTIONS, MODELES ET METHODOLOGIES	Marthe Hurteau & Sylvain Houle
	APPLICATION DU CADRE THEORIQUE DE L'EVALUATION DE PROGRAMME SELON SHADISH, COOK ET LEVITON AU CAS DE L'ENSEIGNEMENT UNIVERSITAIRE	Nathalie Younes
	VERS UN REFERENTIEL D’EVALUATION D’UN CURRICULUM / PROGRAMME DE FORMATION DANS L’ENSEIGNEMENT SUPERIEUR	Maria Palmira Alves & Eusébio André Machado
	ASSURANCE QUALITÉ ET ÉVALUATION DE PROGRAMME	Marie-Françoise Fave-Bonnet

SYMPOSIUM 1 :LA MODÉLISATION DE LA MESURE AU 21E SIÈCLE : NOUVEAUX DÉFIS POUR DE VIEUX PROBLÈMES

Michel Rousseau, Université d'Ottawa

La modélisation de la mesure au 21e siècle : Nouveaux défis pour de vieux problèmes

Cela fait déjà plusieurs années que des modèles de mesure issus de la théorie des réponses aux items (TRI) ont été développés. Ces modèles varient quant au nombre de paramètres d'items considérés et quant au type de données qu'il nous est possible d'analyser. Ces modèles de mesure offrent une grande variété d'applications qui permettent de répondre à plusieurs défis rencontrés lors de l'évaluation en éducation.

Malgré leurs avantages, l'utilisation des nouveaux modèles de mesure par la communauté des chercheurs et des praticiens francophones reste encore limitée. Il peut sembler que leur utilisation ne soit possible qu'au niveau des enquêtes à grande échelle tel que le PISA ou le TIMSS. Ce n'est pourtant pas le cas, et certaines raisons peuvent être avancées pour expliquer l'utilisation limitée qui en est faite. Premièrement, ces modèles sont moins attrayants puisqu'ils comportent des aspects mathématiques plus complexes. De plus, les procédures d'analyses des données à partir de ces modèles n'étant pas implantées dans la plupart des logiciels d'analyse statistique commerciaux, l'accès à des ressources informatiques pour procéder aux analyses des données constitue un obstacle. L'achat de logiciels spécialisés ainsi qu'une formation sur la programmation de ces logiciels sont nécessaires.

Pourtant, la communauté des chercheurs et des praticiens gagnerait à surmonter ces deux obstacles. Le présent symposium vise à répondre à deux questions relatives à l'utilisation des nouveaux modèles de mesure : En quoi de nouveaux modèles de mesure sont-ils nécessaires ? À qui peuvent s'adresser ces modèles?

Ce symposium compte sept présentations. Les deux premières présentations s'attardent à présenter certaines notions de base, essentielles à l'application de la TRI. Par la suite, nous verrons comment ces nouveaux modèles peuvent être jumelés à l'utilisation de l'ordinateur pour effectuer des évaluations. Les avantages que procure la TRI dans le suivi d'un système éducatif font l'objet de la présentation suivante. Finalement, de nouveaux modèles de mesure alliant la TRI et les théories cognitives sont présentés.

De manière plus précise, les présentations individuelles abordent les notions suivantes :

La présentation initiale touche à la notion d'invariance des paramètres d'items et des sujets. La question d'invariance des paramètres constitue le cœur même de la justification de l'utilisation de modèles de mesure issus de la TRI (Bertrand et Blais,

2004). Cette présentation s'attarde plus précisément à montrer l'avantage que possèdent les modèles de la TRI comparativement au modèle classique. La deuxième présentation vise un modèle très particulier, soit le modèle de Rasch. L'application de ce modèle procure des avantages intéressants. Toutefois, il existe des limites à ce modèle. Des exemples pratiques permettront d'illustrer les avantages et les limites de ce modèle de mesure.

Les nouveaux modèles de mesure issus de la TRI peuvent être utilisés à différents niveaux du système éducatif. Dans la troisième présentation, la question de l'utilisation de ces modèles conjointement avec l'utilisation de l'ordinateur pour évaluer la compétence à résoudre des problèmes mathématiques en classe du secondaire est abordée. L'utilisation de l'ordinateur dans l'opération d'évaluation est aussi au cœur de la présentation suivante. Les auteurs se penchent sur l'évaluation dans le domaine du e-learning. L'application de modèles graphiques bayésiens pour une évaluation détaillée et précise de connaissances fait l'objet de cette communication.

L'application de ces modèles peut aussi aider au suivi de système éducatif. La cinquième présentation aborde ainsi la question de la comparaison des résultats d'évaluation à grande échelle d'une cohorte à une autre. Une telle comparaison peut servir à vérifier l'évolution des compétences des élèves dans diverses matières. La méthode d'équilibrage des échelles, qui a recours aux modèles de la TRI, constitue la méthode la plus appropriée pour obtenir des résultats valables.

Enfin, le symposium se termine sur deux présentations qui portent sur des modèles de mesure qui allient la théorie des réponses aux items à des théories cognitives. Ces modèles prennent en compte l'utilisation de divers attributs cognitifs qui sont nécessaires pour résoudre des problèmes. Ils servent à poser un diagnostique plus complet sur les forces et les faiblesses des élèves et peuvent aider à mieux guider l'apprentissage de ces derniers. Des aspects plus précis sur l'application de ces modèles feront l'objet de ces deux communications.

Les présentations effectuées dans le cadre de ce symposium visent à exposer les raisons qui ont mené au développement de nouveaux modèles de mesure et les avantages que procure l'utilisation de ces nouveaux modèles. De plus, nous verrons que l'application de ces modèles peut être faite à divers niveaux du système éducatif et qu'elle permet de répondre adéquatement à de nombreux besoins du milieu éducatif. Le principal objectif du symposium est de faire connaître ces nouveaux modèles de mesure et d'accroître leur utilisation tant au niveau de la salle de classe qu'au niveau des politiques éducatives.

http://www.jemacs.uni.lu/ 2007 volume 1, No 1

DU NOUVEAU DANS LE MONDE DE LA MESURE

Nathalie Loye, Université de Montréal

Mots clés : modèles cognitifs, évaluation diagnostique, classes latentes, attributs cognitifs

Résumé :

Depuis quelques années, les approches cognitives et psychométriques se combinent pour donner de nouveaux modèles de mesure. Ces modèles, souvent appelés modèles cognitifs, sont en pleine évolution grâce aux avancées en psychologie cognitive et à la performance toujours plus grande des ordinateurs et des logiciels.

Les modèles traditionnels, comme ceux liés à la théorie classique des tests (TCT) ou à la théorie de réponse à l'item (TRI) offrent une vision unidimensionnelle du groupe de sujets en estimant l'habileté de chacun sur une échelle continue. Ces modèles permettent également de classer les items selon leurs paramètres de difficulté ou de discrimination. Une telle approche, basée sur une mise en rang, s'accorde relativement mal avec la perspective d'établir un diagnostic des sujets. Ce diagnostic aurait avantage à se baser sur des dimensions multiples. C'est dans cet esprit que les modèles cognitifs cherchent à identifier et inférer les différences de structure des individus dans les processus mis en œuvre pour répondre aux questions.

Ces modèles cognitifs se fondent sur le postulat que la performance des sujets se soumettant à un test dépend d'un ensemble d'habiletés, de compétences ou de caractéristiques latentes, appelées aussi attributs cognitifs, qui ne sont pas directement observables (Gierl, Leighton & Hunka, 2000 ; Tatsuoka, 1983). Ainsi la liste des attributs cognitifs nécessaires et suffisants pour répondre aux items du test doit être élaborée par des experts. Leur travail consiste à identifier les attributs cognitifs et à les relier à chacun des items du test (a) dans une approche déductive lorsque le test est élaboré à partir de la liste des attributs cognitifs ou (b) après la création du test dans une approche inductive. Cette dernière façon de faire offre la possibilité d'exploiter l'information diagnostique d'un examen non initialement élaboré à cette fin. Une caractéristique latente peut être la description de procédures, d'habiletés, de stratégies et/ou de connaissances qu'un sujet doit posséder pour répondre aux questions. Les experts doivent examiner individuellement chaque attribut et chaque item afin de décider si un lien existe ou non entre eux. Le résultat de cette étape primordiale est synthétisé par une matrice, notée Q, dont les lignes représentent les items et les colonnes les attributs. Cette matrice contient la valeur 1 lorsque l'item et l'attribut sont reliés et la valeur 0 sinon (voir Figure 1). C'est le modèle qui permet ensuite d'établir s'il existe ou non un lien entre chaque attribut cognitif et chaque sujet.

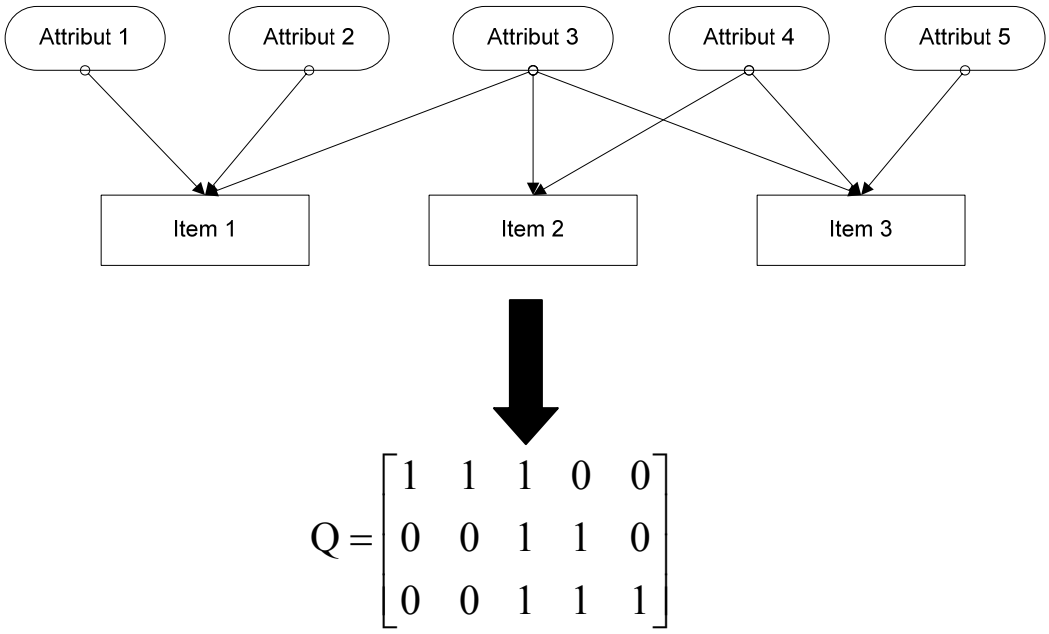


Figure 1 - Exemple de liens entre 5 attributs et 3 items et la matrice correspondante. Les formes arrondies représentent les attributs qui sont des variables latentes. Les rectangles représentent les réponses des sujets qui sont des variables observées.

Il existe de nombreux modèles cognitifs. Les deux modèles présentés ici sont des modèles applicables à des données et dont les paramètres sont estimables par des logiciels et/ou programmes existants. Le modèle *rule space* de Tatsuoka (1983, 1995) est l'un des modèles les plus simples puisqu'il ne tient compte d'aucune source d'erreur et fournit un diagnostic individuel par classement des sujets plutôt que par estimation. Un programme informatique appelé BUGLIB existe sans toutefois être commercialisé. Le modèle RUM (Hartz, 2002) est l'un des modèles les plus complexes car il prend en compte plusieurs sources d'erreur dont il estime les paramètres en plus de ceux des sujets. Le logiciel Arpeggio (ETS) permet de l'appliquer à des données. De plus, depuis peu, une macro nommée CDM et utilisant le logiciel Mplus (Templin, 2006) est également disponible.

Le modèle *rule space* de Tatsuoka (1983, 1995) utilise les scores dichotomiques aux items d'un test alliés à une matrice Q afin de classer les sujets dans des classes latentes qui reflètent leur état de connaissance. Ce modèle suppose qu'un sujet ne peut répondre correctement à un item que s'il maîtrise tous les attributs cognitifs qui lui sont reliés et qu'un sujet ayant un schéma particulier d'attributs maîtrisés va les utiliser pour répondre et non pas deviner ou faire des fautes d'étourderie. Toute déviation par rapport à ce postulat est donc considérée comme une erreur aléatoire. Le *rule space* est défini pour deux dimensions : le paramètre de compétence du sujet noté θ et un indice représentant la façon de répondre noté ζ. À partir de la liste des attributs établie par les experts, l'algèbre de Boole permet de déterminer tous les états de connaissance possibles selon la maîtrise ou non de chacun

des attributs. Le fait que certains attributs soient pré-requis à d'autres permet de minimiser le nombre d'états de connaissance idéaux. Dans un plan cartésien, il est alors possible de représenter graphiquement les couples (θ ,ζ) de chaque sujet et de chaque état idéal (voir Figure 2). La distance de chaque sujet à chaque état idéal permet de décider de son classement dans un état particulier. Si le point caractérisant le sujet est trop éloigné de tous les points idéaux, le sujet peut ne pas être classé. L'état de connaissance est représenté par un vecteur discret de la forme $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$, dans lequel k est le nombre d'attributs et chaque α_i peut prendre la valeur 0 ou 1. Le vecteur est obtenu par classement et l'interprétation est facilitée par la visualisation graphique. Le diagnostic individuel se fait en énonçant quels attributs sont maîtrisés ou non par chaque sujet.

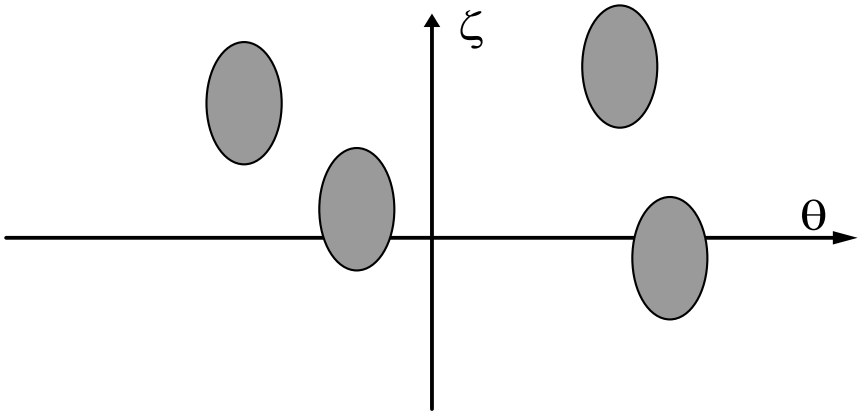


Figure 2 - La représentation graphique les couples (θ ,ζ) de chaque sujet et de chaque état idéal.

Le modèle de fusion (DiBello, Stout & Roussos, 1995), reparamétrisé par Hartz (2002) (voir Figure 3) et appelé RUM, ajoute la modélisation détaillée de la source des déviations systématiques par rapport au schéma idéal. Ce modèle prend en considération le fait que le sujet peut commettre une faute d'étourderie ou deviner la réponse et modélise des données dichotomiques mais aussi polychotomiques dès qu'il est possible d'élaborer une matrice Q. De plus, c'est un modèle non compensatoire puisqu'il suppose que le sujet doit posséder tous les attributs reliés à un item pour y répondre correctement. Un sujet est caractérisé, selon une approche psychométrique, par un trait latent continu θ, et est supposé, selon une approche cognitive, posséder ou pas chacun des attributs, ce qui revient à le caractériser par un vecteur $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$ dans lequel α_i est dichotomique. La probabilité de dévier des schémas établis par les experts est modélisée en fonction de θ et de α. Ce modèle pose le postulat de l'existence d'une structure possédant un certain nombre de classes latentes. L'idée est que dans la réalité, le comportement des sujets sera différent de ce qui a été prévu par les experts. Trois sources de variations sont inclues dans le modèle : (a) l'utilisation des attributs sous la forme d'un paramètre (noté π) estimant

la probabilité de bien utiliser les attributs liés à un item lorsqu'ils sont maîtrisés, (b) le manque d'attributs (*completeness*) : un paramètre (noté c) estime si des attributs non présents dans la matrice Q peuvent intervenir dans le processus de réponse et (c) la pénalité due au fait de ne pas posséder un attribut pour répondre à un item qui lui est relié (discrimination attribut-item, paramètre noté r). L'estimation et l'interprétation de ces paramètres donne la possibilité de porter un regard critique sur la structure cognitive de la matrice Q. Par exemple, une valeur de π supérieur à 0,6 , une valeur de r inférieur à 0,3 ou une valeur de c supérieure à 1,5 sont des indicateurs d'une structure cognitive forte (Templin, 2004, p. 67). Le tableau 1 présente un exemple de paramètres d'items estimés à partir d'un test en mathématique et six attributs cognitifs. De plus, chaque sujet est caractérisé par la valeur de son habileté résiduelle, correspondant à la partie de son habileté due à des attributs non présents dans la matrice Q et estimée par un modèle de Rasch sur une échelle continue. Enfin, le vecteur $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$ est estimé et fournit le diagnostic individuel sous la forme de la liste des attributs maîtrisés et non par chaque sujet.

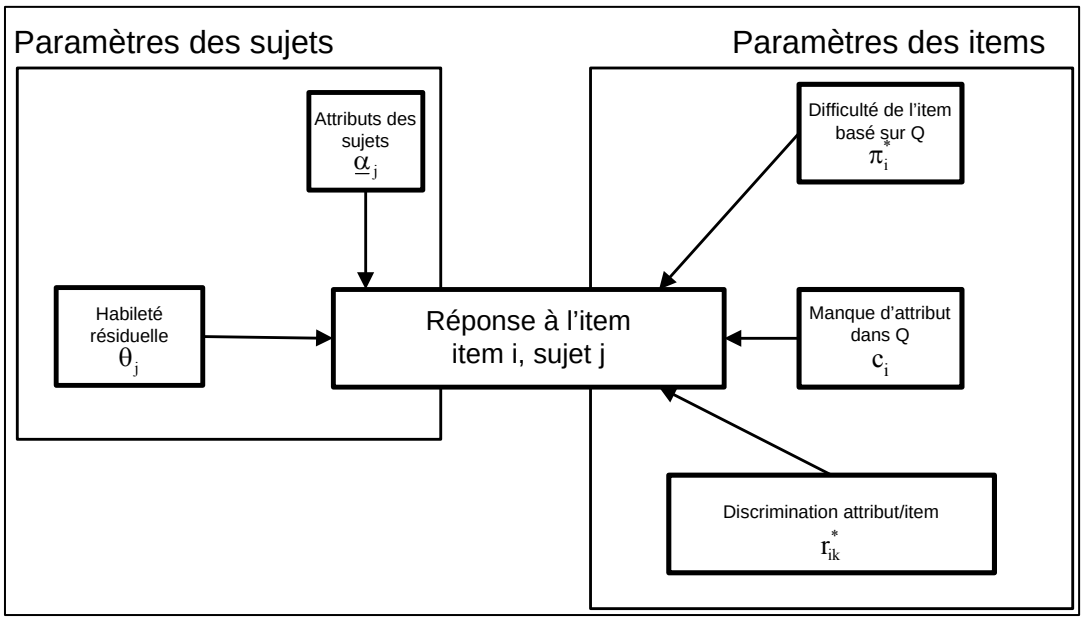


Figure 3 - Le cadre conceptuel du modèle RUM (Hartz, 2002)

Tableau 1.
Exemple de paramètres d'items estimés lors de l'application du modèle RUM à des données

Item	πj	r1	r2	r3	r4	r5	r6	cj
1	0,89	0,75	0	0	0	0	0,63	2,52
2	0,87	0,63	0	0	0	0	0	1,76
3	0,78	0	0,85	0,89	0,89	0	0,90	2,52
4	0,96	0,91	0	0	0	0,73	0	2,56
5	0,87	0,83	0	0,72	0,78	0	0,80	1,49
6	0,38	0,81	0	0	0,75	0	0,85	2,46
7	0,71	0	0	0,73	0	0	0,51	0,32
8	0,73	0,63	0	0	0	0	0	0,62
9	0,99	0,96	0	0	0	0	0,87	1,91
10	0,88	0,93	0	0	0,82	0	0	0,85

Ces deux modèles cognitifs sont présentés puis comparés dans le but de mettre l'accent sur leurs différences de postulats et de fonctionnements mais aussi sur leurs avantages et inconvénients respectifs. Le modèle RUM est illustré par une application à des données. Enfin les choix à faire par le chercheur, les difficultés rencontrées et quelques défis à relever clôturent cette présentation.

Références

DiBello, L.V. , Stout, W.F., & Roussos, L.A (1995). Unified cognitive/psychometric diagnostic assessment likeklihood-based classification techniques. Dans P.D. Nichols, S.F. Chipman et R.L. Brennan (Ed.), *Cognitively diagnostic assessment*, 361-389. Hillsdale, NJ: Erlbaum.

Gierl, M.J., Leighton, J.P., & Hunka, S.M. (2000). Exploring the logic of Tatsuoka's Rule-Space model for test development and analysis. *Educational Measurement : Issues and Practices, 19*(3), 34-44.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Thèse de doctorat non publiée. University of Illinois, Urbana-Champaign.

Tatsuoka, K.K. (1983). Rule-space: an approach for dealing with misconceptions based on item response theory. *Journal of educational measurement*, 20, 345-354.

Tatsuoka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. Dans P.D. Nichols, S.F. Chipman et R.L. Brennan (Ed.), *Cognitively diagnostic assessment*, 305-326. Hillsdale, NJ: Erlbaum.

Templin, J. L. (2004). *Generalized linear mixed proficiency models for cognitive diagnosis.* Thèse de doctorat non publiée, University of Illinois, Urbana-Champaign.

Templin, J. (2006). CDM Program. Disponible gratuitement en ligne à l'adresse http://www.people.ku.edu/~jtemplin/downloads/cdm.html.

UTILISATION DE L'ÉQUILIBRAGE DES ÉCHELLES DANS LE PROGRAMME D'ÉVALUATION DES APPRENTISSAGES DU NOUVEAU-BRUNSWICK

Michel Rousseau, Université d'Ottawa
Richard Bertrand, Université Laval
Christian Arsenault, MÉNB
Robert Laurie, MÉNB

Mots-clés : théorie de la réponse aux items, équilibrage des échelles, évaluation à grande échelle.

Résumé :

Le Ministère de l'Éducation du Nouveau-Brunswick a comme objectif de développer un système d'éducation publique de calibre mondial qui favorise l'excellence et le rendement en tout temps (MENB, 2002). C'est dans ce contexte que le Ministère désire mesurer les progrès des élèves en lecture, en mathématiques et en sciences. Au niveau primaire, des évaluations sont administrées en 2e année (lecture), en 5e année (sciences et mathématiques) et en 8e année (lecture et mathématiques). Au niveau du secondaire, des évaluations en mathématiques et en lecture sont administrées en 11e année. Pour prendre connaissance de l'évolution des apprentissages des élèves, il est visé de comparer les résultats d'une cohorte à une autre et ce à certains niveaux fixes de scolarité. Ce type de devis longitudinal oblige à avoir recours à plus d'une forme d'un même test pour des raisons de sécurité du contenu des épreuves. Devant cette situation, comment s'assurer de la comparabilité des résultats des différentes cohortes à qui nous avons administré différentes épreuves d'évaluation? Le recours à l'équilibrage des échelles permet de faire face à ce défi. Parmi les nombreuses méthodes proposées, l'utilisation de la Théorie des réponses aux items (TRI) dans le cas du design d'items communs avec groupes non équivalents semble être la méthode la plus efficace. Cependant, pour assurer sa validité, certains aspects doivent être pris en compte dès le début du développement du test et un contrôle constant de la qualité se doit d'être appliqué. La présentation porte sur le contexte dans lequel l'équilibrage des échelles sera appliqué, sur la planification des opérations ainsi que sur les futurs développements de recherche possibles.

L'évaluation provinciale des apprentissages au Nouveau-Brunswick

À la recherche d'une prospérité sociale et économique, le gouvernement du Nouveau-Brunswick s'est doté d'un Plan de prospérité à long terme dont l'un des piliers : *Investir dans la population* propose un certain nombre de mesures pour faire du système d'éducation publique un système de calibre mondial. Ces mesures sont

présentées dans un Plan d'apprentissage de qualité, PAQ (MÉNB, 2003) et se traduisent par des cibles dont l'atteinte s'échelonne sur une période de 10 ans. Ce plan prévoit entre autre la publication d'un rapport annuel pour faire état des progrès vers l'atteinte des cibles prescrites (MÉNB, 2003, p.44). Il s'ensuit qu'une attention particulière est portée au rendement académique des élèves.

Programme d'examens provinciaux

Au ministère de l'Éducation, il incombe à la Direction de la mesure et de l'évaluation (DMÉ) de conceptualiser et de mettre en œuvre le programme d'examens provinciaux. Aujourd'hui, le Nouveau-Brunswick mesure le rendement des élèves en lecture, en écriture, en mathématiques, en sciences de la nature ainsi que la compétence à communiquer à l'orale en anglais. À l'exception des examens de mathématiques en 11e année et des entrevues en anglais en 10e année qui sont administrés semestriellement, toutes les autres évaluations sont administrées sur une base annuelle. La séquence des évaluations se présente comme suit : Lecture silencieuse et à haute voix (2e année); Mathématiques et Sciences de la nature (5e année); Mathématiques, Lecture et Écriture (8e année); Entrevue linguistique en anglais (10e année); Mathématiques, Lecture et Écriture (11e année).

L'élaboration d'épreuves uniques est une démarche axée sur la participation active du personnel enseignant, dont le travail est sous la responsabilité du personnel de la DMÉ. Les épreuves contiennent des items à réponse choisi et à réponse construite. Les items à réponse construite sont notés par des enseignants regroupés dans des endroits prévus à cet effet. La validité du processus est assurée par l'utilisation de grilles de notation standardisées, et par une formation donnée aux correcteurs par des responsables de la notation formés et mandatés par la DMÉ. De plus, des copies dites de « calibration » servent à vérifier la fidélité de la correction. Les examens sont renouvelés à chaque administration et leur contenu est gardé confidentiel après l'administration.

L'analyse d'items et l'atténuation statistique

Pour évaluer la qualité des items et de l'examen, on calcule des indices classiques de difficulté et des coefficients de corrélation item-total corrigée. Les items jugés problématiques sont exclus de l'examen et ne sont pas comptabilisés. Pour les examens de 11e année seulement, lorsque le taux d'échec à l'examen est supérieur à 45 %, une mesure d'atténuation statistique est appliquée pour limiter le taux d'échec à 45 %.

Examens semestriels

En mathématiques 11e année, un aspect problématique relève du fait que la compétence de la clientèle des deux semestres présente des différences importantes; les élèves inscrits au 2e semestre ayant pour la plupart déjà échoué un ou plusieurs cours de mathématiques. Comme les deux examens mesurent les mêmes concepts et servent aux mêmes fins de certification, le parallélisme entre les deux formes est difficile à obtenir et à démontrer. L'utilisation de l'un des modèles de la théorie des réponses aux items permettrait d'assurer le parallélisme entre les différentes formes d'un même examen.

Publication de rapports

La direction de la mesure et de l'évaluation publie un rapport statistique provincial des résultats de tous les examens. Le rapport statistique fait état de la moyenne des écoles, des districts scolaires et de la province. En plus du rapport statistique provincial, la DMÉ produit des renseignements statistiques sur les habiletés et les contenus pour chacun des groupes-classes des écoles.

Pour les examens de la 5e et de la 8e année, les résultats sont présentés de deux façons : sur une échelle d'intervalle à l'aide d'un score global et sur une échelle ordinale pour les score d'habileté et les scores de contenu. Le score global présenté sur l'échelle d'intervalle fait l'objet d'une interprétation normative.

Pour chacune des habiletés et chacun des contenus mesurés par l'examen, le score de l'élève est situé sur une échelle ordinale à quatre échelons (profil de compétence), en fonction de seuils de réussite préétablis. Une rubrique décrit la performance des élèves à chacun des échelons. Le rapport fait état du pourcentage d'élèves qui se trouve à chacun des quatre échelons ou niveaux de compétence (Insuffisant, Acceptable, Attendu, Supérieur) et ce, au niveau de l'école, des districts scolaires et de la province. Pour les examens de la 11e année, les résultats sont seulement présentés sur une échelle d'intervalle. Les scores font tous l'objet d'une interprétation normative et le rapport fait état de la moyenne des écoles, des districts scolaires et de la province.

Suivi des résultats

Les districts scolaires, en collaboration avec les directions d'école et le personnel enseignant, sont responsables d'interpréter les résultats aux examens et de se donner un plan éducatif pour améliorer la qualité de l'enseignement et le rendement de leurs

élèves. Comme le suivi longitudinal du progrès des élèves au niveau des apprentissages se résume souvent à comparer directement la moyenne de l'école (ou du district scolaire) d'une année à l'autre, les interprétations qui en découlent sont trop souvent invalides et peuvent menées à des décisions inappropriées.

Les cibles pour l'enseignement

Une des cibles du Plan d'apprentissage de qualité (MÉNB, 2003, p.20) concerne le rendement des élèves : «Nos élèves auront de bons résultats aux examens provinciaux de langue, de mathématiques et de sciences où 85 % des élèves atteindront ou dépasseront un niveau acceptable et notre taux de diplomation demeurera parmi les trois meilleurs au Canada.» Le niveau acceptable est défini comme étant le seuil minimal de réussite à l'examen soit un score de 55 %.

Le suivi, année après année, du pourcentage d'élèves ayant atteint ou dépassé le seuil de réussite des examens provinciaux dans les matières ciblées par le PAQ est un autre aspect problématique. Les taux de réussite aux examens provinciaux sont dépendants du niveau de difficulté des items, de l'habileté des sujets qui passent l'examen et d'une mesure d'atténuation statistique qui limite le taux d'échec à 45 %. Par exemple en mathématiques 11e année, le taux de réussite de l'examen se situe, depuis plusieurs années, aux alentours de 57 % principalement à cause de l'atténuation statistique qui limite le taux d'échec à 45 %. Or, le taux d'échec peut varier considérablement selon le niveau de difficulté des items.

Dans le contexte actuel, le suivi annuel du pourcentage d'élèves ayant atteint ou dépassé le seuil de réussite aux examens provinciaux mène à des interprétations invalides. Pour mesurer le rendement des élèves d'une année à l'autre, et rapporter les différences observées, il importe que les résultats des cohortes d'élèves soient comparables d'une année à l'autre.

La comparaison de résultats tirés d'évaluations différentes

Les résultats tirés d'évaluation des apprentissages constituent un indicateur de choix autant pour la recherche en éducation que pour les ministères de l'éducation qui désirent effectuer un suivi de leur système éducatif. L'augmentation du niveau d'apprentissage des élèves est en effet considérée comme un des principaux objectifs de tout système d'éducation. Les diverses études pouvant être menés prennent souvent la forme d'études comparatives. Il est possible d'effectuer des comparaisons tant transversales que longitudinales du niveau d'apprentissage des élèves.

Les comparaisons transversales, comme celles proposées par le PISA, permettent de mettre dans un contexte mondial des résultats au plan national ou provincial. La

capacité de différents systèmes éducatifs à développer le niveau d'apprentissage de ses élèves peut alors être mise en relation avec certaines caractéristiques des systèmes eux-mêmes ou des caractéristiques des élèves. La comparaison des résultats d'élèves issus de différents pays est alors facilitée par le fait que tous les élèves sont soumis à un même test jugé équivalent. L'unique différence réside alors dans la langue d'administration des évaluations. S'il est possible de postuler que les différentes versions soumises aux élèves de différents pays sont identiques, les différences au niveau de la performance des élèves au test représenteront de réelles différences étant donné que tous les élèves se voient administrer un même test. Des biais causés par des problèmes de traduction ou d'adaptation des tests réduiront la validité des différences qui seront observées.

Les comparaisons longitudinales, quant à elle, ne sont généralement pas influencées par des biais de traduction étant donné que l'intérêt est centré sur une population particulière d'élèves. Deux possibilités de comparaisons longitudinales s'offrent à nous. Premièrement, il est possible de suivre une même cohorte sur une certaine période de temps, des mesures étant prises à intervalles réguliers au cours de la période de suivi. Par exemple, il serait possible de suivre l'évolution d'une cohorte particulière d'élèves de la première année du secondaire à la dernière année. Deuxièmement, nous pouvons aussi suivre l'évolution de différentes cohortes en prenant une mesure lors d'un niveau de scolarité fixe. Cette dernière possibilité est ce qui a été privilégiée par le Ministère de l'éducation du Nouveau-Brunswick. Ainsi, des mesures seront prises sur différentes cohortes d'élèves à des niveaux déterminés de scolarité.

Lorsque nous désirons entreprendre une étude de comparaison longitudinale de résultats d'évaluation d'apprentissage, nous nous butons rapidement à un important défi : rendre équivalents les résultats tirés d'évaluations différentes. Dans un contexte de suivi longitudinal d'une même cohorte, il est non recommandé d'administrer le même test plus d'une fois. Si nous répétons l'administration du même test, les administrations antérieures constitueront des pratiques pour des administrations subséquentes. Les résultats ont alors une grande probabilité de signaler une évolution positive des résultats mais cette conclusion risque d'être plus influencée par l'administration d'un même test que par une réelle évolution des élèves de la cohorte. De plus, comme le contenu du test doit être relié à ce qui est enseigné en classe, comment est-il possible de construire un test de mathématiques, de sciences ou de français qui permettrait de couvrir la matière sur plusieurs années de scolarité? Enfin, il est pratiquement impensable de construire un test présentant un niveau de difficulté semblable d'une année à l'autre.

Ces différents problèmes ne se présentent pas dans le cas où nous comparons l'évolution de différentes cohortes sur des mesures prises lors d'un niveau de scolarité fixe. Cependant, d'autres aspects incitent à ne pas recourir à un même test. En utilisant le même test d'année en année, son contenu cours le risque de devenir connu par différents acteurs du système éducatifs amenant alors des biais dans le processus d'évaluation effectuée. Pour des raisons de sécurité du contenu du test, il est ainsi recommandé d'avoir recours à des tests différents d'année en année.

Le recours à différentes formes de test permet alors de neutraliser certains biais mais amène un problème supplémentaire majeur. Pour que les différences observées d'une cohorte à l'autre reflètent de véritables différences, il importe que les tests soit le plus identique possible. En effet, si nous administrons deux tests différents à deux cohortes d'élèves, comment s'assurer que les différences de performance entre les deux groupes soit attribuables à leur niveau d'apprentissage ou bien à un niveau de difficulté différent des deux tests?

Selon McDonald (1999), la comparaison directe des scores obtenus à partir de deux ou plusieurs formes d'un même test peut être effectuée uniquement si les différentes formes du test constituent des formes strictement parallèles. Pour être considéré comme étant strictement parallèles, chaque sujet qui se voit administrer les formes du même test doit présenter le même score vrai pour chaque forme et une même variance des erreurs de mesures (Bertrand & Blais, 2004).

Il est pratiquement impensable d'atteindre cette condition de parallélisme entre deux ou plusieurs formes du même test. Pour ce faire, il faudrait disposer d'une très grande banque d'items à partir de laquelle nous pourrions constituer des formes parallèles. Tel que le suggère McDonald (1999), pour construire des formes les plus parallèles possible, il importe d'apparier les items de chacune de ces formes en fonction des paramètres d'items, du contenu couvert par l'item ainsi que du type d'item (c.-à-d. item à choix multiple avec le même nombre d'alternatives de réponse). La difficulté de construire des formes parallèles augmente avec le nombre de formes qui doivent être construites.

Devant cette impossibilité de construire des formes strictement parallèles, pour effectuer des comparaisons de résultats issus de différentes formes d'un même test, il est alors préférable de se tourner vers l'équilibrage des échelles (*test equating*). Le but de l'équilibrage est de produire des scores d'échelle qui peuvent être utilisés de manière interchangeable (Kolen & Brennan, 2004). Une mesure est considérée comme étant interchangeable avec une autre mesure si elles mesurent un même concept (la longueur par exemple) et si elles sont situées sur une échelle de valeurs commune (en centimètres par exemple). Les scores peuvent être obtenus à partir de deux versions d'un même instrument, comme par exemple une règle à mesurer qui est souvent graduée en centimètres et en pouces. Si nous mesurons la longueur d'un crayon avec une telle règle à mesurer, nous pouvons facilement convertir la longueur du crayon sur une même échelle de valeurs, soit des centimètres, des pouces, des millimètres, etc. Cette transformation d'échelle de valeurs sert d'équilibrage des échelles. La longueur est le concept qui est mesuré et les centimètres, pouces, millimètres, etc. sont les mesures qui peuvent être équilibrées. En pratique, les instruments de mesure possèdent des limites, la mesure donnée par une règle à mesurer est souvent plus précise en millimètres qu'en pouces. En conséquence, les mesures ne peuvent souvent pas être considérées comme véritablement interchangeables en raison d'une précision inégale (Dorans, 2004).

Dans une situation de mesure des apprentissages, le processus d'équilibrage vise à établir des scores qui soient équivalents sur différentes formes d'un test. Ainsi, si deux formes de test A et B sont équilibrées et qu'un élève se voit administré la forme A, nous pourrons déterminer un score équivalent sur l'échelle de valeur sur laquelle les scores de la forme B sont exprimés. L'opération d'équilibrage oblige aussi qu'il soit possible de passer de l'échelle des scores de la forme B vers la forme A (Lord, 1980).

Crocker et Algina (1986) font une distinction entre l'équilibrage vertical et l'équilibrage horizontal. L'équilibrage vertical est utilisé lorsque nous désirons comparer les résultats de différents tests de niveau de difficulté différent (par exemple les niveaux de scolarité). L'équilibrage horizontal est effectué entre deux formes d'un même test. La situation au Nouveau-Brunswick est un exemple d'équilibrage horizontal. Un équilibrage des formes d'un même test sera effectué après leur administration.

Différentes méthodes pour procéder à l'équilibrage des échelles ont été suggérées telles que l'équilibrage linéaire, l'équilibrage par equipercentile et le recours à la théorie des réponses aux items (Kolen & Brennan, 2004; Lord, 1980).

Équilibrage linéaire

L'équilibrage linéaire est la forme la plus simple de l'équilibrage. Cette méthode postule que les deux formes de mesure à équilibrer diffèrent uniquement au niveau de leurs moyennes et de leurs écarts-types (Woldbeck, 1998). La conversion d'une forme à l'autre est effectuée en additionnant ou multipliant des valeurs constantes dans la forme de l'équation suivante :

$$Y = AX + B$$

Le recours à cette méthode d'équilibrage oblige d'assumer deux postulats. Premièrement, les deux distributions des scores bruts doivent être identiques, à l'exception de la moyenne et de l'écart-type. Deuxièmement, la fidélité des deux mesures doit être la même. Dans bien des situations réelles, ces deux postulats ne peuvent être respectés mettant en péril la validité de notre opération d'équilibrage.

Équilibrage par equipercentile

L'équilibrage par percentile se base sur la distribution de fréquence cumulée du score total de deux formes d'un même test. Les scores sur les deux formes qui correspondent à une valeur de percentile fixe sont considérés comme représentant un

même niveau d'habileté ou de réussite. Les scores sur les deux mesures peuvent alors être appariés en fonction des différentes valeurs de percentiles.

Cette façon de faire exige d'avoir recours à un échantillon de grande taille bien que certaines méthodes de lissage (*smoothing*) peuvent être utilisées lorsque nous travaillons avec des échantillons de taille réduite.

L'équilibrage par équipercentile présente plusieurs limites. Lorsque les distributions de scores sont très différentes, l'équilibrage devient alors non valide (Woldbeck, 1998). De plus, cette méthode est très influencée par les données disponibles rendant très ardu la reproduction des résultats avec un nouvel échantillon (Cook & Petersen, 1987). Enfin, cette méthode, tout comme l'équilibrage linéaire, ne fournit aucune information sur les erreurs attribuables aux scores équilibrés (McDonald, 1999).

Équilibrage par la Théorie des réponses aux items

Plusieurs auteurs reconnaissent que l'utilisation de la théorie des réponses aux items (TRI) pour effectuer l'équilibrage des échelles représente la méthode qui est préférable aux autres (McDonald, 1999). Dans la situation d'équilibrage horizontal, si les différentes cohortes ne sont pas constituées par assignation aléatoire comme dans le cas du cours de mathématiques 11e année, les postulats de l'équilibrage linéaire et par équipercentile risquent fortement de ne pas être respectés (Crocker & Algina, 1986).

Le recours à la théorie des réponses aux items pour effectuer l'équilibrage des échelles oblige à utiliser des items communs à chaque forme du test. L'ensemble de ces items est appelé test ancré (Lord, 1980). Lorsque les items qui composent le test ancré font parti des deux formes du test, le test ancré est appelé interne. Dans le cas contraire, lorsque les items forment un test séparé des deux formes, le test ancré est dit externe (Lord, 1980). Dans le cas du test ancré interne, dans certains cas les items ancrés sont utilisés dans le calcul du score total alors que dans d'autre cas, ils ne sont utilisés uniquement que pour l'opération d'équilibrage.

Bien qu'il existe plusieurs situations où l'équilibrage peut être utilisé, le présent texte se penche sur la situation d'évaluation au Nouveau-Brunswick. La situation correspond au design d'items communs avec groupes non équivalents. Ce design correspond aux deux facteurs énoncés ci-haut soit que les formes du test contiennent des items communs et que les groupes constituent des populations indépendantes non équivalentes.

Afin d'introduire l'idée sous-jacente à l'opération d'équilibrage, il convient de revenir sur les concepts de base de la Théorie des réponses aux items. Cette théorie tente de modéliser les réponses d'individus sur un ensemble d'items comme étant une

fonction des caractéristiques des items (discrimination, difficulté et pseudo-chance) et le niveau d'habileté des individus sur un trait latent (Skaggs & Lissitz, 1986). Pour l'ensemble des différents modèles de la TRI, les caractéristiques des items et le niveau d'habileté sont exprimés sur une échelle de valeurs communes. Habituellement, la valeur 0 de l'échelle des valeurs est fixée de manière arbitraire à partir du niveau moyen d'habileté d'un échantillon ou d'une population particulière.

Dans la situation où un même ensemble d'items est administré à deux populations différentes, puisque la valeur 0 est arbitraire, elle sera fixée séparément pour les deux populations. Pour les deux groupes, le niveau d'habileté moyen sera fixé à la même valeur, soit zéro. Cependant, dans la mesure où les deux populations ne présentent pas un même niveau d'habileté moyen mesuré par le test, cette valeur n'est pas équivalente d'une échelle de mesure à l'autre.

Le processus d'équilibrage revient donc à effectuer une transformation linéaire en utilisant la fonction suivante (Woldbeck, 1998) :

$$\theta^* = \theta + m$$

où θ est le véritable niveau d'habileté des individus, θ^* est le niveau d'habileté transformé et m une constante utilisée pour situer les deux valeurs sur une échelle de mesure commune. Le problème est que nous ne connaissons jamais le véritable niveau d'habileté des individus. Cependant, puisque les caractéristiques des items se retrouvent sur la même échelle de mesure que le niveau d'habileté des individus, il est possible d'effectuer l'opération d'équilibrage en utilisant ces caractéristiques d'items. Et c'est à partir des ces caractéristiques d'items que pourront être mis sur une échelle de mesure commune les valeurs θ obtenues pour chaque forme. L'équilibrage est donc effectué à partir de l'estimation des paramètres des items. Deux façons de procéder sont suggérées, l'estimation concurrente et l'estimation séparée.

Dans l'estimation concurrente, les paramètres de tous les items des deux formes sont estimés simultanément (Hanson & Béguin, 2002). Le fait d'estimer tous les paramètres d'items de manière simultanée procure des estimés de paramètres qui sont situés sur une échelle de mesure commune (Hanson & Béguin, 2002). Le fait que les élèves n'ont répondu qu'à une portion de l'ensemble des items ne constituent pas un problème pour l'estimation des paramètres. Les items non administrés aux individus sont alors considérés comme non-atteints et n'affectent pas la validité des estimés obtenus (Lord, 1980).

Pour ce qui est de l'estimation séparée, les estimés de paramètres d'items sont obtenus à partir de deux phases d'estimation distinctes. Évidemment, les estimés de paramètres pour les deux formes ne seront pas situés sur une échelle de mesure commune. Les estimés des caractéristiques des items ancrés seront alors utilisés pour identifier la transformation des valeurs de l'échelle de mesure qui permettront de situer ces paramètres d'items des deux formes sur une échelle de mesure commune (Kolen & Brennan, 2004).

Les études antérieures ont obtenu des résultats divergents quant à la performance de l'estimation concurrente par rapport à l'estimation séparée (Hanson & Béguin, 2002). D'autres études sont nécessaires pour identifier la méthode la plus performante en fonction des facteurs de la situation d'équilibrage (nombre de sujets, respect de l'unidimensionnalité, nombre d'items totaux, nombre d'items ancrés, etc.).

Les aspects pratiques

Lorsque nous désirons effectuer une opération d'équilibrage, il importe de considérer l'ensemble des aspects reliés au processus d'évaluation. L'équilibrage des échelles requière de porter un jugement quant au développement du test, de la collecte de données, de l'implantation du design, de l'analyse des données et de l'évaluation des résultats (Kolen & Brennan, 2004).

Plusieurs décisions doivent être prises si nous décidons d'utiliser la procédure d'équilibrage. Des critères prédéterminés pour nous permettre de prendre nos décisions ne sont pas disponibles. Bien souvent, l'équilibrage impose de faire des compromis entre plusieurs considérations d'ordre pratique (Kolen & Brennan, 2004). Toujours selon ces auteurs, dans ce sens, une situation idéale d'équilibrage est impossible.

Suivant les recommandations de Kolen et Brennan (2004), les aspects pratiques à considérer sont les suivantes : 1) le développement du test, 2) le design de l'étude, 3) le choix de la méthode d'équilibrage et 4) le contrôle de la qualité. Dans certains cas, le choix du design et de la méthode d'équilibrage sont contraints par la situation d'évaluation à laquelle nous faisons face. C'est le cas dans la situation à l'étude où nous devons avoir recours au design d'items communs avec groupes non équivalents et qu'en conséquence, la méthode d'équilibrage à privilégier est d'avoir recours à la TRI. Qu'en est-il maintenant des autres aspects pratiques?

Le développement du test

Selon Mislevy (1992), le développement d'un test et l'équilibrage sont deux opérations inséparables. La phase de développement est une phase vitale à la procédure d'équilibrage (Kolen & Brennan, 2004). L'équilibrage sera adéquat dans la mesure où chaque forme du test est construite pour représenter un domaine de contenu commun. Ainsi, les spécifications de construction, contenu et type d'items, doivent être clairement identifiées et appliquées de manière rigoureuse. Si les spécifications du contenu sont modifiées de manière significative, les scores obtenus pour les formes construites avant la modification ne pourront être considérés comme interchangeables avec les nouvelles formes.

Il est de plus recommandé de s'assurer que les items de chaque forme présentent des caractéristiques statistiques, telles que la difficulté ou la discrimination, les plus

semblables possibles (Kolen & Brennan, 2004). Il est alors souhaitable de procéder à un pré-test avec toutes nouvelles formes pour vérifier le respect des caractéristiques statistiques mais aussi pour s'assurer que les items qui forment le test ne présentent pas de problèmes statistiques importants.

Dans la situation du design d'items communs avec groupes non équivalents, les caractéristiques des items ancrés prennent une importance particulière. Plusieurs aspects sont à considérer. Premièrement, le nombre d'items doit être assez élevé pour permettre de représenter adéquatement le domaine du contenu à évaluer. Une règle du pouce suggère que le test ancré doit représenter au moins 20 % de la longueur totale d'un test qui contient 40 items ou plus (Kolen & Brennan, 2004). L'hétérogénéité du domaine à évaluer influence directement la longueur nécessaire du test total ainsi que du test ancré.

Les items communs ne doivent pas présenter de fonctionnement différencié d'items (FDI) entre les différentes cohortes. Des items communs présentant un FDI devraient être mis de côté dans l'opération d'équilibrage. Pour prévenir un FDI, il est préférable de situer les items ancrés dans la même position pour les deux formes du test et de ne pas changer l'ordre des réponses. De plus, si un item commun fait partie d'un groupe d'items associés à un même stimulus, comme c'est le cas pour un *testlet*, l'ensemble des items associés doivent se retrouver dans les deux formes. Enfin, les indices de difficulté des items communs doivent avoir des indices de difficulté qui s'étendent sur toute l'échelle des niveaux d'habileté.

Le contrôle de la qualité

Le contrôle de la qualité est un aspect vital de la procédure d'équilibrage et demande souvent beaucoup plus d'effort que les autres aspects (Kolen & Brennan, 2004). Il importe de s'assurer que les spécifications du test ne soient pas modifiées, que le contexte relatif aux items communs soit le même dans chacune des formes et que le libellé des items communs soit exactement le même d'une forme à l'autre.

Les conditions d'administration des tests doivent être standardisées peu importe le lieu ou le moment de l'administration. Certains aspects reliés à la standardisation peuvent avoir un effet significatif sur la qualité de l'équilibrage tels que 1) le nombre d'items pour chaque forme, 2) le temps d'administration, 3) les conditions de motivation et 4) l'aspect graphique du test et des feuilles réponses.

Le contrôle de qualité implique aussi de s'assurer que les procédures statistiques reliées à l'équilibrage soient respectées de manière très rigoureuse.

Défis méthodologiques à explorer

Ces aspects pratiques reflètent une situation d'équilibrage idéale. Dans la pratique, différentes contraintes risquent de limiter le respect de l'application des considérations pratiques. Il en résulte une menace à la validité de l'opération d'équilibrage. Ainsi, il est possible d'identifier certains aspects critiques qui pourront faire l'objet de projet de recherche par l'équipe de chercheurs. Parmi ces aspects qui peuvent être considérés, nous retrouvons notamment l'équilibrage des échelles dans le cas de l'application d'un modèle polytomique, le traitement de la non-réponse particulièrement aux items communs et l'erreur d'équilibrage relatif au choix des items communs.

L'utilisation d'un modèle polytomique pour modéliser les réponses des élèves aux épreuves est rendu nécessaire par les items qui les composent. En effet, plusieurs items partagent une même situation d'évaluation. Pour favoriser le respect de la condition d'unidimensionalité et de dépendance locale, il est préférable de considérer tous les items reliés à une même situation comme étant un *testlet*. En plus de devoir recourir à des modèles polytomiques, ceci réduit le nombre d'items qui composent les épreuves et risque aussi de réduire le nombre d'items communs.

Bien que le traitement de la non-réponse aux items soit un problème connu depuis plusieurs années, aucune méthode de traitement de cette non-réponse ne fait encore consensus. Sa présence peut être encore plus problématique dans le cas des items communs puisqu'un traitement inadéquat risque de remettre en question la validité de l'opération d'équilibrage, particulièrement si la proportion de valeurs manquantes est trop importante.

Enfin, l'erreur-type d'équilibrage relatif au choix des items communs est un problème qui a été soulevé par Michaelides et Haertel (2004). Selon ces auteurs, cet erreur est relative à la généralisation à un grand ensemble possible d'items communs. Comment tenir compte de ce type d'erreur et en évaluer son importance représente une piste de recherche qui sera aussi explorée.

D'autres problèmes relatifs à l'équilibrage des échelles risquent d'émerger au cours des années. Une bonne collaboration entre l'équipe de chercheurs et les intervenants du Ministère de l'éducation permettra leur identification et mènera à d'autres projets de recherche.

Références

Bertrand, R. & Blais, J.G. (2004). *Modèles de mesure : L'apport de la théorie des réponses aux items*. Sainte-Foy, Québec: Presses de l'Université du Québec.

Cook L.L. & Petersen, N.S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied psychological measurement*, 11, 225-244.

Crocker, L.M. & Algina, J. (1986). *Introduction to classical and modern test theory*. Florida : Harcourt Brace Jovanich inc.

Dorans, N.J. (2004), Equating, concordance, and expectation. *Applied psychological measurement*, 28, 227-246.

Hanson, B.A. & Béguin, A.A. (2002). Obtaining a common scale for item response theory parameters using separate versus concurrent estimation in the common-item equating design. *Applied psychological measurement*, 26, 3-24.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.) New-York: Springer.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New-Jersey; Lawrence Erlbaum Associates.

McDonald, R.P. (1999). *Test theory : A unified treatment*. New-Jersey: Lawrence Erlbaum Associates.

Ministère de l'éducation du Nouveau-Brunswick (2003) *Plan d'apprentissage de qualité, Énoncé de politique pour l'enseignement primaire et secondaire : Écoles de qualité, résultats élevés*. Province du Nouveau-Brunswick, CNB, 1894.

Michaelides, M.P. & Haertel, E.H. (2004). *Sampling of common items : An unrecognized source of error in test equating*. Rapport de recherche, The Regents of the University of California.

Myslevy, R. J. (1992). *Linking educational assessments : Concepts, issues, methods, and prospects*. Princeton, NJ : ETS Policy Information Center.

Skaggs, G. & Lissitz, R.W. (1986). An exploration of the robustness of four test equating models. *Applied psychological measurement*, 10, 303-317.

Woldbeck, T. (1998). *Basic concepts in modern methods of test equating*. Paper presented at the Annual meeting of the southwest psychological association. New Orleans, LA, april 1998.

LES IMPACTS DES ERREURS DE SPECIFICATION DE LA MATRICE Q SUR LES ESTIMES DE PARAMETRES ET LE TAUX D'ERREURS DE CLASSIFICATION DU MODELE DINA

André A. Rupp, Humboldt Universität zu Berlin
Jon Templin, University of Kansas

Mots-clés : Théorie de la réponse aux items, modèles cognitifs, erreurs de spécification

Résumé :

Cette présentation porte sur une étude évaluant les effets des erreurs de spécifications de la matrice Q sur les valeurs des estimés de paramètres et sur le taux d'erreurs de classement pour le modèle DINA, qui est un modèle de classe latente restreint permettant d'effectuer des classifications multiples de répondants et utilisé dans le cadre d'évaluations diagnostique. Pour cette étude, une matrice Q spécifique à une évaluation reliant l'ensemble de 16 patrons d'attributs possibles selon quatre attributs indépendants a été modifiée pour créer des erreurs de spécification en changeant une valeur de 0 ou de 1 pour certains items. Cette façon de faire à permis de s'assurer que certaines combinaisons d'attributs soient complètement effacées de la matrice Q et que certaines relations de dépendance incorrectes entre des attributs soient présentes. Les résultats démontrent des effets importants, notamment une surestimation du paramètre d'étourdie (*slipping*) pour certains items quand des attributs sont effacés dans la matrice Q et du paramètre de chance (*guessing*) quand des attributs sont ajoutés à la matrice Q. De plus, un taux élevé d'erreur de classement pour des patrons d'attributs contenant des combinaisons d'attributs qui ont été effacées de la matrice Q est observé.

Les impacts des erreurs de spécification de la matrice Q sur les estimés de paramètre du modèle DINA : Implication pour la construction de la matrice Q et du test

Depuis 1990, il existe un intérêt croissant envers des modèles statistiques qui permettent de dresser un profil des répondants en fonction d'une variété de caractéristique latente appelée *attributs*, *compétences*, *traits* ou *habiletés*; le terme *attribut* est utilisé pour cette présentation. Le « *rule-space methodology* » (e.g., Tatsuoka, 1995), les modèles DINA et NIDA (e.g., de la Torre & Douglas, 2004; Junker, 1999), le « *reparametrized unified model (RUM)* ou *fusion model* » (e.g., Bolt & Fu, 2004; Hartz, 2002; Templin & Henson, 2005, Templin, Roussos, & Stout, 2003) et le « *Bayesian inference networks* » (e.g., Williamson, Almond, Mislevy, & Levy, 2006) sont des exemples de ce type de modèle statistique.

Tous ces modèles exigent que les caractéristiques latentes mesurées par un instrument de mesure soient spécifiées. Cette information peut être représenté selon une *matrice Q* (Tatsuoka, 1983). Dans cette matrice, chaque item est présenté sur une

ligne séparée et chaque caractéristique est présentée sur une colonne séparée. Chaque cellule contient, dans le plus simple des cas, une valeur de 0 ou de 1, indiquant si l'item représenté par la ligne requiert la caractéristique représentée par la colonne pour être réussi. La présente étude porte sur différents types d'erreurs de spécification de la matrice Q et leurs effets sur les estimés de paramètres et le taux d'erreurs de classement pour le modèle DINA.

Méthode

Le modèle DINA

Le modèle «*deterministic inputs, noisy and gate (DINA) »* (e.g., de la Torre & Douglas, 2004; Junker & Sijtsma, 2001) est, comme plusieurs des modèles de classement multiple, un modèle non-compensatoire utilisant une fonction de condensation conjonctive (*conjunctive condensation function).* Pour un répondant, la probabilité d'obtenir une bonne réponse à un item est déterminée par deux différentes probabilités d'erreur dépendante de son groupe d'appartenance parmi l'un des deux groupes que le modèle permet de distinguer. La probabilité d'une erreur de chance (*guessing*; g_j) représente la probabilité d'obtenir une bonne réponse quand au moins un attribut nécessaire est manquant tandis que la probabilité d'une erreur d'étourdie (*slipping*; s_j) représente la probabilité d'obtenir une mauvaise réponse lorsque tous les attributs nécessaires sont présents. La variable latente (ξ_{ij}) est utilisée pour déterminer si un individu présente tous les attributs nécessaires pour répondre à un item. Cette variable prend la valeur de 1 si c'est le cas et de 0 dans la situation contraire.

Ainsi, avec X_{ij} représentant le score dichotomique d'un répondant i à un item j, q_{jk} représentant une valeur binaire dans la matrice Q pour un item j et un attribut k, α_{ik} représentant si un répondant i possède un attribut k et tous les autres termes définis plus haut, l'équation pour le modèle DINA est la suivante :

$$P(X_{ij} = x_{ij}) = \left[s_j^{1-x_{ij}}(1-s_j)^{x_{ij}}\right]^{\xi_{ij}}\left[g_j^{x_{ij}}(1-g_j)^{1-x_{ij}}\right]^{1-\xi_{ij}}$$

Les composantes du modèle sont définies de la manière suivante :

$$\xi_{ij} = \prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}$$

$$s_j = P(X_{ij} = 0 \mid \xi_{ij} = 1)$$

$$g_j = P(X_{ij} = 1 \mid \xi_{ij} = 0)$$

En conséquence, seulement deux paramètres sont estimés pour chaque item soit, s_j and g_j, tandis que pour chaque personne k variable latentes α_{ik} sont estimées (i.e., les répondants sont classés), formant un vecteur d'attribut α_i pour un répondant i. Si le modèle est spécifié de manière adéquate pour une certaine banque de donnée et qu'un item s'ajuste bien aux données, il est attendu que les valeurs des paramètres d'erreurs de chance et d'étourdie seront petites. Si un des deux paramètres présente une valeur élevée, cela indique des erreurs de spécification pour un modèle ou un item spécifique.

Design de simulation

Pour cette étude, la génération des données simulées représente une situation ou un test contenant des items nécessitant tous les patrons d'attributs possibles est administré à un groupe de répondants qui possèdent aussi tous les patrons d'attributs possibles. Quatre attributs sont considérés, ainsi un total de 16 patrons d'attributs peut exister chez les répondants. Le tableau 1 présente les caractéristiques du test et des répondants qui résulte de cette situation avec les valeurs de la matrice Q utilisée pour la génération des données.

Tableau1 - Caractéristiques du test et des répondants utilisés pour la génération des données

Caractéristiques du test							Caractéristiques des répondants						
Item	Matrice Q				s_j	g_j	Patron	Attributs				Fréquence	Score
							1	0	0	0	0	.113	0
1	1	0	0	0	.175	.107	2	1	0	0	0	.046	0
2	0	1	0	0	.111	.014	3	0	1	0	0	.065	0
3	0	0	1	0	.018	.063	4	0	0	1	0	.047	0
4	0	0	0	1	.018	.026	5	0	0	0	1	.048	0
5	1	1	0	0	.051	.076	6	1	1	0	0	.066	1
6	1	0	1	0	.102	.026	7	1	0	1	0	.047	1
7	1	0	0	1	.099	.003	8	1	0	0	1	.044	1
8	0	1	1	0	.124	.055	9	0	1	1	0	.039	1
9	0	1	0	1	.162	.120	10	0	1	0	1	.034	1
10	0	0	1	1	.138	.088	11	0	0	1	1	.028	1
11	1	1	1	0	.244	.110	12	1	1	1	0	.088	4
12	1	1	0	1	.065	.118	13	1	1	0	1	.082	4
13	1	0	1	1	.207	.100	14	1	0	1	1	.060	4
14	0	1	1	1	.001	.129	15	0	1	1	1	.026	4
15	1	1	1	1	.088	.056	16	1	1	1	1	.166	15

Les erreurs de spécifications de la matrice Q étudiées sont divisées en deux groupes : 1) les erreurs de spécification représentant un sous-ajustement, un sur-ajustement et un mauvais ajustement balancé de la matrice Q pour des blocs d'items qui requiert un nombre fixe d'attributs et, 2) des erreurs de spécifications représentant des postulats incorrects de dépendances entre deux attributs. Toutes ces conditions sont présentées au tableau 2.

Tableau 2 - *Conditions d'erreur de spécification de la matrice Q*

Condition	Nombre moyen d'attributs par item	Nombre moyen d'items par attributs	Description des changements à la matrice Q	Nombre total de changement (0-à-1)	Nombre total de changement (1-à-0)
qtrue	2,13	8	Génération des données selon la matrice Q	0	0
Groupe 1:Élimination de blocs d'items					
qmblock4	2,07	7,75	L'item à 4 attributs est changé pour un item à 3 attributs sélectionnés au hasard	0	1
qmblock3	1,87	7	Tous les items à 3 attributs sont changés pour des items à 2 attributs sélectionnés au hasard	0	4
qmblock2	1,73	6,5	Tous les items à 2 attributs sont changés pour des items à 1 attributs sélectionnés au hasard	0	6
qblock1	2,40	9	Tous les items à 1 attributs sont changés pour des items à 2 attributs sélectionnés au hasard	4	0
qblock2	2,53	9,5	Tous les items à 2 attributs sont changés pour des items à 3 attributs sélectionnés au hasard	6	0
qblock3	2,40	9	Tous les items à 3 attributs sont changés pour l'item à 4 attributs	4	0
qmix	2,13	8	Les attributs sont effacés et ajoutés pour balancer le nombre total de changement	7	7
Groupe 2: Relations de dépendance incorrectes					
qmatt4	1,87	7	Quand l'attribut 3 est présent, l'attribut 4 est retiré	0	4
qpatt4	2,40	9	Quand l'attribut 3 est présent, l'attribut 4 est ajouté	4	0
qmatt3	1,87	7	Quand l'attribut 4 est présent, l'attribut 3 est retiré	0	4
qpatt3	2,40	9	Quand l'attribut 4 est présent, l'attribut 3 est ajouté	4	0
Qatt34	2,67	10	Les attributs 3 et 4 sont toujours présent ensembles	8	0

Résultats

Les impacts sur l'estimation des paramètres d'items

Le premier aspect d'intérêt est l'impact de différentes erreurs de spécification de la matrice Q sur les valeurs des estimés de paramètres d'items du modèle DINA (i.e., les paramètres s_j et g_j); Les tableaux A1 et A2 présentent les valeurs des estimés des paramètres d'erreurs de chance et d'étourdie pour toutes les conditions ainsi que les réelles valeurs utilisées pour la génération des données. De plus, le tableau A3 présente la déviation moyenne en valeur absolue (DMA) pour chaque estimé de paramètre en fonction de la valeur de génération des paramètres pour chacun des items.

L'effet dominant des erreurs de spécification des entrées de la matrice Q sur les estimés de paramètre pour le modèle DINA est clairement visible dans les trois tableaux. Lorsqu'une valeur 0 est changée pour une valeur 1 pour un item dans une condition particulière, la valeur du paramètre d'erreur d'étourdie est surestimé alors que le paramètre d'erreur de chance est estimé de manière adéquate. À l'inverse, lorsqu'une valeur 1 est changée pour une valeur 0 pour un item dans une condition particulière, la valeur du paramètre d'erreur de chance est surestimé alors que le paramètre d'erreur d'étourdie est estimé adéquatement. Pour la condition *qmix*, ou les deux changement surviennent, la valeur des DMA augmente pour les deux types de paramètres. Peu importe l'item qui est modifié, les valeurs des estimés des paramètres d'erreurs de chance et d'étourdie des items non modifiés demeurent correctes.

L'impact sur le classement des répondants

Afin de mesurer les impacts potentiels des erreurs de spécification de la matrice Q sur le classement des répondants, une inspection des tables individuelles de classement pour identifier des patrons spécifiques ou les erreurs de classement se produisent sont présentés au tableau 3. Quelques cas sont discutés par la suite.

Tableau 3 - Taux d'erreur de classement pour les patrons d'attributs selon les conditions

Conditions	Patrons d'attributs															
	0000	1000	0100	0010	0001	1100	1010	1001	0110	0101	0011	1110	1101	1011	0111	1111
qtrue	.178	.308	.252	.123	.101	.107	.146	.158	.064	.174	.077	.023	.018	.018	.023	.001
qmblock2	.014	***.978***	***.939***	***.847***	***1.000***	**.991**	**.137**	**.973**	**1.000**	**1.000**	**1.000**	.055	.111	.045	*.962*	.005
qmblock3	.178	.306	.219	.083	.103	.104	.144	.151	.162	.180	.063	**.058**	**.043**	**.037**	**.068**	.001
qmblock4	.178	.308	.252	.123	.101	.107	.146	.158	.067	.172	.070	.021	.018	.018	.068	**.002**
qblock1	.281	**1.000**	**1.000**	**1.000**	**1.000**	.145	.165	.164	.162	.247	.278	.052	.049	.05	.087	.003
qblock2	.178	.251	.281	.169	.164	**.320**	**.326**	**.251**	**.310**	**.186**	**.232**	.034	.080	.035	.068	.002
qblock3	.178	.308	.252	.121	.099	.098	.144	.146	.046	.180	.088	**.066**	**.065**	**.043**	**.080**	.001
qmatt4	.126	.271	.211	***.898***	.101	.096	.146	.158	.197	.180	**.070**	.102	.022	**.015**	**.034**	**.001**
qpatt4	.119	.262	.199	**1.000**	.038	.093	**1.000**	.144	**1.000**	.169	.204	**1.000**	.018	.098	.053	.004
qmatt3	.156	.299	.234	.123	***.914***	.101	.146	.196	.067	.358	**.070**	.023	.122	**.037**	**.030**	**.004**
qpatt3	.152	.297	.229	.100	**1.000**	**1.000**	.144	**1.000**	.051	**1.000**	.211	.021	.091	.086	.045	.002
qatt34	.092	.251	.174	**1.000**	**1.000**	.085	**1.000**	**1.000**	**1.000**	**1.000**	.222	**1.000**	**1.000**	.030	.034	.003
qmix	**.153**	**.943**	**.943**	**.983**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**.045**	**.122**	**.150**	**1.000**	.003

Note : Les patrons d'attributs qui correspondent à des items manipulés sont présentés en caractère gras. Les taux d'erreurs de classement très élevés (i.e., taux supérieur à ,5) qui se produisent plusieurs fois sont présentés en italiques.

Pour la condition *qblock1*, où tous les items à 1 attributs sont effacés du test, aucun répondant n'est en conséquence classé comme ayant un seul attribut. Le taux d'erreur de classement est de 100% pour ces patrons d'attributs. Pour la condition *qmatt4*, où l'attribut 4 est retiré pour un item lorsque l'attribut 3 est nécessaire, le résultat le plus frappant et que 424 répondants (89,93%) qui ont réellement l'attribut 3 mais pas l'attribut 4 (i.e., qui ont le patron d'attribut (0010)) sont classés de manière incorrecte, la plupart (74,58%) étant classé comme n'ayant aucun attribut. Le même impact peut être observé pour la condition *qmatt3* ou l'attribut 3 est toujours retiré pour un item lorsque l'attribut 4 est nécessaire. Pour la condition *qpatt4*, l'attribut 4 est toujours ajouté pour les items dont l'attribut 3 est nécessaire ainsi, tous les items nécessitant l'attribut 3 mais non l'attribut 4 sont effacés de la matrice Q. En conséquence, tous les répondants ayant réellement un patron d'attribut incluant l'attribut 3 mais non l'attribut 4 (i.e., qui ont un patron d'attribut (xx10)) présentent des erreurs de

classement. Le même résultat est observé pour la condition *qpatt3* ou l'attribut 3 est toujours ajouté aux items nécessitant l'attribut 4.

Directions pour de futures recherches

Les recherches futures peuvent explorer certains aspects pas analysés dans cette étude. Par exemple, la même matrice Q peut être utilisée avec d'autres modèles de classement multiple tels que les modèles NIDA ou RUM. De plus, la matrice Q peut être étendue à un plus grand nombre d'attributs pour vérifier si les résultats obtenus dans la présente étude peuvent être généralisés à d'autres contextes, ce qui est attendu étant donné nos résultats. La relation de dépendance représentée dans cette étude peut aussi être élargie pour inclure un plus grand nombre d'attributs ou des dépendances plus complexes. Finalement, il pourrait être utile de spécifier une matrice Q qui ne contient pas tous les patrons d'attributs pour la génération des données et d'estimer ensuite un modèle avec une matrice Q qui contient tous les patrons d'attributs possibles.

Références

Bolt, D., & Fu, J. (2004, April). *A polytomous extension of the fusion model and its Bayesian parameter estimation.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061-1071.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69,* 333-353.

Hartz, S. M. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality*. Unveröffentlichte Dissertation, University of Illinois, Urbana-Champaign, Department of Statistics.

Henson, R. A., & Templin, J. L. (2006). *Implications of Q-matrix misspecification in cognitive diagnosis.* Manuscript submitted for publication.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258-272.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, 13-103. New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*, 327-360. Hillsdale, NJ: Erlbaum.

Templin, J. L., & Henson, R. A. (2005). *The random effects reparametrized unified model: A model for joint estimation of discrete skills and continuous ability.* Manuscript submitted for publication.

Templin, J. L., & Henson, R. A. (in press). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods.*

Templin, J. L., Roussos, L., & Stout, W. (2003). *An extension of the current fusion model to treat polytomous attributes.* Unpublished Technical Report for the ETS External Diagnostic Group.

Williamson, D. M., Almond, R. G., Mislevy, R. J., & Levy, R. (2006). An application of Bayesian networks in automated scoring of computerized simulation tasks. In: D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Hrsg.), *Automated scoring of complex tasks in computer-based testing* , 201-258. Mahwah, NJ: Erlbaum.

A1

Valeur des estimés de s_j pour toutes les conditions de simulation

Item	qtrue	qmblock4	qmblock3	qmblock2	qblock1	qblock2	qblock3	qmix	qmatt3	qpatt3	qmatt4	qpatt4	qatt34
1	.175	.176	.174	.179	**.177**	.187	.175	**.167**	.175	.175	.173	.177	.174
2	.111	.098	.098	.101	**.097**	.099	.101	**.101**	.096	.101	.095	.100	.102
3	.018	.016	.015	.016	**.016**	.019	.015	**.015**	.018	**.014**	.015	.020	**.018**
4	.018	.016	.016	.015	**.016**	.019	.017	**.012**	.015	.019	.017	**.016**	**.019**
5	.051	.055	.052	**.048**	.054	**.057**	.054	**.209**	.053	.054	.053	.054	.054
6	.102	.104	.102	**.261**	.106	**.134**	.103	**.122**	.109	**.099**	.090	.107	**.103**
7	.099	.098	.090	**.105**	.100	**.117**	.096	**.087**	.083	.097	.096	**.096**	**.097**
8	.124	.132	.127	**.368**	.129	**.132**	.132	**.333**	.139	**.130**	.117	.138	**.133**
9	.162	.158	.159	**.346**	.164	**.158**	.161	**.156**	.148	.160	.159	**.155**	**.158**
10	.138	.140	.140	**.434**	.133	**.139**	.141	**.119**	**.411**	.122	**.429**	.130	.129
11	.244	.230	**.351**	.229	.230	.249	**.230**	**.344**	.231	**.226**	.223	.231	**.228**
12	.065	.063	**.293**	.050	.069	.061	**.064**	**.299**	.061	.063	.064	**.062**	**.062**
13	.207	.204	**.447**	.203	.198	.216	**.202**	**.408**	**.441**	.193	**.443**	.198	.198
14	.001	.000	**.337**	.000	.000	.000	**.000**	**.000**	**.282**	.000	**.319**	.000	.000
15	.088	**.207**	.102	.100	.099	.103	.101	.114	**.358**	.097	**.363**	.102	.100

Note : les items manipulés pour chaque condition sont présentés en caractère gras.

A2

Valeur des estimés de g_j pour toutes les conditions de simulation

Item	qtrue	qmblock4	qmblock3	qmblock2	qblock1	qblock2	qblock3	qmix	qmatt3	qpatt3	qmatt4	qpatt4	qatt34
1	.107	.100	.109	.298	**.339**	.080	.102	**.369**	.102	.097	.100	.109	.099
2	.014	.020	.027	.263	**.335**	.026	.021	**.376**	.020	.025	.019	.019	.021
3	.063	.062	.068	.150	**.355**	.047	.061	**.394**	.051	**.355**	.145	.051	**.354**
4	.026	.030	.030	.227	**.229**	.029	.029	**.265**	.111	.028	.032	**.308**	**.310**
5	.076	.075	.074	**.074**	.080	**.253**	.074	**.060**	.077	.075	.078	.075	.075
6	.026	.027	.028	**.024**	.043	**.152**	.028	**.052**	.026	**.182**	.038	.026	**.182**
7	.003	.003	.006	**.004**	.003	**.146**	.003	**.005**	.005	.003	.003	**.148**	**.148**
8	.055	.055	.059	**.055**	.056	**.114**	.054	**.033**	.053	**.183**	.065	.053	**.183**
9	.120	.125	.124	**.175**	.117	**.223**	.124	**.240**	.138	.124	.124	**.224**	**.224**
10	.088	.090	.090	**.093**	.096	**.138**	.090	**.146**	**.085**	.092	**.086**	.093	.093
11	.110	.108	**.106**	.107	.108	.105	**.178**	**.104**	.107	**.178**	.114	.108	**.178**
12	.118	.124	**.124**	.131	.125	.125	**.203**	**.125**	.130	.125	.125	**.203**	**.203**
13	.100	.098	**.100**	.100	.098	.097	**.148**	**.101**	**.099**	.100	**.096**	.099	.099
14	.129	.126	**.125**	.151	.125	.124	**.153**	**.149**	**.122**	.126	**.123**	.125	.125
15	.056	**.057**	.057	.058	.058	.058	.058	.057	**.057**	.058	**.055**	.058	.058

Note : les items manipulés pour chaque condition sont présentés en caractère gras.

A3

Tableau des déviations moyennes en valeur absolue (DMA)

Type	Paramètre	qmblock4	qmblock3	qmblock2	qblock1	qblock2	qblock3	qmix	qmatt3	qpatt3	qmatt4	qpatt4	qatt34
DMA-M	s_j	***.119***	***.228***	***.149***	.005	.035	.005	***.072***	***.265***	.008	***.280***	.004	.005
DMA-NM		.004	.005	.007	.005	.007	.004	.001	.008	.005	.006	.007	.006
DMA-M	g_j	.001	.004	.011	***.262***	***.124***	***.056***	***.104***	.003	***.161***	.003	***.154***	***.158***
DMA-NM		.003	.004	***.085***	.005	.009	.003	.020	.013	.004	.013	.013	.004

Note : DMA-M = Déviation moyenne en valeur absolue du paramètre pour les items manipulé, MAD-NM = Déviation moyenne en valeur absolue du paramètre pour les items non manipulés. Les valeurs élevés de DMA pour chaque condition sont présentés en caractère gras et en italique.

LE MODÈLE DE RASCH : AVANTAGES, LIMITES, APPLICATIONS

Jean-Guy Blais, Université de Montréal
Gilles Raîche, Université du Québec à Montréal
Nathalie Loye, Université de Montréal

Mots-clés : modèle de Rasch, mesure, application

Résumé

Ces vingt dernières années, la scène de la recherche en éducation à vu surgir des développements de la modélisation formelle de la mesure en sciences sociales qui, couplés à ceux de la technologie, ont permis l'émergence d'une façon différente d'aborder la mesure en éducation. Alors que depuis le début du 20e siècle les préoccupations métriques au sujet de la mesure en éducation ont été prises en charge par la théorie classique des tests, qui vise les situations où un instrument est utilisé avec un groupe de personnes, une perspective différente permet d'envisager une situation plus fine où l'accent est mis sur la dynamique item - individu. Les modèles qui encadrent formellement cette perspective se retrouvent sous le parapluie de «La théorie des réponses aux items». Toutefois, tous les modèles réunis sous ce parapluie n'ont pas les mêmes perspectives sur les données et prétentions de celles-ci à la mesure. Parmi ces modèles, un certain nombre se retrouve sous l'appellation «modèle de Rasch». Cette famille de modèles se distingue parce que les préoccupations qui ont mené à sa dissémination sont avant tout métriques plutôt que statistiques. Avec la modélisation de Rasch, ce qui est important c'est que les données s'ajustent bien au modèle et non l'inverse, comme pour la régression multiple par exemple. Si les données s'ajustent au modèle de Rasch, l'échelle des valeurs numériques qui représente le continuum mesuré est une échelle linéaire avec des intervalles égaux. Cette présentation fait le point sur ce modèle, ses propriétés, ses problèmes (ce n'est pas une panacée) et illustre son application dans deux situations de mesure en éducation.

1. Le modèle de Rasch

Les développements récents de la modélisation formelle de la mesure en sciences sociales ont permis l'émergence d'une perspective différente sur la mesure en éducation. Les modèles de mesure qui encadrent cette perspective se retrouvent sous le parapluie de la «Théorie des réponses aux items». Parmi ces modèles, un certain nombre sont héritiers des travaux de G. Rasch et constitue une famille de modèle particulière. Cette famille de modèles se distingue d'abord et avant tout parce que les préoccupations qui ont mené à sa dissémination sont métriques plutôt que statistiques. Ainsi, étant donné les propriétés de la modélisation de Rasch, ce qui est important c'est que les données s'ajustent bien au modèle et non l'inverse, comme pour la régression multiple par exemple. La priorité est donc donnée au construit à mesurer,

plutôt qu'au modèle à ajuster. Ainsi, lorsque les données ne s'ajustent pas au modèle, il faut produire de meilleures données, donc revoir les items et les remettre à l'essai, ou en produire de nouveaux plus susceptibles de rencontrer les exigences de la modélisation. Lorsqu'il y a un problème d'ajustement pour certaines données, cela constitue un signe qu'il y un problème avec le stimulus qui a permis d'observer les données, ou qu'il y a un problème avec la réponse au stimulus. Si les données s'ajustent adéquatement au modèle, elles produiront des estimations des paramètres sur une échelle d'intervalle avec une structure linéaire. En fait, ce que permet de faire la modélisation de Rasch, c'est d'explorer et de tester un construit qui est mesurable au sens classique du terme, c'est-à-dire avec des unités égales et, dans certaines situations où la démonstration empirique est possible, avec la propriété de l'addition.

Le modèle de Rasch se distingue des modèles associés à la théorie classique des tests de plusieurs manières. Premièrement, la modélisation se situe au niveau de l'item et du répondant, plutôt qu'au niveau de l'instrument et du groupe de répondants. Deuxièmement, c'est la probabilité de la réponse observée qui est modélisée plutôt que la réponse elle-même. La modélisation possède des caractéristiques intéressantes pour les applications en sciences humaines et plusieurs avantages par rapport à d'autres approches :

- Des estimations sur une échelle linéaire avec des intervalles égaux,
- Des analyses qui tiennent compte du problème des données manquantes,
- Des estimations sur la précision pour chaque item et chaque répondant,
- Des mécanismes de détection pour les données mal adaptées pour les items et les répondants,
- Les paramètres associés à ce qui fait l'objet de la mesure et ceux associés à l'instrument de mesure sont séparables (une conséquence de l'existence de statistiques suffisantes).
- Le rôle de la population de référence est aboli (propriété d'invariance)

Ainsi, la modélisation proposée par Rasch (1960) permet des mesures individuelles pour des variables continues avec des unités égales, des erreurs de mesure pour chaque score, le tout indépendant des caractéristiques du test et supporté par des indicateurs de qualité pour les items et les répondants. De plus, le modèle de Rasch rencontre certaines conditions d'une structure conjointe additive (Luce & Tukey 1964). Il est en effet possible de démontrer que les données qui s'ajustent aux modèles de Rasch produisent des mesures qui satisfont deux des exigences quantitatives nécessaires à la mesure dite «fondamentale» (Bezruczko & Linacre 2005, Massof 2002, Blais 2003) : les axiomes de l'annulation et de la commutativité.

Rasch (1960) a ainsi proposé un modèle général où c'est la probabilité d'observer une réponse donnée, X_i , pour l'item *i* qui est modélisée en fonction de deux variables latentes, celle de la position du répondant et celle de la position de l'item sur leurs continuum respectifs :

$$\text{Probabilité}\,(X_i = x_i |\ \theta_n,\ \delta_i) = f(\theta_n - \delta_i)$$

La probabilité P pour un répondant *n* de répondre dans la catégorie *x* (0, 1, 2, …) de l'item *I* est donné par une fonction *f* exponentielle où l'exposant est constitué d'une relation additive entre les variables latentes associés aux sujets, aux items et aux catégories. Ainsi, selon la fonction *f* privilégiée, on obtient un modèle pour des scores dichotomique, un modèle pour des catégories ordonnées, ou un modèle à facettes. Le modèle unidimensionnel de Rasch pour les variables dichotomiques peut prendre la forme suivante avec θ comme variable latente associée aux répondants (*n*) et δ comme variable latente associée aux items (*i*):

$$\mathrm{P}\,(X_i = 1 \mid \theta_n,\ \delta_i) = \frac{e^{(\theta_n - \delta_i)}}{1 - e^{(\theta_n - \delta_i)}}\ .$$

Le logarithme *G*(*x*) du rapport de chance (*odds ratio*) est une transformation monotone de la variable *P(X)* qui résulte en une fonction additive des variables latentes θ et δ :

$$G(x) = \log \frac{P(X)}{1 - P(X)} = \theta_n - \delta_i\ .$$

Pour le modèle avec des catégories ordonnées le rapport *G*(*x*) est égal à $(\theta_n - \delta_i - F_x)$, avec une variable latente F_x ajoutée pour les catégories, et pour le modèle à facettes il est égal à $\left(\theta_n - \delta_i - \beta_t - F_x\right)$, avec une variable latente β_t ajoutée pour intégrer des juges (ou des moments, des situations, etc.).

Cependant, l'application du modèle de Rasch ne garantit pas à elle seule la présence d'une mesure, au sens classique du terme (i.e. unités égales, propriété d'addition). En effet, il faut d'abord vérifier si les données s'ajustent adéquatement au modèle. La décision concernant la qualité de l'ajustement repose à l'heure actuelle sur des indices (les statistiques *Infit* et *Outfit*, et leurs équivalents standardisés, présentes dans les logiciels commerciaux, par exemple) qui sont dépendants des données puisqu'ils sont calculés à partir des estimations des paramètres. Les indices peuvent donc être contaminés à différents degrés au même titre que les données et trop s'y fier pourrait mener à des décisions rapides quant à l'exclusion de d'items ne s'ajustant pas bien au modèle. Ensuite, même si le modèle s'apparente à une structure conjointe additive, la démonstration n'a pas encore été faite que les valeurs numériques obtenues avec la modélisation possèdent les propriétés de l'addition et reflètent ainsi les caractéristiques d'une quantité.

Toutefois, le modèle de Rasch semble constituer la modélisation la plus appropriée pour les développements contemporains de la mesure en éducation et en psychologie. Les deux dernières décennies ont en effet vu surgir un nouveau complexe dans le domaine du testing. Ce qui est recherché n'est plus lié à la mise au point d'un test constitué d'un groupe d'items fixes et invariables en contenu et en nombre. Il s'agit plutôt de modéliser la rencontre entre un item et un candidat. Les préoccupations sont ainsi de l'ordre de la mise au point de banques d'items où chaque

item d'une banque est en quelque sorte indépendant des autres items de la banque. En principe donc, n'importe quel item pourrait être extrait de la banque pour être inséré dans un test destiné à un sujet ou à un groupe de sujets. Il s'agit d'ailleurs du principe à la base du développement des tests adaptatifs informatisés. Lors d'un test adaptatif, chaque candidat est susceptible d'être exposé à des items différents, tant en nombre qu'en contenu, et il est donc nécessaire de recourir à une modélisation qui procure pour des candidats différents des estimations de l'habileté qui soient situées sur le même continuum.

2. Deux exemples d'applications

2.1 L'accord entre les correcteurs d'une épreuve de français écrit

L'évaluation de performances complexes exige, le plus souvent, une appréciation réalisée par des correcteurs ou des juges. Ainsi par exemple, pour des épreuves nationales, comme les épreuves de français écrit de la fin du secondaire au Québec ou encore les épreuves du baccalauréat en France, de nombreux correcteurs corrigent un grand nombre de copies, le plus souvent au même endroit mais quelquefois dans de nombreux endroits géographiques, sur des périodes pouvant durer plusieurs heures par jour et ce, pendant plusieurs jours. Certains correcteurs peuvent également exhiber des comportements idiosyncratiques et être systématiquement peu sévères ou trop sévères. La diversité des appréciations dans ces situations soulèvent la question de leur stabilité interne et de la convergence des appréciations selon différents correcteurs (Wolfe, & Kao, 1994 ; Lipkins, Jones, & Halkitis, 1996; Lunz, Stahl, & Wright, 1994).

Dans cette perspective, le modèle unidimensionnel à facettes de Rasch (Linacre 1993, 1994a) a été mis à contribution pour l'étude et le diagnostic des différences entre les correcteurs dans une situation de correction d'épreuves de français écrit. L'application de ce modèle nécessite un réseau d'items ou de rubriques, de candidats et de correcteurs qui se chevauchent. Ainsi, dans le cadre de cette application le modèle à facettes de Rasch a été appliqué à des données recueillies selon un design expérimental particulier nommé *Partially Balanced Incomplete Block semi regular group divisible design* pour des copies en provenance d'élèves de quatre populations et pour quatre équipe de correcteurs (Bose, Clathworthy, & Shrikhande, 1953 ; Braun 1988; Blais 1998). Le *PBIB* retenu propose un design de correction multiple de chaque copie par trois correcteurs différents, chaque correcteur ne lisant une même copie qu'une seule fois et corrigeant en commun avec l'un ou l'autre des autres correcteurs un nombre limité de copies. Le pairage des correcteurs permet de mettre en évidence l'effet correcteur indépendamment de l'effet copie. Les écarts entre les données réelles et les valeurs attendues, étant donné le modèle choisi, permettent de déceler au cas par cas chacun des correcteurs ou chacun des répondants présentant un comportement particulier ou inadéquat et même d'interpréter son comportement.

Les données proviennent d'une étude comparative d'élèves de quatre populations francophones d'Europe et d'Amérique du nord. Cette étude à grande échelle a porté sur des échantillons d'élèves du même âge et du même niveau. Chaque

élève devait produire un texte d'une longueur de 350 à 400 mots présentant son village, sa ville, sa région ou son pays à des lecteurs adultes étrangers (Gagné, & *al*., 1995). Cent-quatre-vingt-douze copies choisies au hasard parmi les copies des quatre échantillons d'élèves (48 x 4) constituent le corpus de base pour cette étude. Pour chaque population participante, une équipe de douze correcteurs a été constituée et la séance de correction a été organisée selon un design *PBIB* bien précis (Blais 1998).

Le logiciel *Facets* (Linacre, 1994b) a été employé pour générer les estimations des paramètres pour le modèle unidimensionnel à facettes multiples. Tout d'abord, quatre analyses correspondant aux 192 copies corrigées par chaque équipe de correcteurs séparément permettent d'obtenir une image des 4 équipes de correcteurs et des 4 groupes d'élèves. Ensuite, les deux équipes de correcteurs présentant des comportements extrêmes ont été sélectionnées et un profil des quatre groupes d'élèves a été dressé en utilisant les scores attribués par les correcteurs de ces deux équipes. Une dernière exploration a consisté en l'analyse des scores obtenus par un seul groupe d'élèves, scores attribués par l'ensemble des quarante-huit correcteurs. Une petite partie seulement des analyses effectuées est présentée ci-dessous afin de donner un aperçu du potentiel de la modélisation pour poser un diagnostic sur l'opération de correction des copies (pour plus de détails sur les analyses, voir Loye 2002).

Les paramètres estimant la sévérité de chaque correcteur de chaque population, rangés dans l'ordre décroissant, permettent la création de la figure 1 qui sert à visualiser l'homogénéité des équipes de correcteurs. La courbe de l'équipe 3 a la plus grande amplitude, cette équipe est la plus hétérogène dans sa façon de noter. La courbe de l'équipe 4 a la plus petite amplitude, cette équipe est la plus homogène dans sa façon de noter.

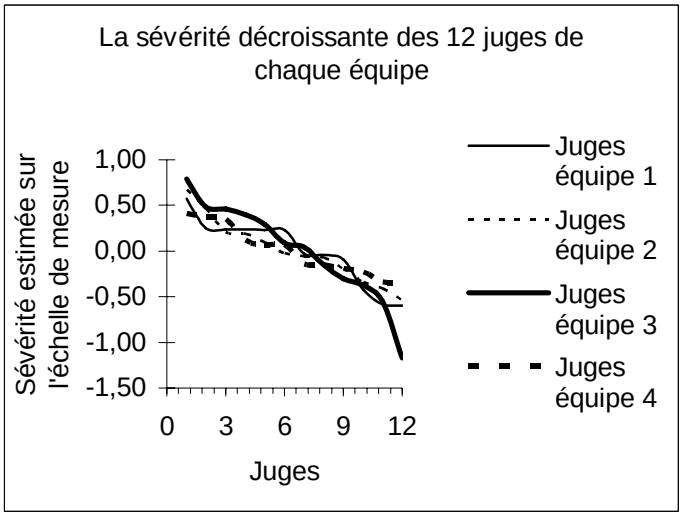


Figure 1 - *Homogénéité des équipes de correcteurs*

La sévérité de chaque correcteur et l'«habileté» de l'élève sont estimées sur la même échelle. De plus, l'analyse permet d'identifier les correcteurs et les élèves dont les données s'ajustent mal au modèle grâce aux statistiques d'adéquation. Le nombre de d'élèves dont les scores présentent un mauvais ajustement au modèle est variable d'une population de correcteurs à l'autre, mais aucun correcteur ne se voit attribuer des valeurs des statistiques d'ajustement très élevées.

Les analyses effectuées ont permis de conclure que les équipes de correcteurs 3 et 4 présentent les comportements extrêmes. En analysant les scores attribués par ces deux équipes de façon séparée, on peut observer (figure 2) la présence de correcteurs particulièrement sévères, indulgents ou inconstants (en traits plus épais)

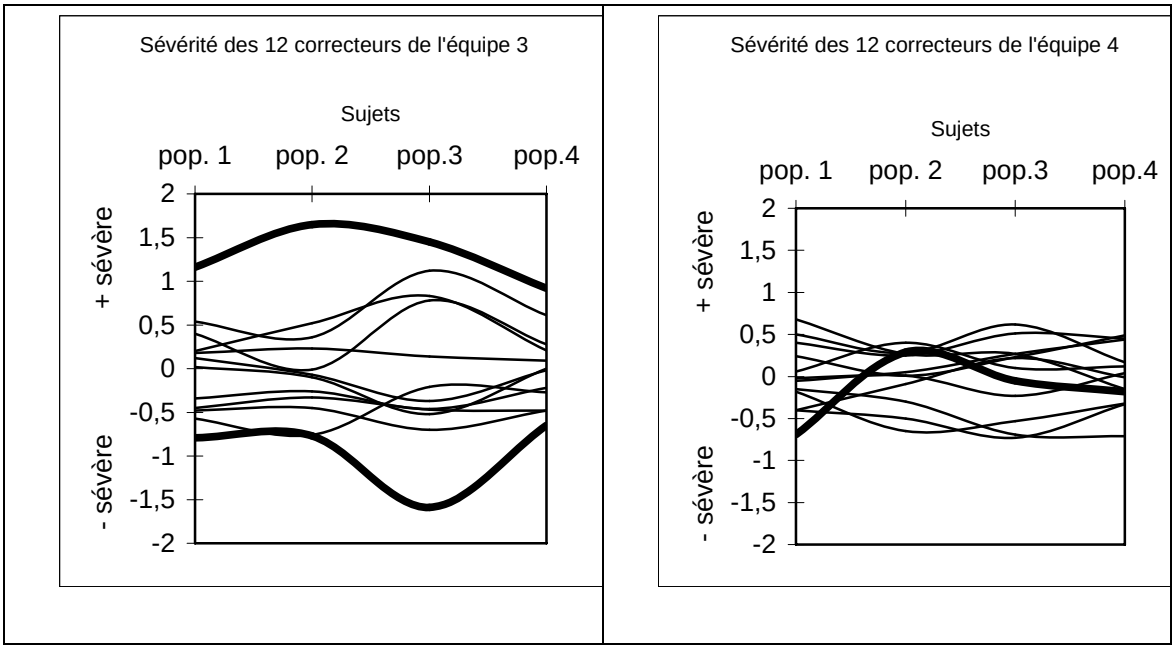


Figure 2 - *Comparaison de la sévérité des correcteurs des équipes 3 et 4 suivant les groupes d'élèves*

2.2 La mise au point d'échelles de niveau de compétence avec le modèle pour des échelles ordonnées[1]

Suite à une réforme des programmes d'études mettant l'accent sur le développement de compétences, le ministère de l'Éducation du Québec a publié une

[1] Cette partie origine d'une recherche menée avec Michel Laurier, professeur à la Faculté des sciences de l'éducation de l'Université de Montréal, en collaboration avec le Ministère de l'éducation du Québec.

politique sur l'évaluation des apprentissages. Dans cette politique, une distinction est faite entre une évaluation continue qui sert essentiellement à aider l'élève dans ses apprentissages et une évaluation qui vise à dresser un bilan à la fin d'une étape scolaire ou d'un cycle. Cette dernière fonction suppose que l'enseignant puisse situer ses élèves sur une échelle de développement de chaque compétence. Au-delà de l'évaluation individuelle à laquelle s'intéresse surtout la politique, l'établissement de descriptions de niveaux facilite le pilotage du système, particulièrement en cours d'implantation d'un nouveau programme, car elle fournit des jalons pour juger du réalisme des cibles de formation définies par ce programme et de l'efficacité des dispositifs qui sont mis en place pour les atteindre. Cette utilisation des échelles de niveaux n'est pas sans les rapprocher des standards de formation qui se généralisent à la fois aux États-Unis (Schmoker & Marzano, 1999; McLeod *et al.*, 1996) et en Europe (Klieme *et al.*, 2004).

2.2.1 Une méthodologie mixte

Pour mettre au point les échelles de niveaux de compétence, nous avons utilisé une méthodologie qui combine les avis d'un groupe restreint d'enseignant experts quant aux manifestations d'une compétence à différents degrés et une approche plus empirique à l'aide du modèle de Rasch pour des catégories ordonnées (*rating scale model*) qui vise à établir les consensus de la communauté enseignante sur l'évolution de la compétence. De plus, des statistiques plus usuelles (moyenne, mode, écart-type) ont aussi été examinées. Il aurait été possible de décider de construire des échelles en s'appuyant principalement sur des théories didactiques et sur les connaissances que nous avons des processus d'apprentissage. Une telle approche mène sans doute à l'élaboration d'échelles cohérentes sur le plan conceptuel, cependant elle comporte des limites sérieuses sur le plan des théories disponibles et fait abstraction du savoir d'expérience des enseignants. Par conséquent, le choix s'est porté vers une approche qui paraît susceptible de dégager les consensus autour de ce qui est effectivement observé par les enseignants chez les élèves à différents niveaux de développement d'une compétence.

La coordination du projet était assurée par le responsable du dossier au service de l'évaluation du MEQ auquel se joignaient une spécialiste de l'enseignement du français et une spécialiste de l'enseignement de la musique. L'encadrement méthodologique était assumé par une équipe de recherche de l'Université de Montréal avec la contribution de deux professeurs spécialistes en mesure et évaluation et d'un agent de recherche. Des groupes de sept enseignants experts ont été formés pour chaque compétence, avec la tâche de formuler, avec l'aide de l'équipe d'encadrement, une série d'indicateurs pur chacun des cinq niveaux de compétence pressenti. Les indicateurs se présentent comme des énoncés qui décrivent une action qu'il est possible d'observer chez les élèves dans la réalisation d'une tâche complexe et qui est susceptible de témoigner de l'atteinte, à la fin du 1er cycle du secondaire, d'un certain niveau dans le développement d'une compétence. À titre d'exemple, au terme de cet exercice, les enseignants experts avaient formulé 108 indicateurs en français et 119 en musique. Ces indicateurs ont été intégrés dans deux questionnaires destinés à des groupes plus nombreux d'enseignants et ceux-ci

devaient estimer le niveau de compétence requis pour réaliser l'action décrite par l'indicateur. Afin d'assurer une présentation homogène des échelles d'une discipline à l'autre, on a retenu une présentation en cinq niveaux de compétence où le niveau 3 correspond à un niveau seuil.

Différentes techniques ont été utilisées afin de dégager les indicateurs les plus révélateurs d'un niveau de compétence. Les réponses ont été placées sur un plan cartésien (les répondants en abscisses et les indicateurs en ordonnées) de façon à composer un diagramme de Guttman qui permet de vérifier visuellement à quel point l'unidimensionnalité recherchée est atteinte et d'identifier rapidement les indicateurs ainsi que les répondants qui présentent des schémas de réponses très atypiques. Ensuite, les données ont été modélisées à l'aide du modèle de Rasch à catégories ordonnées. Étant donné que l'objectif était d'identifier les convergences au sein de la communauté des enseignants, les répondants dont les configurations de réponses paraissaient atypiques ont été systématiquement retirés, dans un processus itératif, de même que les indicateurs pour lesquels les répondants divergeaient d'opinion quant au niveau auquel la réalisation de l'action décrite devait être associée. Des indices statistiques classiques ont également été utiles pour porter des jugements sur la qualité des indicateurs. Ainsi, la dispersion des réponses telle que révélée par la variance et l'étalement des choix tel que visualisé par l'histogramme de la figure 3 permet de reconnaître des indicateurs où les répondants divergent, et ce, même si les indices d'adéquation restent acceptables. La corrélation des réponses pour un indicateur avec l'ensemble des indicateurs retenus pour l'analyse permet de confirmer ce jugement.

Par ailleurs, l'analyse visait également à positionner les indicateurs sur l'échelle, c'est-à-dire à déterminer à quel niveau de développement de la compétence une action particulière peut être associée. Idéalement, l'échelon qui est mentionné le plus souvent par les répondants devrait correspondre au niveau anticipé par les enseignants experts au moment de la formulation des indicateurs. Un écart trop important témoigne d'une différence entre la vision des experts, en grande partie inspirée par la logique du programme, et celle des enseignants, davantage conditionnée par un savoir d'expérience. On peut également juger du niveau des indicateurs à partir de leur rang quintile et à partir de la valeur du paramètre de difficulté estimé par l'analyse de Rasch. Ce dernier indice permet de situer le mode sur une échelle de valeurs variant entre -4 et 4, la moyenne se situant autour de 0.

Ainsi, même si ce n'est pas une description exhaustive des niveaux de compétence qui était recherché, la démarche conduisant aux descriptions finales de niveaux devaient s'aligner le mieux possible sur les principes suivants :

- Des indicateurs pertinents
- Des indicateurs observables
- Des descriptions conformes au curriculum
- Des descriptions claires pour les personnes concernées
- Une cohérence intra et inter niveaux
- Des niveaux qui permettent la discrimination

Les échelles de niveaux de compétence ne reposent pas sur une théorie qui prédirait la séquence de développement. Elles sont plutôt issues d'une approche visant à établir un consensus dans une communauté de pratique à partir de comportements observables considérés typiques pour un niveau de compétence. Ainsi, la prise en compte l'expérience des enseignants devrait faciliter leur jugement sur les performances des élèves dans le nouveau programme de formation.

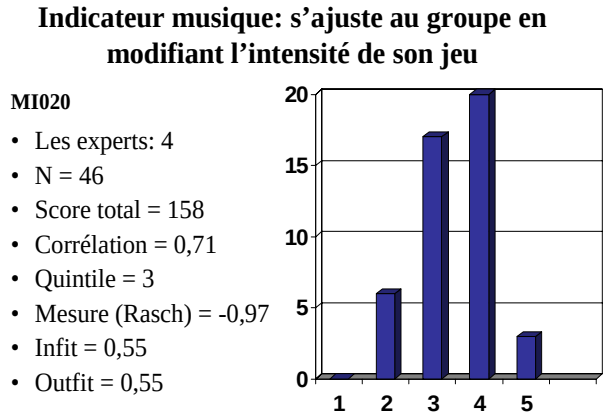


Figure 3 - Exemple des informations utilisées pour analyser les indicateurs

3. Références

Bezruczko, N. & Linacre, J.M. (2005). Measurement theory foundations. Dans N. Bezruczko (dir.) *Rasch measuement in health science*, 8-34, Maple Grove : JAM Press.

Blais, J.G. (1998). Estimating essay reliability in a large-scale international assessment: combining experimental design and raters' resampling. Dans Y. Xuewei (dir.) *The effects and related problems of large scale testing in educational assessment*, 24-36, Beijing : Foreign Language Teaching and Research Press.

Blais, J.-G. (2003). Synthèse sur les fondements de la mesure et sur la contribution de la modélisation moderne. Dans J.-G. Blais et G. Raîche (dir.), *Regards sur la modélisation de la mesure en éducation et en sciences sociales*, 9-32, Québec : Presses de l'Université Laval.

Bose, R. C., Clathworthy & W. H., Shrikhande, S. S. (1953). *Tables of Partially Balanced Designs with two associate classes.* A.N.C. State College publication, Chapel Hill NC.

Braun, H.I. (1988). Understanding scoring reliability : Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*, 1-18.

Gagné et al. (1995). *Savoir écrire au secondaire, projet Diepe.* De Boeck Université.

Klieme, E. *et al.* (2004). *Le développement de standards nationaux de formation : une expertise* (trad. française par le CDIP). Bonn : Ministère fédéral de l'éducation et de la recherche (Allemagne).

Linacre, J. M. (1994a). Constructing measurement with a many-facet Rasch model. Dans M. Wilson (dir.), *Objective measurement, theory into practice, vol. 2*, 129-144, Norwood, N.J. : Ablex Publishing.

Linacre, J.M. (1994b). *A user's guide to Facets*. Rasch Measurement Computer Program: Mesa Press, Chicago.

Lipkins, R. H., Jones, J. P. & Halkitis, P. N (1996). *On the trail of the deviant rater : influence of item effects and method on consistency*. Texte présenté lors de la rencontre de l'American Educational Research Association, New York, Etats-Unis, avril.

Loye, N. (2003). *Une étude de l'accord et de la fidélité inter juges utilisant les approches de Longford et de Linacre*. Mémoire de maîtrise non publié, Université de Montréal.

Luce, R.D. & Tukey, J.W. (1964). Simultaneous conjoint measurement : a new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1-27.

Lunz, M.E., Stahl, J.A. & Wright, B.D. (1994). Inter-judge reliability and decision reproducibility. *Educational and psychological measurement, 54*, 913-925.

Massof, R.W. (2002). The measurement of vision disability. *Optometry and vision science*, 79(8), 516-552.

McLeod, D. B., Stake, R. E., Schappelle, B. Mellisinos, M. & Gierl, M. J. (1996). Setting the standards : NCTM's role in the reform of mathematics education. In S. A. Raizen & E. D. Britton (dir.), *Bold Ventures : US Innovations in Sciences and Mathematics Education. Vol 3: Cases in Mathematics Education*, 13-132. Dordrecht : Kluwer.

Rasch, G. (1960). *Probabilistic models for intelligence and attainment tests*. Copenhague: Danish Institute for Educational Research (Édition revue de 1980. Chicago: University of Chicago Press)

Schmoker, M. & Marzano, R. J. (1999). Realizing the promise of standard-based education. *Educational Leadership 56* (6), 17-21.

Wolfe, E. W. & Kao, C.-W., (1996). *Expert/novice differences in the focus and procedures used by essay scorers.* Texte présenté lors de la rencontre de l'American Educational Research Association, New York, avril.

COMPARAISON EMPIRIQUE DES PARAMÈTRES D'ITEMS/SUJETS DE LA THÉORIE DES RÉPONSES AUX ITEMS ET DE LA THÉORIE CLASSIQUE DES TESTS

Eric Frenette, Université Laval

Richard Bertrand, Université Laval

Pierre Valois, Université Laval

Marc Dussault, Université du Québec à Trois-Rivières

Marie-Hélène Hébert, Université Laval

Mots-clés : Invariance, Théorie des réponses aux items, Théorie classique des tests.

Résumé

Un des grands enjeux des programmes d'évaluation internationale comme l'*International Assessment of Educational Progress* (IAEP ; Lapointe, Mead, & Askew, 1992), le *Programme International pour le Suivi des Acquis Scolaires* (PISA ; OCDE, 2000) et le *Third International Mathematics and Science Study* (TIMSS ; Beaton *et al*, 1996) consiste à fournir des comparaisons justes et équitables de la réussite des élèves, et ce, en dépit des différences culturelles des pays participants. La présente recherche de type méthodologique portera sur la validité des résultats obtenus dans les évaluations internationales utilisant un devis d'échantillonnage d'items complet (où les élèves répondent à tous les items). Deux modèles de mesure, soit la Théorie des réponses aux items (TRI) et la Théorie classique des tests (TCT), peuvent être utilisés afin d'analyser les données recueillies dans les évaluations à devis d'échantillonnage d'items complet. Cependant, peu d'études ont permis de comparer les caractéristiques de ces deux modèles de mesure lors de l'utilisation d'un devis d'échantillonnage d'items complet (Fan, 1998). Lorsque ces études existent, elles portent généralement sur des données simulées. Lors de la présente recherche, nous utiliserons des données réelles provenant du test en mathématiques recueillies dans le cadre de l'IAEP. En premier lieu, le respect de la condition d'application qu'est l'unidimensionalité de la mesure sera examiné. L'unidimensionalité sera vérifiée à partir d'une analyse factorielle dans le cas de la TCT et selon les approches à la base des logiciels DIMTEST et TESTFACT dans le cas de la TRI. Par la suite, la propriété d'invariance des estimés des paramètres d'items et du paramètre d'habileté de la TCT et des modèles logistiques unidimensionels à un, deux et trois paramètres de la TRI sera examinée.

Problématique

Au cours des dernières décennies, les évaluations internationales en éducation ont connu un essor considérable. Entre autres, les programmes *International Assessment of Educational Progress* (IAEP ; Lapointe, Mead, & Askew, 1992), *Programme International pour le Suivi des Acquis Scolaires* (PISA ; OCDE, 2000) et *Third International Mathematics and Science Study* (TIMSS ; Beaton *et al*, 1996) évaluent les compétences scolaires des étudiants de différents pays, en plus de recueillir de l'information sur un ensemble de données sociodémographiques. Leur but, surtout d'ordre politique, concerne le système d'éducation des pays participants. La planification d'une évaluation comportant plusieurs pays nécessite des règles très strictes au plan de l'élaboration de tests, du devis d'échantillonnage, de la collecte de données, de l'analyse des données et de l'interprétation des résultats. Et seulement si ces règles sont respectées, des comparaisons justes et valides qui transcendent les différences culturelles de chacun des pays seront alors permises.

Deux modèles de mesure, soit la Théorie des réponses aux items (TRI) et la Théorie classique des tests (TCT), peuvent être utilisés afin d'analyser les données recueillies lors de l'utilisation d'un devis d'échantillonnage d'items complet. Peu d'études ont permis lors de l'utilisation d'un tel devis de comparer les caractéristiques de la TCT et de la TRI (Fan, 1998). En premier lieu, le respect de la condition d'application qu'est l'unidimensionalité de la mesure sera évalué en fonction de la définition propre à chacune des théories de la mesure. Par la suite, nous vérifierons la propriété d'invariance pour les deux théories de la mesure. Bien que cette propriété a été illustrée théoriquement (Hambleton & Swaminathan, 1985), peu de recherches ont été effectuées sur la propriété d'invariance des paramètres d'items et d'habileté de la TCT et de la TRI avec des données réelles (Anderson, 1999; Cook, Eignor, & Taft, 1988; Fan, 1998). Ses fondements attribuent à la TRI des avantages théoriques importants par rapport à la TCT. Il semble donc raisonnable de penser obtenir des différences appréciables sur le plan des estimés des paramètres des items et d'habileté entre la TCT et la TRI.

Cadre conceptuel

Propriété d'invariance

La tendance dans les évaluations internationales est passée de l'utilisation de la TCT à la TRI. Le principal avantage de la TCT repose sur des conditions d'application relativement souples (Hambleton & Jones, 1993). Une des principales limites de la TCT est appelée dépendance circulaire : (a) le paramètre d'habileté (c.-à-d. le résultat observé) est dépendant de l'échantillon d'items et, (b) les paramètres d'items (c.-à-d. la difficulté et la discrimination de l'item) sont dépendants de l'échantillon de sujets. Cette dépendance circulaire soulève des difficultés théoriques dans certaines situations de mesure en TCT (p. ex. les tests adaptatifs informatisés et l'appariement de tests). L'avantage premier de la TRI sur la TCT repose sur la propriété d'invariance des paramètres d'items et d'habileté (Engelhard, 1992; Rudner, 1983). Selon la théorie, les modèles de la TRI fournissent des paramètres d'items qui

sont indépendants de l'échantillon de sujets et le paramètre d'habileté est indépendant de l'échantillon d'items administré (Fan, 1998). La propriété d'invariance permet de résoudre des problèmes inaccessibles avec les modèles de la TCT : le calcul du résultat global dans les évaluations internationales utilisant un devis matriciel, les tests adaptatifs informatisés et l'appariement de tests (Bertrand & Boiteau, 2003; Frenette, 2004; Hambleton, Swaminathan, & Rogers, 1991). Cependant, en tant que pierre angulaire de la TRI, l'importance de la propriété d'invariance des paramètres des modèles de la TRI ne peut être exagérée, parce que sans cette propriété cruciale, l'utilisation des modèles complexes de la TRI peut difficilement être justifiée tant sur des bases théoriques que pratiques. Afin de pouvoir utiliser ces deux théories de la mesure (TCT et TRI), une condition d'application, nommée unidimensionalité, doit être respectée.

Unidimensionalité

Le défi des évaluations internationales en éducation tels l'IAEP, le TIMSS ou le PISA est d'assurer des comparaisons justes du niveau de rendement scolaire, et ce, pour toutes les juridictions participantes. Pour ce faire, les résultats obtenus prennent sens si et seulement si un concept dominant est nécessaire à l'étudiant afin de bien réussir. Selon Humphreys (1985, 1986), l'idée d'unidimensionalité fait référence au concept dominant pour lequel le test est conçu. La TCT et la TRI ne possèdent pas une définition commune de l'unidimensionalité. Selon l'approche basée sur la TCT, l'unidimensionalité est définie en fonction d'un seul facteur commun, tandis que selon la TRI, elle l'est en fonction d'un seul trait latent nécessaire afin de satisfaire la condition d'application d'indépendance locale (Batley, 1988). Étant donné la difficulté de rencontrer la condition d'application d'indépendance locale, et par conséquent celle d'unidimensionalité, des chercheurs (Humphreys, 1984; Lord & Novick, 1968; McDonald, 1979; Stout, 1987) ont suggéré de la remplacer par la condition d'application «faible» d'indépendance locale. Selon McDonald (1979, 1981), la forme « faible » de la condition d'application ne crée pas de changement important dans l'interprétation des résultats, elle assouplit seulement la définition d'indépendance locale entre les items.

Question de recherche

Cette recherche porte sur la validité des résultats dans les évaluations internationales utilisant un devis d'échantillonnage d'items complet. Nous tenterons de trouver réponse aux questions suivantes : Est-ce que la condition d'application qu'est l'unidimensionalité de la mesure est respectée tant pour la TCT que pour la TRI ? Est-ce que la propriété d'invariance des estimés des paramètres d'items et d'habileté attribue un avantage important aux modèles logistiques unidimensionels à un, deux et trois paramètres de la TRI par rapport à la TCT ?

Méthode

Un échantillon de 175 000 élèves de 9 et 13 ans provenant de 20 pays et de 10 provinces canadiennes a répondu à des tests de mathématique, de sciences et de géographie (Lapointe *et al.*, 1992). Pour notre étude, seulement les étudiants de 13 ans

qui ont répondu au test de mathématique ont été retenus. Le test de mathématique contient 76 items, dont les trois quarts sont à choix multiple et le reste à réponse brève. Ces items sont répartis selon cinq contenus : 28 items portent sur les nombres et les opérations; 13 sur la mesure; 11 sur la géométrie; 15 sur l'algèbre et les fonctions et enfin 9 items concernent les probabilités, les statistiques et l'analyse de données. Chaque étudiant devait répondre aux 76 items du test de mathématique, d'où le devis d'échantillonnage d'items complet. Les pays retenus pour les analyses sont la France, l'Écosse et les États-Unis ainsi que les provinces canadiennes suivantes : l'Alberta anglophone, l'Ontario anglophone et le Québec francophone. Ces pays et ces provinces ont été retenus, car leur échantillon était de taille suffisante (plus de 1 000 étudiants) pour effectuer les analyses choisies.

Unidimensionalité

Analyse factorielle linéaire

Afin de vérifier la condition d'application d'unidimensionalité d'un test (présence d'un facteur commun dominant) sur le plan de la TCT, nous avons eu recours à l'analyse des valeurs propres d'une analyse factorielle linéaire sans rotation à partir des corrélations tétrachoriques. L'approche proposée par Hambleton et Murray (1983), ainsi que par Reckase (1979) pour vérifier l'unidimensionalité d'un test porte sur le ratio entre les valeurs propres d'une analyse factorielle. La présence d'une différence entre la première valeur propre et la deuxième valeur propre qui est 10 fois supérieure à la différence entre la deuxième et la troisième valeur propre indiquera que le test de mathématique est unidimensionel.

DIMTEST

La première approche basée sur la TRI pour vérifier la condition d'application d'unidimensionalité d'un test provient des travaux de Stout (1987, 1990). Cette approche est basée sur l'indépendance essentielle, qui est une opérationnalisation du faible principe d'indépendance locale. Cette approche non paramétrique est utilisée dans l'algorithme du logiciel DIMTEST (Stout *et al*., 1991) et comprend quatre étapes. Deux statistiques sont présentées : la statistique *T* du DIMTEST (Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar, 1993; Stout, 1990) et la statistique *T'* (Nandakumar, 1993). Mentionnons qu'à la première étape, un sous-ensemble d'items doit être choisi parmi les items du test pour qu'ils soient les plus homogènes possibles entre eux et les plus dimensionnellement distincts des autres items. Deux stratégies ont été retenues : (1) analyse factorielle linéaire sans rotation à partir des corrélations tétrachoriques et (2) le jugement d'experts. En ce qui a trait à l'analyse factorielle, cinq versions aléatoires de cinq items seront choisies à partir de cinq sous-échantillons d'élèves différents. Le nombre d'élèves retenus pour les analyses factorielles de l'IAEP est de 330. Pour l'analyse d'experts, cinq regroupements d'items qui représentent les cinq contenus du test de mathématique seront utilisés.

TESTFACT

La deuxième approche basée sur la TRI pour vérifier la condition d'application d'unidimensionalité d'un test est appelée *Full Information Item Factor Analysis* (Bock, Gibbons, & Muraki, 1988). Elle est basée sur le principe d'indépendance locale et est incluse dans le logiciel TESTFACT (Version 2.12.1; Wilson, Wood, & Gibbons, 1991). L'analyse factorielle d'items est basée sur le modèle de facteurs multiples de Thurstone qui permet un examen détaillé et compréhensible de la dimensionalité des items d'un test. Selon cette approche, un ensemble d'items est considéré unidimensionel lorsque le changement de la statistique du khi-deux (différence de la statistique G2) entre le premier facteur et le second facteur n'est pas trouvé statistiquement significatif. Deux choix d'analyses sont proposés avec TESTFACT : (1) l'inclusion d'une valeur fixe du paramètre de pseudo-chance ($c \neq 0$), c.-à-d. l'inverse du nombre de choix de réponses ($c = 1/5$) et (2) lorsque le paramètre de pseudo-chance est fixé à zéro ($c = 0$). La deuxième stratégie est considérée plus restrictive dans l'acceptation de l'hypothèse d'unidimensionalité que la première.

Invariance

Sur le plan de la TCT, les estimés des paramètres d'items, de difficulté et de discrimination (corrélation point-bisériale item-total corrigée) seront calculés à l'aide du logiciel TESTFACT. L'estimé du paramètre d'habileté des élèves est simplement le nombre de bonnes réponses au test de mathématique. Sur le plan de la TRI, le logiciel BILOG (Mislevy & Bock, 1990) permet d'obtenir les estimés des paramètres d'items et d'habileté pour les modèles dichotomiques de la TRI à un, deux et trois paramètres (difficulté, discrimination et pseudo-chance). Les corrélations seront utilisées afin de vérifier la propriété d'invariance des paramètres d'items à partir de différents échantillons de sujets et du paramètre d'habileté à partir de différents échantillons d'items. Fan (1998) a obtenu des corrélations de l'ordre de 0.90 entre les différents estimés de paramètres d'items pour des banques de données qui comportaient à peu près le même nombre d'étudiants que celui de l'IAEP. Dans notre recherche, nous considérerons donc que les corrélations entre les estimés des paramètres d'items obtenus à partir des différentes banques de données doivent atteindre le seuil de .90 pour conclure qu'il y a un bon degré d'invariance. Nous utiliserons aussi ce critère pour les corrélations entre les estimés des paramètres d'habileté des étudiants.

Afin d'obtenir les estimés des paramètres d'items, nous avons choisi les dichotomies suivantes : (1) l'échantillon total est divisé en deux groupes aléatoires d'environ 50 % des étudiants, (2) l'échantillon total est divisé en deux sous-échantillons aléatoires d'exactement 50 % des étudiants, (3) garçons-filles, et (4) le sous-échantillon d'étudiants avec un degré d'habileté se trouvant dans le tiers faible des résultats versus ceux dans le sous-échantillon d'étudiants dans le tiers fort. Afin d'obtenir l'estimé des paramètres d'habileté, nous avons choisi les dichotomies suivantes : (1) les groupes d'items pairs et impairs, (2) des sous-échantillons d'items aléatoires d'exactement 50 %, (3) les items de début (tiers des items) et de fin (tiers des items), et (4) le tiers des items avec le degré de difficulté le plus faible et le tiers des items avec le degré de difficulté le plus élevé.

BILOG

Le logiciel BILOG (Mislevy & Bock, 1990) permet d'estimer les paramètres des modèles logistiques à un, deux et trois paramètres de la TRI. Une distribution log normale (M = 0, ET = 0,5) pour le paramètre de discrimination est utilisée afin de prévenir le cas Heywood (Mislevy & Bock, 1990). Une distribution Bêta est requise pour les estimés du paramètre de pseudo-chance afin d'obtenir des valeurs positives pour les items faciles. Enfin, la distribution normale (M = 0, ET = 2) permet de conserver les valeurs des estimés du paramètre de difficulté dans un intervalle plus restreint lors des cycles d'estimation.

Résultats

Le tableau 1 présente le nombre d'étudiants ayant participé à l'étude de l'IAEP pour chacun des pays et chacune des provinces retenus. Le nombre d'étudiants de chacun des sous-échantillons utilisés lors de la vérification de l'invariance des paramètres d'items est aussi inclus.

Tableau 1 -. Différents échantillons utilisés dans les analyses

	Alberta	Franc e	Ontario	Québec	Écoss e	États-Unis
Total	1 423	1 768	1 428	1 563	1 564	1 407
Aléat11	719	913	691	751	755	710
Aléat12	704	855	734	812	809	697
Aléat13	712	884	714	782	782	703
Aléat14	711	884	714	781	782	704
Filles	683	884	757	795	776	744
Garçons	740	884	671	768	788	663
Faibles	457	571	462	513	578	475
Forts	395	591	459	512	546	457

Unidimensionalité

Analyse factorielle linéaire

La figure 1 présente le graphique des éboulis (de la valeur propre la plus élevée à la plus faible) d'une analyse factorielle linéaire sans rotation sur les corrélations tétrachoriques des 76 items du test de mathématique de l'IAEP pour les étudiants de l'Écosse. À la vue de ce graphique, nous constatons la présence d'un facteur commun dominant selon la TCT, c.-à-d. que la différence entre la première valeur propre et la deuxième valeur propre soit 10 fois plus grande que la différence entre la deuxième et la troisième valeur propre. Des résultats similaires sont obtenus pour les autres pays et provinces.

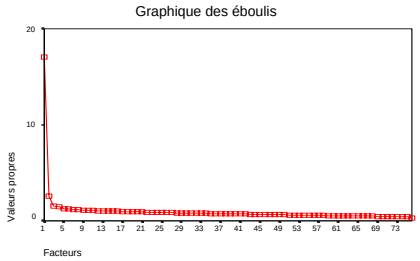


Figure 1 -. Graphique des éboulis

DIMTEST

Un total de 68 analyses (35 analyses factorielles et 33 analyses par jugement d'experts) ont été effectuées sur les banques de données de l'IAEP. Le taux de rejet total de l'hypothèse d'unidimensionalité est de 22 % (15 rejets) pour la statistique *T* et de 29 % (20 rejets) pour la statistique *T'*. Au niveau de l'analyse factorielle, le taux de rejet est de 20 % (7 rejets) pour *T* et de 28,6 % (10 rejets) pour *T'*. Quant à l'analyse par jugement d'experts, le taux de rejet est de 24,2 % (8 rejets) pour *T* et de 30,3 % (10 rejets) pour *T'*. Au tableau 2, nous présentons le taux de rejet pour chacun des pays et chacune des provinces de l'IAEP. Ces résultats sont basés sur les dix analyses DIMTEST effectuées (5 analyses factorielles et 5 analyses par jugement d'experts). Nous remarquons que deux pays à savoir l'Écosse, la France, de même que la province de l'Ontario présentent des taux élevés de rejet. Les autres banques de données peuvent être considérées comme satisfaisant le principe d'indépendance essentiel à un facteur selon les analyses DIMTEST.

Tableau 2 -. Taux de rejet de l'hypothèse d'unidimensionalité sur la base des statistiques *T* et *T'* du DIMTEST

Pays ou provinces	*T*	*T'*
Alberta	10 %	20 %
Écosse	30 %	40 %
France	30 %	30 %
Ontario	30 %	40 %
Québec	0 %	10 %
États-Unis	10 %	10 %

TESTFACT

L'examen des résultats présentés au tableau 3 révèle que pour le test de mathématique les banques de données de l'Alberta, de l'Ontario, du Québec et des États-Unis ne sont pas unidimensionelles lorsque le paramètre de pseudo-chance est fixé à zéro ($c = 0$). Nous sommes en mesure de croire en la présence d'un facteur dominant pour le test de mathématique de l'IAEP pour les banques de données de la France et de l'Écosse. Lorsque le paramètre de pseudo-chance a une valeur différente de zéro ($c \neq 0$), seul le test de mathématique de la banque de données des États-Unis n'est pas considéré unidimensionel. De plus, pour le test de mathématique de la banque de données de l'Ontario, le niveau de signification calculé ($p = 0.058$) est légèrement supérieur au seuil fixé.

Tableau 3 -. Statistique G2 de TESTFACT en fonction de la valeur c

Valeur *c*	Alberta	France	Ontario	Québec	Écosse	Rép. Sov.	États-Unis
Avec c c = 0	DNI p = 0.00	DNI p = 1.00	p = .058 p = 0.00	DNI p = 0.00	DNI p = 1.00	DNI p = 0.00	p = .00 p = 0.00

DNI = Does not increase

Invariance

L'estimation des paramètres d'items pour le sous-échantillon d'étudiants avec des résultats dans le tiers inférieur et le sous-échantillon d'étudiants avec des résultats dans le tiers supérieur pose problème. De fait, le logiciel BILOG affiche une erreur de calcul (*Floating point Error*) lors de la phase 2. Aucune indication n'est fournie dans le guide d'utilisation du logiciel qui permettrait de corriger le problème. D'autres tentatives de regroupements d'élèves ont été faites. Ces tentatives n'ont pas permis d'obtenir les estimés des paramètres d'items pour les groupes d'étudiants faibles et forts.

Paramètres d'items

Les tableaux 4 à 7 présentent les corrélations entre les estimés des paramètres d'items en fonction des dichotomies d'étudiants retenues, et ce, pour la TCT et pour les trois modèles de la TRI, respectivement. Les corrélations obtenues entre les estimés du paramètre de difficulté pour la TCT et pour les trois modèles de la TRI sont généralement supérieures à .90. Dans des conditions similaires, Fan (1998) a obtenu des corrélations légèrement supérieures à celles obtenues dans notre recherche. Sur le plan de la TRI, le modèle à deux paramètres (2P) présente des corrélations entre les estimés du paramètre de difficulté légèrement inférieures à celles du modèle à un paramètre (1P) pour chacune des dichotomies d'étudiants. De même, pour chacune des dichotomies d'étudiants, les corrélations entre les estimés du facteur de difficulté pour le modèle à trois paramètres (3P) sont légèrement inférieures aux modèles à deux paramètres (2P). Ces résultats démontrent que l'ajout d'un paramètre au modèle 1P est marqué d'une diminution de la valeur des corrélations entre les estimés du paramètre de difficulté. Fan (1998) a obtenu des résultats montrant aussi une diminution des corrélations entre les estimés du facteur de difficulté avec l'ajout d'un paramètre au modèle. Enfin, les corrélations obtenues pour le paramètre de difficulté selon la TCT sont généralement plus élevées que celles de la TRI.

La plupart des corrélations obtenues entre les estimés du paramètre de discrimination pour la TCT et pour les trois modèles de la TRI sont inférieures à .90. Fan (1998) a obtenu des corrélations du même ordre de grandeur à celles obtenues dans notre recherche. Les corrélations entre les estimés du paramètre de discrimination sont plus élevées pour le modèle 2P que pour le modèle 3P, et ce, pour chacune des dichotomies d'étudiants. Dans le cas du modèle 3P, les corrélations entre les estimés du paramètre de discrimination sont parfois inférieures à .70. L'ajout du paramètre de pseudo-chance amène une diminution de la valeur des corrélations entre les estimés du paramètre de discrimination. Les corrélations entre les estimés du paramètre de pseudo-chance sont toutes inférieures à .90. La corrélation la plus faible est de .529.

Enfin, les corrélations obtenues pour le paramètre de discrimination selon la TCT sont généralement plus élevées que celles de la TRI.

Tableau 4 -. Corrélations entre les estimés des paramètres d'items pour le modèle de la TCT

Pays et provinces	Aléat11-Aléat12		Aléat13-Aléat14		Garçons-Filles	
	diff	discr	diff	discr	diff	discr
Alberta	.994	.898	.994	.879	.966	.874
France	.994	.923	.994	.941	.974	.932
Ontario	.993	.925	.995	.915	.978	.831
Québec	.995	.876	.993	.915	.969	.845
Écosse	.996	.927	.994	.943	.978	.916
États-Unis	.994	.921	.994	.934	.973	.911

Tableau 5 -. Corrélations entre les estimés des paramètres d'items pour le modèle à un paramètre

Pays et provinces	Aléat11-Aléat12	Aléat13-Aléat14	Garçons-Filles
	b	b	b
Alberta	.991	.990	.960
France	.994	.993	.972
Ontario	.992	.994	.973
Québec	.994	.992	.968
Écosse	.996	.994	.974
États-Unis	.994	.992	.972

Tableau 6 -. Corrélations entre les estimés des paramètres d'items pour le modèle à deux paramètres

Pays et provinces	Aléat11-Aléat12		Aléat13-Aléat14		Garçons-Filles	
	b	a	b	a	b	a
Alberta	.975	.822	.975	.830	.931	.795
France	.990	.909	.980	.931	.968	.925
Ontario	.984	.886	.981	.890	.962	.732
Québec	.958	.861	.982	.876	.943	.811
Écosse	.991	.917	.981	.901	.968	.906
États-Unis	.987	.906	.985	.909	.957	.915

Tableau 7 -. Corrélations entre les estimés des paramètres d'items pour le modèle à trois paramètres

Pays et provinces	Aléat11-Aléat12			Aléat13-Aléat14			Garçons-Filles		
	b	a	c	b	a	c	b	a	c
Alberta	.974	.822	.745	.955	.836	.614	.917	.739	.529
France	.983	.898	.867	.868	.978	.879	.948	.925	.789
Ontario	.978	.672	.836	.974	.907	.868	.940	.939	.786
Québec	.966	.766	.807	.881	.921	.802	.942	.647	.684
Écosse	.983	.858	.850	.881	.959	.754	.962	.782	.745
États-Unis	.977	.732	.796	.969	.991	.722	.852	.997	.808

Paramètre d'habileté

Les tableaux 8 à 11 présentent les corrélations obtenues entre les estimés des paramètres d'habileté en fonction des différentes dichotomies d'items, et ce, pour les banques de données de l'IAEP selon la TCT et les trois modèles de la TRI respectivement. La plupart des corrélations entre les estimés des paramètres d'habileté pour les sous-échantillons d'items sont inférieures aux résultats attendus (.90). Mentionnons que les corrélations entre les estimés des paramètres d'habileté du modèle 3P pour chacune des dichotomies sont légèrement supérieures aux corrélations obtenues pour les modèles à un et deux paramètres. L'ajout de paramètres aux modèles 1P a pour but d'obtenir un estimé plus précis de l'habileté des étudiants. Cependant, la faible augmentation des corrélations entre les estimés des paramètres d'habileté est également associée à une moins bonne estimation des paramètres de discrimination et de pseudo-chance. Les corrélations obtenues pour le paramètre d'habileté de la TCT sont légèrement inférieures à celles du modèle 1P, mais supérieures à celles des modèles 2P et 3P.

Tableau 8 -. Corrélations entre les estimés des paramètres d'habileté pour le modèle de la TCT

Pays et provinces	Début-Fin	Pair-Impair	Aléat21-Aléat22	Facile-Difficile
Alberta	.861	.884	.884	.705
France	.874	.903	.900	.724
Ontario	.818	.854	.849	.636
Québec	.827	.874	.867	.681
Écosse	.867	.923	.918	.757
États-Unis	.868	.905	.907	.727

Tableau 9 -. Corrélations entre les estimés des paramètres d'habileté pour le modèle à un paramètre

Pays et provinces	Début-Fin	Pair-Impair	Aléat21-Aléat22	Facile-Difficile
Alberta	.864	.882	.882	.666
France	.867	.895	.893	.733
Ontario	.817	.855	.846	.632
Québec	.824	.871	.864	.674
Écosse	.874	.895	.888	.734
États-Unis	.860	.898	.900	.716

Tableau 10 -. Corrélations entre les estimés des paramètres d'habileté pour le modèle à deux paramètres

Pays et provinces	Début-Fin	Pair-Impair	Aléat21-Aléat22	Facile-Difficile
Alberta	.867	.886	.881	.685
France	.876	.902	.895	.748
Ontario	.834	.869	.854	.661
Québec	.834	.877	.859	.682
Écosse	.882	.902	.891	.745
États-Unis	.873	.906	.903	.738

Tableau 11 -. Corrélations entre les estimés des paramètres d'habileté pour le modèle à trois paramètres

Pays et provinces	Début-Fin	Pair-Impair	Aléat21-Aléat22	Facile-Difficile
Alberta	.868	.890	.880	.697
France	.880	.903	.896	.763
Ontario	.835	.869	.857	.671
Québec	.839	.882	.859	.688
Écosse	.883	.902	.893	.761
États-Unis	.876	.909	.906	---

-- n'a pu être calculé

Conclusion

Les résultats de la présente recherche démontrent que, lors de l'utilisation d'un devis d'échantillonnage d'items complet, la TCT présente des résultats similaires,

sinon meilleurs que les modèles 1P, 2P et 3P de la TRI. En effet, la condition d'application qu'est l'unidimensionalité de la mesure est respecté pour la TCT, mais elle ne l'est que rarement pour la TRI. De plus, les deux approches basées sur la TRI (DIMTEST et TESTFACT) n'identifient pas les mêmes banques de données comme unidimensionelles. Sur le plan de la propriété d'invariance, la TCT présente des résultats similaires, sinon meilleurs que la TRI. Voici, en résumé, d'autres faits saillants de notre recherche :

- Les résultats de notre recherche laissent croire qu'une seule analyse DIMTEST serait insuffisante pour vérifier l'hypothèse d'unidimensionalité dans le contexte des évaluations internationales. Il est donc recommandé d'effectuer plusieurs analyses afin d'espérer obtenir un consensus au plan des statistiques.
- Tel qu'attendu, les résultats des analyses TESTFACT sont plus restrictifs dans l'acceptation de l'hypothèse d'unidimensionalité lorsque le paramètre de pseudo-chance est égal à zéro ($c = 0$) que lorsqu'on lui attribue une valeur fixe ($c \neq 0$).
- Des réserves sont émises concernant la statistique *G2* de TESTFACT et ont amené les auteurs de TESTFACT à retirer cette statistique de la nouvelle version du logiciel. Les effets de regroupements d'items peuvent faire hausser artificiellement la valeur du khi-deux et ainsi tendre à faire surévaluer le nombre de facteurs dans les enquêtes à grande échelle (Bock *et al.,* 1988; Berger & Knol, 1990 ; De Champlain & Gessaroli, 1991, 1998).
- L'ajout des paramètres de discrimination et de pseudo-chance au modèle à un paramètre (1P) permet d'obtenir de meilleurs résultats au plan de l'invariance des paramètres d'habileté. Ce résultat rencontre les fondements théoriques des modèles de la TRI (Hambleton & Murray, 1983). Cependant, cette amélioration va de pair avec une diminution du degré d'invariance des paramètres de discrimination et de pseudo-chance. Enfin, l'inclusion du paramètre de pseudo-chance apporte peu d'amélioration à l'invariance des paramètres d'habileté. De plus, on constate un manque d'invariance au plan du paramètre de pseudo-chance. L'inclusion du paramètre de pseudo-chance pose souvent problème lors de l'estimation des paramètres d'items (Thissen & Steinberg, 1984).
- Les banques de données identifiées comme étant non unidimensionnelles avec les logiciels DIMTEST et TESTFACT ne présentent pas de différence significative sur le plan de la propriété d'invariance par rapport aux banques de données considérées unidimensionelles. Il importe donc de vérifier en premier lieu l'unidimensionalité du test avant de vérifier la propriété d'invariance. Il nous serait difficile d'imaginer que nous pourrions évaluer l'invariance des paramètres d'items et d'habileté avec un modèle unidimensionel lorsque plus d'une habileté est nécessaire pour rendre compte du patron de réponses des étudiants. Comme le mentionne Blais (1987), l'émergence d'une seconde dimension qui discrimine autant que la dimension visée par le test rend caduque la propriété d'invariance.

Références

Anderson, J. O. (1999). Does complexes analysis (IRT) pay any dividends in achievement testing? *The Alberta Journal of Educational Research, 45,* 344-352.

Batley, R. (1988). *The effects on parameter estimation of correlated dimensions and differentiated ability in a two-dimensional, two-parameter item response model.* Unpublished doctoral dissertation, University of Ottawa, Canada.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics assessments in the middle school years : IEA's Third International Mathematics and Science Study (TIMSS).* Chesnut Hill, MA : Boston College.

Berger, M., & Knol, D. (1990). *On the assessment of dimensionality in multidimensional item response theory models.* Research Report. Netherlands, Twente University.

Bertrand, R. & Boiteau, N. (2003). Comparing the stability of IRT-based and non IRT-based DIF methods in different cultural contexts using TIMSS data. Paper presented at the Annual Meeting of American Educational Research Association, Chicago, USA.

Blais, J. (1987). *Effet de la violation du postulat d'unidimensionalité dans la théorie des réponses aux items.* Unpublished doctoral dissertation, University of Montreal, Canada.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameters estimates. *Journal of Educational Measurement, 25,* 31-45.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

De Champlain, A., & Gessaroli, M. (1998). Assessing the dimensionality of item response matrices with small sample size and short test lengths. *Applied Measurement in Education, 11,* 231-253.

Englehard, G. Jr (1992). Historical views of invariance: evidence from measurement theories of Thorndike, Thurstone, and Rasch. *Educational and Psychological Measurement, 52,* 275-291.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Journal of Educational and Psychological Measurement, 58,* 357-381.

Frenette, E. (2004). La pertinence d'utiliser les modèles complexes de la Théorie des Réponses aux Items dans les évaluations internationales selon un devis d'échantillonnage d'items complet ou matriciel. Unpublished doctoral dissertation, Université Laval, Québec, Canada.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications in test development. *Educational Measurement: Issues ans Practices, 12,* 38-47.

Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Eds.) *Applications of item response theory.* Vancouver, BC : Educational Research Institute of British Columbia.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Priciples and application.* Boston: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20*, 1-14.

Humphreys, L. G. (1984). *A theoretical and empirical study of the psychometric assessment of psychological test dimensionality and bias* (ONR Reseach Proposal). Washington, DC: Office of Naval Research.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Eds.), *Handbook of intelligence* (pp. 201-224). New York: Wiley.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction contexte. *Journal of Applied Psychology, 71*, 327-333.

Lapointe, A. E., Mead, N. A., & Askew, J. A. (1992). *Learning mathematics*. Princeton, NJ: The International Assessment of Educational Progress, Educational Testing Service.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thmpson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 159-168). Greenwich, CT: JAI.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale , NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of general theory. *Multivariate Behavioral research, 14*, 21-38.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3 : Item Analysis and Test Scoring with binary logistic models* [Computer program]. Chicago, IL : Scientific Software, Inc.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 17*, 29-38.

OCDE (2000). *Connaissances et compétences : des atouts pour la vie : Premiers résultats du PISA 2000*. Rapport international. Ottawa, Canada : Statistiques Canada.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor test: Results and implication. *Journal of Educational Statistics, 4*, 207-230.

Rudner, L. M. (1983). A closer look at latent trait parameter invariance. *Educational and Psychological Measurement, 43*, 951-955.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Stout, W. F., Nandakumar, R., Junker, B., Chang, H. H., & Steindinger, D. (1991). *DIMTEST and TESTSIM* (Computer software). Urbana Champaing: University of Illinois, Department of Statistics.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT. Test scoring, item statistics and item factor analysis*. Mooresville, IN : Scientific Software.

L'ÉLABORATION DE MODÈLES DE L'APPRENANT À PARTIR D'UNE APPROCHE BAYÉSIENNES

Michel C. Desmarais, École Polytechnique de Montréal
Jean-Guy Blais, Université de Montréal

Mots clés : modèles graphiques, réseaux bayésiens, diagnostic

Résumé :

Des applications émergentes dans le domaine du *e-learning* font usage de modules pour l'évaluation des acquis et de la connaissance. Cependant, elles impliquent le plus souvent des exigences fortes quant à la granularité et la précision pour servir de base à l'évaluation. Elles exigent également un niveau élevé d'efficience pour un diagnostic de compétences. Dans cette présentation, nous abordons l'application de différents modèles graphiques bayésiens pour l'évaluation détaillée et précise des connaissances.

1. Présentation générale

Plusieurs applications émergentes dans le domaine de l'apprentissage à distance via le réseau Internet demandent l'élaboration de modules pour l'évaluation des acquis et des connaissances. C'est le cas, par exemple, des guides d'étude (Falmagne, Cosyn, Doignon, & Thiéry, 2006; Heller, Steiner, Hockemeyer, & Albert, 2006). Ces applications adaptent le contenu et les recommandations en fonction du diagnostic des compétences de l'apprenant. Toutefois, elles imposent souvent des exigences très fortes quant au niveau de granularité et de précision pour l'évaluation. En outre, elles exigent un haut niveau d'efficience pour poser le diagnostic de compétences.

Ainsi, s'il est fréquent pour un étudiant de se soumettre à une épreuve d'une à plusieurs heures dans le cadre de l'évaluation sommative à la fin d'un cours, cette pratique serait inacceptable pour un utilisateur qui cherche simplement à bénéficier

des conseils et instructions du système. L'évaluation étant perçue par l'utilisateur, au mieux, comme une étape nécessaire et, au pis, un moment désagréable et improductif à ses yeux. Le système doit donc en arriver à un diagnostic des difficultés précis tout en limitant au maximum le volet formel de l'évaluation pour établir ce diagnostic.

Dans le cadre du *e-learning*, les modèles graphiques comme les réseaux bayésiens sont, depuis plus d'une décennie, reconnus comme une approche très prometteuse vers l'objectif de fournir un diagnostic cognitif précis et efficient (Almond & Mislevy, 1999; Desmarais & Pu, 2005). D'une part, ils possèdent l'avantage d'une très grande flexibilité pour la modélisation des habiletés. Ils peuvent intégrer, dans un schème probabiliste cohérent et unificateur, un grand nombre d'éléments de nature directement observables, notamment des réponses aux items et des comportements, et des variables comme des concepts et des conceptions erronées. Le modèle graphique permet de représenter les relations entre ces éléments sous la forme d'une hiérarchie de concepts et de conceptions erronées. On retrouve généralement à la base de cette hiérarchie des «nœuds» représentant des indices observables, i.e. les items (Vomlel 2004; Millan, Trella, Pérez-de-la-Cruz, & Conejo 2000).

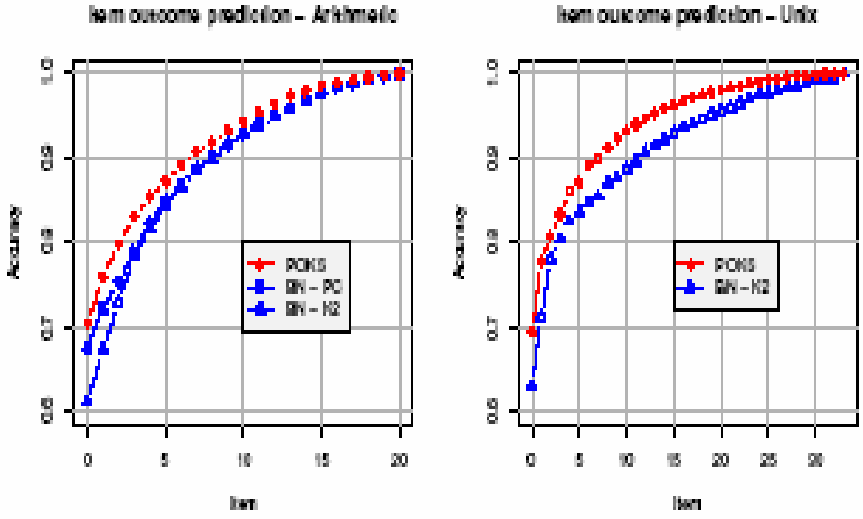


Figure 1 - La prédiction de la performance selon les items. À gauche, la prédiction moyenne pour l'épreuve de mathématiques. À droite, la prédiction moyenne pour l'épreuve Unix (chaque ligne représente la moyenne pour les six échantillons simulés).

Nous étudions ainsi l'application de différents modèles graphiques bayésiens pour l'évaluation détaillée et précise des connaissances. Plusieurs méthodes ont été évaluées. Une première méthode consiste à induire les relations entre les items eux-mêmes. Cette méthode ne comporte donc aucune variable latente (au contraire de la théorie des réponses aux items par exemple) et elle s'inspire des travaux de Doignon et Falmagne (1999) sur la théorie des espaces de connaissances. Cette théorie met de l'avant l'idée que la maîtrise des items progresse dans un ordre partiellement défini. Cet ordre est représenté par un graphe dirigé acyclique qui fournit la base du modèle probabiliste pour évaluer les acquis.

Une simulation effectuée avec deux ensembles de données, un test sur la connaissance de UNIX et un test de mathématiques sur les fractions, démontrent que cette méthode permet de prédire efficacement la connaissance des apprenants. Les résultats sont illustrés à la figure 1. Toutefois, les résultats diffèrent entre les deux

tests et selon l'algorithme utilisé. Un algorithme dit «naïf» affiche une meilleure performance que des algorithmes plus génériques, malgré le fait qu'il repose sur des hypothèses statistiques fortes et nécessite beaucoup moins de calculs. On attribue cet avantage à la nature des espaces de connaissances, mais cette hypothèse reste à valider.

Cette première expérience démontre l'efficacité des méthodes bayésiennes pour la prédiction du succès aux items. Une seconde méthode consiste à utiliser un réseau bayésien comportant des noeuds latents qui représentent des concepts et des conceptions erronées pour modéliser et inférer les connaissances de plus haut niveau (Vomlel, 2004). Il est possible de combiner ce modèle avec la méthode basée uniquement sur les items pour en arriver à une performance combinée qui dépasse celle de chaque méthode individuelle.

2. Références

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 223–237.

Desmarais, M. C., & Pu, X. (2005). Computer adaptive testing: Comparison of a probabilistic network approach with item response theory. *Proceedings of the 10th International Conference on User Modeling (UM'2005)*, 392–396. Edinburg: Springer.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin: Springer-Verlag. Falmagne, J.-C., Cosyn, E., Doignon, J.-P., & Thiéry, N. (2006). The assessment of knowledge, in theory and in practice. Dans R. Missaoui, & J. Schmid (dir.), *ICFCA*, Vol. 3874 of *Lecture Notes in Computer Science* (pp. 61–79). New York: Springer.

Heller, J., Steiner, C., Hockemeyer, C., & Albert, D. (2006). Competence–based knowledge structures for personalised learning. *International Journal on E–Learning*, *5*(1), 75–88.

Millan, E., Trella, M., Pérez-de-la-Cruz, J.-L., & Conejo, R. (2000). Using Bayesian networks in computerized adaptive tests. In M. Ortega, & J. Bravo (Eds.), *Computers and education in the 21st century*, 217–228. New York: Springer.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. New Jersey: Prentice Hall.

Vomlel, J. (2004). Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 12* (Supplementary Issue 1), 83–100.

INTEGRATION DE LA THEORIE DES REPONSES AUX ITEMS A L'EVALUATION DE LA COMPETENCE A RESOUDRE DES PROBLEMES MATHEMATIQUES

Nancy Boiteau, UQAR (Campus de Lévis)
Dominic Voyer, UQAR (Campus de Lévis)
Michel Rousseau, Université d'Ottawa

Mots-clés : Testing adaptatif, TRI, TIC, compétences, résolution de problèmes mathématiques

Résumé:

L'activité de résolution de problèmes fascine encore les chercheurs et pose toujours beaucoup de difficultés dans les classes, notamment au chapitre de l'évaluation de la compétence à résoudre des problèmes mathématiques. Le rôle nouveau que la réforme des curriculums fait jouer à la résolution de problèmes au Québec, nous force maintenant à questionner les méthodes d'évaluation traditionnelles qui rendent difficile l'évaluation de la progression des apprentissages des élèves en lien avec cette compétence. Certains chercheurs ont reconnu les problèmes de précision et de validité des interprétations associés aux méthodes d'évaluation traditionnelles. Parmi les alternatives, le testing adaptatif par ordinateur (TAO) s'avère être un moyen permettant d'obtenir des informations dotées d'une grande précision concernant la performance d'un individu grâce à un arrimage des modèles de la TRI et des TIC. En considérant que l'évaluation de la compétence à résoudre des problèmes mathématiques doit tenir compte à la fois de la maîtrise des savoirs, savoir-faire et stratégies, et à la fois du processus de représentation que l'élève met en place pour comprendre le problème, nous visons la création d'outils d'évaluation informatisés permettant de dégager des informations pertinentes par rapport à cette compétence. L'objectif de cette présentation est d'amorcer une réflexion sur la nécessité de combiner les modèles de la TRI, les TIC et la didactique des mathématiques dans l'élaboration d'outils d'évaluation informatisés visant à appuyer l'inférence de la compétence à résoudre des problèmes mathématiques, tel que la définissent les programmes actuels.

Intégration de la théorie des réponses aux items à l'évaluation de la compétence à résoudre des problèmes mathématiques

Pour quiconque s'intéresse de près ou de loin au monde de l'éducation, le mot « compétence » et l'expression « résolution de problèmes » sont très présents dans le jargon quotidien. Dans l'enseignement des mathématiques, l'activité de résolution de problèmes est présente depuis déjà bien longtemps. Il est possible de remonter aussi loin que dans les programmes de 1861 pour voir la présence de la résolution de problèmes[1] dans les programmes de mathématiques (Bélanger, Gauthier & Tardif, 1993). En 1945, George Pólya a publié un ouvrage marquant sur la résolution de problèmes et ses vertus pour l'apprentissage : How to Solve It. En 1980, le NCTM[2] publiait « An Agenda for Action » qui réclamait une intégration de la résolution de problèmes dans les curriculums. Depuis, dans ses rapports de 1989 et de 2000, le NCTM a réitéré la nécessité d'accorder un rôle majeur à la résolution de problèmes dans les programmes de formation. Au Québec, l'une des trois compétences transversales qui chapeautent l'ensemble des programmes de formation en vigueur au primaire et au secondaire est la compétence à résoudre des problèmes (MELS, 2001; 2003). Dans ce contexte, la résolution de problèmes assume maintenant un double rôle dans les programmes de formation québécois. Cette activité est vue à la fois comme un objet d'apprentissage, c'est-à-dire comme une compétence à développer, et comme un moyen d'acquisition de nouvelles connaissances. Malgré la valeur que l'on accorde à la résolution de problèmes et qu'elle soit implantée depuis longtemps dans les programmes de mathématiques, cette activité fascine encore les chercheurs et pose toujours beaucoup de difficultés dans les classes, notamment au chapitre de l'évaluation de la compétence à résoudre des problèmes mathématiques.

Pour parler de résolution de problèmes, il faut clarifier ce que l'on entend par PROBLÈME. Plusieurs chercheurs se sont attardés à formuler une définition de ce

[1] Ici l'expression « résolution de problèmes » est utilisée dans un sens très large, incluant la résolution de problèmes routiniers, qui se veut davantage un exercice d'application de procédures apprises précédemment.

[2] National Council of Teachers of Mathematics

qu'est un problème et un consensus semble s'établir autour de deux éléments importants. Le premier veut qu'il y ait problème lorsque nous ne connaissons pas *a priori* la procédure pour obtenir une solution, ce qui rejoint la définition de Hayes (1981) : « Lorsqu'il y a un écart entre où vous êtes et où vous voulez être et que vous ne savez pas comment trouver une façon de combler l'écart, vous avez un problème » (traduction personnelle). Le deuxième élément a trait à la possibilité d'agir sur la situation : certains auteurs utilisent l'expression *défi raisonnable* pour mettre en évidence le caractère accessible du problème. En effet, pour qu'il s'agisse d'un véritable problème, la solution ne doit pas être évidente, mais le sentiment qu'il est possible d'intervenir doit être présent. La solution nécessite une recherche appuyée par le réinvestissement de connaissances acquises (ressources disponibles). Ainsi, un problème pour les uns n'en sera pas un pour les autres.

La résolution de problème, considérée dans les nouveaux programmes québécois à la fois comme une compétence à développer et un objet d'apprentissage en soi, entraîne un réaménagement dans la méthodologie liée à l'évaluation des apprentissages. Dans une perspective de renouvellement scolaire, l'évaluation n'est plus considérée étrangère au processus pédagogique. Elle fait partie intégrante de l'enseignement et de l'apprentissage parce qu'elle offre la possibilité de prendre des décisions relatives au processus d'apprentissage de l'élève et à la pédagogie de l'enseignant. Vue ainsi, l'évaluation doit contribuer aux vertus d'une communication efficace et constructive entre les enseignants et les élèves, elle doit véhiculer un potentiel métacognitif visant la régulation des apprentissages. Plusieurs approches d'évaluation sont possibles. Elles peuvent s'appuyer sur des démarches intuitives et informelles, mais peuvent tout aussi bien tendrent vers des démarches instrumentées, approfondies et rigoureuses (grilles d'appréciation, tests, examens, liste de vérification, etc.)

Faire appel aux démarches instrumentées dans un processus d'évaluation de la compétence à résoudre des problèmes mathématiques exige que les outils de mesure et d'évaluation soient structurés de façon à répondre à la demande comtemporaine de l'évaluation des apprentissages. Si l'on souhaite faire de l'évaluation un véritable

instrument d'aide à l'apprentissage, il est primordial de s'assurer que les tâches évaluatives constituant nos outils de mesure véhiculent de l'information cohérente et pertinente au développement de la compétence. Pour atteindre cette visée, les outils doivent être structurés en fonction de cadres théoriques et méthodologiques biens définis.

Objectif de la présentation

Le rôle nouveau que la réforme des curriculums fait jouer à la résolution de problèmes nous force maintenant à questionner les méthodes d'évaluation traditionnelles qui rendent difficile l'évaluation de la progression des apprentissages des élèves en lien avec cette compétence. L'objectif de cette présentation est d'amorcer une réflexion sur l'utilité de combiner les modèles de la TRI, les TIC et la didactique des mathématiques dans l'élaboration d'outils d'évaluation informatisés visant à appuyer l'inférence de la compétence à résoudre des problèmes mathématiques, tel que la définissent les programmes actuels.

Résoudre un problème : une mobilisation de ressources disponibles

La compétence à résoudre des problèmes mathématiques s'articule autour de plusieurs ressources cognitives (savoirs, savoir-faire et stratégies) dont la présence est déterminante pour son développement. Bien que le processus d'évaluation de cette compétence doive s'appuyer sur la réalisation de tâches complexes contextualisées et réalistes (situations problèmes) qui permettent d'évaluer la capacité d'un élève à mobiliser et combiner l'ensemble de ces ressources, il faut aussi s'assurer, dans une certaine mesure, de la disponibilité de ces ressources dans le répertoire cognitif de l'élève (Scallon, 2004). Ainsi, l'évaluation de la compétence à résoudre des problèmes mathématiques ne doit pas porter exclusivement sur la réalisation de situations problèmes complexes.

Le niveau de maîtrise des ressources cognitives d'un élève est un facteur fondamental dans l'évaluation de sa compétence à résoudre des problèmes. Être compétent à résoudre des problèmes signifie être en mesure de mobiliser et d'utiliser efficacement un ensemble de ressources (MELS, 2001; 2003). Évaluer la compétence dans ce contexte requiert, dans un premier temps, de s'assurer de la disponibilité des ressources cognitives nécessaires à la résolution du problème. En omettant de vérifier les ressources cognitives maîtrisées par l'élève, une mauvaise inférence peut être faite par l'enseignant qui jugera l'élève incompétent à résoudre un problème, c'est-à-dire incapable de mobiliser les ressources nécessaires à la résolution, alors que ces ressources ne sont peut-être pas disponibles dans son répertoire cognitif. Dans un tel cas, l'élève n'est pas nécessairement incompétent à résoudre le problème, il peut simplement ne pas avoir les bonnes ressources disponibles dans son répertoire cognitif.

C'est ainsi que les ressources cognitives à mobiliser doivent être objets de vérification avec des tâches évaluatives variant en termes de complexité, c'est-à-dire des tâches qui se distinguent non seulement par le contenu qu'elles véhiculent mais aussi, par le degré d'autonomie qui est laissé à l'élève dans leur réalisation (contenu plus simple avec degré d'autonomie plus faible chez les élèves vers un contenu plus élaboré avec degré d'autonomie plus élevé). C'est sur cette base qu'il convient de fonder une démarche d'évaluation de la compétence à résoudre des problèmes visant l'aide à l'apprentissage (Scallon, 2004). La liste des ressources maîtrisées et non maîtrisées présente un intérêt certain pour ce qui est de guider la progression de l'élève dans le développement de la compétence à résoudre des problèmes mathématiques. Cette liste devrait idéalement offrir la possibilité de saisir aussi précisément que possible ce qui pourrait entraîner des difficultés éventuelles à l'élève au moment de la résolution d'un problème. En effet, si les situations-problèmes qu'on lui présente sollicite la mobilisation de ressources non maîtrisées par l'élève, ce dernier aura inévitablement de la difficulté à résoudre le problème. L'intérêt de connaître le niveau d'habileté en mathématiques lié à la maîtrise des ressources cognitives (savoirs, savoir-faire, stratégies) est d'offrir un éclairage lié à la

performance de l'élève pour la résolution de problèmes mathématiques qui sollicitent la mobilisation de ces ressources.

En mathématiques comme ailleurs, la plupart des outils d'évaluation des ressources cognitives ont été élaborés de façon à ce que tous les individus à qui on les administre répondent en même temps à un même groupe de tâches évaluatives sous la forme papier-crayon et ce, dans l'optique d'obtenir une estimation de leur niveau d'habileté pour un domaine donné. Pour témoigner du niveau d'habileté en mathématique liés à la maîtrise des ressources cognitives, ces outils d'évaluation traditionnels ont une portée très limitée en ce qui a trait à la validité des interprétations qui en résulte, étant donné les problèmes de précision qui leur sont associés (Hambleton, Swaminathan & Rogers, 1991; Wainer, 2000). En termes d'efficacité de la mesure, si l'on veut attester de l'atteinte d'un certain niveau de maîtrise avec une faible marge d'erreur, on peut penser que des tâches évaluatives dont le niveau de difficulté se trouve autour du niveau d'habileté de l'individu seront préférables (Lord, 1980).

Les TIC en support à l'évaluation

L'évolution technologique des dernières années associée au développement des nouvelles théories de la mesure ont permis aux chercheurs de proposer une alternative afin de remédier aux problèmes de précision liés à l'utilisation des outils traditionnels papier-crayon; il s'agit du testing adaptatif par ordinateur[3] (TAO) (Hambleton, Jones & Rogers, 1990; Kim & Plake, 1993; Lord, 1980; McBride, 1997; Wainer, 2000; Weiss, 1985; Weiss & Betz, 1973). L'une des caractéristiques les plus importantes du TAO est de permettre d'assigner, aux élèves, des tâches évaluatives dont le niveau de difficulté correspond à leur niveau d'habileté. Un des avantages du TAO est son efficacité. Il permet d'obtenir des estimations dotées d'une grande précision concernant la performance ou l'habileté des élèves et ce, en utilisant moins de tâches évaluatives et en moins de temps que ne le requièrent les outils traditionnels papier-crayon (Weiss, 1982; Bock & Zimowski, 1998; Sands & Waters, 1997; Wainer,

2000). Selon McBride (1997), plusieurs simulations ont montré la très grande efficacité des tests adaptatifs informatisés : l'utilisation d'estimés de paramètres basés sur des modèles issus de la théorie des réponses aux items (TRI) ne serait pas étrangère à ce constat. Hambleton, Jones et Rogers (1990) affirment que les écrits supportent grandement l'utilisation de ces modèles lors du développement de tests adaptatifs. La TRI est, parmi toutes les théories des tests existantes, la plus utilisée dans le cadre du testing adaptatif par ordinateur.

Les modèles de la TRI offrent des cadres théorique et méthodologique puissants pour structurer des outils de mesure permettant de dégager de l'information précise et de qualité concernant la performance des individus. La TRI offre des modèles logistiques de réponses aux items permettant d'obtenir une estimation la plus adéquate possible de la performance des individus à partir de leurs réponses aux différents items (tâches évaluatives) en plus de permettre l'estimation de la qualité des items eux-mêmes (Hambleton & Swaminathan, 1985; Hulin, Drasgow & Parsons, 1983; Lord, 1980). Cette théorie possède une caractéristique qui lui confère son avantage sur les autres théories soit celle de l'invariance des paramètres d'items. En fait, la TRI fournit un puissant cadre mathématique pour calibrer les items (estimer les paramètres d'items) d'une banque tout en les situant sur une échelle commune et ce, indépendamment des sujets (tirés d'une même population) utilisés pour estimer les paramètres d'items. Étant donné que l'objet d'analyse en TRI est la relation entre l'habileté de l'individu et la réponse à chacun des items (probabiltié de réussite), une mesure de précision est disponible pour chaque item à chaque niveau d'habileté et il devient possible de structurer des outils de mesure ayant une précision désirée. De plus, il est possible d'obtenir un niveau de performance qui se situe sur la même échelle de mesure que le niveau de difficulté des items de la banque. Cette correspondance métrique permet de sélectionner les items (tâches évaluatives) les mieux appariés au niveau d'habileté des individus. En d'autres termes, elle permet de structurer des outils sur mesure et adaptés au niveau de performance des individus.

[3] Résulte de la traduction du terme « Computerized Adaptive Testing » (CAT).

Sachant qu'au plan théorique, la modalité TAO permet d'obtenir des informations dotées d'une grande précision concernant la performance d'un individu grâce à un arrimage des modèles de la TRI et des TIC, nous pouvons entrevoir les possibilités d'implanter des outils d'évaluation informatisés visant l'évaluation de la maîtrise des ressources cognitives (savoirs, savoir-faire et stratégies) pour témoigner de façon précise du développement de la compétence à résoudre des problèmes en mathématiques (Boiteau, 2004).

La compréhension du problème

Au-delà du niveau de maîtrise des ressources cognitives en lien avec la résolution de problèmes mathématiques, la compétence à résoudre un problème est liée à la compréhension que l'élève se construit du problème. La compréhension d'un problème mathématique peut se définir comme la construction d'une succession de représentations mentales (internes) du problème (Andre, 1986; Hegarty & Mayer, 1992; Hinsley, Hayes & Simon, 1977; Janvier, 1987). Le processus de compréhension mis en œuvre par les élèves pour se représenter une situation-problème jouent un rôle important dans la réussite en résolution de problèmes. Cummins, Kintsch, Reusser et Weimer (1988) font ressortir de leur étude que la réussite en résolution de problèmes dépend en premier lieu d'une bonne représentation interne du problème. Lorsque nous définissons le processus de compréhension d'un problème mathématique comme une construction successive de représentations mentales, il faut traduire cela comme l'évolution de l'organisation mentale des données du problème faite par le sujet. Il ne fait aucun doute que la résolution de problèmes, par sa complexité, exige davantage de l'élève que la maîtrise des savoirs, savoir-faire et stratégies mathématiques en lien avec le problème. Ce constat alimente l'idée que l'évaluation de la compétence à résoudre des problèmes mathématiques doit tenir compte à la fois de la maîtrise des savoirs, savoir-faire et stratégies, et à la fois du processus de représentation que l'élève met en place pour comprendre le problème.

Dans le cas d'un problème écrit, la compréhension du problème passe par la compréhension du texte de l'énoncé et cette compréhension peut être influencée par le type d'information contenu dans l'énoncé. En sus des données essentielles à la résolution, un énoncé de problème écrit peut contenir certaines informations non essentielles. Ainsi, les énoncés peuvent comprendre des éléments situationnels permettant de mieux situer le problème mathématique dans son contexte et des éléments explicatifs servant à préciser les liens entre les données. Les chercheurs qui se sont penchés sur le type d'informations contenues dans les énoncés de problèmes ont surtout cherché à savoir si la présence d'éléments situationnels et explicatifs peuvent avoir un effet sur la compréhension du problème et sur la réussite des élèves. Récemment, des études ont mis en relation la présence de tels éléments (situationnels et explicatifs) dans un énoncé et l'habileté en mathématiques des élèves. Moreau et Coquin-Viennot (2003) et Voyer (2006) ont montré que l'habileté en mathématiques avait une incidence sur les représentations que les élèves se construisent pour comprendre un problème et plus spécifiquement sur le rapport qu'ils entretiennent avec les différents types d'information présents dans un énoncé. L'étude de Voyer (2006) montre que les élèves faibles en mathématiques réussissent plus facilement les problèmes dont les énoncés ne contiennent pas d'éléments situationnels. Ainsi, on pourrait penser qu'un élève faible qui ne réussit pas à résoudre un problème dont l'énoncé comprends des éléments situationnels aurait pu réussir le même problème si son énoncé avait été formulé différemment, en autant que les ressources cognitives à mobiliser par le problème soient maîtrisées par l'élève en question. Le type d'énoncé de problème est donc également un facteur à considérer lorsque vient le temps d'attester de la compétence d'un élève en résolution de problèmes à partir de problèmes présentés sous forme écrite.

Conclusion

Le choix des situations problèmes doit donc refléter cette préoccupation majeure de l'évaluation de la compétence. Avant même de s'engager dans la création de situations problèmes propices à l'évaluation de la compétence à résoudre des problèmes mathématiques, il faut examiner de près les situations dans lesquelles les élèves auront à mobiliser et utiliser à bon escient leurs ressources cognitives (savoirs,

savoir-faire et stratégies). En plus d'être complexes et dotées d'une certaine forme de réalisme (authenticité), ces situations doivent véhiculer une solution qui n'apparaît pas évidente pour les élèves tout en représentant un défi raisonnable. Le niveau d'habileté lié à la maîtrise des ressources cognitives des élèves s'avère un moyen intéressant pour nous permettre de présenter aux élèves des situations problèmes qui constituent de vrais problèmes et pour lesquelles ils ont la capacité d'agir sur la situation. L'apport de la TRI et des TIC dans l'estimation du niveau d'habileté en mathématiques des élèves par l'entremise du testing adaptatif par ordinateur se présente comme un moyen efficace pour obtenir des estimés d'habileté reflétant précisément le niveau de maîtrise des ressources cognitives des élèves (Boiteau, 2004). Évidemment, pour atteindre la perspective de l'évaluation visant l'aide à l'apprentissage, la structure et le contenu des tâches évaluatives incluses dans un TAO doivent être élaborés en fonction de principes didactiques bien définis. Si l'on souhaite accompagner l'estimation du niveau d'habileté d'informations substantielles pouvant témoigner des forces et des faiblesses des élèves, il faut que la structure et le contenu des tâches véhiculent cette information. Les tâches doivent correspondre à ce que l'on cherche à évaluer si l'on souhaite formuler des inférences de qualité concernant la progression dans l'apprentissage des élèves.

Toujours dans un esprit de progression de l'apprentissage, même si l'élève faible n'est pas confronté au départ à des situations problèmes contextualisées, il demeure qu'il aura du moins la capacité d'agir sur celles qu'on lui demandera de résoudre. La réalisation de ces situations problèmes pourra lui donner l'occasion de structurer de nouveaux savoirs, de consolider ceux déjà acquis pour ainsi progresser dans ses apprentissages et parvenir à améliorer son niveau de maîtrise des ressources cognitives et ainsi, résoudre des situations problèmes plus contextualisées et plus complexes. En fait, il aura l'occasion d'apprendre à résoudre des problèmes mathématiques.

Références

Andre, T. (1986). Problem solving and education. In G. D. Phye & T. Andre (Eds.), *Cognitive classroom learning: Understanding, thinking, and problem solving*, 169-204. Toronto: Academic Press Inc.

Bélanger, N., Gauthier, C. et Tardif, M. (1993). *Évolution des programmes de mathématiques: de 1861 à nos jours*. Sainte-Foy: Université Laval.

Bock, R. D., & Zimowski, M. F. (1998). *Feasibility studies of two-stage testing in large-scale educational assessment : Implications for NAEP*. Chicago : National Opinion Research Center (NORC).

Boiteau, N. (2004). *Comparaison d'une modalité optimale de testing à deux niveaux et de la modalité de testing adaptatif par ordinateur*. Thèse de doctorat inédite, Université Laval, Québec, Canada.

Cummins, D. D., Kintsch, W., Reusser, K. et Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*(4), 405-438.

Hambleton, R. K., Jones, R. W., & Rogers H. J. (1990). *Influence of item parameter errors in test development*. Paper presented at the Annual Meeting of APA, Boston.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory : Principles and applications*. Boston : Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA : Sage.

Hayes, J. R. (1981). *The complete problem solver*. Philadelphia: Franklin Institute Press.

Hegarty, M. et Mayer, R. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology, 84*(1), 76.

Hinsley, D., Hayes, J. R. et Simon, H. A. (1977). From words to equations. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension*, 89-106. Hillsdale, NJ: Erlbaum.

Hulin, C. L ., Drasgow, F., & Parsons, C. K. (1983). *Item response theory : Application to psychological measurement*. Homewood, IL : Dow Jones-Irwin.

Janvier, C. (1987). Representation and Understanding: The Notion of Function as an Example. In C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics*, 67-71. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Kim, H., & Plake, B. S. (1993). *Monte carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ : Lawrence Erlbaum Associates.

McBride, J. R. (1997). Technical perspective. Dans W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing : From inquiry to operation*, 29-44. Washington, DC : American Psychological Association.

Ministère de l'Éducation, du Loisir et du Sport (2001). *Programme de formation de l'école québécoise, enseignement primaire*. Québec, Gouvernement du Québec, 350 pages.

Ministère de l'Éducation, du Loisir et du Sport (2001). *Programme de formation de l'école québécoise, enseignement secondaire, premier cycle*. Québec, Gouvernement du Québec, 575 pages.

Moreau, S. et Coquin-Viennot, D. (2003). Comprehension of arithmetic word problems by fifth-grade pupils: Representations and selection of information. *British Journal of Educational Psychology, 73*(1), 109-121.

National Council of Teachers of Mathematics. (1980). An Agenda for Action: Recommendations for school mathematics of the 1980s. Reston, VA: NCTM.

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics: NCTM.

National Council of Teachers of Mathematics. (2000). Principles and Standards for School Mathematics. Reston, VA: NCTM.

Polya, G. (1945). *How to Solve it*. Princeton: Princeton University Press.

Sands, W. A., & Waters, B. K. (1997). Introduction to ASVAB and CAT. Dans W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing : From inquiry to operation*, 3-9. Washington, DC : American Psychological Association.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Montréal, Québec : Éditions du Renouveau Pédagogique.

Voyer, D. (2006). *L'influence de facteurs liés à l'élève ou à l'énoncé sur la compréhension en résolution de problèmes écrits d'arithmétique*. Thèse de doctorat inédite, Université Laval, Québec, Canada.

Wainer, H. (2000). *Computerized adaptive testing : A primer*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 6, 774-789.

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement : Conventional or adaptative?* (Research Report No. 73-1). Minneapolis : University of Minnesota, Department of Psychology, Psychometric Methods Program.

SYMPOSIUM 2: ETAT DES LIEUX ET SPECIFICITE DE L'EVALUATION ARTISTIQUE ET SPORTIVE AU 21ÈME siecle. QUELLES SONT LES NOUVELLES FORMES, MODELISATIONS ET PRATIQUES QUE LES DIDACTIQUES ARTISTIQUES ET SPORTIVES APPORTENT ?

Maugars Cédricia, Observatoire Musical Français, Paris Sorbonne

Résumé

La création et le mouvement sont des besoins essentiels à l'épanouissement et au développement de l'humain. Cependant, les programmes scolaires européens actuels donnent peu de temps à ces disciplines. Elles se trouvent souvent considérées comme « secondaires » ou optionnelles dans les curriculums. En dépit des écrits des spécialistes de l'Education, l'Ecole offre peu de solutions concrètes pour améliorer cette situation. Les savoirs artistiques peuvent-ils continuer à exister dans le cursus scolaire ?

Les innovations et les écrits en langue française sont rares concernant l'évaluation artistique et sportive. Si la problématique de l'évaluation est celle de la quête du sens, on pourrait se demander quel sens prend l'enseignement artistique et sportif au 21eme siècle et quels outils peuvent aider les enseignants et les apprenants dans le bilan de leurs progressions ou de leurs actions.

Pour la première fois, une équipe regroupant des chercheurs suisses et français partagent leurs résultats sur la thématique « évaluation et enseignements artistiques et sportifs ». Ce symposium réunira des spécialistes en Sciences de l'Education, Arts plastiques, Education musicale et Education physique et sportive qui ont souhaité confronter leurs travaux de recherches et expériences formatives.

Les communications seront regroupées autour des questions suivantes : quelles sont les outils en évaluation artistique et sportive ? Quels sont les enjeux et les implications de ces outils pour la pratique artistique et sportive ? Dans quel contexte évolutif s'inscrivent les innovations présentées ?

Tout d'abord, René Rickenmann proposera les résultats d'un dispositif de formation à l'« évaluation formative des élèves » en Arts plastiques comme instrument de conduite des activités d'enseignement/apprentissage. Adrián Cordoba transposera ce même dispositif de recherches aux rapports entre l' « évaluation formative des élèves » et sa mobilisation dans la conduite des activités d'enseignement/apprentissage en fonction des outils disciplinaire (sciences de l'éducation,; formations sportives ou Ecole des Beaux Arts). La recherche est centrée sur la transcription écrite des leçons d'arts plastiques ayant été vidéoscopées,
En EPS, le recours à un logiciel nommé *DartTrainer* ouvrirait des perspectives novatrices dans les pratiques d'évaluation formative notamment en impliquant de manière plus intense l'élève. Dans un contexte de formation, Pierre Monney et Benoît Gisler montreront l'impact de cet outil utilisé par les étudiants. Le feedback vidéo

immédiat a permis aux étudiants de formuler de manière précise une évaluation de leur prestation.
Dans le domaine musical, l'usage du portfolio permettrait aussi d'articuler les deux temps de l'évaluation (régulation et certification) tout en restant dans un paradigme compréhensif. Le process-folio et le portfolio seraient des moyens efficaces dans le cadre d'une évaluation formative d'une leçon d'instrument. Selon Pierre-Francois Coen, ils favoriseraient le passage d'une évaluation normative à une évaluation compréhensive en formation instrumentale.

L'usage de l'enregistrement audio dans l'évaluation des performances vocales des élèves en classe de musique permettrait-il d'aider les enseignants à s'ouvrir aux pratiques d'auto-évaluation ? A partir d'entretiens d'enseignants en musique, Cédricia Maugars décrit les représentations d'enseignants de musique en formation ayant pour la première fois utilisée cet outil.

Enfin, un dispositif pour des cours de piano en ligne est-il pertinent ? Quels sont réellement les avantages dans l'évaluation par l'exécutant de sa propre prestation musicale ? Quelles sont les possibilités réelles d'auto-évaluation pour l'apprenant ? Ce type d'enseignement via Internet est-il compatible avec une conception de l'évaluation hors de la mesure et du contrôle ? Telles sont les questions que Jacques Audran posera au cours de sa communication

ETAT DES LIEUX ET SPECIFICITE DE L'EVALUATION MUSICALE

Cedricia Maugars, Observatoire Musical Français, Paris Sorbonne

Résumé :

La recherche aborde la question de l'évaluation dans une perspective descriptive et interprétative des pratiques et des représentations d'enseignants en musique.

Les écrits spécifiques à l'évaluation musicale sont rares en langue française[1]. Nous avons transposé au domaine musical les concepts généraux de l'évaluation. Aussi, nous avons cherché le sens spécifique que l'évaluation pouvait prendre en Education musicale. Nous avons étudié les questions spécifiques de l'évaluation musicale dans le cadre des travaux anglo-saxons (Mills, 1991 ; Hunter, 1999 ; Hewitt, 2003).

Les questions de recherche de cette enquête sont centrées sur la construction du jugement des performances vocales. L'objectif de l'expérience est de décrire plus spécifiquement les procédures d'évaluation et de notation en classe de collège utilisées par des enseignants de musique.

Les questions de départ

- Comment les enseignants "construisent "-ils leurs jugements évaluatifs ?
- Les professeurs jugent-ils de la même façon leurs propres enregistrements et ceux de leurs collègues ?

Les objectifs opérationnels de cette enquête sont

- de créer des conditions d'évaluation proches de la réalité de terrain : l'évaluation collective du chant au collège, d'analyser les différentes formes d'évaluation pratiquées par les enseignants de musique et d'identifier comment ils notent.
- d'étudier les relations entre les pratiques réelles d'évaluation en classe et ce que les enseignants disent sur les activités musicales, de recueillir leurs

1 LE DIAGON-JACQUIN, L., L'évaluation en cours d'Education musicale, *L'éducation musicale*, Mai 1999./ TERRIEN, P., L'évaluation en question dans l'Education musicale, *L'éducation musicale*, sept-oct 2000./ BEGUE C., SISCA C., Comment sensibiliser les élèves à l'auto-évaluation dans les différentes activités du cours d'Education musicale en 6°?, Académie de Rouen, Mémoire de PLC2, Education Musicale, 1999/2000./ Le magazine *Infomusicollège* n°8, mai 2000.

appréciations et leurs préférences.

- d’examiner les changements de représentations pendant des interviews en miroir en relation avec des variables spécifiques des activités musicales en classe de collège; comparer ces appréciations et les classements au cours de deux "phases " avec l’évaluation du travail vocal en classe.

La méthodologie

Deux campagnes d’entretiens espacées de six mois ont permis de mieux analyser les déclarations, l’entretien s’est avéré être un outil incontournable parce qu'il oblige le stagiaire à verbaliser son comportement, à réfléchir sur sa propre attitude, à saisir la donnée émotionnelle. La difficulté de trouver des mots justes pour

traduire les pratiques d'évaluation en musique et le manque de recul par rapport à la réflexion pédagogique sont des raisons qui amènent à vérifier les données recueillies au cours d'entretiens approfondis.

Les résultats : l'évaluation musicale reste centrée sur la notation.

Dans le cadre d’une expérience sur l’enregistrement d’un chant Pygmée, les enseignants de musique interrogés ont eu une pratique d'évaluation sommative et pratiquaient la notation dans toutes les activités musicales en classe sans véritablement la remettre en cause. Les enseignants interrogés ont jugé de façon sommative, le "résultat final sonore" qu'ils justifiaient dans une notation chiffrée. L'expérience a montré que les enseignants de musique utilisaient différentes formes de notations (/20, /10, ou des "plus").

Nous avons découvert que des professeurs de musique ont des critères d'évaluations et des objectifs pédagogiques similaires. L'évaluation des enseignants de musique interrogés fut imprécise, souvent ils ne savaient pas expliquer quels sont leurs critères de notation. Cette notation semble arbitraire, subjective, s'appuyant sur une impression générale.

Au cours des enquêtes, deux profils d'évaluation basés sur un modèle sommatif/produit et formatif/processus sont apparus dans les groupes d'enseignants:

Les "profils centrés sur le produit final" ont un jugement évaluatif sévère. Ils prônent la valeur inaliénable de la notation exacte. Ces enseignants ont évalué la production musicale finale.

Les "profils centrés sur le processus" qui associent l'élève et le processus d'enseignement dans l'évaluation, sont rares. Le jugement est indulgent, ces enseignants sont sceptiques pour ce qui concerne à la notation exacte. Ils prennent en compte les attitudes, le comportement des élèves dans la notation du travail musical. Les critères d'évaluation du profil centré sur le processus correspondent à des objectifs plus généraux comme l’attitude des élèves, la spontanéité, le plaisir. Les progrès réalisés par les élèves : *"évaluer la progression des élèves"*.

A ces deux types dévaluation, plusieurs approches ont été identifiées dans l'élaboration du jugement en chant:

1- une approche holistique (adoptée surtout par les enseignants sans objectifs pédagogiques précis). Les enseignants utilisent des stratégies holistiques en privilégiant des appréciations globales : *"le caractère d’ensemble et l’impression générale","satisfaisant", "bon ensemble"*.

2- une approche critériée (plutôt utilisée par les enseignants qui ont réalisé une préparation du chant Pygmée en classe). Les critères techniques musicaux de "justesse de l’octave et de la quinte" et de "mise en place polyphonique" sont très précis dans les appréciations de ce groupe.

3- Une approche mixte dans les appréciations qui mêlent les critères d’évaluation : *"Beaucoup de précision rythmique et tout était cohérent, la partie soliste était très belle, c'était bien chanté"*.

L'écart entre les représentations "idéales" et "réelles" des pratiques ont révélé que les critères d'évaluation en chant privilégiaient la justesse et le respect de la précision du texte musical. Il est apparu dans les discours évaluatifs que les enseignants bien souvent jugent autre chose que la réalisation musicale en elle-

même : le comportement de l'élève, l'attitude en classe interfèrent dans les paramètres évaluatifs.

Cette recherche montre la nécessité de contextualiser et de multiplier les occasions d'évaluation musicale; une notation hors contexte ne signifie rien sans donner un sens

aux critères d'évaluation. Le jugement professoral se modifie selon les perceptions individuelles: la personne « évaluée » ou « évaluant » peut partager des observations identiques mais les jugements sont variés et les décisions finales sont rarement faciles et unanimes.

Les résultats de cette recherche ont mis en lumière certaines variables significatives dans le processus évaluatif en musique. A partir de la description des procédures d’évaluation sur les performances vocales utilisées par des enseignants de musique en deuxième année de formation à l’I.U.F.M., il ressort que les enseignants jugent de façon sommative en utilisant la notation.

QUELLE MODELE D'EVALUATION « EN LIGNE » AU SERVICE DE L'AMELIORATION DE L'EXPRESSION DE L'APPRENANT ?

Jacques Audran – jacques.audran@uha.fr
Laboratoire Interuniversitaire
des Sciences de l'Education et de la Communication (Lisec)
Cufef - Université de Haute-Alsace

Résumé :

Qu'il soit question d'évaluer un apprentissage ou d'offrir des possibilités d'auto-évaluation, l'évaluation des enseignements « en ligne » sur Internet semble presque toujours se centrer autour de la mesure (Audran, 2001, pp.156-175). Les possibilités de l'informatique en matière de métrologie y sont sans doute pour beaucoup. En conséquence, le modèle d'évaluation dominant semble de l'ordre d'une évaluation-contrôle qui ne s'applique pas toujours forcément de façon efficace aux enseignements qui visent à améliorer l'expression de l'apprenant, tout particulièrement lorsqu'il s'agit d'expression artistique.

Pour pointer de façon plus précise les nécessités et particularités de dispositifs « en ligne » d'évaluation de l'amélioration de l'expression, nous avons étudié, un peu en dehors des sentiers battus, les caractéristiques d'un enseignement visant l'amélioration de l'expression artistique sur Internet. Il s'agit de cours de piano destinés à des personnes désirant améliorer leur pratique pianistique sans pour autant s'inscrire à des cours particuliers jugés sans doute trop contraignants (Glickman, 2000). Ces cours viennent compléter des leçons diffusées par le canal d'une revue traditionnelle diffusée par la presse grand public.

Evaluation et pratiques « en ligne » sur internet :

La remise en cause de l'existence des pratiques d'évaluation « en ligne » sur Internet est liée la plupart du temps à la mise en cause de la légitimité de l'usage normatif des outils technologiques (par exemple, comment savoir, lors d'un examen ou un test effectué à distance si la personne qui opère le fait dans des conditions acceptables du point de vue de la validité de l'évaluation, c'est-à-dire sans aides ou ressources cachées ?).

Mais on peut estimer qu'un autre modèle de l'évaluation « en ligne » est peut-être plus légitime, c'est celui de l'évaluation comprise comme la prise de repères à des fins d'amélioration de la pratique (si test il y a, il est alors « formatif », générateur d'indicateurs à des fins d'autoévaluation, comme peut l'être par exemple un test de positionnement). Dans ce deuxième cas la question qui reste posée est : « Selon quelles modalités concrètes, grâce à quels critères, l'évaluation de pratiques sociales éducatives peut-elle s'opérer lorsque Internet constitue le théâtre principal de ces pratiques, ou intervient dans les dispositifs éducatifs ? ».

Un double rapport à l'évaluation

A ce niveau il faut bien distinguer, semble-t-il, deux types de rapports que peut entretenir Internet avec la sphère évaluative :

- soit, dans une posture objectivante et analytique de mesure, on confère aux systèmes informatiques le statut d'objets-témoins des pratiques humaines à évaluer. Il faut néanmoins s'interroger sur la pertinence des indications produites, en quoi elles nous renseignent sur les pratiques. Ainsi nous devons nous interroger sur ce que nous disent vraiment les mesures de réalisation ou de réussite à des items, les calculs délivrant les pourcentages d'activités réalisées, le comptage des clics de souris, ou du nombre de connexions. Le risque est bien sûr de surinterpréter cette mesure, d'aller au-delà de la signification qu'elle véhicule. Mais l'évaluateur peut trouver ces indicateurs suffisants (notamment du fait des recoupements possibles), même s'il ne dispose pas de toutes les informations métrologiques nécessaires, pour juger de l'avancée d'un travail sans aller y voir de trop près. L'évaluation sera ainsi centrée sur la performance ou l'efficacité d'un certain type d'activités jugées représentatives de pratiques de référence ou d'une typologie des usages (évaluation sur la base de la consultation d'un document croisée avec la réalisation d'un test, par exemple) ;

- une autre évaluation s'intéresse à la « profondeur » ou à l'effectivité des pratiques sociales dans différents contextes d'usage. L'usager est alors à la fois *évaluateur* et *évaluataire* et participe d'une activité évaluative « interne » socialement constituée[1] (Chardenet, 1999, pp.93-103). Le système informatique en réseau sert alors à la fois d'instrument d'évaluation pour les acteurs tout en participant de leurs pratiques, dans la mesure où il véhicule les interactions spécifiques à l'activité évaluative[2] (il peut s'agir par exemple d'un document à produire ou en groupe, à partager, à socialiser). Ce principe d'évaluation est de permettre à chaque acteur d'extraire une réalité sociale complexe, y compris subjective, où le système informatique permet d'observer les échanges et autorise du même coup, par la « capture » informatique, un effet de mémoire au bénéfice des participants (autoévaluation des productions, des interactions, portfolio électronique etc.). Le réseau Internet, en permettant de rendre compte avec une certaine intelligibilité des pratiques, en affichant et mémorisant des traces verbales, (production des acteurs, dialogues sur les forums témoignant de situations, de postures, d'orientations choisies, de points de vues défendus), entre dans une logique évaluative plus formative et qualitative.

Dans les deux cas l'évaluation porte sur des produits de l'activité et non sur l'activité elle-même, ce qui est la plupart du temps le cas en évaluation. Dans le premier cas c'est le système qui produit des indicateurs déjà pré-interprétés (dont la lecture métrologique, par un observateur souvent extérieur, est alors quasi-immédiate) ; dans

[1] Chardenet (1999, p.25) précise « L'évaluation peut être considérée comme une activité sémiotique dont les actes se traduisent par des formes verbales particulières et intentionnelles. ». Chardenet situe l'évaluation au sein des pratiques langagières (Chardenet, 1999, pp.93-103).

[2] L'évaluation n'est pas « toute » l'activité de communication, mais une communication particulière.

le second cas les productions des acteurs sont observées sans système normatif d'interprétation et les possibilités d'évaluation dépendront alors de la précision des consignes et des outils d'analyse mobilisables. Compte tenu du primat de l'univers technique, toute centration excessive sur le produit évalué peut entraîner une certaine perte de vue de la dimension politique du référentiel autorisant l'évaluation. Mais même si analyser n'est pas évaluer, l'analyse constitue une démarche importante dans la construction d'un référentiel de pratiques.

C'est dans cette perspective qu'il s'agit de voir si un environnement dédié au perfectionnement de l'expression musicale peut tirer profit de la dimension formative d'Internet.

L'évaluation dans les cours d'expression pianistique sur Internet

Dans le cadre de cet étude qui ne prétend pas cerner toutes les possibilités offertes par la toile, six cours[3] diffusés « en ligne » via Internet ont été étudiés d'un point de vue ergonomique et sémiotique. Il s'agit donc d'une étude de cas qui ne peut être en aucune façon généralisée, mais qui peut apporter un éclairage sur les représentations de l'acte évaluatif chez les concepteurs de dispositifs de formation en ligne sur Internet.

L'environnement des cours en ligne

Les cours en ligne étudiés et se présentent sous la forme d'écrans de type Web (Figure 1). Le recours à l'image animée y est fréquent au moyen de technologies « macromédia flash ». Une fois passée une page d'accueil animée, l'internaute entre sur le site du magazine. Un « espace abonné » permet d'accéder aux cours. Le site Web est donc partagé entre une partie à l'accès public et une partie privée réservée aux abonnés. C'est la partie privée réservée aux abonnés qui nous intéresse ici. L'internaute est amené à choisir entre une connexion haut-débit et une collection bas-débit. La première permet de bénéficier d'un accès à des fichiers multimédias correspondant aux différents cours. Chaque cours correspond à un numéro du magazine.

[3] Il s'agit des cours 34 à 39 proposés par pianiste magazine (www.magazine.com) entre le mois d'octobre 2005 et le mois de juin 2006. D'autres cours du même type pourraient être analysés (http://www.tousaupiano.com)

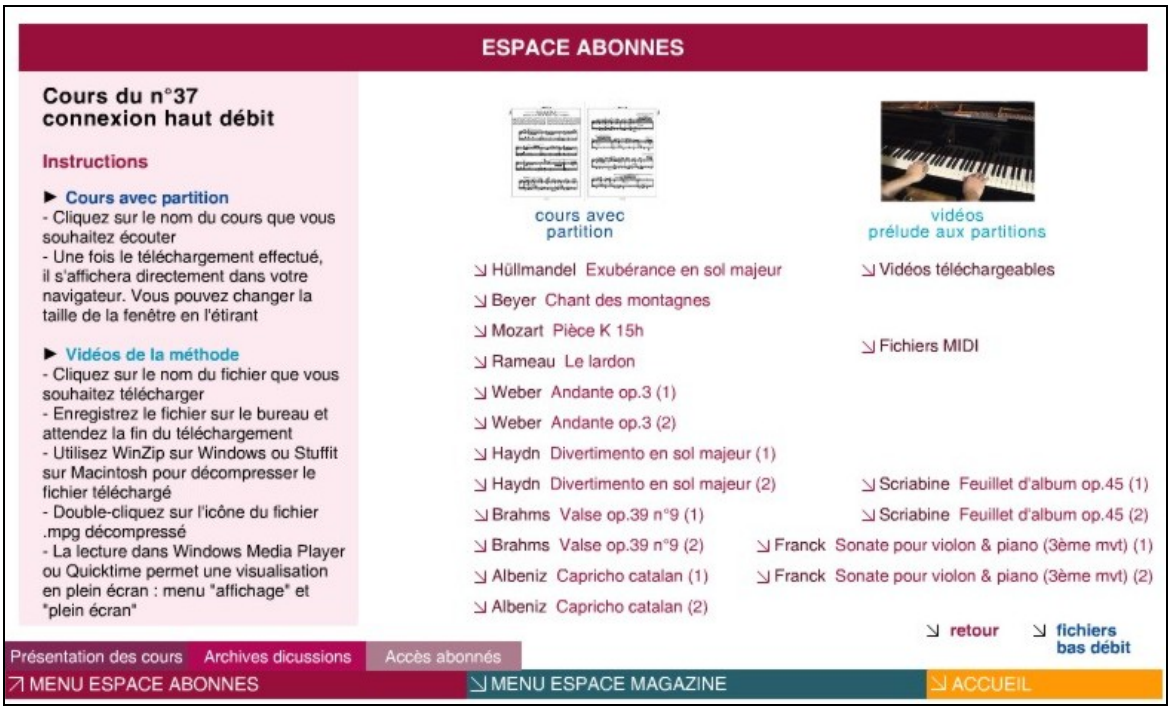


Figure 1 – Ecran d'accueil de l'environnement en ligne

Un cours est un ensemble de morceaux (Figure 1) dont la partition est téléchargeable sous la forme de fichiers « Macromédia flash », de vidéos courtes montrant la disposition des mains sur le clavier et de fichiers midi à télécharger.

Le cours en flash

La pièce maîtresse du cours est constituée par les fichiers Macromédia flash qui correspondent aux morceaux présentés dans le magazine. Les fichiers vidéo et les fichiers midi sont plutôt des pièces complémentaires au cours en flash.

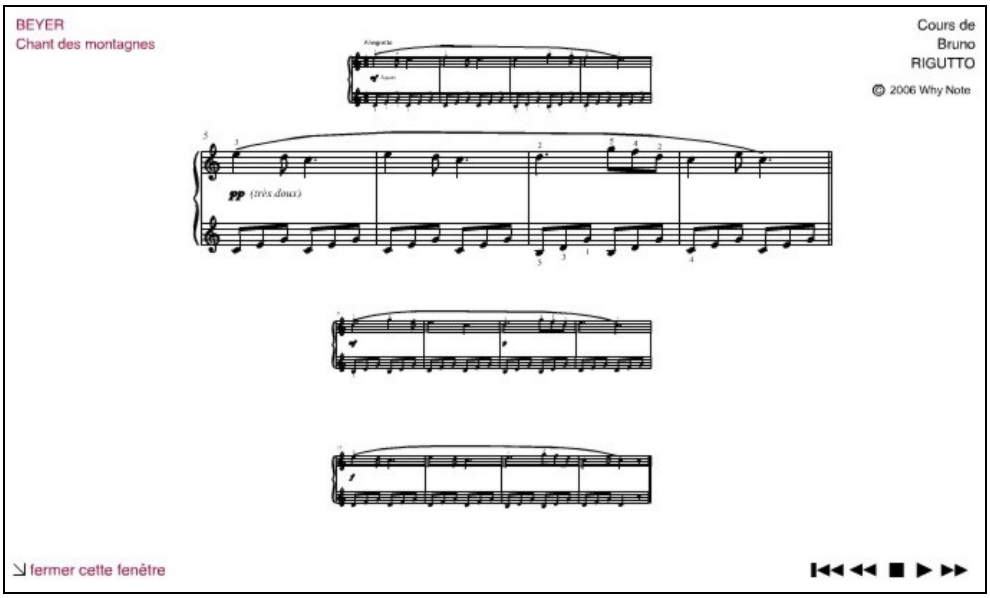


Figure 2 – Ecran d'étude et d'audition de commentaires

Un cours est donc une partition (Figure 2) commentée sous forme sonore par un enseignant qui s'adresserait à un élève virtuel. Le professeur donne les indications nécessaires à l'exécution de morceaux sans jamais s'interrompre (mais aussi dans le même cours sonore des indications historiques, des commentaires personnels). L'internaute à la possibilité de revenir en arrière dans cette présentation en utilisant des touches de type magnétoscope (Figure 2). L'internaute manque malheureusement de repères pour naviguer dans le morceau. Il aurait pourtant été facile d'intégrer un certain nombre de pointeurs permettant de se repérer sur la partition. Le grossissement de la ligne commentée est très utile pour se repérer dans la partition.

Sur le plan sonore, comme dans un cours traditionnel, la pertinence des conseils varie d'un enseignant à l'autre. Les commentaires du maître font parfois sourire : « je suis sûr que si on mettait cette musique dans une boîte à musique pour s'endormir ce serait merveilleux » (espérons que ce ne sera pas le cas des les élèves en ligne), et cultivent un petit côté « radio scolaire » d'autrefois. Les conseils sont souvent extrêmement techniques et pointus, à des recommandations moins précises mais plus centrées sur l'expression. Il est par ailleurs impossible d'obtenir des conseils supplémentaires, aucun programme de communication ne permettant d'interroger l'enseignant.

- sur le plan de l'évaluation cybernétique, ce cours ne laisse donc aucune place à l'interaction médiatisée. Les vidéos et les fichiers midi reprennent le même principe : l'internaute doit les télécharger et les visionner sur son ordinateur. Rien n'est prévu sur le site pour déposer des productions personnelles (des fichiers midi ou mp3 réalisés par l'étudiant, par exemple) qui pourraient être comparées à des documents de

référence en vue de réaliser une évaluation. Il n'est pas prévu non plus de tests d'exécution des passages difficiles.

- sur le plan de l'évaluation socialisée on le trouve pas non plus d'outils de communication qui permettrait décharger entre pairs des conseils des trucs et astuces de morceaux de fichiers, toutes sortes d'objets qui permettraient de perfectionner une pratique. Tout au plus peut-on noter qu'une discussion de type « chat » est organisée une fois par mois avec l'enseignant sur la partie publique du site. Alors que les forums et les blogs se développent un peu partout sur Internet, ce cours reste centré sur la mise en ligne de contenus.

Alors que le cours en lui-même utilise des technologies assez sophistiquées sensées être au service de l'expression pianistique de l'élève en ligne, il apparaît que l'évaluation n'est finalement présente qu'à travers un système de guidage procédural et assez statique, somme toute peu présent si on le compare au soin apporté aux techniques à orientation transmissive principalement centrées sur les conseils d'exécution de pianistes chevronnés.

Au-delà de ces critiques, la réalisation informatique proprement dite du site est irréprochable. Des technologies récentes sont à l'oeuvre, la qualité visuelle et sonore des cours en flash. Mais on a impression que la représentation transmissive d'un cours de piano est difficile à dépasser.

L'absence d'évaluation, témoin d'une carence didactique ?

On note donc le faible souci de donner les moyens à l'apprenant de repérer les indices de ses propres progrès, d'échanger des pratiques d'expérience avec des pairs, de « publier » des phrases musicales personnelles à des fins de critique toutes choses qui constitueraient la dimension anthropologique de l'affaire (Ravestein, sous-presse) et que les technologies semblent pouvoir rendre possible aujourd'hui.

Alors qu'il semblait que l'évaluation par l'exécutant de sa propre prestation musicale devait être au centre de la préoccupation des concepteurs de ce complément numérique, il apparaît que l'évaluation et *a fortiori* l'auto-évaluation, passent largement au second plan, un peu comme si le guidage de l'apprenant grâce aux conseils du professeur se suffisait à lui-même. Cet état de fait révèle à la fois la reproduction technologique d'une pédagogie fortement marquée par la tradition, et confirme que l'imaginaire informatique reste très centré sur les aspects de diffusion technologiques des savoirs, même lorsqu'il s'agit d'expression musicale instrumentale.

À défaut de viser le perfectionnement des pratiques grâce à l'interaction entre pairs, le modèle qui sous-tend ces cours est réalisé dans une perspective typiquement transmissible fondée sur l'exemple et s'inscrit dans une logique que j'appellerai « marchande » dans la mesure où il est question que de délivrer des contenus informatifs. Ce modèle de l'information marchande se développe sur la toile depuis 1990 (Dimech, Turpin, 1998, p.26) et son importance s'est considérablement accrue durant la dernière décennie. Il faut noter que l'influence du modèle marketing n'est pas propre à Internet et touche toutes les sphères sociales y compris celle de

l'éducation (Gaziel, Warnet, 1998) et que ce modèle cohabite sur Internet avec les formes d'expression les plus diverses. Il ne représente pas non plus un modèle viable pour penser l'ensemble des enjeux éducatifs de l'Internet où l'infrastructure pèse lourdement sur les investissements (Dimech, Turpin, 1998, pp.23-66). Ce modèle opère selon le modèle du *pull* (l'usager va rechercher l'information dont il a besoin, et seulement celle-là). L'évaluation dans ce cas sert un système « offre-demande » qui, ayant détecté un certain nombre de tendances, vise à les renforcer encore dans un esprit proche des formes de l'enseignement programmé.

Conclusion

Pour ne pas rester sur le constat pessimiste on peut observer qu'un certain nombre d'initiatives prennent forme sur la toile. Bizarrement, le plus souvent ce ne sont pas des offres strictement commerciales. L'exemple de Piano Web[4] en est une illustration. Ces espaces d'échange comprenant à la fois des cours, un forum, offrant un ensemble de services aux musiciens amateurs (à ce stade, on ne peut plus les appeler sites artisanaux) ne survivent sans doute que grâce à l'énergie des passionnés qui les animent et à la publicité qui accompagne leurs écrans. Internet deviendra-t-il un instrument d'évaluation des pratiques entre pairs et, peut-être permettra-t-il l'élaboration de processus d'auto-questionnement des musiciens sur leurs pratiques, grâce à des formes électroniques de confrontations et construction sociales (qu'il reste toutefois à identifier plus précisément) ? Encore faut-il que les dispositifs conçus s'y prêtent et il semble qu'on ne le constate pas souvent actuellement.

Références

Audran, J. (2001). *Influences réciproques relatives à l'usage des NTIC par les acteurs de l'école*. Thèse de doctorat, Université de Provence.

Blumer, H. (1969). *Symbolic Interactionism : perspective and method*. Englewood Cliffs: Prentice-Hall

Chardenet, P. (1999). *De l'activité évaluative à l'acte d'évaluation*. Paris : L'Harmattan.

Clot, Y. (1999). *La fonction psychologique du travail*. Paris : PUF.

Dimech, P., Turpin, E. (1998). « Le réseau Internet est-il économiquement viable ? », *Internet une nouvelle donne économique, Communications et stratégies n°32*, Montpellier, IDATE, 23-66.

Gaziel, H., Warnet, M. (1998). *Le facteur qualité dans l'école de l'an deux-mille*. Paris : PUF.

Glickman, V. (2000). *Des cours par correspondance au e-learning*. Paris : PUF.

Ravestein, J. (sous-presse). *Contribution à une anthropologie didactique des technologies de l'information et de la communication en éducation*. HDR. Université de Provence.

4 A titre d'exemple, http://pianoweb.free.fr/ est un site semi-commercial très complet. De son côté, www.pianonanny.com/ est un site qui propose un parcours gratuit avec un astucieux carnet de notes.

L'ANALYSE DES PRATIQUES ET LE PILOTAGE DES CONDUITES D'ENSEIGNEMENT : UNE ETUDE COMPARATIVE DES DISPOSITIFS D'EVALUATION EN RECHERCHE ET EN FORMATION EN DIDACTIQUES DE L'EDUCATION PHYSIQUE ET DES ARTS PLASTIQUES

Adrián Cordoba , René Rickenmann, Université de Genève

Mots-clés : autoconfrontation, gestes professionnels, évaluation formative

Résumé :

L'évaluation constitue une véritable problématique dans le cadre de l'analyse des pratiques visant la formation aux gestes professionnels de conduite des séquences d'enseignement/apprentissage dans le champ des didactiques des arts plastiques et/ou de l'éducation physique. Elle tient autant des conceptions générales du rôle de l'école dans le développement des enfants que des aspects pragmatiques liés aux significations et à la « manière » de conduire les activités d'enseignement/apprentissage.

Notre travail analyse un dispositif de recherche sur les rapports entre les éléments d'intelligibilité de la pratique enseignante qui émergent à travers un dispositif d'«*analyse et d'évaluation des leçons* », d'une part, et leur mobilisation (ou leur non mobilisation) dans la *conduite des leçons* par les enseignants titulaires (généralistes) et des enseignants spécialistes amenés à co-gérer une séquence d'enseignement, d'autre part.

Les données sont issues de deux mises en œuvre de ce dispositif auprès de deux situations contrastées :

- Dispositif appliqué à la formation initiale des enseignants dans le cadre de la gestion en co-responsabilité (par des stagiaires) d'une séquence didactique en arts plastiques ;
- Dispositif appliqué à une recherche sur la gestion en co-responsabilité (enseignants généraliste et spécialiste) d'une séquence d'enseignement en éducation physique.

Du point de vue méthodologique, l'analyse est centrée sur la transcription écrite des leçons d'une séquence d'enseignement gérée en co-responsabilité par les deux types d'enseignants, ayant été vidéoscopées. Nos catégories d'analyse sont issues du modèle descriptif de l'action enseignante proposé par Mercier, Sensevy & Schubauer-Leoni (2000), ainsi que des concepts de l'action enseignante proposés dans la théorie des situations didactiques (Brousseau, 1996). Du point de vue méthodologique, le dispositif s'inspire des propositions sur l'autoconfrontation croisée dans le champ de l'analyse du travail (par exemple, Clot & Faïta, 2000), mais avec les adaptations mises en oeuvre par plusieurs chercheurs dans le champ de la formation initiale et continue des enseignants (Goigoux, sous presse ; Lenoir & *ali*, 2002).

Introduction

Dans ce texte, nous souhaitons explorer les liens qui peuvent être établis entre l'évaluation formative en tant qu'outil institutionnel de régulation des processus d'apprentissage et des concepts didactiques (dévolution, régulation enseignante) tels abordés par la *théorie des situations didactiques* (Brousseau, 1996) ou la *théorie anthropologique du didactique* (Chevallard, 1999, 1995). L'objectif principal est celui d'interroger les rapports qui peuvent être établis entre les modèles, les concepts et les méthodologies de la recherche en didactiques, d'une part, et les démarches et dispositifs de formation des enseignants, d'autre part.

La question centrale est celle de savoir à quelles conditions et avec quels résultats les concepts et modèles construits dans un but de recherche descriptive du fonctionnement didactique peuvent-ils être mobilisés dans les dispositifs de formation d'enseignants, tout en évitant le piège de l'« applicationnisme » pédagogique et didactique.

Cette étude se construit dans la conjonction des situations didactiques en tant que systèmes articulés d'enseignement/apprentissages -entrée par l'analyse des pratiques et les théories de l'activité humaine- et, parallèlement, elle s'élabore sur la base des gestes professionnels de programmation, de gestion et d'évaluation de l'enseignant -l'activité enseignante comme analyseur des situations didactiques-.

De fait, de nombreux développements sont issus des didactiques disciplinaires telles que celle des mathématiques ou celle des langues, qui de par leur reconnaissance institutionnelle, ont connu une forte croissance depuis le milieu des années 1980. Ces travaux envisagent l'analyse des pratiques enseignantes dans une perspective de didactiques comparées (cf. Revue Française de pédagogie, 2002), dans le but de cerner la généricité de certains aspects liés à l'enseignement dans une discipline. Moins nombreux sont les travaux issus de didactiques disciplinaires dont la participation aux démarches de l'analyse descriptive du didactique ordinaire est plus récente, telles que la didactique des arts plastiques et/ou celle de l'éducation physique. Ces dernières développent actuellement des démarches d'*ingénierie* intégrées aux démarches de recherche (Brousseau, 1996) et font de l'*activité enseignante* un *analyseur* des systèmes didactiques ordinaires.

Problématique

Diverses modélisations du fonctionnement des systèmes didactiques ont émergé avec le passage, opéré durant les années 1980, d'une didactique prescriptive vers une didactique descriptive des activités d'enseignement (Schubauer-Leoni, 1998), notamment celle de la *théorie des situations didactiques* (Brousseau, 1986) basée sur la théorie des jeux, celle de l'*approche anthropologique du didactique* (Chevallard, 1991) et celle plus récente, de la *théorie de l'action enseignante* (Mercier, Sensevy & Schubauer-Leoni, 2000). Ces approches ont en commun une conception systémique et dynamique du fonctionnement des systèmes didactiques et tentent de produire des modèles capables de décrire et d'articuler les dimensions *structurelle* et *fonctionnelle* de l'activité didactique.

Dans le cadre d'une analyse des situations éducatives, la démarche descriptive en didactiques cherche à analyser les systèmes formels d'enseignement/apprentissage en lien avec les pratiques effectives qui se déroulent dans les institutions.

Cette approche met essentiellement en avant la nature dynamique des dispositifs d'enseignement/ apprentissage en ce qu'elle postule, dans le cadre de toute activité didactique, la mise en place et transformations d'un *système de déterminations mutuelles* :

enseignant ←→ tâches d'enseignement/apprentissage ←→ groupe-classe/élève.

Une des caractéristiques centrales des travaux récents issus de ces approches est sans doute le développement d'une conception résolument collective et dynamique de l'*activité* (didactique). Dans cette conception, les situations didactiques sont considérées comme résultantes de l'activité conjointe enseignant←→ élève(s) (Leutenegger & Ligozat, 2006 ; Mili & Rickenmann, 2004 ; Rickenmann, 2006a). Cependant, il ne s'agit pas de considérer que ces agents partagent les mêmes buts, mais de se donner les moyens de penser l'*interdépendance de leurs actions* à l'intérieur d'un système dans lequel,

- l'action de l'élève est en partie *contrainte* par les choix de l'enseignant (choix des contenus d'enseignement, du type de tâche, du milieu didactique, de l'organisation sociale, etc…) ;
- l'action de l'enseignant est *orientée* en fonction des conduites potentielles et effectives de ses élèves.

Cette conception systémique et dynamique de l'*activité conjointe*, peut être décrite à partir de trois grandes familles de processus mis en évidence par Mercier, Sensevy & Schubauer-Leoni (2000), qui sont à l'œuvre dans les systèmes didactiques ordinaires :

- Processus mésogénétiques : relatifs à la manière dont l'enseignant aménage un milieu d'apprentissages avec lequel les élèves interagissent et qu'ils font évoluer ;
- Processus topogénétiques : relatifs aux postures et rapports aux savoirs de l'enseignant et des élèves et leur évolution durant l'activité ;
- Processus chronogénétiques : relatifs à l'avancement du temps didactique et aux rapports entre les temps d'enseignement et ceux de l'apprentissage.

Du point de vue de la recherche, les concepts de *meso, topo et chrono*-genèse présentent l'avantage de produire des descriptions dynamiques des phénomènes didactiques observables, c'est-à-dire, qui articulent l'« état » d'un élément du système didactique (enseignant, élèves, objets de savoir) par rapport aux autres, à un certain moment du déroulement de la tâche d'enseignement/apprentissage.

Comme nous tenterons de le montrer dans ce texte, sous couvert d'un *dispositif d'évaluation formative*, ces concepts pourraient également constituer un apport déterminant du point de vue de la formation aux gestes professionnels d'enseignement.

L'étude descriptive de l'activité d'enseignement

La recherche sur les gestes professionnels de l'enseignant dans la perspective d'une didactique *descriptive* plutôt que prescriptive est relativement récente (Calderhead, 1995 ; Chevallard, 1995 ; Goigoux, 2000 ; Mercier, Sensevy & Schubauer-Leoni, 2001 ; Portugais, 1992). Elle constitue le prolongement d'une longue tradition de recherches en pédagogie expérimentale, aussi bien aux EEUU (Biddle, Good & Goodson, 1997) qu'en Europe (Crahay & Lafontaine, 1989).

En sciences de l'éducation, ce sont essentiellement trois entrées qui abordent actuellement l'articulation des déterminations mutuelles des composantes des systèmes didactiques à partir de la notion d'*activité* -didactique-.

Une première entrée, *praxéologique*, propose un modèle théorique qui articule la tâche comme activité aux gestes afférents que les agents effectuent pour l'accomplir (Chevallard, 1995). Cette entrée constitue le développement du projet scientifique de fonder les didactiques disciplinaires, principalement la didactique des mathématiques, autour d'une réflexion sur les *conditions de possibilité* de la transmission des savoirs, en ce que ces derniers constituent des « objets vivants » évoluant à la fois à l'« extérieur », dans la culture et à l' « intérieur » de l'institution scolaire (Brousseau, 1986).

Une deuxième entrée, celle des *communautés de pratiques* (Lave & Wenger, 1991), propose une approche environnementaliste de la production collective et distribuée des savoirs. Moins sensible à la spécificité des objets de savoirs que la première, cette entrée fournit cependant un modèle descriptif des processus socio-cognitifs centré sur les interactions, négociations et régulations de l'action entre agents au sein des situations d'enseignement/apprentissage. Dans cette entrée, ces dernières ont tendance à être considérées comme une version « micro » des situations socioculturelles -adidactiques- de production et de circulation des savoirs.

Empruntant la perspective historico-culturelle de Vygotski, les travaux d'une troisième entrée, *sémiopragmatique* (Sensevy, 2001 ; Mili & Rickenmann, 2004 ; Rickenmann, 2001 et 2005), tentent d'articuler les apports des deux premières à partir de l'étude de *l'objet enseigné* d'une part et, d'autre part *des* gestes d'enseignement, en suivant l'évolution systémique et dynamique des dispositifs didactiques. Sensible au caractère propre de chaque discipline d'enseignement[1], cette entrée prend en compte les phénomènes liés à la transposition interne des savoirs dans les situations didactiques. Sensible, aussi, à la nature co-construite, distribuée et conjointe des processus cognitifs, cette entrée considère également les *formats didactiques*, c'est-à-dire les types et modalités d'organisation des tâches d'apprentissage propres à l'institution scolaire et les effets qu'ils ont sur l'activité des agents -enseignant/élève-s-, sur la construction des objets effectivement enseignés (Wirthner & Schneuwly, 2004).

[1] Pour une revue analytique sur l'historique et constitution des différentes démarches adoptées par les didactiques disciplinaires cf. Bronckart, 2005 ; Sieber-Thévenaz, 2005.

Les travaux centrés sur les deux premières entrées ont permis de développer, particulièrement, les connaissances relatives aux composantes savoir(s)/tâches d'enseignement et sur les processus d'apprentissage. En ce qui concerne l'action enseignante, ce sont les travaux de la troisième entrée qui abordent la description des systèmes didactiques « ordinaires » à partir d'une étude des gestes de l'enseignant en situation d'enseignement effectif (Dévolvé, 2000 ; Mercier, Sensevy & Schubauer-Leoni, 2000 ; Salin, 1997).

L'action enseignante comme analyseur des situations didactiques

Dans le cadre de cette démarche, les didactiques AP et EP ont recours aux conceptualisations et démarches méthodologiques sur l'action enseignante produites, notamment, par les travaux en didactiques des mathématiques et des langues.

Ces travaux proposent, d'une part, des catégories fonctionnelles de l'action enseignante, relatives à son rôle fondamental de médiateur des processus d'enseignement/apprentissage : *définir, dévoluer, réguler et instituer* se présentent ainsi en tant que catégories favorables à l'identification des différentes *actions* de l'enseignant et des *fonctions* qui s'y rapportent. Ces catégories facilitent ainsi le découpage séquentiel du continuum de l'activité enseignante, ainsi que son analyse en fonction du caractère finalisé -transmettre des savoirs-, de la situation didactique. Articulées aux concepts de mésogenèse, de topogenèse et de chronogenèse, ces catégories permettent également d'analyser le rôle, les fonctions et effets de l'action enseignante aux prises avec le système didactique effectivement à l'œuvre et puis, son évolution.

La focalisation sur l'enseignant pour l'étude des systèmes didactiques, en particulier en ce qui concerne la gestion de l'activité, les décisions qu'il prend tout au long de l'évolution de la tâche et les différents *topos* qu'il assume, rend compte de l'influence mutuelle entre les types de savoirs, les formats d'activité didactique et l'agir interdépendant de l'enseignant et des élèves.

Méthodes cliniques d'étude de l'action enseignante

Parmi les méthodes auxquelles ont recours les chercheurs pour l'analyse de l'action enseignante, le recours aux techniques d'*autoscopie* ont souvent permis de mettre en évidence les dimensions de l'activité didactique liées à sa mise en œuvre effective. Issues, notamment, des techniques d'autoconfrontation croisée de l'analyse du travail (Clot & Faïta, 2000), différents dispositifs d'autoscopie ont actuellement cours dans les travaux de recherche, le plus souvent complémentaires des dispositifs plus classiques de l'analyse didactique tels que la vidéoscopie des séquences d'enseignement et leur mise en protocole écrit.

Premier dispositif (EP) : Autoconfrontations croisées, recherche et régulation

À travers le premier dispositif que nous allons vous présenter, nous souhaitons montrer quelques moments d'un processus conjoint d'auto-évaluation des enseignants. Pour une question de place, nous nous centrerons ici sur l'évaluation globale qu'ils effectuent de l'ensemble de l'unité d'enseignement-apprentissage sur le basket-ball (5P), démarche qui les amène à évoquer les apprentissages des élèves et à questionner la pertinence de leur

programmation en analysant le travail de *construction et de mobilisation des savoirs* effectué par les élèves (des petits jeux à la situation de match) et les *moments de verbalisation* sur lesquels repose une partie de la programmation.

Cette étude fait partie d'une recherche en cours qui a comme but la description et la compréhension de l'action didactique des maîtres dans l'enseignement de l'éducation physique à Genève. Les caractéristiques particulières du fonctionnement de cette discipline dans ce canton nous ont amenés à nous intéresser à une double perspective d'investigation. D'une part, la relation didactique qu'entretiennent le maître et le(s) élève(s) au cours des séances ordinaires (Leutenegger 1999), ce qui nous a conduits à observer et à filmer plusieurs leçons en salle. Nos préoccupations rejoignent ici le questionnement développé en France depuis une vingtaine d'années à propos de l'action de l'enseignant (Amade-Escot et al., 1998) et tout particulièrement des travaux récents en didactique de l'éducation physique (Amade-Escot, 2002, 2003, Loquet et al., 2002, Thepaut, 2003).

D'autre part, nous nous intéressons à la relation didactique que construisent le maître spécialiste (MS) et le maître généraliste (MG) dans l'enseignement conjoint et en alternance[2] (co-responsabilité) de l'EP et à la progression du savoir au sein de l'unité d'enseignement/apprentissage.

Pour tenter de mieux comprendre l'action de ces enseignants, nous avons adapté un outillage méthodologique utilisé avec d'autres catégories professionnelles et sensé « créer les conditions d'un rapport dialogique » (Faita et Viera, 2003) entre les deux acteurs. En effet, à l'issue des séances filmées, les deux enseignants concernés ont participé à un entretien d'autoconfrontation croisée. C'est à partir de cet entretien, et par le regard posé par chaque maître sur sa propre pratique et sur celle de son collègue et, concurremment, sur leur participation/articulation aux apprentissages des élèves, que nous tentons de comprendre la dynamique des acteurs et leur action dans le système didactique.

Au niveau méthodologique, soulignons simplement que les séquences visionnées lors de l'entretien ont été choisies par le chercheur, en assumant ainsi la responsabilité d'extraire des séances les dimensions pertinentes de l'action des enseignants[3], et préalablement visionnées par les maîtres chez eux[4].

---

[2] A l'école primaire Genevoise, le maître spécialiste a la fonction de « méthodologue ». En tant que tel, il a la responsabilité de la programmation annuelle et une place centrale dans le choix des contenus et des tâches à aborder avec les élèves. C'est lui le responsable de la 1ère séance de l'unité d'enseignement/apprentissage. Ensuite, le MG est sensé poursuivre le travail du spécialiste durant deux séances, avant que ce dernier revienne dans la classe pour donner une autre leçon. En moyenne, les unités sont composées de 6-7 séances (2 pour le MS, 4-5 pour le MG).

[3] C'est au « chercheur, partiellement, parfois totalement, d'envisager le choix de moments et séquences de l'activité en fonction de cette activité – plus exactement de la perception qu'il en a – plus que de la représentation discursive que le collectif est disposé à en fournir. » (Faita & Viera, op. cit. ; Rickenmann, 2006b)

[4] Quinze jours avant l'entretien, nous avons donné aux maîtres un DVD avec les différentes séquences. Sur papier, nous leur avons fourni une explication du choix effectué ainsi que 4 points pouvant être pris en considération lors du visionnement (1. Selon moi, de

De même, et cela constitue une des particularités méthodologiques de la recherche, la structuration temporelle des leçons a été modifiée en plaçant *au début* du DVD la situation de jeu réel (basket dans tout le terrain), situation qui clôturait la dernière séance de chaque maître. Notre choix de confronter en premier lieu les enseignants au visionnement de la tâche finale, qu'ils considèrent comme étant la situation de jeu la plus complexe et potentiellement mobilisatrice des savoirs abordés précédemment, a produit des effets, comme nous allons le voir plus bas, dans la réflexion et le discours des maîtres.

Les autres tâches du support respectent la chronologie des séances et ont été retenues au regard de trois critères : a- un lien *évident* avec les contenus propres au basket (nous n'avons pas retenu, par exemple, les moments d'échauffement sans ballon ou le travail d'adresse individuelle), b- des moments pouvant être considérés comme centraux dans la démarche d'enseignement-apprentissage des maîtres (ex. : les verbalisations), c- la présence de la même tâche chez les deux maîtres.

---

quelle manière sont organisées les différentes tâches et leçons de l'unité d'enseignement/apprentissage ? 2. Qu'est-ce que je peux dire par rapport à l'articulation de mon enseignement avec celui de ma collègue ? 3. Quelles questions est-ce que j'aimerais poser à ma collègue pour mieux comprendre son enseignement ? 4. Si je devais préparer un autre cycle de basket, comment m'y prendrais-je ?)

L'évaluation des apprentissages comme déclencheur de l'évaluation de l'enseignement

Le premier extrait que nous présentons provient du début de l'entretien. Le MS se positionne par rapport aux séquences choisies par le chercheur et formule une première évaluation de l'unité d'enseignement-apprentissage.

29- C : Est-ce que vous souhaitez dire quelque chose avant de passer au visionnement ? Toi, tu avais déjà réagi par rapport au choix…

30- MS : Comme tu avais mis le match d'abord

31- C : oui (en accompagnement du discours)

32- MS : et après on voit les séquences, je me disais, du coup euh, en fait, j'ai fait une mise en place, donc quand j'ai construit mon truc,

33- C : tout à fait (en accompagnement du discours)

34- MS : j'ai commencé par les jeux, il faut faire ceci et après cela, pour arriver au match. Donc, du coup, c'est vrai qu'on voit la situation finale, mais du coup on sait pas, ben, moi, ce qui m'a gêné c'est que…on ne savait pas tout ce que j'avais mis en place pour arriver au match, donc c'était juste comme ça.

35- C : Et donc, si tu devais dire la progression que tu as mise en place, comment elle est faite cette suite

36- MS : Oui, alors, alors moi, ce qui m'a frappé, je le disais à Nadia, c'est qu'on a mis des choses en place, mais en fait, finalement dans le match, les enfants ils sont pas fait le lien du tout avec tout ce qu'on avait fait. Ils ont joué, en fait, on aurait pu faire le match sans faire tout ce qu'on a fait avant (sourire) et moi j'ai l'impression qu'ils n'auraient pas plus mal joué, quoi, ou mieux joué.

Le MS ne s'y retrouve pas dans les choix effectués par le chercheur et évoque la logique de progression de leur enseignement, celle qui a animé les 6 séances. Il constate que, par l'absence des tâches initiales qu'il considère comme essentielles, cette logique n'est pas prise en considération ou reconnue en tant que savoir professionnel (*« …ce qui m'a gêné c'est que…on ne savait pas tout ce que j'avais mis en place pour arriver au match… »*). Ce visionnement *à l'envers* (*« on voit la situation finale mais du coup on ne sait pas »*) produit, en même temps, une prise de distance chez l'enseignant qui l'amène à porter un regard global sur l'ensemble des séances et à évaluer le travail réalisé. La situation de match devient ainsi révélatrice des acquisitions des élèves et, du même coup, de leur propre enseignement.

37- C : Mais, qu'est-ce qu'il leur manque dans la situation de match ?

38- MS : Ils arrivent pas à, à

39- MG : (simultanément) Mettre en (XXXX)

40- MS : à mettre en pratique ce qu'on a fait avant, pour eux c'était autre chose. Je remarque ça dans la natation, quand on leur dit de nager librement, ils, ils nagent n'importe comment

41- MG : oui, oui (simultanément)

42- MS : tu leur fais faire des exercices pour corriger la respiration, ils font très bien et après, tout à coup, tu leur dis bon, nagez…ils nagent comme des petit chiens

43- MG : On a le même problème en en, par exemple en français, euh, dans les didactiques. Des exercices séparés, ils les font justes, mais après en pratique lorsqu'ils doivent écrire un texte, ça, ils ne font pas

44- MS : Je ne sais pas si c'est l'âge, si c'est…je ne sais pas

45- MG : Moi je pense que ces exercices, quand même de, de technique qui sont importants, que ça soit pour (phrase interrompue)

46- MS : On a fait peu de technique, on a justement fait des, des, des, des mises en situation, j'ai l'impression après qu'ils transposeraient plus facilement, mais ce que je remarque c'est qu'ils ne transposent rien du tout

(…)

81- MS : Cath : Mais, bon ça, je crois que c'est euh comme, si tu dis qu'en français c'est la même chose, je le remarque en natation, c'est vraiment flagrant

82- MG : Mais je crois qu'au bout d'un moment ça fait le déclic, mais peut- être pas au premier match. Mais je crois que euh, au bout d'un moment, à force de jouer, ils se souviennent des exercices, des petites choses qu'ils ont faites et puis euh…ça leur permet de progresser.

Le MS attribue l'absence de transposition à l'incapacité des élèves à (ré) utiliser les savoirs abordés dans les différents « exercices » préparatoires. En même temps, il voit qu'une action de sa part pour maintenir la mémoire didactique pourrait rendre plus efficace les moments de verbalisation (TdP 51 : *« Alors, moi ce que je ferais si je refais l'année prochaine, enfin, la prochaine fois que je refais du basket, je pense que je leur dirai…vous vous rappelez de ce jeu, qu'est-ce qu'on cherchait dans ce jeu ? »*). Pour le MG, qui considère l'assemblage des parties *(« des exercices séparés »*) dans un tout comme relevant d'une problématique plus générale et semblable aux autres didactiques , (*« On a le même problème en en, par exemple en français… »*), le temps didactique expliquerait le niveau atteint par les élèves (TdP 82).

Cette interprétation, bien qu'intéressante par la prise en considération du temps qui serait nécessaire aux transformations des conduites motrices, va clore le débat sur ce sujet et nous paraît, par son caractère restrictif, empêcher les maîtres de questionner les objets effectivement traités, le type de tâches agencées et leur rapport avec les compétences nécessaires pour faire du basket.

L'analyse a priori des tâches que nous avons réalisée montre que la plupart d'entre elles ont un rapport de contextualisation très faible[5] (ex. : absence d'adversaire ou de terrain orienté) et que la transposition attendue repose, en grande partie, sur un savoir procédural construit lors des moments de verbalisation, moments qui structurent et jalonnent l'ensemble des séances.
Pour montrer cette démarche de prise de conscience et de verbalisation instaurée par les enseignants, nous allons présenter des extraits tirés de la 1ère leçon du MS et de l'entretien cité plus haut. Ils nous paraissent représentatifs des techniques topogénétiques et mésogenétiques qu'ils ont utilisées pour solliciter, valider et instituer des règles d'action (savoirs à mobiliser dans le jeu réel) et permettent, là aussi, d'appréhender la particularité de leurs rapports au savoir.

Les dispositifs d'enseignement de sollicitation réflexive

Lors de la première leçon et après l'échauffement, le MS *définit* les règles des trois jeux auxquels les élèves vont jouer *(TdP 41 : « Sur ce terrain on va dire que ce sont les rouges en premier qui sont les attaquants, donc ils auront le ballon et ils vont essayer de se faire des passes »)* et *dévolue* la tâche tout en leur proposant des pistes pour réussir.

> 53- MS : Pas le droit de marcher, avec le ballon mais sans le ballon, évidemment, tu vas devoir tout le temps, tout le temps bouger, parce que vous n'allez pas pouvoir rester comme ça (il se place immobile, pieds écartés), la même chose est-ce que ça vaut la peine (il se dirige vers le tapis) pour un défenseur (il est sur le tapis) de rester tout le temps sur ce tapis?
>
> 54- Abdel: Non
>
> 55- MS: Vous allez voir.

Ce « vous allez voir » peut être considéré comme une *clôture topogénétique* (Sensevy et al. P. 290). *L*'enseignant dit aux élèves, sans le développer explicitement, qu'il existe une stratégie optimale et qu'à partir de ce moment, c'est à eux de trouver la solution. En même temps, sa formulation nous semble renforcer la procédure transmise aux élèves par questionnement (pour défendre, on ne reste pas tout le temps sur le tapis) et évoquer implicitement les contraintes de la tâche auxquelles ils seront confrontés (le nombre de tapis est plus grand que le nombre de joueurs). Pour le MS, les contraintes revêtent la

[5] Lors d'une autre réunion, nous avons présenté au MS le contenu des interactions didactiques maître-maître C'est à ce moment qu'il a pu porter un regard différent sur la nature des tâches de la programmation et comprendre autrement les résultats de leur action. Manifestement, il éprouvait des difficultés à s'éloigner de son rapport habituel au savoir et aux déterminations institutionnelles, mouvement facilité par l'apport de savoirs didactiques qu'il n'avait pas à disposition.

fonction de déclencher le processus adaptatif et créatif des élèves. En disant « vous allez voir », on prévoit également une future validation de la part de l'enseignant par de *traits pertinents* (TdP 53) et une action d'instituer une règle d'action que nous considérons, en regard des matériaux récoltés (transcriptions des entretiens et séances vidéoscopées) comme un des contenus principaux visés.

La situation de jeu est suivie d'un moment de verbalisation[6] qui permet au maître de savoir si les élèves ont *bien vu*. Ce moment débute par la sollicitation de l'enseignant d'une position réflexive des élèves et cela par un questionnement implicite (MS : *« Si vous deviez donner un conseil (…) pour aider vos camarades à bien jouer tout de suite »*). Durant tout l'échange, il adopte une posture de *surplomb topogénétique* en évaluant et statuant la pertinence de ce qui est dit, ceci lui permet de faire désigner aux élèves (El : *« il faut marquer son joueur »*) ou de désigner lui-même *ce qui est vrai* et d'instituer, en les rendant publiques, les repères (règles d'action) que les élèves doivent prendre en considération lors des situations de jeu qui vont suivre (MS : *« …ça paraît plus difficile de surveiller les tapis que surveiller chacun son joueur »*). Ce processus, qui s'apparente à la structure générale dialogique et tripolaire évoquée par Sensevy et al. (op. cit., p. 282) est emblématique du fonctionnement des deux enseignants.

Lors de l'entretien, ils y reviennent en faisant référence aux situations de *jeu réel* (match de basket en grand terrain) durant lesquelles les maîtres ont demandé aux élèves d'adopter, quand ils sont sur les bancs, un topos différent : l'élève n'observe/parle plus de *sa* pratique mais est amené à analyser, en adoptant la posture d'observateur (journaliste), la façon de jouer des équipes sur le terrain *(« Donc vous vous allez regarder que ça dans le jeu, les verts comment ils défendent, vous les rouges, comment ils se font des passes, comment ils s'organisent pour aller à l'attaque, ça marche?*).

Les moments de verbalisation qui suivent les matchs commencent aussi par une sollicitation topogénétique (ex. : *« Messieurs, mesdames les journalistes qu'avez-vous observé* ? ») qui constitue le début du processus dialogique évoqué plus haut. Les enseignants vont, en validant la pertinence des avis émis, ou en reformulant l'énoncé des élèves, définir le projet d'action pour les parties suivantes (ex. : *« vous choisissez dès le début de la partie, le joueur que vous allez marquer »*).

Durant l'entretien, les enseignants analysent cette partie du dispositif. Leurs avis se rejoignent dans l'évaluation de la qualité des observations/interventions des élèves (*« leurs remarques, elles étaient, je les ai trouvées bonnes »*), mais divergent au moment de comprendre l'impact au niveau des apprentissages effectifs.

[6] L'organisation des séances a induit chez les élèves une sorte « d'attente » de moments de verbalisation. Il suffisait à l'enseignant de siffler la fin d'un jeu pour que les élèves lèvent automatiquement la main, sachant à l'avance qu'ils seraient questionnés.

469- MS : Je trouve qu'ils avaient bien vu le jeu. Mais où j'étais déçu c'est qu'ils…ils voyaient bien, mais après, quand ils jouaient, ils arrivaient pas à, à tenir compte de ce qu'ils avaient pourtant vu, quoi. Moi je trouve que ça serait plus utile ces moments euh pour jouer.

470- C : Donc, verbaliser, dire, ça serait utile pour jouer.

471- MS : Oui, moi je pensais que ça serait, le fait de verbaliser, d'avoir observé, d'avoir vu les fautes des autres, je pensais que ça leur serait utile à eux pour le jeu.

472- MG : En même temps, c'est super difficile, parce qu'une fois que toi tu es dans le jeu, je veux dire

473- MS : (simultané) Oui

474- MG : tu joues donc le, le fait de réfléchir aux erreurs des autres et de tenir compte

475- MS : (simultané) Oui, oui

476- MG : de ces erreurs pour les mettre en pratique…même nous, en tant qu'adultes je, je ne suis pas sûr que ça marche vraiment

477- MS : Donc, justement, est-ce que c'est finalement euh

478- MG : (simultané) Je crois qu'il faut le répéter, le répéter, le répéter.

479- MS : Parce qu'à ce moment-là c'est vraiment si utile que ça, mais moi c'est aussi parce que, comme les élèves sont un moment immobiles sur un banc, moi je n'ai pas envie qu'ils passent leur temps à rigoler entre eux et à pas s'occuper de ce qui se passe, puisque je pense qu'on apprend quand même en observant

480- MG : (simultané) Je crois que c'est utile quand même

481- MS : je crois que c'est aussi leur donner quelque

482- MG : Quelque chose à faire

483- MS : à faire et pas rester et pas rester

484- MG : Je crois que c'est vraiment utile mais sur euh, euh, sur le long terme

485- MS : Oui

486- MG : pas sûr euh, par sûr, d'un match à l'autre, en fait. Mais je pense qu'à force de répéter ces choses…

Dans cet extrait, le MS revient aux situations de match et, en les considérant comme des moments d'actualisation des savoirs évoqués précédemment (*« mais après, quand ils jouaient »*), il questionne l'intérêt des verbalisations. Le MS constate que la capacité à *bien* voir le jeu (topos observateur) et à « verbaliser » (savoir procédural) n'est pas corrélative,

pour autant, d'une transposition sous forme de savoir en action (*« ils arrivaient pas à, à tenir compte de ce qu'ils avaient pourtant vu »*). Son analyse le conduit à envisager de transformer le temps de verbalisation en temps de jeu (*« Moi je trouve que ça serait plus utile ces moments euh pour jouer »*») et à dévoiler d'autres raisons concernant la mise en place du dispositif comportant des observateurs, celle-ci plus générique et en lien avec la gestion du groupe-classe (TdP 479, 480/81).

La position du spécialiste contraste avec l'avis du MG qui considère que le niveau atteint par les élèves et, par conséquent, la transposition qu'ils ont réalisée, sont à comprendre au regard de la complexité du passage du savoir procédural au savoir en action *(« c'est super difficile », « même nous, en tant qu'adultes »*) et au besoin d'un temps pour apprendre plus important que celui de l'unité mise en place (TdP 478, 484, 486). Les références des enseignants se distinguent et montrent à nouveau des assujettissements différents en termes de rapport au savoir.

Pour conclure provisoirement

L'entretien d'autoconfrontation croisée que nous avons mis en place devait permettre, à travers le rapport dialogique des acteurs, incité par le visionnement de séquences choisies, de favoriser l'émergence des réflexions concernant l'enseignement conjoint et en alternance de l'éducation physique. Les extraits présentés nous paraissent montrer une certaine distanciation des enseignants envers leur pratique et le potentiel formateur du dispositif, sans que l'on puisse pour autant présager, à partir de ce retour réflexif, des changements dans leur manière de concevoir et de faire les leçons.

Nous constatons l'assujettissement des maîtres au fonctionnement habituel de la dyade qu'ils forment. La confrontation, dans ce sens, nous a paru être influencée par les repères partagés, limitant sûrement ainsi le questionnement de leurs dispositifs et modèles didactiques. Ceci interroge l'absence (voulue) dans le dispositif, de propositions ou de *références extérieures* au système étudié (introduites par le chercheur), repères par exemple didactiques concernant les contenus d'apprentissage qui seront (voir par exemple Sensevy & Mercier, 2003) ou ont été abordés durant l'unité. Ces références nous semblent être porteuses potentielles d'un autre niveau de décentration et d'analyse et être plus enclines à répondre aux problématiques qu'émergent des démarches de formation des enseignants.

Avec le deuxième dispositif que nous vous présentons, nous poursuivons notre réflexion et nous tentons, dans un contexte bien différent (formation initiale) de comprendre les effets que produit dans la construction de gestes professionnels chez les étudiants, un dispositif d'autoconfrontation outillé par des connaissances didactiques sur l'action de l'enseignant et sur le contenu d'enseignement (Peindre à la manière de Pollock).

Deuxième dispositif (AP) : Vidéoscopie et régulations de l'activité d'enseignement

Comme nous venons de le montrer, les dispositifs basés sur l'autoconfrontation croisée recèlent potentiellement des processus d'explicitation et de distanciation analytique de la pratique enseignante, qui peuvent être mobilisés également dans le cadre de la formation professionnelle des enseignants.

Dans le cadre de la formation des enseignants du primaire à Genève, notre équipe de recherches mène depuis deux ans un travail d'analyse de ce type de dispositif, que nous utilisons, en tant que formateurs en didactiques des disciplines, pour instrumenter les démarches d'*évaluation formative* de nos dispositifs de pratiques sur le terrain scolaire. Dans le cadre de ce texte, nous souhaitons présenter quelques éléments résultant de ce travail en soulignant, toutefois, qu'il ne s'agit ici d'avancer que quelques premiers éléments d'une problématique émergente dans le champ de la recherche en didactiques (cf. Goigoux, 2000).

Comme nous l'avons annoncé au début, notre dispositif de formation se base sur l'étude de l'action enseignante comme *analyseur* des pratiques de formation. La vidéoscopie des leçons et l'autoconfrontation des stagiaires à leurs pratiques de terrain passe par un travail d'instrumentation didactique pour la formation qui nous semble répondre à quelques-unes des limites du dispositif tel qu'utilisé pour la recherche et soulignées précédemment.

La formation des enseignants à Genève est basée sur une structure d'alternance entre les semaines universitaires (cours, séminaires et ateliers) et des semaines de terrain durant lesquelles les étudiants sont amenés à prendre en charge, en dyades, l'enseignement des disciplines des modules didactiques. Résolument tourné vers une intégration des apports théoriques et de divers outillages de gestion et d'analyse des pratiques, notre dispositif cherche à travailler autant la dimension de planification (séquences didactiques de 2 à 4 séances de 50 min sur deux semaines), que celle de gestion en responsabilité des enseignements. Durant les semaines universitaires qui précèdent les terrains, demandons à nos étudiants de produire, en dyades, des séquences planifiées autour d'un sujet appartenant au programme de la classe qui les accueille, en accord avec l'enseignant titulaire qui joue le rôle central de formateur de terrain. Outre la planification des leçons, les dyades doivent filmer entièrement leurs interventions, avec répartition de rôles : un étudiant mène une ou deux leçons alors que l'autre se limite à le filmer et à observer la pratique du premier. La semaine de terrain qui suit, les rôles sont inter-changés à l'intérieur de la dyade de manière à ce que chaque étudiant ait l'opportunité de travailler les deux postures.

Le dispositif d'autoconfrontation croisée intervient entre la deuxième et la troisième semaine de terrain. Les étudiants sont invités à analyser leur pratique en cherchant à identifier des phénomènes liés à l'un des trois processus méso-topo ou chrono/génétiques ; puis, ils sont incités à investir les résultats de cette analyse dans la suite des pratiques de la troisième semaine. L'objectif du dispositif est que les étudiants arrivent à développer ou, le cas échéant, à remédier, certains de leurs gestes d'enseignement.

Deuxième dispositif - Comparaison de éléments de structuration dans la planifications GT et GSta

Séquence du groupe témoin (GT)	Séquence du groupe de stagiaires (GSta)
1. **Organisation en phases et tâches d'apprentissage** **Leçon 1** Activité principale (expérimentale) pour les élèves de 4ème primaire: produire un tableau avec la technique du dripping **(20 min.)** Activité secondaire, pour les élèves de 5ème primaire: produire un dessin à main levée de leurs camarades au travail. Le professeur donne une très brève consigne concernant les règles du dripping (ne pas toucher la feuille, laissez couler la peinture) (20 secondes). Pour l'autre tâche, il donne pour consigne de dessiner un camarada au travail (10 secondes) **Leçon 2** Echange autour des productions des élèves **(10 min)** Echange assez libre sur les rapports entre les productions d'élèves et quelques reproductions des tableaux de Pollock **(20 mn.)** **2. Milieu didactique et organisation sociale** Dans la cour de l'école, sur une surface recouverte de plastique protecteur (20m2), sept élèves produisent individuellement leurs peintures. Ils partagent les pots de peinture situés au centre de la surface. Les choix des couleurs est limité (variable didactique qui ne sera pas mise en évidence lors de la deuxième leçon). La deuxième leçon a eu lieu dans la salle de classe.	1. **Organización en fases y tareas de aprendizaje** **Leçon 1** Première tâche: Observer des reproductions des tableaux de Pollock disposés au centre d'une table autour de laquelle sont situées les élèves; constater qu'elles ne sont pas figuratives; constater qu'elles sont des « traces des gestes de l'artiste » ; suggérer le type d'effet que ces tableaux produisent chez celui qui les regarde. (**15min.**) Deuxième tâche (tâche principale): produire une peinture " à la manière de Pollock" à partir des thèmes proposés par les élèves comme interprétation des tableaux dans la première tâche. Les thèmes mettent en évidence les effets sur le récepteur plutôt que l'expression. (**20 min.**) Dans la consigne, les stagiaires insistent sur la nécessité d'adopter différentes postures corporelles et gestuelles et à utiliser tout l'espace de la feuille (tourner autour de la feuille). **Leçon 2** Première tâche: Echange critique sur les productions des autres groupes. L'objectif est de mettre en évidence des critères qui permettent de valider les productions par rapport au thème et à la fonction que remplissent les éléments techniques (geste et couleurs) dans cette validation. La consigne insiste sur le fait que la validation doit être basée sur "la manière de peindre" (**20 min.**) Deuxième tâche: démonstration des gestes picturaux qui ont été validés durant la première tâche (institutionnalisation) (**10 min.**) Troisième tâche: les élèves qui le désirent peuvent donner leur avis sur leur propre expérience esthétique (**5 min.**) **2. Milieu didactique et organisation sociale** Dans la salle de classe, plusieurs surfaces (4m2) couvertes avec du plastique permettent aux élèves de produire le tableau par groupes de trois élèves. Le groupe doit choisir un thème et le mettre en rapport avec les gestes techniques du dripping, qui n'ont pas été encore nommés, mais qui ont été décrits durant la première tâche (leçon 1). Les élèves de chaque groupe doivent choisir trois pots de peinture parmi quatorze mis à disposition du groupe classe. Entre la leçon 1 et 2, plusieurs organisations sociales entrent en jeu: travail individuel, groupe-classe, petits groupes. De même, sont sollicités diverses postures d'apprentissage inspirées des pratiques sociales de référence : produire, critiquer, observer, décrire, …

Un exemple : la modification de la gestion topogénétique dans une leçon « Faire à la manière de Pollock » (3P et 5P)

Les étudiants ayant choisi un sujet assez fréquent dans l'enseignement des arts plastiques (Peindre à la manière de J. Pollock et/ou peindre avec la technique du dripping), nous avons pu constituer un petit dispositif de recherche, dont voici les principales caractéristiques :

- Constitution d'un « groupe témoin » (GT) à partir de la vidéoscopie des pratiques d'un enseignant spécialiste ayant mené une séquence de trois séances auprès d'un groupe d'élèves de double degré (4ème et 5ème primaire). Les trois séances ont été planifiées et gérées sans intervention des chercheurs et constituent, dans le cadre de notre recherche, un groupe témoin représentatif des pratiques de l'enseignement ordinaire.
- Quatre groupes expérimentaux avec des élèves de 1ère (1 gr.), troisième (1gr.) et quatrième (2 grs.) primaire, auprès desquels quatre[7] dyades ont réalisé leurs pratiques de terrain (ci-après, groupes de stage ou GSta).
- Un travail de planification des séquences autour du thème « peindre à la manière de… », en dyades selon la démarche explicitée plus haut ;
- Une première phase d'autoconfrontations entre la deuxième et troisième semaines de terrain, avec une durée minimale d'une heure. Cette séance a été instrumentée à l'aide des trois catégories de phénomènes didactiques (méso, topo ou chrono/génétiques), que la consigne du formateur universitaire les a invité à mobiliser pour l'analyse (analyse de leur pratique à partir de l'un des trois types de processus) ;
- Un retour sur le terrain, avec des régulations concernant la planification et/ou la gestion de l'enseignement, basées sur la séance d'autoconfrontation. Ces pratiques ont été également vidéoscopées ;
- Une deuxième séance d'autoconfontations, cette fois en présence du chercheur, à partir d'extraits, parmi l'ensemble des leçons données, considérés comme significatifs par les étudiants.
- Dans le cadre de la recherche, les étudiants ont été invités à regarder également la séquence du GT[8] entre la deuxième et troisième semaine de terrain.

Du point de vue de la structure des leçons, les étudiantes qui ont comparé, lors de la deuxième séance d'autoconfrontations, les leçons Gsta avec la leçon GT, ont constaté quelques différences significatives du point de vue des gestes professionnels d'enseignement.

Une différence, identifiée par les étudiantes comme plutôt négative du point de vue de leur propre travail, consiste en une certaine « rigidité » dans la structure de la séquence GSta par rapport à celle du GT (voir tableau en fin de texte). Cette rigidité a été constatée, notamment, par rapport à une division beaucoup plus évidente dans la leçon Gsta entre les phases appartenant au travail de l'enseignant et celles considérées comme des phases dévolues au travail de l'élève. Un des observables utilisés par les étudiantes a été leur analyse de la première tâche de la leçon 1, dont la fonction était celle de faire « un petit apport théorique sur J. Pollock et sa peinture ». Cette première tâche était conçue par les étudiantes comme relevant essentiellement de l'enseignement, avec une activité des élèves réduite à une écoute attentive. En ce sens, un des effets positifs de l'autoconfrontation a été la proposition de rémédiation réalisée durant la troisième semaine de terrain, avec l'a production d'une « première tâche », ayant gardé la fonction d'introduction à la peinture de Pollock, mais travaillant davantage sur les *topos* des élèves. C'est ainsi que la deuxième étudiante a posé davantage de questions, a invité ses élèves à observer et à comparer des

[7] Faute de place, nous ne présenterons que les éléments d'analyse d'une seule des dyades.
[8] Cette dernière étape ne fait pas partie du dispositif de formation habituel

reproductions de Pollock ; durant cette première tâche, les informations sur Pollock ont davantage été produites à partir d'informations proposées par les observations des élèves.

Une autre différence soulignée par les étudiantes, cette fois-ci comme positive, a été le constat d'une richesse topogénétique beaucoup plus grande dans leur propre séquence que dans celle du GT. En effet, dans le séquence du GT, la posture d'apprentissage privilégiée a été celle de « l'artiste producteur », accompagnée par une deuxième posture de verbalisation (troisième leçon) ; par contraste, la leçon GSta développait une variété de postures plus importante, basée sur une vision plus large des pratiques sociales de référence (voir tableau en fin de texte): posture d'observateur (tâche 1, leçon 1), posture de production mais articulée à une modalité de travail collaboratif (tâche 2, leçon 1), posture critique (activité 1, leçon 2), posture de verbalisation (tâches 1 et 3, leçon 3),…

Du point de vue de la gestion des leçons effectives, les étudiantes ont choisi de mener une analyse de quelques phénomènes topogénétiques relatifs aux tâches de la deuxième semaine de terrain.

Les résultats de cette analyse ont été centrés sur les gestes et/ou actions de « dévolution et régulation » de l'activité des élèves. Par rapport aux leçons GT, les étudiantes ont constaté leur difficulté à produire des gestes de dévolution effective. L'autoconfrontation a permis de mettre en évidence le fait qu'elles n'avaient pas « vraiment écouté» les réponses d'élèves aux questions émergeant de l'interaction enseignant-élève ; ainsi que le fait qu'elles ont eu tendance à fournir souvent les réponses à la place des élèves, dans un souci de faire avancer la leçon. Par exemple, durant la deuxième tâche (leçon 1), les étudiantes ont constaté qu'au bout de cinq minutes seulement, l'une d'entre elles avait déjà « poussé » ses élèves à produire l'ensemble des gestes picturaux attendus (se déplacer autour de la feuille, changer les positions du corps, contrôler l'amplitude du geste, etc…). Autrement dit, elles ont constaté que, sous la pression de leur propre contrat de formation[9], la stagiaire n'avait pas laissé le temps aux élèves pour interagir avec le milieu didactique proposé.

Fruit de ces analyses, nous avons pu observer des effets positifs lors de la reprise de la séquence en troisième semaine de terrain. Un des observables est le changement sensible relativement au temps de parole pris par les stagiaires par rapport au temps laissé aux élèves pour agir et/ou s'exprimer. Alors que durant la deuxième semaine de terrain le temps de parole enseignante avoisinait les 60%, durant les deux tâches de la première leçon de la troisième semaine ce temps a été réduit à 48% environ. Ce changement quantitatif a été également accompagné de changements qualitatifs : augmentation des gestes de dévolution (consignes plus brèves, davantage de questions adressées aux élèves, régulations moins fréquentes et plus ciblées), ainsi que davantage de temps et d'opportunités laissés aux élèves pour interagir et avec le milieu didactique et dans chaque sous-groupe.

Quelques éléments de discussion

[9] Nous avons déjà analysé ailleurs (Schubauer & Rickenmann, REF-2003) quelques problèmes qui se posent aux stagiaires du fait d'être soumis, à la fois, au contrat de formation (posture d'apprenant) et au contrat didactique avec leurs élèves (posture d'enseignant).

L'utilisation des dispositifs d'autoconfrontation pour la formation constitue un apport dans le cadre de la recherche sur les dispositifs de formation par alternance. Même si le dispositif de recherche présenté n'est qu'à ses débuts, nous avons déjà quelques éléments qui nous incitent à réfléchir sur nos dispositifs de formation professionnelle en termes de « situations didactiques », dans le sens que leur donne Brousseau (1996). De même que pour les situations scolaires, il est possible de concevoir un *milieu didactique* en interaction avec lequel les étudiants puissent produire de nouvelles connaissances à partir des incertitudes et des obstacles qui émergent de leur pratique effective. Comme pour les situations scolaires, l'interaction des stagiaires avec et dans le milieu didactique peut être *instrumentée* par les formateurs universitaires et/ou de terrain dans le but, d'une part, de leur fournir des outils de transformation et d'analyse et, d'autre part, de leur permettre de réguler/modifier activement leurs pratiques.

Le travail sur les séquences EP analysé nous a permis de dégager les potentialités des dispositifs méthodologiques d'autoconfrontation pour la prise de distance analytique, ainsi que comme moyen pour faire émerger *l'implicite* des pratiques. Les exemples analysés nous ont permis également de montrer la possibilité de surmonter quelques limites et obstacles de ce type de dispositif. En ce sens l'intervention du chercheur dans le choix de structuration des séances vidéoscopées s'avère être un puissant catalyseur des processus potentiellement liés à l'autoconfrontation. Cette intervention rend manifeste la nécessité de ne pas *naturaliser* les méthodes, notamment dans le cadre des démarches cliniques de la recherche didactique (Leutenegger, 1999).

L'analyse des séquences AP nous a permis d'illustrer cette nécessité d'instrumentation conceptuelle, mobilisée pour l'occasion dans un but formatif. Les concepts sur l'action professorale ont en ce sens une fonction de grille de lecture qui médiatise les lectures de la pratique. Les catégories de l'action professorale ont permis aux étudiants de « voir » une structure plus complexe que la seule organisation par phases que les étudiants ont tendance à produire spontanément (introduction du maître, tâches de l'élève, bilan ou évaluation). Les fonctions de l'action professorale (définir, dévoluer, réguler, institutionnaliser) leur permettent de valider la structuration initialement proposée en même temps qu'elles incitent à approfondir leur regard sur chaque phase. L'analyse a montré que cette médiation permet aux étudiants de les « re-trouver » à un niveau plus micro, dont ils n'ont souvent pas conscience comme nous l'avons montré avec l'analyse d'autoconfrontation des maîtres dans les séquences EP.

Un deuxième niveau d'instrumentation est basé sur l'organisation du travail étudiant en dyades, qui leur permet de capter en même temps deux dimensions fortement liées à la pratique enseignante (posture d'enseignement et posture d'observation). Dans ce cas, la mise à disposition des catégories d'analyse (méso-topo et chrono/génétiques) produit des effets de distanciation et permet d'adopter une conception plus systémique de l'action professorale.

Notons, pour conclure, qu'il ne suffit pas de mettre à disposition des catégories et modèles conceptuels pour aider à la lecture des pratiques, mais qu'il est aussi nécessaire de prévoir des « moments » de formation qui ne soient pas toujours liés aux « urgences »

de la pratique. L'autoconfrontation comme partie intégrante d'un *milieu didactique de formation* (mésogenèse), de même que le temps qui est mis à disposition pour observer, analyser les pratiques et pour modifier les planifications (chronogenèse), constituent des actions de dévolution de la responsabilité associées à une diversification des *postures d'apprentissage* (topogenèse).

Références

Amade-Escot, C., Barrué, J.-P., Bos, J.-C., Duffor, F., Dugrand, M., Terisse, A., (1998). Didactique de l'EPS. Grenoble : La Pensée sauvage.

Amade-Escot, C. (2002). Étude du travail de l'enseignant d'éducation physique dans la classe : contribution des recherches didactiques à l'analyse des pratiques effective. In J.-F. Marcel (Éd).), *Les sciences de l'éducation : des recherches, une discipline*, 53-78. Paris : L'harmattan, coll. Savoir et formation.

Amade-Escot, C. (2003). La gestion interactive du contrat didactique en volley-ball : agencement des milieux et régulations du professeur, 255-278. In : *Didactique de l'éducation physique.* (Dir.) Amade-Escot, Paris : Ed. EPS.

Biddle, B.J., Good, T.-L., Goodson, I.F. (eds.)(1997). *International Handbook of teachers end teaching* (vol 3). Dordrecht: Kluwer.

Bronckart, J.-P. (2005). Vingt-cinq ans de didactique de l'expression écrite. Eléments de bilan et perspectives d'avenir. *Revue Suisse des Sciences de l'Education*, 27 (3), 361-380.

Brousseau, G (1986). Fondements et méthodes de la didactique des mathématiques. *Recherches en didactique des mathématiques*, La pensée Sauvage, Vol 7, no 2, 33-115.

Brousseau, G. (1996). L'enseignant dans la théorie des situations didactiques. *Actes de la 8e école d'été de didactique des mathématiques*. IREM : Clermont-Ferrand.

Calderhead, J. (1996). Teachers: Beliefs and knowledge. In D.C. Berliner & R.C. Calfee (Eds.), *Handbook of educational psychology* , 709-725. New York: Macmillan

Chevallard, Y. (1995). La fonction professorale. Esquisse d'un modèle didactique. In R. Noirfalise (éd.), *Actes de la 8ème école d'été de didactique des mathématiques*, 83-122.

Chevallard Y., 1999, L'analyse des pratiques enseignantes en théorie anthropologique du didactique. *Recherches en Didactique des Mathématiques*, 19 (2), pp 221-265, La Pensée Sauvage, Grenoble.

Chevallard, Y (2004). La place des mathématiques vivantes dans l'éducation secondaire : transposition didactique des mathématiques et nouvelle épistémologie scolaire. *3e Université d'été Animath*, Saint-Flour (Cantal), 22-27 août.

Clot, Y., Faïta, D (2000). Genres et styles en analyse du travail. Concepts et méthodes. *Travailler*, 4, 7-42.

Crahay, M., Lafontaine, D. (eds) (1989). *L'art et la science de l'enseignement.* Paris : PUF.

Devolvé, N. (2000). Pour une approche systémique du métier d'enseignant, le point de vue de l'ergonomie. *Approches systémiques et recherches en Sciences de l'Education*, revue du CREFI, No 3.

Goigoux, R. (2000). Enseigner le lecture à l'école primaire. Dossier d'Habilitation à diriger des recherches, U.F.R. de Sciences de l'éducation, Université de Paris VIII.

Faïta, D. Vieira D. (2003). Réflexions méthodologiques sur l'autoconfrontation croisée. Internet :http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502003000100005

Lave, J., Wenger, E. (1991). *Situated Learning : Legitimate peropherical participation*. Cambridge: University University Press.

Leutenegger, F. (1999). *Contribution à la théorisation d'une clinique pour le didactique*. Thèse de Doctorat en sciences de l'éducation. Université de Genève.

Leutenegger, F., Ligozat, F. (2006). Analyse de l'action conjointe professeur-élèves dans le traitement d'un problème d'agrandissement de distances. *Séminaire National de DDM.*.

Leutenegger, F. (2000). Construction d'une « clinique » pour le didactique. Une étude des phénomènes temporelles de l'enseignement. Recherches en Didactique des mathématiques, Vol. 20/2, 209-250.

Loquet M., Garnier A. & Amade-Escot C. (2002) Transmission des savoirs en activités physiques, sportives et artistiques dans des institutions différentes : enseignement scolaire, entraînement sportif, transmission chorégraphique. Revue française de pédagogie, *« Vers une didactique comparée »*, 141, 99-109.

Mercier, A, Sensevy, G., Schubauer-Leoni, M.-L. (2000). Vers un modèle de l'action didactique du professeur. A propos de la course à vingt. *Recherches en didactiques des mathématiques*, 20(3), 263-304.

Mili, I., Rickenmann, R. (2005). Approches didactiques dans l'enseignement artistique: une revue de littérature sur la construction de l'expertise culturelle en classe. *Revue Suisse des Sciences de l'Education*.

Mili, I., Rickenmann, R. (2004). La construction des objets culturels dans l'enseignement artistique et musical. In C. Moro & R. Rickenmann (éds), *Situation éducative et significations*, 165-196, Bruxelles : DeBoeck.

Portugais, J. (1992). *Didactique des mathématiques et formation des enseignants*. Berne : Peter Lang.

Revue française de Pédagogie. (2002) *Vers une didactique comparée, oct. nov. déc.141*.

Rickenmann, R. (2001). Sémiotique de l'action éducative. In J.-M. Baudouin & J. Friedrich (éd.), *Théories de l'action et éducation,* 225-254, Bruxeles : DeBoeck

Rickenmann, R. (2005). Didactics in art education and appropriation of cultural objects in classroom. A semiotical analysis from a teaching/learning perspective. Communication dans le 1st International Congress of the ISCAR, Sevilla, Spain, Sept 20-24.

Rickenmann, R. (2006a). El rol de los artefactos culturales en la estructuración y gestión de secuencias de enseñanza-aprendizaje. In Conférence invitée, Actes du 1er simposio internacional de educación y formación docente, Universidad de Antioquia, Medellín (Colombia), agosto 29 – septiembre 1ro 2006.

Rickenmann, (2006b, sous presse). Metodologías clínicas de investigación en didácticas y formación del profesorado: un estudio de los dispositivos de formación en alternancia, *Revista educación*, Universidad Nacional de Antioquia

Salin, M.H. (1997). Contraintes de la situation didactique et décisions de l'enseignante. In C. Blanchard – Laville (éd.), *Variations sur une leçon de mathématiques*. Paris : L'Harmattant. 31-57.

Sensevy, G (2001). Théories de l'action et action du professeur. In J.-M. Baudouin & J. Friedrich (éd.), *Théories de l'action et éducation,* 203-224, Bruxelles : DeBoeck

Sensevy, G., Mercier, A.., Schubauer-Leoni. M.-L. (2000) Vers un modèle de l'action didactique du professeur à propos de la couse à 20. *Recherches en Didactique des Mathématiques*, Vol. 20, n°3, 263-304.

Sensevy, G., Mercier, A. (REF-2003 sous presse) Logiques d'action du professeur et du chercheur. Un essai d'analyse. In *Actes du symposium Recherche et intervention en didactiques,* Réseau Education Formation (REF), Genève.

Schubauer, R., Rickenmann, R., (REF-2003, *sous presse)*. Un dispositif expérimental de formation aux gestes professionnels. In *Actes du symposium Recherche et intervention en didactiques*, Réseau Education Formation (REF), Genève.

Schubauer-Leoni, M.-L, (1998). Les sciences didactiques parmi les sciences de l'éducation : l'étude du projet scientifique de la didactique des mathématiques. In R. Hofstetter & B. Schneuwly (Ed.), *Le pari des sciences de l'éducation* (Raisons éducatives N°1998, 1-2, 329-352). Bruxelles : DeBoeck.

Sieber, P., Thévenaz, T. (2005). Didactique des disciplines et didactique comparée : entre continuité et réorientation. *Revue Suisse des Sciences de l'éducation.*

Thepaut A. (2003). Echec scolaire et éducation physique et sportive à l'école élémentaire. Etude des interaction maître-élèves dans la construction des savoirs. L'exemple de l'apprentissage de la passe en basket-ball. Thèse de Doctorat. Université de Rennes II, France.

Wirthner, M., Schneuwly B. (2004). Variabilité et contrainte dans la construction des significations d'un objet d'enseignement. In C. Moro & R. Rickenmann (Eds), *Situation éducative et significations,* 107-133, Bruxelles : DeBoeck.

EVALUATION EN EPS : L'APPORT DE LA VIDEO TEL QUE PERCU PAR DES ETUDIANTS DE LA HEP FRIBOURG

Pierre Monney, professeur d'EPS, Haute école pédagogique de Fribourg (CH)
MonneyPi@edufr.ch
Benoît Gisler, professeur d'EPS, Haute école pédagogique de Fribourg (CH)
Gislerb@edufr.ch
Pierre-François Coen, responsable du service de la recherche, Haute école pédagogique de Fribourg
pierre-francois.coen@edufr.ch

Problématique

Le recours aux technologies (à la vidéo notamment) comment moyen d'aider les apprentissages n'est pas récent. Des travaux datant des années 70 démontrent toute la pertinence d'une approche pédagogique recourrant à ce type de technologie. Le développement considérable de ces dernières années tant sur le plan logiciel que matériel permet d'envisager des usages souples et adaptés aux besoins des enseignants. Ainsi le dispositif *DartTrainer* développé par la firme *Dartfish* donne à l'enseignement toute liberté de concevoir des séquences d'apprentissage en intégrant la vidéo particulièrement pour donner des feedbacks aux élèves. Actuellement, de nombreuses recherches (Boyce, 1996 ; Guadagnoli, 2002) démontrent que l'usage de la vidéo permet d'améliorer de manière significatives les performances des élèves novices. Ce type d'outil ouvre des perspectives novatrices dans les pratiques d'évaluation formative notamment en impliquant de manière plus intense l'élève. Ainsi l'accès à la trace via les technologies donne à l'apprenant l'opportunité de mieux comprendre ce qu'il fait (Coen, 2006) et par là de mieux agir sur les différents paramètres en jeu. Ces pratiques d'évaluation très répandues dans le domaine de la compétition sportive sont encore peu présentes dans les classes et pourtant des moyens abordables et faciles à utiliser sont disponibles sur le marché.

Dans un contexte de formation (la Haute école pédagogique de Fribourg forme de futurs enseignants généralistes), il est apparu intéressant de confronter les étudiants à un dispositif d'apprentissage incluant cet aspect. Premièrement pour les confronter à des situations dans lesquelles il leur était possible de développer une posture autoévaluative et réflexive et deuxièmement pour démontrer les avantages d'une démarche instrumentée à l'aide d'un logiciel utilisable dans un contexte d'apprentissage.

Cette recherche vise à répondre à différentes questions ?

- est-ce que l'usage du dispositif *DartTrainer* permet une prise de conscience plus précise des mouvements que l'étudiant effectue ?
- après avoir visualisé sa performance, l'étudiant sait-il mieux sur quoi il doit porter son attention pour s'améliorer ?
- les corrections qu'il apporte sont-elles opportunes et en lien avec les difficultés identifiées ?

- quelles sont les représentations des étudiants sur la pertinence du dispositif, son utilité pour les apprentissages et sa possible mise en place dans la classe ?

Méthodologie et instrumentation

Le dispositif expérimental a impliqué une vingtaine d'étudiants (N=18) de la HEP Fribourg. Chaque étudiant a dû exécuter 5 fois une figure (appui renversé) en bénéficiant après chaque présentation d'un feedback vidéo immédiat donné par le logiciel *DartTrainer*. Après chaque exécution, l'étudiant devait produire un commentaire de sa prestation et proposait par écrit des améliorations possibles et des stratégies pour y parvenir. En parallèle et de manière indépendante, un professeur d'EPS expert procédait aux mêmes démarches : constats liés à la prestation et pistes d'amélioration, cependant, à aucun moment, les constatations du professeur expert n'ont été communiquées à l'étudiant. Après l'expérience, chaque étudiant s'est entretenu une quinzaine de minutes avec un expérimentateur neutre sur la validité de la démarche (impressions générales, avantages, inconvénients) en donnant ses réactions « à chaud ».

Résultats

Le traitement des données permet de mettre en évidence les différents regards sur la performance selon le point de vue du professeur et celui de l'élève. On constate que les étudiants identifient de manière pertinente les points faibles de leur prestation et que ceux-ci sont le plus souvent en adéquation avec le regard du professeur expert. En outre, les étudiants jugent pertinent le recours au feedback vidéo immédiat pour améliorer leur prise de conscience. Ainsi, ils disent être plus conscients de ce qu'ils font et plus en mesure de formuler de manière précise une évaluation de leur prestation. Par contre, ils estiment que cette démarche d'autoévaluation les laisse relativement démunis par rapport aux moyens de s'améliorer et estiment importante l'intervention d'un professeur externe pour leur donner des pistes susceptibles de se corriger.

Références

Boyce, A. B., Markos, N. J., Jenkins, D. W., & Loftus, J. R. (1996). How should feedback be delivered? *Journal of physical education, recreation and dance, 67*(1), 18-22.

Coen, P.-F. (2006). Les technologies, des aides précieuses pour développer la réflexivité des apprenants. *Formation et pratiques d'enseignement en questions, 3*, 149-160.

Guadagnoli, M. A., Holcomb, W., & Davis, M. (2002). The efficacity of video feedback for learning golf swing. *Journal of sports sciences, 20*, 615-622.

Meers, G. (1972). Effects of various feedback media on psychomotor achievement. University of Michigan.

PASSER D'UNE EVALUATION NORMATIVE A UNE EVALUATION COMPREHENSIVE EN FORMATION INSTRUMENTALE.

Pierre-François Coen, responsable du service de la recherche, Haute école de musique de Fribourg pierre-francois.coen@edufr.ch

Résumé :

La création et le mouvement sont des besoins essentiels à l'épanouissement et au développement de l'humain. Cependant, les programmes scolaires européens donnent peu de temps à ces disciplines. Elles se trouvent souvent considérées comme « secondaires » ou optionnelles dans les curriculums. En dépit des écrits des spécialistes de l'Education, l'Ecole d'aujourd'hui offre peu de solutions concrètes pour améliorer cette situation.

Les tentatives de terrain sont encore trop rares ainsi que les écrits spécifiques à l'évaluation artistique et sportive en langue française. Les savoirs artistiques peuvent-ils continuer à exister dans le cursus scolaire ? Si oui, quels outils peuvent-ils être proposés ?

La problématique de l'évaluation est celle de la quête du sens, on pourrait se demander si l'enseignement artistique et sportif a encore un sens en classe et quels outils peuvent aider les enseignants et les apprenants dans leurs actions.

Pour la première fois, une équipe de chercheurs suisses et français partagent leurs résultats sur la thématique évaluation et enseignements artistiques et sportifs. Ce symposium de l'ADMEE réunira des spécialistes en Sciences de l'Education, Arts plastiques, Education musicale et Education physique et sportive qui ont souhaité confronter leurs travaux de recherches et expériences formatives.

Les communications seront regroupées autour des questions suivantes : quelles sont les outils en évaluation artistique et sportive ? Quels sont les enjeux et les implications de ces outils pour la pratique artistique et sportive ? Dans quel contexte évolutif s'inscrivent les innovations présentées ?

Tout d'abord, René Rickenmann proposera les résultats d'un dispositif de formation à l'*« évaluation formative des élèves »* en Arts plastiques *comme instrument de conduite des activités d'enseignement/apprentissage*.
De même, Adrián Cordoba exposera ce dispositif de recherches sur les rapports entre l' *« évaluation formative des élèves »* et sa mobilisation dans la *conduite des activités d'enseignement/apprentissage* comment l'ancrage disciplinaire (sciences de l'éducation ou études pédagogiques pour les titulaires ; formations sportives ou Ecole des Beaux Arts, pour les spécialistes) :

La recherche est centrée sur la transcription écrite des leçons d'arts plastiques ayant été vidéoscopées
Puis, le recours à un logiciel nommé *DartTrainer* ouvrirait des perspectives novatrices dans les pratiques d'évaluation formative notamment en impliquant de

manière plus intense l'élève. Dans un contexte de formation, Pierre Monney et Benoît Gisler montreront l'impact de cet outil utilisé par les étudiants en situation. Le feedback vidéo immédiat permis aux étudiants de formuler de manière précise une évaluation de leur prestation.
Dans le domaine musical, l'usage du portfolio permettrait d'articuler les deux temps de l'évaluation (régulation et certification) tout en restant dans un paradigme compréhensif. Le process-folio et le portfolio seraient des moyens efficaces dans le cadre d'une évaluation formative d'une leçon d'instrument. Selon Pierre-Francois Coen, ils favoriseraient le passage d'une évaluation normative à une évaluation compréhensive en formation instrumentale.

L'usage de l'enregistrement audio dans l'évaluation des performances vocales des élèves en classe de musique permettrait-il d'aider vers une meilleure auto-évaluation ? A partir d'entretiens d'enseignants en musique, Cedricia Maugars décrit les représentations des enseignants de musique en formation ayant pour la première fois utilisée cet outil.
Enfin, un dispositif des cours de piano en ligne sera discuté : quels sont réellement les avantages dans l'évaluation par l'exécutant de sa propre prestation musicale ? Quelles sont les possibilités réelles d'auto-évaluation pour l'apprenant ? Ce type d'enseignement via Internet est-il compatible avec une conception de l'évaluation hors de la mesure et du contrôle ? C'est que Jacques Audran nous questionnera dans sa communication.

Problématique

Les conceptions de l'évaluation des futurs professeurs d'instrument révèlent un modèle normatif dans lequel le professeur juge la prestation de l'élève en termes positifs et/ou négatifs par rapport à l'image de ce qu'il estime être bien de faire. Une analyse plus détaillée des pratiques (des leçons d'instrument) démontre que cette image du produit attendu n'est que très rarement explicitée et que ce n'est que progressivement que l'élève parvient à la construire et à ajuster sa production en fonction de ce que lui dit le professeur.

Partant de ces constats (révélés par la collecte de données empiriques), nous proposons un cadre conceptuel faisant apparaître deux pôles sur lesquels des pratiques d'évaluation innovantes peuvent se greffer. Le premier – inscrit dans un paradigme instructiviste – est un pôle *normatif* où le professeur apprécie et juge la prestation de l'élève ; le second – inscrit dans un paradigme constructiviste – est un pôle *compréhensif* où le professeur essaie de comprendre ce que fait l'élève. Le premier est basé sur l'idée qu'il faut apprendre *à* l'élève, le second qu'il faut apprendre *de* l'élève. En ce sens, le premier envisage l'apprentissage comme une accumulation de savoirs et de savoir-faire alors que le second considère l'apprentissage comme une intégration de différents apports qu'il est nécessaire de ré-assembler. Si dans le premier, l'excellence est signifiée par la parfaite restitution, dans le second c'est la re-création qui prime.

Outre l'exposition de ces deux perspectives, la communication vise à présenter de manière concrète deux outils susceptibles de répondre aux exigences du deuxième paradigme. Ainsi l'usage du process-folio et du portfolio transposer dans le contexte de la leçon d'instrument permet d'articuler les deux temps de l'évaluation (régulation et certification) tout en restant dans un paradigme compréhensif. Ces outils donnent au professeur et à l'élève les moyens d'expliciter les produits attendus, de se centrer sur le processus d'apprentissage sans perdre de vue la maîtrise de compétences et de ressources à maîtriser.

SYMPOSIUM 3 : L'EVALUATION AU CŒUR DES ACTIVITES DE VALIDATION DES ACQUIS DE L'EXPERIENCE.

Emmanuel Triby,Université louis Pasteur
Marie-Christine Presse, Cueep Université de Lille1 France
Isabelle Cherqui-Hout, Université Henri Poincaré
Brigitte Pagnani, CUCES - Universités
Pascal Tisserant, Université Paul Verlaine Metz - SCFC
Martine Morisse, Université Charles-de-Gaulle, Lille 3

Résumé

Dans la démarche de validation des acquis de l'expérience, tendue par la double injonction de déclarer ces acquis valides au regard des attendus d'un diplôme et celle de les certifier par ce même diplôme, la pratique de l'évaluation est à la fois omniprésente et masquée. Trois domaines de questionnement structurent la problématique :

- *les objets de l'évaluation.* Que s'agit-il d'évaluer ? Des connaissances et des aptitudes**,** si l'on en croit les décrets applicables à l'enseignement supérieur, des compétences pour les décrets applicables dans d'autres ministères. Comment articuler ces trois notions dans une même pratique d'évaluation, comment en fait certifier que les acquis sont effectifs ? Ces questions ne peuvent s'énoncer indépendamment de celle des finalités propres au dispositif d'évaluation : uniquement rapportée à l'évaluation des compétences professionnelles, l'évaluation des acquis de l'expérience ne revient-elle pas à une forme d'évaluation du travail dans laquelle l'institution validante jouerait le rôle d'une instance extérieure ? Portant de manière indifférenciée sur "l'expérience", l'évaluation ne risque-t-elle pas de faire porter le jugement sur la personne elle-même ? Ces risques sont d'autant plus concevables que *le rapport d'étrangeté* est grand entre l'institution validante et les milieux professionnels, avec tous les effets réciproques de fascination ou d'incompréhension que cela comporte.

- *les instruments de l'évaluation.* Si le dossier de validation apprécié par le jury constitue un incontestable support d'évaluation, et l'entretien une situation assez commune d'évaluation, ces deux éléments ne sont pas l'expression des acquis eux-mêmes, mais leur mise en forme au terme d'un processus de construction en général *accompagné* : en ce sens, cela touche directement le travail de l'accompagnateur et celui accompli par le candidat, ainsi que leurs capacités respectives à intégrer les stratégies évaluatives du jury. Quant aux référentiels (de formation/de compétences /d'activités) qui, quand ils existent – notamment à l'université - pourraient constituer des cadres de référence, ils sont peu utilisés. Lorsqu'ils le sont, c'est avec un succès mitigé car ils ont été conçus à d'autres fins que celle de l'évaluation des acquis expérientiels.

- *la nature de l'activité d'évaluation elle-même.* Cette question se pose d'abord pour le candidat et sa capacité d'*autoévaluer* ses possibilités de validation

tout au long du processus, et donc des outils dont il dispose pour y parvenir. Par ailleurs, cette question se pose différemment suivant l'acteur concerné, à chacune des étapes de la procédure : "l'expert" (ou le groupe-expert) jugeant de la recevabilité de la demande, l'accompagnateur aidant le candidat à construire son dossier, les membres du jury, professionnels ou enseignants. Elle se pose également quand il s'agit de savoir ce qui va être valorisé et ce qui ne le sera pas, dans le croisement des représentations et des attentes des candidats et des valideurs. Elle se pose enfin lorsqu'il s'agit de décider de valider, totalement ou partiellement ; dans ce dernier cas, quand il s'agit de décider des modalités du comblement des manques repérés.

Les réponses apportées à ces différentes questions détermineront en grande partie les *effets* de la procédure : sur le candidat, ses attentes et sa *reconnaissance*, sur l'institution validante, la valeur de ses diplômes et son rapport aux savoirs, sur les organisations productives intéressées, leur mode de classification des ressources humaines et d'évaluation des salariés.

Nous invitons les chercheurs et les praticiens du champ de la validation des acquis de l'expérience à s'emparer de l'une ou l'autre de ces questions selon leur domaine de pratique et/ou d'investigation, afin d'alimenter la réflexion, le débat et le travail de conceptualisation que cette pratique exige.

L'ACTIVITE D'ACCOMPAGNEMENT EN VAE SAISIE PAR L'EVALUATION

Patricia CHAMPY-REMOUSSENARD, Emmanuel TRIBY, LISEC,
Universités de Nancy2 et Strasbourg 1 (ULP)

Mots clés : activité, accompagnement, VAE

Résumé :

La démarche de validation des acquis de l'expérience (VAE) relève nettement du champ de l'évaluation ; elle n'est pourtant pas souvent interrogée dans toutes les occurrences de cette évaluation. La recherche présentée (développée dans le cadre d'un projet FSE) porte sur l'analyse de l'activité des différents protagonistes de la démarche VAE à l'université. Dans cette contribution, elle est centrée sur l'activité d'accompagnement. L'investigation dont les résultats sont ici exploités est nourrie de l'analyse de contenu d'entretiens semi-directifs conduits avec des accompagnateurs aux statuts fort divers.

Cette investigation vise d'abord à faire émerger la teneur de l'activité d'accompagnement mise en œuvre dans des contextes divers de l'enseignement supérieur, ainsi que les différentes stratégies mises en œuvre par l'accompagnateur pour assurer la réussite du candidat, les différents temps ou étapes du processus d'accompagnement. L'objet et le mode d'évaluation constituent une préoccupation commune, bien que seconde, à tous les accompagnateurs ; elle se trouve déclinée diversement dans les différentes dimensions de leur activité. L'enquête a également permis d'interroger la relative spécificité de l'activité d'accompagnement du candidat dans l'enseignement supérieur, au regard des pratiques des jurys dans ce contexte particulier, démarches marquées par le caractère central de ce qu'on peut qualifier de mise en mots de l'expérience. Un des enjeux d'une telle recherche est donc de comprendre ce que permet d'évaluer ce processus de mise en mots débouchant sur un dossier destiné au jury et comment cela permet de l'évaluer. La pratique d'accompagnement confère par ailleurs une portée formative au « travail » engagé par le candidat dans la VAE et donne accès à la construction d'une forme de réflexivité dont on peut, dans ce contexte, se demander si elle se trouve évaluée par les jurys.

Problématique

La démarche de validation des acquis de l'expérience (VAE) relève nettement du champ de l'évaluation, tant du point de vue du jugement qui est formulé que du point de vue de l'estimation de soi (le candidat tout au long du processus) et des acquis de l'expérience (des savoirs, des aptitudes…) qui y est pratiquée ; elle est aussi appréciation de la valeur des certifications autant que des institutions validantes elles-mêmes, à travers toute la procédure, des ressources qu'elle mobilise jusqu'à ses issues économiques et sociales. Cette activité d'évaluation multiforme n'est pas toujours

reconnue comme telle dans les institutions de validation. Il nous paraît important de l'interroger en nous intéressant notamment à la manière dont les différents protagonistes de la VAE *bricolent* cette activité : les candidats entre attentes et projets, les accompagnateurs qui aident à construire le dossier de validation, les jurys de validation qui apprécient ce dossier lors d'un entretien. Pour ces différents protagonistes, la question doit porter sur les objets qui sont effectivement évalués, les outils qu'ils mobilisent et les pratiques d'évaluation qu'ils développent. Dans cette contribution, nous nous intéressons particulièrement aux accompagnateurs et à la place de l'évaluation dans leur activité d'accompagnement.

Démarche et matériaux

Notre démarche consiste à mobiliser et croiser trois types de matériaux :

- un travail de conceptualisation à partir de l'analyse de l'activité et ses "traductions" écrites ainsi que de l'évaluation des dispositifs de formation
- le fonctionnement d'un dispositif universitaire de validation : réflexions sur l'expérience de l'analyse de l'expérience
- surtout, un corpus d'enquêtes par entretiens approfondis auprès d'accompagnateurs en VAE, réalisées dans le cadre d'un projet de recherche FSE.

Résultats et questions

L'analyse des entretiens

L'analyse des entretiens permet de distinguer dans chacun des cas les stratégies mises en œuvre par l'accompagnateur pour que le candidat réussisse, les différents temps ou étapes du processus d'accompagnement : positionnement dès l'entrée dans la VAE, suivi pour aboutir au dossier, préparation en vue de la rencontre avec le jury, suivi des effets de la VAE sur le candidat ; finalement des logiques dominantes de la pratique émergent et se distinguent. Au sein de cette singularité de chaque style d'accompagnement, les prémices d'un genre professionnel commencent à poindre. C'est par conséquent également la légitimité et la légitimation de l'utilité de l'activité d'accompagnement qui se trouve questionnée alors même que les praticiens sont en cours de construction d'une déontologie de leurs activités.

Inscrite dans le processus qui aboutit potentiellement à la délivrance d'un diplôme par un jury, l'activité d'accompagnement est traversée par la question de l'évaluation. Ce qu'il s'agit d'évaluer et comment constitue une préoccupation commune à tous les accompagnateurs et déclinées dans les différentes dimensions de leur activité. L'enquête a également permis d'interroger la relative spécificité de l'activité d'accompagnement du candidat dans l'enseignement supérieur, au regard des pratiques des jurys dans ce contexte particulier, de la culture évaluative des acteurs de l'enseignement supérieur, des conceptions et usages de la VAE qui sont les leurs.

Questions en suspens

Les modalités de validation proposée par la loi donnent lieu à des démarches marquées par le caractère central de ce qu'on peut qualifier de mise en mots de l'expérience. Un des enjeux d'une telle recherche est donc de comprendre ce que permet d'évaluer ce processus de mise en mots débouchant sur un dossier destiné au jury et comment cela permet de l'évaluer. Le rôle de l'oral, l'écrit, dans un processus d'accompagnement de l'explicitation est un des axes de l'activité de

l'accompagnateur dont il s'agit de comprendre les formes. Comment se déclinent les modalités de questionnement, de reformulation, d'aide à la formulation, à l'écriture et à la réécriture dans le travail de VAE ? Quelles formes de travail à distance, quelle activité autonome, quelles formes d'auto évaluation du candidat dans ces processus d'accompagnement ? Quel usage est-il fait du référentiel dans l'accompagnement et plus généralement d'autres ressources à disposition des acteurs ?

Enfin la pratique d'accompagnement confère-t-elle une portée formative au « travail » engagé par le candidat dans la VAE. Si elle donne accès à la construction d'une forme de réflexivité, cette dernière se trouve t-elle évaluée par les jurys de VAE de l'enseignement supérieur ?

LA MAÎTRISE DE LA LANGUE : OBJET CENTRAL DE L'EVALUATION DANS LA PROCEDURE VAE ?

Marie-Christine Presse, Université de Lille 1. Laboratoire Trigone

Mots clés : Expérience**,** Langage, Idéologie

Résumé:

La communication porte sur la fonction du dispositif de VAE : s'agit-il d'un dispositif qui facilite la promotion sociale et la montée en qualification des personnes actives non diplômés ou s'agit d'un dispositif sélectif. L'hypothèse est la suivante : en tension entre deux logiques, celle de la promotion sociale et celle de la sélection sociale, le dispositif de VAE est davantage un outil de reproduction sociale qu'un outil de transformation sociale. Le travail de recherche qui s'est effectué par entretiens et par recherche action a permis d'identifier les obstacles majeurs de ce dispositif : il pousse à l'extrême le processus d'individualisation et de responsabilisation individuelle, sous couvert d'autonomie, l'un des facteurs qui permet de contrôler les effets de ce dispositif sur la montée en qualification de ces personnes se situe dans le rapport à langue des candidats et des professionnels. L'étude approfondie de ces difficultés effectuée auprès de publics en situation d'illettrisme fait apparaître que les compétences que les candidats doivent savoir mobiliser afin que la valeur des savoirs de l'expérience puisse « se superposer » à la valeur des savoirs « objectivés » , objet des formation formelles.

Ces résultats permettent d'entrer dans une réflexion méta analytique sur le rôle du dispositif VAE dans la construction européenne. Si le diplôme ou le titre joue le rôle de filtre ou de signalement à titre individuel, la montée en certification de toute la population active ne peut viser une simple amplification de ces rôles. L'intérêt implicite est de faire émerger des compétences mobilisées dans l'activité économique, construites par l'expérience mais non référencées dans les certifications existantes bien que participant directement à la productivité. L'ensemble de ces savoirs d'expérience constitue le capital expérientiel collectif qu'il est nécessaire d'identifier.

Pour ce faire et sous couvert d'une logique de transparence des compétences, la VAE repose sur une idéologie de transparence du langage.

Problématique

La Validation des Acquis de l'Expérience en France, au vu des résultats nationaux donnés par les différents ministères ne produit pas les résultats escomptés en matière de montée en qualification de la population active, non diplômée. S'agit d'un effet « de jeunisme du dispositif » qui se cherche et cherche comment évaluer au mieux les acquis de l'expérience sans disposer à l'heure actuelle de moyens had-hoc, ou s'agit des effets d'un dispositif de certification, qui s'inscrit dans les mêmes logiques que

les autres dispositifs de certification, sélectif et socialement discriminant ? L'hypothèse est la suivante : en tension entre deux logiques, celle de la promotion sociale et celle de la sélection sociale, le dispositif de VAE est actuellement davantage un outil de reproduction sociale qu'un outil de transformation sociale. Il s'appuie sur le rapport à la langue des candidats, dont la maîtrise est un signe de distinction sociale.

Afin de comprendre les enjeux sous jacents de ce dispositif le travail de recherche s'est effectué en deux temps qui ont conduit à une réflexion d'ordre méta analytique.

1) Etre autonome : une condition minimum

Une étude régionale (Presse, MC, Delache, D, 2003,) portant sur les obstacles de la procédure de Validation des acquis, a conduit à porter un regard particulier sur le rôle de la langue dans ces processus. L'identification des difficultés, réalisées par l'analyse d'entretiens de candidats (n=18), permet de dire que les principales difficultés se situent dans l'articulation activités décrites – compétences apparentes- diplôme visé (Presse, 2004). Cette première recherche permet de dire, d'une part, que le dispositif de VAE est un dispositif de certification parmi d'autres et qu'il procède selon des logiques similaires à celles de tout dispositif de certification, tout en poussant à l'extrême les processus d'individualisation et de responsabilisation individuelle, sous couvert d'autonomie des candidats. Autonomie, responsabilité, motivation projection dans le futur, ce discours, qui préside au dispositif de la VAE voudrait donc convaincre que le déroulement de la procédure dépend de l'investissement de chacun, et laisse penser que le système offre à chacun la possibilité proportionnelle à son investissement de se faire reconnaître ainsi que celle de contrecarrer les inégalités dont chacun a pu être antérieurement victime, discours de l'égalité des chances qui repose donc de fait sur l'idéologie d'une autonomie subie et non d'une autonomie libératrice.. Les inégalités sont alors conçues comme la conséquence de la compétition entre les individus, égaux entre eux, inégalités découlant de leurs activités (Dubet, 2000). Ce discours « s'impose comme un discours rationnel en soi qui prétend révéler celui de la réalité sociale empirique Lefort (1986)». Hors, les travaux de recherche de Serge Paugam (2000) montrent que cette autonomie attendue, au travail, n'est profitable au salarié que lorsque les conditions sont réunies pour lui permettre de réaliser un programme pour lequel il a participé à l'élaboration. « Dans le cas contraire l'autonomie a pour conséquence de consacrer ses défaillances, voire ses défaillances, voire ses incompétences. Elle crée alors non pas de la satisfaction au travail mais de la frustration et des formes de détresse psychologique. Cette analyse peut se projeter dans le dispositif de la VAE ou, bien que le souhait de voir son expérience transformée en titre ou diplôme soit celui des candidats, le parcours qui l'attend est loin d'avoir été choisi par celui-ci et bien souvent, non seulement, il ne s'y attend pas mais de plus il n'est pas en mesure de prendre en charge ce parcours « du combattant » en solitaire. Ces attentes en matière d'autonomie peuvent permettre à certains « d'atteindre l'excellence et d'en être valorisés, mais elles conduisent en même temps à la disqualification sociale des salariés les moins aptes à relever le défi qu'on attend d'eux,

-d'autre part, que l'un des facteurs qui permet de contrôler les effets de ce dispositif sur la montée en qualification de ces personnes se situe dans le rapport à

langue des candidats et des professionnels. Ainsi certaines expériences professionnelles peuvent être transformées en qualification, d'autres non. Des compétences spécifiques au dispositif sont requises chez le candidat : être motivé, avoir un projet personnel, savoir parler de son activité et écrire sur celle-ci avec recul.

Afin d'affiner la compréhension de ce rapport à la langue le deuxième temps de recherche a porté sur l'étude des conditions nécessaires à la « réussite des parcours de Validation d'Acquis de l'Expérience » pour des publics en situation d'illettrisme, non qualifiés et donc concernés explicitement par le dispositif (Presse, 2005).

2) Savoir mobiliser les compétences gigognes : des pré requis supposés acquis

Les analyses montrent, que pour pouvoir répondre aux attentes du dispositif, il est présupposé que les candidats disposent de méta compétences, socialement acquises, signes de distinction sociale (Bourdieu, 1979, 1982), éliminant ainsi , les « in-compétents ».

Cependant cette mise en mots se heurte à des difficultés, de différents niveaux, qui ont été identifiées (Presse, 2006 a).

1) se sentir compétent

2) savoir utiliser le genre adapté à la situation

3) prendre de la distance et adopter une posture réflexive

4) mobiliser des compétences gigognes : compétences conatives, cognitives indispensables à la formalisation des connaissances, compétences habiletés et aptitudes, formalisation requise la transformation d'une expérience en titre ou diplôme.

Si ces compétences s'emboîtent les unes dans les autres, l'une d'entre elles, la compétence à utiliser le genre approprié, apparaît comme déterminante du processus de VAE.

En effet, on peut se sentir compétent et ne pas savoir le faire valoir, ne pas savoir choisir les mots appropriés, même lorsque l'on adopte une posture distanciée voire réflexive, et dès lors ne pas parvenir à formaliser de manière had hoc ses connaissances, compétences, expériences et autres composantes de la vie professionnelle et sociale, faute d'avoir su ou pu mobiliser le genre approprié.

3) Maîtriser le genre linguistique approprié : une condition décisive

Le processus d'évaluation est conditionné, au sein de ce dispositif, par le langage utilisé. Pour faire reconnaître ses compétences il ne suffit de s'auto déclarer compétent. Dans le cadre de la validation des acquis de l'expérience, un genre langagier spécifique est attendu. Il s'agit de produire un discours issu d'un processus d'analyse réflexive mis en œuvre par le candidat, processus donc de conceptualisation, qui devrait aboutir à une décontextualisation du savoir-faire, et permettrait de considérer la compétence -comme étant - adaptable et transférable à d'autres situations. (Pastré, 1999).

La situation de VAE est donc une situation d'évaluation au sein de laquelle le candidat est appelé à utiliser le genre linguistique approprié lui permettant au minimum de décrire les activités professionnelles. Cette description doit rendre les composantes de l'activité compréhensibles même par un professionnel de la VAE qui ne connaît pas l'activité. Un candidat qui a des facilités à l'oral, même s'il ne parvient pas à formuler et formaliser immédiatement son expérience, permettra au professionnel de la VAE de trouver des points d'accroche pour engager l'échange et faciliter la formalisation. Alors que cette possibilité est réduite lorsqu'un candidat synthétise son expérience en une formule : « j'fais d'la peinture » et qu'il répond, à la question de l'accompagnateur sollicitant quelques précisions, « ben d'la peinture vous voyez quoi ! » et ce en montrant un mur peint.

Cet exemple montre que le discours peut être chargé d'implicites et permet de préciser la notion du genre attendu.

4) Le genre utilisé au cœur de la démarche d'évaluation

Le langage est selon les situations plus ou moins chargé d'implicites. Dans une situation d'échange avec une personne affectivement un proche les implicites peuvent être nombreux et cela n'empêche pas la compréhension, permise par la connaissance réciproque des interlocuteurs les échanges, mais dans cette situation d'évaluation, afin d'être évalué à la hauteur « de sa valeur », être explicite, réflexif et utiliser le je s'imposent : « une des choses les plus difficiles à obtenir est que nos interlocuteurs parlent de leur travail à la première personne, qu'ils décrivent vraiment leur travail réel et non ce qui se passe en général sur des postes analogues à celui qu'ils ont tenu ».

Cependant :

- la capacité à utiliser le langage explicite dans une situation formelle est socialement déterminée comme Basil Bernstein(1975) l'a montré en mettant en évidence les codes restreints et élaborés,
- la capacité à construire un rapport réflexif au langage, attendue dans le dispositif Vae est, socialement, inégalement répartie (Lahire, 1995),
- « l'utilisation du je pour un OS, toujours selon les travaux de Basil Bernstein (1975), est quasi impossible lorsqu'il décrit son activité de travail alors qu' un cadre pourra aisément conceptualiser son activité et la mettre en discours sur ce mode » (Boutet, Gardin, 2001) .

La production d'un discours explicite et répondant aux attentes –implicites- n'est donc pas donnée en soi.

La distinction existante, sur le plan linguistique, entre le genre premier et le genre second (Bronckart, 1996) apporte un éclairage quant à la compréhension des difficultés que rencontrent les personnes les moins qualifiées, celles qui souvent ont développés des expériences de travail manuel.

Les genres premiers sont structurés par des activités non langagières (en actes) auxquelles ils s'articulent. Certaines de ces activités peuvent disposer d'un langage spécifique, qu'il peut donc être judicieux d'utiliser pour faire valoir sa connaissance du métier, mais fait courir le risque de ne pas être compris, ou d'être évalué comme

étant « trop collé » à son activité, discours ne permettant pas d'identifier des compétences généralisables à une classe de situations.

A l'inverse les genres seconds sont détachés des activités (rapport distancié) et font l'objet d'une structuration autonome. Ce genre second est celui qui est attendu dans la procédure.

Il s'agit donc d'opérer une transformation du discours tenu dans le genre premier sur le lieu de travail (lorsqu'il y a discours, ce qui n'est pas toujours le cas) en discours détaché des activités du genre second, transformation qui est le frein essentiel de la transformation de l'expérience en certification par la voie de la VAE chez ces personnes qui ne parlent pas de leur expérience en terme de compétences mais en terme d'activités réalisées.

Traduire l'expérience en mot pose donc la question des moyens (langage et pensée) dont chacun dispose pour la reconstruire, inégalement distribués sur un plan sociologique, mais pose également la question des modalités d'évaluation et des attentes implicites des interlocuteurs, qui déterminent « le niveau de diplôme envisageable », en fonction du genre linguistique utilisé. Ainsi pour l'obtention d'un Cap il est « peu tenu compte du niveau de maîtrise de français » disent les accompagnateurs… ce qui se traduit par « ce niveau de français ne vaut pas au delà du CAP » ou encore « puisque ce n'est qu'un Cap on ne tient pas compte du niveau de français… », le candidat, qui ne maîtrise pas « correctement » la langue française ne peut donc viser au delà de ce niveau de diplôme et ce quelque soit le niveau d'expertise dans l'activité professionnelle ou sociale développée…

Dès lors on peut s'interroger sur la volonté de promotion sociale de ce dispositif, n'a y –t-il pas enfermement dans une situation sociale et professionnelle dont il est difficile de sortir, pour certains candidats ceux pour lesquels au départ le dispositif était pensé?

5) Verbal et non verbal dans la VAE : prendre sa place dans un rapport social de communication

Si on considère que le langage est une action, dans le cadre de la situation d'évaluation VAE l'action langagière va dépendre du contexte à un double niveau :

-celui dans lequel s'est réalisé l'expérience, qui mobilise un genre discursif spécifique lié à la place occupée dans les rapports de production (un travailleur manuel n'occupe pas la même place qu'un travailleur intellectuel, ne possède pas le mêmes moyens linguistiques pour parler de son activité, ce qui amène certains des candidats à faire part de leur désarroi et de leur sentiment d'incompétence : « je ne trouve pas les mots ! », « comment dit-on cela ? », « j'aurai voulu qu'on m'aide ! ».

-celui de la situation dans laquelle s'expose cette activité, situation d'évaluation qui au delà de l'expérience évalue la personne, à travers le genre linguistique utilisé (premier ou second, et qui , dans les situations orales présentes dans tout le processus depuis le premier accueil jusqu'à l'entretien devant le jury, est associé à un ensemble d'indicateurs non verbaux , interprétés par les professionnels en fonction du rapport social de communication (Presse 2000, Todorov, Bakthine 1981). Dès lors en fonction de l'ensemble de leurs représentations les professionnels convoquent

l'interlocuteur à la place qui lui socialement assignée (Flahaut, 1979) . Le langage utilisé sera donc le reflets des rapports sociaux réunissant les interlocuteurs.

Penser que " l'écriture sur " ou le " parler sur " ses pratiques professionnelles engagent

d'emblée les deux interlocuteurs dans un « agir communicationnel » orienté vers le même but, ce qui permettrait très rapidement de s'accorder sur le sens de l'expérience décrite et sur les compétences mobilisées serait oublier que les systèmes de valeurs et de références portés par les protagonistes peuvent être antagoniques, puisqu'ils sont socialement différenciés. Hors cet agir communicationnel n'est possible que lorsqu'il y a, sinon partage des valeurs, du moins explicitation des valeurs de chacun.

Cet échange langagier est donc source d'évaluation en soi et, par delà le discours tenu, se mesure le niveau de maîtrise de la langue, et bien souvent, et peut être inconsciemment, le niveau social de la personne le tout déterminant le niveau de diplôme ou titre au quel il peut prétendre. Quel intérêt y a-t il alors à faire parler de l'expérience les travailleurs au-delà de la reproduction sociale que ce dispositif engendre ?

6) En guise de conclusion : La théorie sous jacente du capital expérientiel collectif

Dans le cadre de la construction européenne, le capital humain est considéré comme source de richesse potentielle. L'Europe s'est donné pour objectif de « rendre l'économie de la connaissance européenne la plus compétitive et la plus dynamique du monde avant 2010 ». La VAE entre dans le cadre du développement européen. Cependant ce dispositif ne provoque pas d'accroissement des connaissances et compétences par la seule identification des compétences maîtrisées, qui sont dans la théorie du capital humain considérées comme facteur de productivité.

La VAE exerce une première fonction de capitalisation des savoirs d'expérience non référencés, qui pourront être transmis par formation ensuite et venir enrichir le capital humain collectif. Mais ce dispositif profite davantage à ceux qui maîtrisent les prérequis implicites, méta-compétences permettant la formalisation, incontournables de la procédure. Hors ces méta compétences sont soient acquises préalablement par la formation soit acquises par imprégnation dans un milieu social qui le permet.

Si le diplôme ou le titre joue le rôle de filtre ou de signalement à titre individuel, la montée en certification de toute la population active ne peut viser une simple amplification de ces rôles. Elle a un intérêt affiché : permettre la mobilité des travailleurs. Le second intérêt, implicite, est faire émerger des compétences mobilisées dans l'activité économique, construites par l'expérience mais non référencées dans les certifications existantes. Or celles-ci participent directement à la productivité. L'ensemble de ces savoirs d'expérience constitue donc un capital individuel qu'il faut mettre au service du collectif. Une fois répertorié, il pourra donc être transmis sur le mode formel. L'ensemble constitue le capital expérientiel collectif.

Mais et parce qu'il est nécessaire d'identifier ce capital expérientiel collectif utile à la productivité, cela exclut de fait de la démarche tous ceux qui ne présente pas un intérêt au regard de cette productivité, les moins diplômés, les non diplômés ou les

personnes en situation d'illettrisme, les moins compétents donc et ceux qui sont de fait les plus éloignés de la formation, et qui seront les moins à même de profiter de ce nouveau capital. En conclusion de ce travail on peut donc dire que sous couvert donc d'une logique de transparence des compétences, la VAE repose sur une idéologie de transparence du langage, qui dissimule cet emboîtement de compétences et fait abstraction des rapports sociaux. Logique de « communication qui nourrit la fiction d'une parole vivante » dirait Claude Lefort, et « la représentation d'une démocratie accomplie ou la parole circulerait sans obstacle » ce qui conduit à l'exclusion de tous les « socio-incompétents ». (Presse 2006 b).

Références Bibliographiques

Bernstein, B. (1975). *Langages et classe sociales*. Paris : Editions de minuit.

Bourdieu P (1982). *Ce que parler veut dire*. Paris : Fayard.

Bourdieu P. (1979). *La distinction*. Paris : Editions de minuit.

Boutet Josianne, Gardin (2001), Bernard. Une linguistique au travail. In, Anni Borzeix et Béatrice Fraenkel (dir.) *Langage et travail. Communication, cognition, action*. Paris, CNRS Éditions, 95.

Bronckart, J.-P. (1996). Genres de textes, types de discours et opérations discursives. *Enjeux* n°37-38, 31-47. Namur.

Dubet, F. (2000). *Les inégalités multipliées*. Editions de l'Aube.

Flahaut, F. (1979). *La parole intermédiaire*. Paris : Seuil.

Lahire, B. (1995*). Tableaux de famille*. Paris : Gallimard.

Lefort C. (1986). *Essais sur le politique XIX XX siècles*. Paris :Seuil .

Lefort, C. (1978) *Les formes de l'histoire ; esquisse d'une genèse de l'idéologie.* Paris Gallimard.

Pastré, P. (1999). Travail et compétences : un point de vue de didacticien. *Formation Emploi, n° 67.*

Paugam, S. (2000). *Le salarié et la précarité. Les nouvelles formes de l'intégration professionnelle*. Paris PUF, Coll le lien social,

Presse, MC. (2000). *Non verbal et communication : pédagogie différenciée ou différenciation sociale ?* Lille : Septentrion, 2000, coll thèse à la carte.

Presse, MC., Delache, D. (2003), La validation d'acquis professionnel : des obstacles à franchir, des propositions à construire. *Cahier du CUEEP, n°53.* Université de LILLE1

Presse, MC (2004 d). Entre intention et réalité : les obstacles à la validation des acquis. *Education Permanente, n°158,* Les acquis de l'expérience, 141-152.

Presse, MC (2005).*VAE et imaginaire : logique instituée ou logique instituante ?* REF Montpellier. Paris : L'Harmattan, à paraître

Presse, MC. (2006 a). *Formaliser son expérience dans une démarche de validation : une question de compétences ?* Groupe RVAE Colloque Admee. Luxembourg septembre 06. 14 p. A paraître

Presse, MC. (2006 b). *Compétences, ethnocentrisme et exclusion.* Colloque CNAM, mars, Les usages sociaux de la notion de compétences. www.ensieta.fr/jecompetence/communication/Presse

Todorov T, Bakthine M (1981). *Le principe dialogique*. Paris : Seuil

LES PRATIQUES EVALUATIVES DES JURYS VAE DANS LES DISPOSITIFS DE L'ENSEIGNEMENT SUPERIEUR : QUELS PROCESSUS DE DECISION ? QUELS REFERENTS ?

Isabelle Cherqui-Houot, Université Henri Poincaré, Nancy
Paul Nkeng, Université Louis pasteur**,** Strasbourg, LISEC

Mots clés : jurys, VAE, pratiques d'évaluation

Résumé:
La démarche de validation des acquis de l'expérience (VAE) impose aux jurys universitaires non seulement une composition élargie à des professionnels externes mais aussi et plus encore la nécessité de repenser leurs pratiques évaluatives.

Dans cette contribution, nous nous intéresserons particulièrement aux pratiques d'évaluation développées par les membres de jurys (universitaires et professionnels) de VAE, les objets qu'ils définissent, les indicateurs de performance qu'ils relèvent, les référents qu'ils mobilisent, les critères qu'ils sont amenés à construire, la manière dont sont prises les décisions.

Les matériaux ont été recueillis à l'occasion de la tenue de cinq jurys : enregistrement intégral des jurys et entretiens post-jurys.

Les premiers résultats font état de disparités dans les pratiques dont on peut penser qu'elles sont liées, à la fois à des conceptions d'acteurs et des mises en œuvre institutionnelles des procédures sensiblement différentes, et à des processus de décision fondés sur des registres originellement distincts mais susceptibles de fortes variations provoquées par la démarche particulière que constitue la VAE.

Problématique

La démarche de validation des acquis de l'expérience (VAE) impose aux jurys universitaires non seulement une composition élargie à des professionnels externes mais aussi et plus encore la nécessité de repenser leurs pratiques évaluatives.

Car si l'évaluation d'acquis développés dans un parcours formel de formation universitaire et l'évaluation d'acquis issus d'expériences diverses développées au long d'un parcours professionnel et personnel sont a priori deux processus qui portent sur les mêmes objets (c'est à dire des acquis issus d'un apprentissage : des connaissances, des aptitudes, des qualités générales et professionnelles, ce sont en revanche les conditions de "formation" (au sens large du terme) de ces mêmes objets ainsi que leurs formes apparentes qui sont divergentes et appellent ainsi à des processus évaluatifs relativement inédits.

L'évaluation d'acquis à l'issue d'une formation formelle s'opère à partir d'une performance instituée, définie et contrôlée par l'enseignant (examen, contrôle, entretien ...) : l'*épreuve*. Elle est menée en référence à des critères et/ou des normes liés à des objectifs d'apprentissage. On présume, à partir des résultats obtenus, de la

capacité de l'étudiant à exercer ses connaissances et aptitudes dans d'autres situations que celles de l'épreuve.

L'évaluation d'acquis issus de l'expérience s'opère à partir de l'observation de *preuves* fournies par la personne évaluée et qui constituent des traces formalisées de ses connaissances et compétences acquises en d'autres lieux. Elle est menée en référence à des critères liés à la mobilisation effective d'aptitudes en situation. On présume, à partir d'un "feed-back" sur le parcours de la personne, des connaissances et compétences effectives de la personne.

Dans cette contribution, nous nous intéresserons particulièrement aux pratiques d'évaluation développées par les membres de jurys (universitaires et professionnels) de VAE, les objets qu'ils définissent, les indicateurs de performance qu'ils relèvent, les référents qu'ils mobilisent, les critères qu'ils sont amenés à construire, la manière dont sont prises les décisions.

Démarche méthodologique et matériaux

La recherche a été menée dans cinq établissements universitaires (universités et institut polytechnique). Elle repose sur le constat suivant : alors que ces établissements affichent des procédures de validation des acquis de l'expérience communes ou très proches, les pratiques évaluatives y apparaissent cependant sensiblement différentes d'un établissement à un autre. Au sein des établissements, de la même manière, les jurys semblent généralement peiner à formaliser les contours et les attendus de leurs pratiques, les procédures et critères d'évaluation apparaissent mouvants. Cela peut supposer des pratiques très différentes d'un jury à un autre.

Que produisent comme effets différenciés les modes d'organisation et de fonctionnement des jurys ? En quoi et comment ces derniers orientent-ils les décisions qui sont à prendre (attribuer un diplôme, en totalité ou partiellement)?

Nous faisons l'hypothèse que ces effets tiennent autant aux manières différentes qu'ont eu les différents acteurs impliqués dans la VAE (conseillers VAE, administratifs, enseignants…) de concevoir, s'approprier, et partager les procédures générales qui ont été proposées, qu'à la grande diversité des référents potentiellement mobilisables par les jurys pour élaborer et formuler leur décision.

Il s'agit de montrer comment se structurent les pratiques d'évaluation dans les dispositifs de VAE propres à chaque établissement, les activités sur lesquelles elles reposent tout au long du processus de prise de décision, les objets pris en compte, les référents mobilisés, les instruments utilisés.

Il s'agit également de mettre en évidence les événements qui scandent la lecture du dossier et le moment de l'entretien du jury avec le candidat et ceux qui influent particulièrement sur le processus de décision.

Il s'agit enfin de d'observer les signes des représentations, et surtout des heurts et mouvements dans les représentations induits par la pratique d'évaluation : les conceptions du savoir académique, des savoirs d'expérience…

Les matériaux ont été recueillis à l'occasion de la tenue de cinq jurys (un par établissement). Chaque entretien du jury avec le candidat a fait l'objet d'un enregistrement intégral (à l'exception du moment des délibérations) et a été suivi d'entretiens avec trois de ses membres (le président de jury, un enseignant, un professionnel).

L'enregistrement de l'entretien entre le candidat et les membres du jury permet de recueillir les échanges qui président à l'évaluation du candidat : les questions posées par le jury et les réponses données par le candidat. L'objectif est de mettre en évidence la manière dont procède le jury pour interroger les acquis du candidat. La nature des questions posées et la façon dont elles sont formulées peuvent permettre de déterminer comment le jury cherche à évaluer ces acquis et les critères retenus pour le faire. Les réponses du candidat permettent quant à elles de montrer comment il justifie de ses acquis. Les interactions verbales entre membres des jurys permettent de rendre compte de la dynamique de l'évaluation conduite.

Les entretiens qui succèdent ont une fonction d'explicitation, il s'agit de mettre en évidence, la manière dont chacun de ces membres décrit et justifie a posteriori le jugement qu'il a porté sur les acquis du candidat (depuis son premier contact avec la candidature jusqu'à la délibération) ainsi que la décision qui a été prise collectivement.

Premiers résultats

Nous observons d'ores et déjà des pratiques de jury très différentes selon les établissements où le candidat dispose dès le début de la réunion du jury de quelques minutes pour présenter son dossier (façon début de soutenance de mémoire) suivies par une discussion "à bâtons rompus" avec les membres du jury et d'autres, où le candidat est questionné par les membres du jury après un examen approfondi de son dossier hors sa présence. Les différents temps et temporalités de la prise de décision et de sa négociation à l'aune individuelle et à l'aune du collectif (à la lecture du dossier, en présence ou hors présence du candidat, en présence ou hors présence des pairs) apparaissent comme des déterminants importants. De même l'attention portée, notamment lors de l'entretien avec le candidat, aux projets professionnels du candidat introduisent dans le jugement porté cette autre dimension du temps qu'est l'anticipation (le futur au présent).

Outre les éléments nommément pris en compte et qui se rapportent au dossier : les réalisations, le parcours, les méthodes, les connaissances, les compétences … dont l'évocation sera à explorer de manière approfondie, d'autres points saillants retiennent dés maintenant l'attention des observateurs : l'importance extrême accordée à "l'investissement personnel" des candidats ainsi que la demande récurrente qui leur est faite de s'auto évaluer, deux constantes dont il convient d'interroger les fondements.

Enfin, le "retour réflexif" sur les délibérations menées et les décisions prises que suscitent les entretiens proposés, met au jour les registres (académiques, professionnels …) à partir desquelles les membres des jurys fondent leurs interventions, laisse apparaître tant des pesanteurs (les habitus liés aux fonctions occupées notamment) que des variations parfois inattendues (il est fréquent que le professionnel se "mette à la place" de l'enseignant universitaire et vice versa).

REGARDS CROISES SUR L'EVALUATION DANS LE PROCESSUS VAE : POINTS DE VUE DE CANDIDATS ET DE CONSEILLERS VAE A L'UNIVERSITE

Brigitte PAGNANI, Frédérique FEBVRE-BUISSON, Université Henri Poincaré Nancy (CUCES)

Mots clés : Espace(s) d'évaluation- professionnalisation – compétences

Résumé:

La VAE, porteuse d'enjeux multiples dans un environnement socio-économique en mutation, constitue une pratique sociale naissante qui, à ce titre, ne manque pas d'interroger l'ensemble des acteurs. Elle les questionne sur le registre de l'évaluation qui s'impose tout au long du processus et sur celui de la professionnalisation, du "professionnel-conseiller VAE" comme du "professionnel-candidat VAE", qui s'inscrivent l'un et l'autre dans une démarche de professionnalisation finalisée par la production de compétences que l'on peut qualifier de "nouvelles" soit parce qu'elles n'auront pas été jusque-là reconnues soit parce qu'à la faveur de ces activités, surgissent des compétences inédites.

Il importe aujourd'hui de faire aboutir une large réflexion, initiée dans le cadre de notre recherche, sur ce que nous nommons un "espace d'évaluation" parce qu'une multitude d'acteurs y est impliquée et engagée. Notre expérience de"conseiller VAE" à l'université nous a permises d'accéder à certaines "parcelles" de cet espace en friches dans lequel circule le candidat lui-même.

Un corpus de matériaux d'enquête doit bientôt être constitué et sera présenté en partie aux congressistes. L'exploitation fine de ces données collectées et l'analyse qui s'en suivra devront permettre de mettre au jour les éléments constitutifs de cet espace d'évaluation au fil du déroulement du processus et de la procédure VAE. L'analyse de contenu des entretiens doit pouvoir aussi répondre à un certain nombre de questions liées à l'instrumentation de la VAE. Elle doit enfin s'ouvrir sur des perspectives d'outillage méthodologique des acteurs en présence dans une démarche de co-production de compétences individuelles et collectives si l'on s'oriente, dans le cas de la VAE, vers une culture de la professionnalisation.

Présentation

La VAE, porteuse d'enjeux multiples dans un environnement socio-économique en mutation, constitue une pratique sociale naissante qui, à ce titre, ne manque pas d'interroger l'ensemble des acteurs qui contribue aujourd'hui à son développement.

Notre pratique professionnelle de conseiller VAE à l'université et nos activités de recherche en parallèle nous amènent à penser, à la fois de l'intérieur et de l'extérieur,

la construction de cette nouvelle pratique sociale et à porter un regard nouveau sur la question de l'évaluation qui y est centrale.

Que l'on se place du point de vue du candidat, de celui des accompagnateurs, des membres de jury, de l'institution certificatrice ou bien du milieu professionnel et de l'environnement, l'acte d'évaluation est une constante et s'impose tout au long du processus pour l'obtention ou l'attribution du diplôme *in fine*.

Nous proposons ici une réflexion dans la double posture du praticien-chercheur qui réfléchit sur son activité en même temps qu'il la produit, en charge ici de proposer au candidat un exercice de remémoration et de formalisation de ses activités (antérieures et en cours) sous une forme censée permettre l'acte d'évaluation pour la délivrance du diplôme visé.

A notre sens, qu'il s'agisse du "professionnel-conseiller VAE" ou du "professionnel-candidat VAE", ce qui est en jeu au cœur de la VAE est précisément la démarche de professionnalisation inscrite dans une logique de production de compétences que l'on peut qualifier de "nouvelles" soit parce qu'elles n'auront pas été jusque-là reconnues soit parce qu'à la faveur de ces activités, surgissent des compétences inédites. Du fait qu'elle valide des compétences construites dans des contextes et des environnements qui sont extérieurs au périmètre des institutions certificatrices, la VAE se situe explicitement dans une culture de la professionnalisation dont la référence centrale est la compétence (J.-M. BARBIER, 2005).

Le procès d'évaluation se complexifie singulièrement d'autant que les règles ne sont pas pré-établies au sein de l'institution certificatrice ou, si elles le sont, certaines apparaissent déjà inopérantes. Elles s'élaborent donc pas à pas, généralement dans des transactions qui peuvent être difficiles, parfois inopinément, et dont le candidat, malgré tout, s'il n'en fait pas les frais peut réussir à en tirer parti de façon presque hasardeuse.

Ne serait-ce que par souci d'équité entre les candidats, il importe de faire aboutir une large réflexion, initiée dans le cadre de notre recherche, sur ce que nous nommons un "espace d'évaluation" parce qu'une multitude d'acteurs y est impliquée et engagée. Notre expérience de conseiller VAE à l'université nous a permises d'accéder à certaines "parcelles" de cet espace en friches dans lequel circule le candidat lui-même.

Un corpus de matériaux d'enquête doit bientôt être constitué. Il s'appuie sur des entretiens semi-directifs approfondis qui vont être conduits auprès d'une quinzaine de candidats à la VAE ayant validé partiellement ou totalement leurs acquis d'expérience. L'exploitation fine de ces données collectées et l'analyse qui s'en suivra devront permettre de mettre au jour les éléments constitutifs de cet espace d'évaluation au fil du déroulement du processus et de la procédure VAE, de repérer les situations, les lieux et les moments dans lesquels l'évaluation s'est posée et se pose de façon critique du point de vue du candidat. L'analyse de contenu des entretiens doit pouvoir aussi répondre à un certain nombre de questions liées à l'instrumentation de la VAE (cf. le dossier), à la position et à la posture du candidat dans les différentes situations dans lesquelles il est exposé et auxquelles ils s'exposent tout au long du procès d'évaluation (et de formation quand il y a prescriptions de formation par le jury). Elle doit enfin s'ouvrir sur des perspectives d'outillage

méthodologique des acteurs en présence dans une démarche de co-production de compétences individuelles et collectives si l'on s'oriente, dans le cas de la VAE, vers une culture de la professionnalisation.

LA QUESTION DE L'EVALUATION DANS UN DISPOSITIF UNIVERSITAIRE DE TRAITEMENT DES DEMANDES DE VAE.

Pascal TISSERANT, Cynthia SELLIC , Université Paul Verlaine Metz

Mots clés : VAE, structure organisationnelle, enseignement supérieur, enquête

Résumé

Depuis la promulgation de la loi de modernisation sociale en janvier 2002, le législateur a laissé aux universités françaises une autonomie importante dans la conception du dispositif de traitement des demandes de VAE relevant de leur établissement. La question de l'évaluation est au cœur de ces dispositifs qui organisent les activités de validation des acquis de l'expérience.

Cette communication proposera une analyse d'un dispositif en particulier, celui de l'université pluridisciplinaire Paul Verlaine – Metz, à partir de trois enquêtes réalisées après 3 années de pratique de VAE:

- une série de 20 entretiens semi-directifs réalisés auprès de responsables politiques de l'université et auprès des principaux acteurs de terrain en matière de VAE : cette première enquête portait sur les représentations du dispositif, ses dysfonctionnements et ses perspectives d'évolution ;
- un questionnaire à destination de tout salarié de l'université ayant été impliqué au moins une fois dans le dispositif VAE : 38 acteurs sur les 98 concernés ont répondu à cette enquête cherchant à analyser les représentations et les pratiques de la VAE dans l'établissement ;
- un questionnaire à destination des candidats ayant saisi une demande de pré-dossier VAE : 67 candidats sur 206 personnes contactées ont répondu à cette enquête de satisfaction.

L'analyse de ces données s'arrêtera plus particulièrement sur la façon dont l'évaluation est organisée, pratiquée, perçue ou encore souhaitée tout au long du dispositif de traitement des demandes de VAE, en tenant compte, en particulier, du statut des répondants: les candidats (en fonction du stade d'avancement ou d'abandon dans la procédure), les évaluateurs (lors de l'avis pédagogique, dans l'accompagnement mais aussi au moment du jury) et les responsables politiques de l'université.

Introduction

L'instauration de la VAE avec la loi dite de ''modernisation sociale'' du 17 janvier 2002 s'inscrit dans la poursuite de la loi de 1984 et du décret 1985 permettant à un candidat qui en fait la demande, d'accéder à un diplôme de l'enseignement supérieur, en faisant valider ses acquis professionnels ou personnels (VAP ou VAPP). Avec le décret de 1993, une partie du diplôme pouvait être obtenue sur la base de cette validation, obligeant tout de même, dans le cas le plus favorable, le candidat à se

présenter aux évaluations traditionnelles organisées par les enseignants intervenants dans l'unité manquante. En donnant la possibilité de délivrer totalement un diplôme, la VAE introduit un changement dans les pratiques d'évaluation des certificateurs et bouleverse d'une certaine façon l'ensemble de l'organisation.

Nous proposons dans ce texte, de décrire le contexte national de la VAE en montrant que la mise en conformité de ces pratiques d'évaluation n'a pas été définie précisément par le législateur qui fixe à partir de quelques décrets les grandes lignes du dispositif à concevoir pour traiter les demandes de VAE. Les acteurs de la formation continue, le plus souvent minoritaires dans les universités, sont devenus une ressource pour beaucoup d'établissements dans la mise en œuvre de la procédure reposant sur trois phases offrant des disparités d'une université à l'autre. Dans un deuxième temps, nous ferons le choix d'observer la VAE sous l'angle de la structure organisationnelle adoptée pour gérer la VAE dans ces établissements d'enseignement supérieur. Après avoir mentionné les principaux facteurs de disparités à ce sujet, nous nous attarderons alors sur le cas de l'Université Paul Verlaine – Metz. La troisième et dernière partie, rapporte les résultats d'une enquête menée dans cette université après trois années de fonctionnement. Elle montre qu'au delà d'une satisfaction majoritairement partagée d'importantes difficultés subsistent à la fois pour les candidats et chez le personnel.

1) Le contexte national de la VAE

1.1) L'expérience de la formation continue au service du développement de la VAE dans les universités

Le décret principal concernant les universités est le n°02-590 du 24 avril 2002[1] fixant *''les conditions de validation des acquis de l'expérience par un établissement d'enseignement supérieur''* (article 1). Après avoir donné les conditions d'éligibilité des publics, les articles suivants de ce décret abordent l'évaluation en évoquant, d'une part un dossier à réaliser et, d'autre part, en fixant quelques règles de composition du jury en spécifiant qu'il doit s'entretenir avec le candidat. Le législateur n'a pas souhaité préciser davantage les contours du dispositif d'évaluation à concevoir et laisse le soin *''au conseil d'administration ou à l'instance qui en tient lieu de fixer les règles communes de validation des acquis de l'expérience par l'établissement et de constitution des jurys de validation''* (article 5). Par conséquent, avec la VAE, les universités ont conservé leur autonomie pour faire face à l'obligation du changement à introduire sans bénéficier pour autant de moyens supplémentaires de la part de l'Etat.

Dans la plupart des universités, la mission de mise en œuvre de la VAE est revenue à des acteurs de la formation continue, comme le démontre le premier ouvrage dans ce domaine consacré à l'enseignement supérieur (Cherqui Houot, 2001). Ceci s'explique pour deux raisons principales. La première concerne l'association relativement étroite du monde de la formation continue à l'élaboration de la loi et des décrets relatifs à la VAE. La seconde touche à l'expérience des personnels des

[1] Décret paru au Journal Officiel du 26 avril 2002

universités, habitués à gérer les dossiers de validation des acquis relatifs aux décrets de 1985 et 1993. La VAE a constitué le cœur de la réflexion de plusieurs colloques annuels de la conférence des directeurs de services de formation continue universitaire qui assure une mise à jour de l'information sur la VAE dans les universités [2]. Précédent de quelques années la promulgation de la loi de modernisation sociale, cette instance a organisé une coordination nationale dédiée à la VAE qui est devenue une force de propositions pour la mise en œuvre de la VAE, proposant par ailleurs et jusqu'à une date récente, des formations sur plusieurs jours et dans toute la France, à destination de différents acteurs de la VAE, comme les accompagnateurs ou encore les membres de jury. Ces moyens de communication avaient pour objectif de dépasser le peu d'engouement suscité par la validation des acquis relatives aux décrets de 1985 et 1993. En effet, contrairement à certains certificateurs dont l'essentiel de l'activité concerne la formation continue, dans les universités françaises le public en reprise d'études constitue une minorité face aux jeunes étudiants de formation initiale pour qui le système a été conçu. En changeant les pratiques d'évaluation pour s'adapter à un type de public particulier, c'est l'ensemble de la structure universitaire qui est touchée par la VAE. Les acteurs de la formation continue universitaires ont donc souvent joué un rôle important dans ce changement qui laisse néanmoins subsister une diversité de dispositifs sur le territoire national, allant des pratiques d'accueil ou de jury, en passant par la tarification. Parallèlement à ce type de regroupement, on peut signaler l'organisation d'une réflexion par filière centrée sur les objets et les outils de l'évaluation, appliquée à l'informatique à propos des DUT ou, plus récemment, à la psychologie concernant le grade de licence.

1.2) La disparité dans le fonctionnement de la VAE dans les universités

Au delà de la diversité des dispositifs et des structures organisationnelles les supportant, on peut distinguer dans la procédure VAE trois phases mises en œuvre par les universités pour traiter les demandes de VAE : l'accueil, l'accompagnement et le jury.

1. – L'accueil: il renvoie généralement à trois types d'activités menées par les personnels des universités :

(i) évaluer le projet du candidat (déclaré ou potentiel) pour l'informer sur la spécificité de VAE

(ii) évaluer de façon exhaustive, mais non approfondie, la trajectoire socio-professionnelle du candidat pour lui conseiller le diplôme le plus pertinent

(iii) évaluer de façon plus précise des éléments de la trajectoire socio-professionnelle pour estimer la faisabilité du projet de VAE, souvent appelée avis de recevabilité.

Contrairement aux deux premières activités, cette dernière évaluation est souvent pratiquée par des enseignants, dans certains cas le responsable du diplôme seul et dans d'autres établissements par des commissions. Les deux premières activités d'évaluation sont couramment réalisées par du personnel œuvrant dans les services de

[2] Ces informations sont accessibles à l'adresse : www.fcu.fr et nous nous appuierons en partie sur ces données dans notre analyse de la VAE dans les universités.

formation continue, parfois des services d'information et d'orientation, voire des scolarités. Elles ont créé la fonction de conseiller VAE, le plus souvent en complément d'autres fonctions exercées par l'agent. Le conseiller peut pratiquer ces activités dans le cadre d'entretiens individuels mais aussi de séances collectives. Elles peuvent également avoir lieu à l'extérieur de l'université, au sein du réseau de points relais conseil (PRC). En 2004, 70 000 personnes se sont ainsi adressées aux PRC débouchant dans moins de la moitié des cas sur une orientation vers une VAE (Cereq, 2006). Dans tous les cas, et dès cette première phase, la VAE a obligé les universités à mettre en place des outils (plaquette, site Internet, …) et/ou des lieux d'accueil, pour tenter de répondre à ce type de demande renforçant ainsi le besoin de compétence du personnel des universités dans le domaine de l'information et du conseil en orientation.

2. L'accompagnement : cette phase, non prévue par le législateur, revêt dans les universités un caractère facultatif entretenant souvent une forte proximité avec le caractère obligatoire. En effet, après la phase d'accueil, le candidat souhaitant poursuivre son projet de VAE doit désormais, comme le spécifie le décret du 24 avril 2002, rédiger *''un dossier qui doit expliciter par référence au diplôme postulé, les connaissances, compétences et aptitudes qu'il a acquises par l'expérience''* (article 4). La complexité de cette activité conduit les universités à proposer au candidat un accompagnement pour mener à bien cette tâche débouchant sur l'entretien avec les membres du jury *''sur la base du dossier présenté ''* (article 6). La fonction d'accompagnateur revient à un enseignant du diplôme ou à un conseiller ayant acquis une expertise dans le domaine de l'explicitation de l'expérience ; elle peut-être également partagée entre ces deux types de personnel. Selon les universités et les circonstances, par exemple dans le cas de plusieurs candidats à un même diplôme, cette activité d'accompagnement peut se dérouler en groupe. Le recours à des outils tels que le référentiel de compétences du diplôme, quand il existe et lorsqu'il a été conçu pour cette utilisation (Pons Dessoutter, 2005), peut faciliter cette activité d'évaluation nouvelle et parfois éphémère, pour beaucoup d'accompagnateurs. Pour dépasser ces difficultés, certains auteurs (Aubret, 2003) préconisent le recours a des techniques d'évaluation de l'expérience, comme l'entretien d'explicitation (Vermersch, 1996) ou encore l'analyse clinique de l'activité (Clot, 1999).

3. Le jury : il constitue l'aboutissement d'une procédure ayant mobilisée en moyenne une dizaine d'acteurs de l'université et souvent vécue par les candidats comme étant *''complexe, longue, à l'issue incertaine''* comme le titre un rapport d'une enquête, réalisée par le Cereq (2005) à propos de l'ensemble des certificateurs. La composition du jury est grossièrement fixée par le législateur qui laisse aux établissements le soin d'arrêter plus précisément les modalités de cette dernière évaluation débouchant sur une prise de décision quant à la délivrance totale ou partielle du diplôme et, dans ce dernier cas, sur les prescriptions à suivre pour obtenir le diplôme dans sa totalité ultérieurement. En 2004, les candidats à la VAE avaient en moyenne une chance sur deux d'obtenir le diplôme dans sa totalité, pour tous les certificateurs confondus et avec de fortes variations entre eux . Dans ce rapport du Cereq (2006), avec 28% de réussite, les universités apparaissent les certificateurs les plus exigeants, notamment au regard des 61% réalisés par l'enseignement secondaire.

Par ailleurs, les auteurs de ce rapport, notent de fortes différences d'une université à l'autre, illustrant, d'une certaine façon, la diversité des pratiques d'évaluation et des dispositifs dans lesquels elles s'opèrent au-delà de ces trois phases communes.

Une analyse des dispositifs VAE des universités du grand ouest (Benmoussi, 2005), relève de fortes disparités tout au long de ces trois phases, concernant en particulier, le degré de formalisation de l'avis de recevabilité, le niveau de structuration du dossier, le soutien apporté lors de l'accompagnement et le niveau d'exigence des jurys au moment de l'entretien avec le candidat mais également lors des délibérations ou encore dans les prescriptions en cas de validation partielle.

Ces disparités ne s'arrêtent pas là comme le démontre le dernier recensement réalisé par la conférence des directeurs de formation continue universitaire à propos de la tarification. Le prix moyen d'une VAE (hors inscription à l'université) se situe aux alentours de 700 euros pour l'ensemble de la procédure avec une variation allant de 75 euros à 1620 euros pour 55 universités aux données exploitables. Cette tarification fluctue également en fonction de certains critères retenus par certaines universités : le statut du candidat (demandeur d'emploi, individuel ou entreprise), la prise en charge (par les OPCA ou encore la Région) et dans quelques cas, le volume de la certification demandée avec l'exemple de 10 euros par ECTS à l'Université de Grenoble 1. La diversité des procédures ne suffit pas à expliquer la diversité des tarifs appliqués. Ce constat appelle avant tout une question: celle du coût de la VAE, en particulier celle de l'évaluation et des modalités de sa prise en charge. Autrement dit, au-delà de la performance des instruments de l'évaluation existants ou à développer, les certificateurs ont-ils toujours les moyens humains et financiers de réaliser des évaluations de qualité ?

Enfin, une des dernières, et probablement plus importante disparité à souligner dans le domaine de la mise en œuvre de la VAE, concerne la structure organisationnelle permettant de répondre à cet ensemble de nouvelles activités qui s'imposait de par la loi, aux universités sans qu'elles bénéficient en contrepartie de moyens humains, matériels ou encore financiers supplémentaires, à l'exception des financements FSE en réponse à des appels à projets et dont le remboursement aujourd'hui des sommes engagées n'est pas encore acquis.

2) Le dispositif VAE comme structure organisationnelle : le cas de l'UPV – Metz

2.1) les facteurs de disparité dans les structures organisationnelles en charge de la VAE

La VAE dans l'université ne concerne pas seulement le responsable d'un diplôme en particulier ou encore une personne chargée de l'accueil des publics en reprise d'études mais un ensemble d'acteurs administratifs, enseignants et enseignants-chercheurs allant du premier accueil téléphonique jusqu'au Président. Son introduction dans ces établissements pose donc la question de la structure organisationnelle mise en place pour faire fonctionner la VAE.

S'appuyant sur les travaux de Mintzberg (1982), Louche (2000) caractérise les structures à partir d'une série d'éléments pouvant être résumée ici en quatre facteurs

dont la plupart permettent de rendre compte des disparités dans le fonctionnement de la VAE dans les universités.

1. La spécialisation : le degré de spécialisation, des personnes mais aussi des services en charge de la VAE varie d'une université à l'autre. Dans certaines, une cellule à été spécialement créée, parfois volontairement éloignée des services de formation continue, prenant en charge les trois phases de la VAE, de l'accueil de toute personne souhaitant faire une VAE, jusqu'à l'organisation des jurys. Dans d'autres universités, le service de formation continue, parfois avec l'appui du service d'information et d'orientation, a été réorganisé donnant lieu par exemple, à des bureaux REVA (reprise d'études et validation d'acquis) où la spécialisation des postes de travail n'est pas centrée sur la VAE mais plus globalement sur le public en reprise d'études.

2. La standardisation et la formalisation des règles et des procédures: la latitude laissée aux conseils d'administration par le législateur en matière de définition des règles, donne lieu ici aussi, à une forte disparité d'une université à l'autre. Certaines ont précisé quelques points du décret laissant aux acteurs de la VAE une marge de manœuvre quant aux interprétations possibles et d'autres ont standardisé, puis avec le temps formalisé, des procédures précisant, par exemple les modalités de prise de décision des jurys (vote à bulletin secret, à main levée, à la majorité, à l'unanimité, etc…), ont édité des documents précisant le fonctionnement de la VAE (fiche de poste des personnes en charge de la VAE, procédure administrative des modalités financières, etc..°). En l'absence de données précises sur l'étendue de cette disparité, on peut tout de même faire l'hypothèse que la standardisation et la formalisation se soient davantage développées dans les universités retenues pour des financements FSE qui les obligent à justifier très précisément les sommes engagées au regard des actions entreprises.

3. Le regroupement en unités: l'introduction de la VAE a réactivé une opposition classique dans les universités : les acteurs en charge de la délivrance des diplômes et ceux dont l'activité repose sur le public en reprise d'études. Par conséquent, les logiques de regroupement d'unités se sont opérées en partie autour de cette tension, parfois dans la rupture avec les unités spécialisées avec ce type de public minoritaire dans les universités ou, et à l'opposé, comme certains le dénoncent, *''pour une prise de pouvoir des services de formation continue''* (Agulhon, 2006). Entre ces deux alternatives radicales, des universités ont réalisé des regroupements adaptés aux spécificités locales, favorisant parfois des rapprochements entre des types d'acteurs qui s'ignoraient jusque là.

4. La prise de décision, la planification et le contrôle : le rôle du conseil d'administration dans la définition des règles et celui du chef d'établissement dans la notification finale au candidat spécifiés dans le décret, respectent d'une certaine façon, le fonctionnement de l'université en termes de prise de décision. Cependant, la disparité à propos de la standardisation, conduit à une plus ou moins grande autonomie des composantes, des équipes pédagogiques ou encore des acteurs chargés de délivrer les diplômes par la VAE. Ce niveau d'autonomie ou de standardisation à l'échelle de l'université dépend, en grande partie de l'implication du Président et/ou des Vice Présidents du CA (conseil d'administration) ou du CEVU (conseil des études et de la vie universitaire). Une forte implication de leur part assure un minimum de centralisation concernant la prise des décisions et une relative cohérence du système à l'échelle de l'université. Cette différence de niveau d'implication des Présidents et/ou Vice

Présidents se traduit aussi par une disparité en matière de planification et de contrôle, certaines universités fonctionnant sans véritable politique et projets en matière de VAE, alors que d'autres deviennent des acteurs de premier plan, au niveau local, régional, voire national. Souvent précurseur dans le domaine de la validation des acquis, ces universités où les Présidents et/ou Vice Présidents associent les actes à leurs discours en matière de VAE, constituent des relais importants pour les autres certificateurs.

2.2) le cas de l'Université Paul Verlaine - Metz

2.2.1) Le contexte local de développement de la VAE :

Avec 15000 étudiants, l'université Paul Verlaine – Metz est une université pluridisciplinaire organisée aujourd'hui autour de trois instituts universitaires de technologies (IUT de Metz, de Thionville et de Moselle-Est) et six unités de formation et de recherche :

- Droit, économie et administration : UFR DEA
- Etudes en Supérieures en Management : UFR ESM
- Lettres et langues : UFR LL
- Mathématiques, informatique, mécanique et automatique : UFR MIM
- Sciences fondamentales et appliquées : UFR SciFA
- Sciences humaines et arts : UFR SHA

L'arrivée de la VAE en 2002 a été anticipée dès 1999 par la réponse à un appel à projet du ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche qui cherchait à développer la formation continue dans les universités. Quatre postes de maître de conférences dédiés à la formation continue ont été obtenus, dont un chargé d'harmoniser et de développer la pratique de la validation d'acquis. En effet, à cette époque, le service commun de formation continue était embryonnaire et peu en capacité de fédérer des projets à l'échelle de l'université. L'expérience en formation continue reposait sur quelques composantes actives dans ce domaine (principalement l'IUT, l'UFR ESM et l'UFR MIM) qui s'étaient organisées, chacune à leur manière, pour faire appliquer les décrets de 1985 et 1993, ignorés ou mal utilisés par les autres composantes. En février 2000, le recrutement de l'enseignant-chercheur chargé du développement de la validation des acquis, pour une partie de son service, a coïncidé avec les premières réunions consacrées à la VAE, organisées par la direction de l'enseignement supérieur puis par la conférence des directeurs de services de formation continue universitaire. Ce chargé de mission a eu pour objectif de réunir au sein d'un groupe de travail des acteurs de la formation continue de l'université souhaitant construire un dispositif commun de VAE à partir de l'expérience et des spécificités des composantes engagées.

Parallèlement à ce rapprochement en interne d'acteurs de la formation continue, la réponse commune à l'appel à projets financés par le fond social européen (FSE) a renforcé la collaboration avec les trois autres universités de la région (l'Université Henri Poincaré, l'Université Nancy 2 et l'INPL : Institut National Polytechnique de Lorraine). Ce projet a permis non seulement soutenir le financement de postes ou d'actions visant le développement de la VAE mais également d'instituer un comité de pilotage en quête de standardisation régionale. Le développement d'un site commun

et d'un logiciel d'aide au suivi des candidats[3] mais également d'un logiciel d'aide à la décision quant au choix du diplôme en amont de la procédure, les tentatives de rapprochement en matière de procédure ou encore de tarification ou encore des actions de communication communes ont contribué au rapprochement des services de formation continue de ces quatre universités.

Après quelques VAE expérimentales menées par une UFR (l'ESM), le 18 novembre 2002, le conseil d'administration de l'université adopte la structure organisationnelle et la procédure proposée par ce groupe de travail qui devint par la suite une commission. Durant les trois premières années de fonctionnement, 38 dossiers ont été présentés à des jurys ayant délivrés le diplôme dans sa totalité à 28 candidats et des prescriptions aux autres, dont 3 qui les ont réalisées sur cette période. Ce taux de réussite important (82%) explique probablement, le faible nombre de jurys organisés, comparé au volume de l'activité globale de la VAE dans l'enseignement supérieur avec, par exemple, 4560 dossiers présentés pour 1277 diplômes obtenus en 2004 (soit 28% de réussite d'après les chiffres du Cereq, 2006). Ce chiffre va également dans le sens de la diminution importante du nombre de candidats tout au long de la procédure étant donné que l'on estime à 700 le nombre personnes par an qui prend contact avec l'université au sujet de la VAE. La procédure mise en place à l'UPVM peut-être un des facteurs expliquant cette situation. Elle opérerait une forme de ''sélection par dissuasion'' où les candidats qui s'engageraient alors dans la rédaction du dossier auraient plus de chances qu'ailleurs d'obtenir leur diplôme. La similitude des statistiques régionales concernant les universités lorraines qui partagent une procédure relativement semblable et des structures organisationnelles différentes confortent cette hypothèse. En effet, en 2005, sur 67 jurys organisés dans les universités lorraines, 20 l'ont été à l'UPVM.

2.2.2) La structure organisationnelle et la procédure VAE à l'UPVM :

La structure organisationnelle mise en place pour gérer les trois phases de la procédure VAE (accueil, accompagnement et jury) a peu évolué durant ces quatre premières années de fonctionnement. Elle trouve son équilibre entre des fonctions centralisées et d'autres décentralisées dans les composantes :

La centralisation assure essentiellement trois fonctions : la coordination des activités, la prise de décision et le contrôle à l'égard du dispositif impliquant de nombreux acteurs, en particulier:

- le Président qui signe les conventions et notifie au candidat les résultats du jury, comme l'exige le décret
- le Vice Président du CEVU, garant du dispositif, est présent (ou représenté) à tous les jurys et au comité de pilotage régional
- le CEVU puis le CA qui décide des règles et procédures à mettre en œuvre

[3] Voire le site LUREVA : Lorraine université reprise d'études et validation d'acquis à l'adresse : www.lureva.org

- le service commun de formation continue (SCFC) avec son directeur, le chargé de mission VAE, le coordonateur VAE et l'assistante du service qui assurent des fonctions de veille, de contrôle, de développement d'outils et de projets, de coordination entre les différentes composantes, de communication en interne et à l'extérieur ou encore de bilan pour les candidats en recherche du type de reprise d'études le plus pertinent (VAP / VAE, choix du diplôme, …°). Le service gère également la première étape de la première phase d'accueil avant de transmettre le pré-dossier vers la composante concernée. Enfin, il gère également la partie financière de ce fonctionnement en relation avec les services comptables et financiers de l'université.

La décentralisation dans les composantes s'est traduite par la création de la fonction de conseiller VAE, supportée par des personnels administratifs, appartenant au service de formation continue de la composante ou dans les services de scolarité quand l'activité en matière de formation continue relative aux diplômes visés était peu développée. A l'exception d'un IUT revendiquant une autonomie de fonctionnement où le travail du conseiller se limite à la gestion de la première phase, tous les autres conseillers VAE sont les organisateurs de l'ensemble de la procédure, parfois du premier accueil, et dans tous les cas, de l'avis de recevabilité, de l'accompagnement et du jury. Ils deviennent l'interlocuteur privilégié du candidat ainsi amené à rencontrer différents acteurs de la VAE, en particulier les enseignants du diplôme. L'animation du groupe des conseillers par le coordonateur assure la standardisation de la procédure quelque soit la composante. De la même façon, une liste de diffusion réunissant tout personnel intervenu à un moment de la procédure, remplit une fonction de coordination entre des acteurs de plus en plus nombreux. Le conseiller VAE travaille dans certaines composantes avec un enseignant référent VAE, parfois impliqué dans d'autres activités en matière de formation continue dans sa composante.

La procédure votée par le conseil d'administration en novembre 2002 a permis de préciser le déroulement de la procédure et les fonctions des différents acteurs devant intervenir pour traiter de façon satisfaisante les candidatures :

- L'accueil : avant toute pré-inscription, tout candidat qui en fait la demande peut être reçu au SCFC ou au sein d'une composante en vue d'un conseil ou une orientation pour une reprise d'études. Le premier engagement formel du candidat dans la démarche VAE se traduit ensuite par la saisie de son pré-dossier d'inscription sur le site LUREVA, commun aux quatre universités de Lorraine, l'obligeant à passer par ce portail unique pour saisir sa pré-demande. Celle-ci est ensuite traitée par le SCFC avant d'être transmise au conseiller VAE de la composante concernée qui instruit l'avis de recevabilité, tout d'abord administratif, puis pédagogique en lien avec le responsable du diplôme et le coordonateur VAE.
- L'accompagnement : après réception de l'avis de recevabilité, le candidat peut poursuivre la démarche VAE en s'inscrivant à l'université. Un accompagnateur est alors systématiquement nommé et propose ses services au candidat en vue de l'aider à compléter le dossier dans lequel il doit faire état de ses ‘'*connaissances, compétences et aptitudes*''. A l'exception de l'UFR ESM où le conseiller VAE est également

accompagnateur, cette fonction est toujours assurée par un enseignant (ou enseignant – chercheur) du diplôme dans les autres composantes.

- Le jury : deux semaines minimum avant la soutenance, chaque membre du jury reçoit le dossier du candidat. Le jury comprend six membres au pouvoir décisionnel : le VP CEVU, le directeur de la composante concernée, trois enseignants du diplôme visé et un professionnel. La présence systématique du VP CEVU, et plus particulièrement son représentant (dans la majorité des cas le coordonateur VAE) a permis de standardiser le déroulement des jurys, quelque soit le diplôme, sans pour autant formaliser cette aspect de la procédure : après s'être mis d'accord sur son président, le jury entend pendant une vingtaine de minutes le candidat qui répond ensuite aux questions qui lui sont posées pendant une durée moyenne d'une heure. Les membres du jury délibèrent, prennent une décision, la plupart du temps sur la base du consensus, dans quelques cas en recourant au vote à la majorité à main levée, puis font entrer de nouveau le candidat pour lui faire part de l'obtention totale du diplôme ou des prescriptions à suivre. L'accompagnateur est présent pendant toute la durée du jury et répond aux questions qui peuvent lui être posées sur sa perception du candidat ou la façon dont s'est déroulé l'accompagnement. Seul un IUT, sur les neuf composantes de l'université s'écarte de ce fonctionnement avec un jury VAE organisé sous la forme d'une commission qui transmet sa proposition au jury de diplôme.

Un calcul du coût de la VAE avait estimé à plus de 1600 euros la procédure, conduisant à un tarif de lancement en 2002 à 1100 euros pour les candidats dont le financement était pris en charge et 600 euros pour les autres. Les tentatives d'harmonisation de la procédure, des outils et de la tarification avec l'Université de Nancy 2 et l'Université Henri Poincaré, ont conduit l'UPVM à demander, depuis 2005, 900 euros plus 25 euros en début de procédure au moment du dépôt du pré-dossier, réduisant ainsi le nombre de candidats, qui après examen de leur pré dossier, ne poursuivaient pas la démarche. Cette réduction du nombre d'évaluations réalisées lors de la première phase, permettait non seulement de diminuer le coût global de fonctionnement de la VAE mais aussi de réduire la démotivation des évaluateurs et de l'ensemble des personnels intervenants lors de cette première phase de la procédure. Le tableau 1 rend compte de ce changement avec un taux d'accompagnement qui est passé de 15% des avis positifs émis lors de l'analyse de la recevabilité en 2004 à 67% en 2005.

	Année 2004	Année 2005
Nombre de pré-dossiers déposés	201	94
Nombre d'avis de recevabilité positifs	161 Soit 80% des pré-dossiers déposés	55 Soit 59% des pré-dossiers déposés
Nombre de candidats en accompagnement	25 Soit 15% des avis positifs	37 Soit 67% des avis positifs
Nombre de validations (jurys)	9 dont 1 partielle	20 dont 5 partielles

Tab. 1 : Répartition du nombre de candidats à la VAE à l'université Paul Verlaine – Metz, sur les différentes phases de la procédure, pour 2004 et 2005.

Le tableau 1 rend compte aussi de l'évolution globale du nombre de VAE où la ''sélection par dissuasion'' a été renforcée, diminuant le volume d'activité en entrée tout en augmentant le nombre de candidats qui s'engagent réellement dans une VAE avec l'accompagnement. Pour compléter cette analyse financière du problème, il faut également ajouter la question de la rémunération des intervenants, en particulier celle des membres de jury qui étaient indemnisés jusqu'en 2005 provoquant depuis, une source d'insatisfaction à l'égard de la VAE dans plusieurs composantes.

3) Le vécu du dispositif VAE par ses utilisateurs

3.1) Le contexte de l'enquête

L'année 2006 marque la fin des financements FSE pour l'ensemble des projets qui avaient permis depuis 2002 de créer des postes et de financer des projets pour développer la VAE. La réalisation d'un état des lieux de ces trois années de fonctionnement de VAE a été décidée par le SCFC afin de proposer d'éventuels changements en termes de structure organisationnelle et/ou de procédure.

Trois enquêtes ont alors été réalisées par le chargé de mission, aidé d'une stagiaire de master en sciences humaines et sociales pour les enquêtes par questionnaires :

- une série de 20 entretiens semi-directifs ont concerné les responsables politiques de l'université et les principaux acteurs de terrain en matière de VAE : cette première enquête portait sur les représentations du dispositif, ses dysfonctionnements et ses perspectives d'évolution ;
- un questionnaire à destination de tout salarié de l'université ayant été impliqué au moins une fois dans le dispositif VAE : 38 acteurs (dont 27 enseignants ou enseignants – chercheurs) sur les 98 inscrits dans la liste de diffusion ont répondu à cette enquête cherchant à analyser les représentations et les pratiques de la VAE dans l'établissement ;
- un questionnaire à destination des candidats ayant saisi une demande de pré-dossier VAE : 67 candidats sur 206 personnes contactées ont répondu à cette enquête de satisfaction. Le profil moyen du répondant est un homme de 35 ans, exerçant une profession intellectuelle supérieure.

Chacune de ces enquêtes comportait une série de questions sur le fonctionnement général de la VAE avant de demander des précisions sur le vécu des différentes phases de la procédure. Nous présentons ci-dessous les principaux résultats obtenus à ces trois enquêtes menées simultanément et dans la complémentarité.

3.2) Principaux résultats

L'ensemble des répondants (personnel et candidats) a plutôt bien accueilli cette enquête, permettant pour certains d'exprimer des insatisfactions vécues par ces utilisateurs du dispositif VAE et pour d'autres, en particulier des candidats, de témoigner leur remerciement pour l'accueil ou pour l'ensemble de la réalisation comme l'écrit un candidat:*''merci à toute l'équipe pédagogique de m'avoir accompagné dans cette démarche''*.

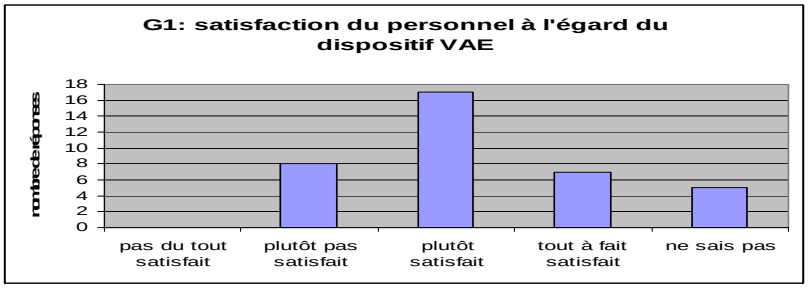
3.2.1) Des candidats majoritairement satisfaits de la procédure et des difficultés liées principalement à l'accompagnement

Plus les candidats ont avancé dans la démarche, plus ils se montrent satisfaits, expliquant ainsi quelques insatisfactions exprimées en cas de notification négative de l'avis de recevabilité. 66% d'entre eux n'ont pas connu d'évolution professionnelle grâce aux démarches entreprises mais 50% de ceux qui ont obtenu le diplôme ont évolué.

La plupart des indicateurs de satisfaction mesurés (recherche d'informations, clarté de la procédure, prise de contacts, accueil, aide apportée par le conseiller ou encore décision du jury) donnent une fréquence moyenne de satisfaction des candidats oscillant entre 62% et 89%. L'accompagnement, avec la difficulté du dossier à compléter (pour 61% des accompagnés), l'importance de la charge de travail liée à la rédaction du dossier (pour 75% d'entre eux) et le sentiment d'avoir été bien préparé au jury (pour seulement 43% d'entre eux) constitue la principale zone d'insatisfaction démontrant ainsi la complexité de cette phase. L'estimation du temps moyen passé au contact de l'accompagnateur est de 6h20mn (soit la moitié du temps que les répondants accompagnateurs estiment consacrer à l'accompagnement) et celui du temps de rédaction du dossier est de 181 heures pour une durée moyenne de 13 mois et demi de démarches avec l'université. Côté tarification, 64% estiment que la procédure est chère et, 4 des 9 répondants pensant le contraire, ont demandé un diplôme en sciences de gestion, un secteur en concurrence, en particulier avec les écoles de commerce qui pratiquent effectivement une tarification en moyenne plus élevée que les universités. Dans tous les cas, pour la plupart des candidats dépassant l'avis de recevabilité, la VAE est un lourd investissement : elle est couteuse en temps et financièrement.

3.2.2) Un personnel majoritairement satisfait de la VAE et du dispositif mais en manque de reconnaissance

Côté personnel de l'UPVM, comme le montre le graphe 1, la satisfaction domine également.



7 des 8 ''plutôt pas satisfaits'' sont enseignants ou enseignants – chercheurs et 4 proviennent d'une même UFR. Autrement dit, le statut et, probablement plus encore, la composante de rattachement du personnel, constituent des facteurs agissant sur la représentation du dispositif qu'ils tentent de faire fonctionner.

Les entretiens réalisés avec les responsables politiques de l'université et les principaux acteurs de terrain traduisent également ce sentiment de satisfaction relatif. Le bilan historique, avec l'organisation mise en place pour développer la VAE ou encore le dispositif actuel qui pour un directeur de composante *''allie rigueur et souplesse''* donne entière satisfaction. Aux yeux de ces responsables, souvent éloignés du fonctionnement concret de la VAE, ce dispositif apparaît satisfaisant, éventuellement *''pas toujours très clair''*, voire *''préoccupant''* du point de vue de son coût pour l'université.

Pour les acteurs de terrain, en particulier les conseillers, la VAE reste une activité minoritaire. Dans les composantes où le volume d'activité est le plus important, le conseiller déclare que la VAE lui prend au maximum 25% de son temps de travail. Par conséquent, dans leur service, la VAE ne constitue pas toujours la priorité et elle est vécue comme une charge de travail supplémentaire. Certains soulignent également, la difficulté de mobiliser les enseignants et enseignants – chercheurs tout au long des trois phases, en particulier pour les avis pédagogiques.

Ce constat peut s'expliquer par le problème de la non rémunération des membres de jury mais d'une façon plus générale celle de la non reconnaissance du travail suscité par la VAE, de la première évaluation du candidat, lors de l'avis de recevabilité, jusqu'à celle réalisée en tant que membre de jury. L'auto évaluation des temps moyens réalisés pour les trois phases de la procédure traduit effectivement une activité non négligeable avec 2h10mn pour un avis pédagogique, 12 heures pour un accompagnement (en recourant à l'entretien, au courrier électronique de façon systématique et, dans une moindre mesure au téléphone) et 4h20mn pour un jury (évaluation du dossier et participation au jury). De plus, les accords moyens obtenus aux questions *''je pense que la VAE est une bonne chose pour…''* le candidat (95%), l'entreprise (73%), l'université (66%) se stabilisent à 59% pour les enseignants et enseignants – chercheurs, les 8 répondants en désaccord avec cette dernière affirmation appartenant à cette catégorie de personnel. D'une façon générale, c'est le manque de moyens qui est reproché, avec seulement 4 répondants pour dire que l'UPVM bénéficie de suffisamment de moyens humains et aucun répondant en accord avec une suffisance des moyens financiers. Ce manque de reconnaissance provient avant tout de l'institution universitaire avec seulement 5 répondants estimant que la VAE est portée par un véritable projet politique de l'université.

Conclusion

En fixant le minimum de règles et en donnant le minimum de moyens, le législateur et l'Etat ont obtenu des établissements d'enseignement supérieur qu'ils se transforment pour faire face ce changement. Trois ans plus tard, d'importantes disparités d'ordres très différents traversent le paysage de la certification française, en particulier dans les universités. Ces disparités concernent les structures organisationnelles supportant le fonctionnement de la VAE, les procédures ou encore les évaluations pratiquées tout au long de la démarche. S'il est tout à fait pertinent de d'interroger les objets, les instruments ou encore la nature de l'activité d'évaluation, il est tout aussi important de s'arrêter sur les moyens humains permettant de conduire au mieux cette évaluation, tout au long de la démarche, et pour tous les candidats.

Le cas de l'Université Paul Verlaine – Metz traduit à ce sujet un dispositif qui a permis de fédérer une dynamique, prenant appui sur l'expérience en formation

continue, et qui semble s'être adapté aux contraintes de coûts financiers et humains (en termes de motivation) en opérant une sélection en amont de la procédure. Si les candidats et le personnel interrogés se montrent majoritairement satisfaits de ce système, l'accompagnement avec l'aide à apporter au candidat reste une des difficultés à surmonter, justifiant l'intérêt des réflexions actuelles sur l'évaluation lors de cette deuxième phase particulièrement complexe. Cette aide supplémentaire, et probablement inconnue dans la forme qu'elle prendra aux yeux des accompagnateurs actuels, ne pourra pas se faire si l'institution à laquelle ils appartiennent ne la reconnaît pas davantage, comme l'ensemble des autres activités engendrées par la VAE. Quelque soit la forme de cette reconnaissance, elle provoquera un coût supplémentaire pour l'université, posant, en retour, la question de sa prise en charge : qui doit supporter la qualité des évaluations réalisées, gage de la validité des certifications délivrées ? Le candidat ? L'Université ? Les OPCA ? La Région ? L'Etat ? Une série de réflexions à ce sujet mérite également d'être menée.

Références

Agulhon, C. (2006.) La VAE à l'université, levier pour une prise de pouvoir des services de formation continue.

Aubret, J. (2003.) Comment évaluer et valider les acquis de l'expérience ? In C. Lévy- Leboyer, M. Huteau, C. Louche et J.P. Rolland (Eds). (2003). *La psychologie du travail.* Paris: Editions d'organisation.

Benmoussi, L. (2005.) Validation des acquis de l'expérience 202-2005: un premier bilan. *Communication au Colloque de la conférence des directeurs de formation continue universitaire*. Poitiers, les 1-3 juin 2005.

Cereq (2005.) Bref n°224.

Cereq (2006.) Bref n°230.

Cherqui Houot, I. (2001). *Validation des acquis de l'expérience et universités : quel avenir ?* Paris : L'Harmattan.

Clot, Y. (1999). *La fonction psychologique du travail*, Paris : PUF.

Pons-Dessoutter, M. (2005.) *Réalisation et usages de référentiels « compétences » dans les universités et INP*.

Vermersch, P. (1996). *L'entretien d'explicitation*, Paris : ESF.

Mintzberg (1982). *Structures et dynamique des organisations*, Paris, Editions d'Organisation.

Louche (2000). Structure et structuration des organisations. In J.L. Bernaud & C. Lemoine, (Eds.) (2000). *Traité de psychologie du travail et des organisations*, Paris: Dunod.

ANALYSE DE L'ACTIVITE EVALUATIVE DES MEMBRES DE JURY DE VAE.

Matine Morisse , Université de Lille 3

Résumé :

La communication présente les résultats d'une recherche[1] qui a pour objectif l'analyse des activités évaluatives des membres de jurys. C'est à partir d'un travail d'investigation et de recueil de données[2] auprès de trois jurys, dans trois domaines professionnels différents[3], que nous analyserons les activités des jurés, dans leur recherche de la preuve des compétences des candidats et dans leur décision d'attribuer une partie ou la totalité du C.Q.P. (certificat de qualification professionnelle). Quatre catégories d'analyse se dégagent des données recueillies, permettant d'identifier des éléments consensuels et des zones de tension entre les jurés. Dans leur attitude, on observe également une grande hétérogénéité. Elle fait apparaître des postures évaluatives différentes que nous présenterons dans notre communication. Ces éléments d'analyse nous conduiront à interroger le sens de l'évaluation pour ces jurés, et leur rôle dans une démarche de VAE.

Ces catégories d'analyse et les postures évaluatives que nous avons identifiées, à partir de l'observation et des discours des membres des jurys, posent la question du sens qu'ils donnent à l'évaluation et à leur rôle de juré. Qu'est-ce qu'ils évaluent et comment ? C'est ce que nous développons dans cette communication.

1) Contexte de la recherche

Cette communication s'inscrit dans le cadre d'une recherche conduite par l'équipe Théodile, de l'université de Lille 3, commanditée par le F.A.F.S.E.A. (Fonds d'Assurance Formation des Salariés des Entreprises et Exploitations Agricoles). Elle porte sur l'analyse de l'activité de jurys en V.A.E. dans le cadre des C.Q.P. (Certificat de Qualification Professionnelle) mis en place par le F.A.F.S.E.A. et la C.P.N.E. (…….) dans le secteur agricole. Les jurys observés concernent trois domaines professionnels :

- CQP « ouvrier qualifié de l'exploitation viticole ».
- CQP « ouvrier spécialisé de services de remplacement ».

[1] Cette recherche a été commanditée par le FAFSEA (fonds d'assurance formation des salariés es entreprises et exploitations agricoles), OPCA de branche, impliqué dans le dispositif de validation des acquis de l'expérience comme financeur du congé de VAE, mais aussi comme organisme technique d'appui.

[2] Celui-ci s'est déroulé sur la base d'observations d'entretiens avec les candidats et de réunions de bilans des jurés, et à partir d'interviews menés avec chaque membre des jurys.

[3] Les trois jurys ont eu lieu dans trois villes différentes (Nancy, Narbonne, Perpignan), concernant trois domaines professionnels : ouvrier qualifié de l'exploitation viticole, ouvrier spécialisé en production légumière, agent des services de remplacement

- CQP « ouvrier spécialisé en productions légumières ».

Les membres de jurys de VAE sont composés de professionnels uniquement, provenant de syndicats salariés et patronaux de la profession. Il s'agit de leur première expérience en tant que membres de jury.

Tableau 1: caractéristiques des trois jurys

Caractéristiques Jurys	CQP	Membres	Candidats	Résultats
Jury N	Ouvrier qualifié de l'exploitation viticole	7 membres dont 4 salariés et 3 employeurs	5 candidats	3 CQP complets 2 obtenus partiellement
Jury L	Agent de service de remplacement	9 membres dont 5 salariés et 4 employeurs	6 candidats	4 CQP complets 1 obtenu sous réserve de compléter le dossier 1 obtenu partiellement
Jury P	Ouvrier spécialisé en productions légumières	4 membres dont 2 salariés et 2 employeurs	3 candidats	2 CQP complets 1 obtenu partiellement

Les modalités de recueil de données sont les suivantes :

Des enregistrements vidéo des trois phases de l'activité du jury :

- la 1ère phase de réunion du jury, c'est-à-dire la phase de l'examen collectif des dossiers des candidats.
- la 2ème phase qui comprend la réalisation des entretiens avec les candidats.
- la 3ème phase qui comprend le bilan et la prise de décision. Ce bilan se fait en partie juste après le départ du candidat et de façon récapitulative après la fin de tous les entretiens.

Des enregistrements audio d'entretiens individuels effectués avec quatre membres du jury pour chaque CQP, et avec quatre candidats du CQP « ouvrier spécialisé de services de remplacement » et deux candidats du CQP « ouvrier spécialisé en production légumières ».

2) Questionnement autour de la notion de « posture évaluative »

Les objectifs des observations et des entretiens que nous avons réalisés étaient d'analyser l'activité du jury VAE dans ce secteur à partir des activités énonciatives, du type de questions posées pendant les entretiens, de l'analyse de contenu des commentaires qui ont précédé et suivi les entretiens, des outils utilisés, de la place donnée au dossier au cours des commentaires et de la prise de décision.

L'analyse du discours des entretiens avec les candidats a mis en évidence que les jurés adoptent deux types de « postures évaluatives » : une posture impliquée et une

posture d'explicitation[4] .Ces deux postures déterminent un genre de discours particulier que les jurés adoptent pour se référer aux trois dimensions qui sous-tendent l'activité d'évaluation:

- la dimension pédagogique qui renvoie au sens donné à l'évaluation dans le cadre de la VAE pour le développement du candidat ;
- la dimension professionnelle liée à la définition de l'expertise du métier ;
- la dimension institutionnelle qui renvoie à la position hiérarchique à laquelle les membres du jury se situent par rapport au candidat.

Pour analyser, de façon plus approfondie, les logiques orientant les activités énonciatives des membres du jurys, nous avons procédé à l'analyse des commentaires qu'ils avaient effectués lors de l'examen des dossiers et des bilans. Ce qui nous a conduit à élargir la notion de « posture » en utilisant les travaux de Dominique Bucheton (1997, 1999). La notion de posture fait partie aujourd'hui de ces concepts fréquemment utilisés que certains auteurs ont contribué à faire émerger à travers l'étude des postures d'écriture (Bucheton, 1997 ; Bucheton et Bautier, 1997) et des postures de lecture (Bucheton, 1999). « Les postures sont des schèmes préconstruits de pensée-langage-action pour traiter de problèmes langagiers et non langagiers, des manières de s'emparer d'une tâche ou d'un problème pour le résoudre, ou simplement pour le réaliser ». Si la posture est inscrite dans l'histoire scolaire, personnelle et professionnelle, l'auteur ajoute qu'on dispose de plusieurs postures pour négocier une tâche, elle est relative aussi à la tâche. Cette notion est donc du côté du sujet et de l'objet, elle désigne tout à la fois: « le rapport du sujet aux tâches, aux savoirs, au contexte (ce qu'il en fait, ses logiques de travail) et en même temps son mode d'implication et son rapport au langage » (2002, pp.221-222).

L'usage de cette notion a conduit à nous interroger également sur le rapport des membres de jurys à la tâche qui est la leur, leur rapport aux savoirs, à l'expérience mobilisés chez les candidat, leur rapport au langage utilisé, support privilégié de l'évaluation, lors des entretiens ou des bilans. Si, selon les textes officiels, les candidats sont évalués à partir d'un certain nombre d'indicateurs, explicités et répertoriés dans les référentiels remis aux membres de jury, on peut se demander s'il existe d'autres indicateurs, implicites, plus insidieux, que les jurés vont mobiliser, de façon non consciente et non volontaire, pour valoriser ou non l'activité des candidats. Sur quoi reposent-ils alors, sur des items non partagés, des références à une identité culturelle du métier non prise en compte dans les textes officiels, des glissements de sens dans les consignes qui sont données, des ressources du candidat non explicitées? La question est de voir alors quelle cohérence s'établit entre le sens de l'évaluation, que donnent les membres de jurys, les indicateurs qu'ils utilisent pour évaluer les candidat et les consensus auxquels ils parviennent pour prendre leurs décisions finales de validation de la totalité ou d'une partie seulement du C.Q.P.. ?

C'est à partir de ce questionnement que nous avons analysé les résultats de l'une des étapes du processus de validation des candidats, les commentaires sur les dossiers de V.A.E et les bilans des jurys.

---

[4] Voir à ce propos la première partie du rapport : Pagoni M., Morisse M. (2006), Analyse de l'activité du jury en VAE dans le secteur agricole : enjeux de construction d'une « posture évaluative », Rapport de Recherche, Université Lille III.

3) Le sens donné à l'évaluation par les membres de jurys de V.A.E.

Plusieurs dimensions de l'évaluation se dégagent de l'analyse des commentaires des membres de jurys conduisant à nous interroger sur le sens même de l'évaluation pour ces acteurs de la V.A.E.

3.1. La dimension identitaire de l'évaluation et la « culture de métier » en question

Il ressort un fonctionnement consensuel entre les membres du jury, qu'ils soient salariés ou employeurs, pour l'analyse de l'activité du jury et le processus de prise de décision. Nous pouvons, d'ores et déjà, identifier les valeurs partagées par les membres de jurys et un certain nombre d'indicateurs à partir desquels les jurés vont apprécier ou non les compétences des candidats:

- La confiance en soi est un indicateur important qui traduit le professionnalisme du candidat. C'est presque une valeur en soi. L'un des candidats qui n'est pas à l'aise lors de l'entretien est parfois suspecté de vouloir dissimuler quelque chose, son manque d'expérience ou son incompétence. « *Ne pas être à l'aise* » ou « *ne pas très être sûr de lui* » laisse alors supposer, dans certains cas, que le candidat est novice dans l'un des domaines d'activités considérés (le traitement par exemple).
- La spontanéité du candidat, son corollaire, est jugée positivement par les membres du jury. Tous les membres de l'un des jurys s'accordent pour dire de l'un des candidats est « spontané », parce qu'il n'a pas d'hésitation dans les réponses qu'il apporte aux membres du jury, il répond « *sans réfléchir* », de plus « *il est honnête* » en disant ce qu'il sait faire et ne sait pas faire. La spontanéité apparaît alors comme une marque de franchise et sera valorisée lors de la prise de décision.
- Le discours, valorisé par les membres du jury, est qualifié également de « *pratico pratique* », c'est-à-dire sans ambiguïté, sans apports théoriques, un discours également dans lequel chacun des professionnels pourra se reconnaître, un discours jugé très positivement par la majorité des membres du jury.
- La valeur du travail, l'investissement et la motivation du candidat, sa volonté à vouloir travailler, y compris pour exercer les tâches les plus difficiles, telles que la cave, génère un fonctionnement consensuel entre les membres de jurys très favorable à celui-ci. L'un des signes distinctifs porte sur les mains du candidat et sa façon de s'exprimer (pour le jury N) : « *les mains d'un bosseur* » ou « *les mains d'un caviste* », opposés à celles d'intellectuels, dira l'un d'eux.

Dans le bilan des jurés, on constate un consensus sur des façons d'être au travail (motivé, autonome, courageux), des façons d'être avec les autres (honnête, franc,

spontané), un « genre de discours »[5] (à l'aise, concret, pratico pratique). Il s'agit d'« allant de soi » qui servent d'indicateurs, produits de façon non consciente et non volontaire le plus souvent, à partir desquels certains jugés vont apprécier ou non les aptitudes des candidats. C'est ce qui fait partie de la culture de métier. Elle crée les conditions d'une identification au discours du candidat, à son activité, dans une demande de reconnaissance de la culture de métier, plus que dans une démarche d'évaluation. Ce qui conduit parfois certains jurés à répondre à la place des candidats (jury N). L'un des salariés dira : « *j'ai retrouvé mon métier là dedans* » (jury L). Un autre, employeur, souligne: « *C'est très fidèle à ce que moi professionnel, je demande à mes ouvriers* » (jury P).

3.2 le glissement de sens entre acquis de l'expérience et expérience, à partir de la prise en compte de la trajectoire des candidats

La durée et la diversité de l'expérience du candidat et la régularité de l'activité sont des indicateurs mobilisés dans la prise de décision des membres de deux jurys N et L. Ils traduisent la richesse du candidat, son autonomie, sa force de caractère, éléments faisant partie de la dimension identitaire du candidat. Par exemple, le candidat (jury N) dispose de 10 années d'expérience dans le secteur agricole, après avoir passé cinq années dans l'armée, avec un diplôme de boulanger. La durée de l'expérience et sa diversité conduisent les jurés à valoriser le fait que le candidat « *s'est formé tout seul* », tout en occultant la formation de 28 heures qu'il a pourtant suivie. A l'inverse, la régularité d'une activité sur une longue durée est perçue comme un indicateur de compétences. La traite des animaux, que le candidat (jury L) effectue quotidiennement, apparaît aux jurés comme un élément de preuve de ces acquis dans ce domaine: « *C'est un mec qui traie pratiquement tous les jours - C'était au moins son domaine de compétences* ». C'est la mise en perspective de l'expérience, plus que l'identification des acquis de l'expérience, qui influe alors sur les prises de décision des jurés.

Dans la prise en compte de l'expérience des candidats, il y a un glissement de sens entre « expérience » et « acquis de l'expérience ». Dans les entretiens, les membres de jurys cherchent à vérifier la réalité, la véracité, la diversité de l'expérience des candidats, plutôt qu'à faire émerger les connaissances issues de l'expérience. Ils tendent alors à occulter les acquis, en tant que résultante d'un apprentissage, potentiel, « définissable en terme d'effets durables (Aubret et Gilbert, 1994, p.12)[6]. On parlera alors de connaissance au sens large, en nous reportant à la définition des acquis de l'expérience proposée par I. Cherqui Houot. « Toute trace héritée d'un apprentissage, quels qu'en soient les modalités et les lieux d'acquisition, et qui a modifié durablement et significativement la façon de connaître, de se connaître ou d'agir sur le monde d'une personne donnée » (2001, 31). La durée et la régularité d'une expérience sont des facteurs favorables à la construction d'acquis, mais ils ne sont pas

[5] Le « genre de discours » renvoie à «des stocks d'énoncés attendus, prototypes de manières de dire ou de ne pas dire dans un « espace temps socio-discursif » d'après Clot (1999).

[6] Aubret J., Gilbert P. (1994), Reconnaissance et validation des acquis, Paris, P.U.F., Que sais-je.

suffisants. Il faudrait également analyser les conditions de transformation d'une expérience en apprentissage (ref. Kolb, 1984).

La formation des candidats est, quant à elle, convoquée de façon très inégale dans les entretiens et peu commentée dans les bilans. Elle renvoie à une position paradoxale et controversée des membres de jurys: elle produit de la méfiance à l'égard de candidats jugés trop théoriques et, en même temps, elle fait partie de préconisations pour l'obtention du C.Q.P. D'après certains membres du jury, le recours à la théorie viendrait plutôt justifier un manque de pratique : « *Il connaît tout théoriquement et il a pas l'expérience de tout ce qu'il a appris, donc c'est ça qui le mettrait mal à l'aise…* » dira l'un d'eux. Le manque de formation traduit un certain malaise chez les membres de jurys, il remettrait en cause son professionnalisme. On observe également une tendance dominante à occulter le temps de formation des candidats, pour valoriser son « intelligence pratique », et une difficulté à se mettre d'accord sur les formations à dispenser pour les candidats qui n'ont obtenu qu'une validation partielle du C.Q.P. Voici un échange entre jurés qui illustrent bien les difficultés de préconisation de formation d'un candidat: *« Le travail en équipe quand même c'est pas un mauvais travail, il est habitué à travailler avec 7 ou 8 personnes … - Il faudrait qu'il fasse un complément de formation pour acquérir l'entretien et la surveillance des cultures … - Alors sur le travail en équipe tu lui donnerais quoi ? – C'est dur aussi de lui dire de se former pour apprendre à travailler en équipe dans le cadre du MC3… »*.

3.3. La dimension langagière et la difficile articulation entre le dire/l'écrire/le faire/le savoir

Le dire, l'écrire, le faire, le savoir sont sollicités dans l'analyse du dossier et les commentaires des bilans. L'articulation entre eux créent des zones de tension chez les candidats, exprimée lors des entretiens individuels[7]. Ceux-ci corroborent des niveaux d'attente différents de la part des membres de jurys :

- Le rapport entre le dire et l'écrire

L'écrit prépare les jurés pour l'entretien oral. Ensuite, lors de bilans, il y a une tendance à se reporter au dossier pour comparer ce qui est dit avec ce qui est écrit. L'oral vient alors confirmer ou compléter l'écrit, modifier parfois le jugement porté au candidat. L'écrit devient alors un élément de preuve, en faveur ou non de la validation des acquis du candidat . En cas de litige, il tend à faire jurisprudence par rapport à l'oral. *« si tu juges sur dossier, c'est pas faux là »* dira l'un d'eux.

- Savoir écrire et savoir dire

Le dire et l'écrire jouent un rôle important dans la démarche de V.A.E. Les capacités langagières des candidats sont mobilisées dans cette démarche. Plusieurs jurés découvrent des difficultés d'expression écrite de candidats et expriment leur gêne, parfois leur compassion. Les difficultés à écrire sont significatives chez

[7] L'analyse des activités d'écriture des candidats devrait faire l'objet d'un travail de recherche plus approfondi. Nous ne disposons pas de données suffisantes pour modéliser ce travail. Il s'agit d'hypothèses explicatives au regard des commentaires recueillis et des travaux de recherche déjà réalisés sur l'écriture sur sa pratique professionnelle (Morisse, 2006).

plusieurs candidats ayant suivi un cursus scolaire court et elles peuvent les pénaliser au moment de la lecture des dossiers.

- Dire et écrire le faire

C'est une écriture dans laquelle le candidat n'a pas de distance entre ce qu'il écrit et ce qu'il fait. Il écrit ce qu'il fait, de façon impliquée, imbriquée parfois. Il restitue des éléments de son expérience sans effectuer un travail d'objectivation, c'est-à-dire sans parvenir à placer l'objet dont il parle à l'extérieur de soi, à l'extérieur du monde qui est le sien. Il ne parvient pas à se mettre à distance de l'objet dont il parle, pour le réfléchir. Il procède le plus souvent par énumération, pour décrire une activité dont il se saisit par écrit, ajoutant parfois des commentaires personnels.[8]

- Dire et écrire sur le faire

La sélection et l'exposition des activités, dans le dossier du candidat, vont le contraindre à gérer les écarts qui vont apparaître, de façon consciente ou non, entre ce qu'il fait et ce qu'il pense faire, entre ce qu'il fait et ce qu'on lui demande de faire (les méthodes de travail sont souvent différentes selon les exploitants), entre ce qu'il fait et ce qu'il aimerait faire, si cela était possible, si les circonstances le permettaient[9]. Les tensions vécues par le candidat dépassent ainsi largement l'activité scripturale, dans le sens linguistique du terme (préoccupations dominantes chez les candidats et les accompagnateurs). Elles traduisent une activité cognitive qui mobilise à la fois la mémoire du sujet, sollicitant son expérience passée pour sélectionner les situations qui lui semblent les plus significatives par rapport aux compétences attendues des membres du jury - en mettant en évidence le déroulement de chacune d'elles selon un ordre, une chronologie, une logique qui lui est propre, à partir des consignes données - et les connaissances qui vont se dégager de ces activités.

- Dire et écrire ce qu'il a appris du faire

Ecrire ce qu'il a appris du faire implique d'« écrire sur » sa pratique, pour la dépasser. C'est-à-dire de faire émerger les connaissances qui auront modifié, de façon durable et significative, sa façon d'être et d'agir sur le monde. Elles constitueront ainsi un capital qu'il pourra mobiliser et transférer ensuite à d'autres situations, dans d'autres circonstances. Ces connaissances sont mobilisées, lors des entretiens de V.A.E., à partir de l'analyse de situations réelles ou hypothétiques soumises au candidat dans l'un des jury (jury L). Elles ne sont pas sollicitées dans le dossier et, cependant, elles sont valorisées dans certains jurys.

L'activité langagière des candidats, lors de l'écriture du dossier ou au cours de l'entretien, est renforcée, accompagnées, stimulée différemment par les membres de jury, en fonction des représentations qu'ils ont de leur tâche et de l'évaluation dans une démarche de V.A.E.

[8] Un constat déjà observé dans l'écriture sur sa pratique professionnelle, chez des personnes en situation d'illettrisme (Morisse, 1995).

[9] Des écarts mis en évidence par plusieurs auteurs de l'analyse du travail.

3.3.Les ressources sollicitées chez les candidats, non explicitées dans le référentiel

On peut identifier des grandes tendances et des aspects dominants, plus valorisés que d'autres, dans les commentaires consensuelles des jurés. Les plus commentées d'entre elles portent sans doute sur les ressources relationnelles de candidats[10]. Les jurés attendent du candidat qu'il ne soit pas « un exécutant » mais qu'ils prennent des initiatives, qu'il soit autonome. Par exemple, le candidat L6 est perçu comme un « bon » candidat qui sait, non seulement, « *maîtriser la situation* » mais aussi « *négocier* » avec son employeur : « *Il sait à qui il a à faire tout de suite et il sait s'il peut négocier ou s'il peut pas*» dira l'un d'eux. Quand le candidat se trouve en difficulté dans son activité professionnelle, il peut faire appel à des

personnes ressources, des personnes de son entourage qui pourront lui donner la réponse ou lui apporter la solution[11]. Ces compétences sont sans doute les plus récurrentes et les moins explicitées du référentiel.

La capacité d'adaptation et de prise d'initiatives s'inscrivent également dans les compétences attendues des candidats au C.Q.P. Celles-ci pourraient s'évaluer à partir de solutions apportées aux candidats pour faire face à des situations difficiles ou d'urgence ou en leur soumettant des situations réelles ou hypothétiques. Dans le premier cas, voici un extrait de commentaires de jurés : *« Il n'a pas été suffisamment affronté à des situations difficiles. Ce n'est pas sa faute mais là je suis dans l'incapacité à l'évaluer, j'aurai bien voulu qu'il me dise bah là j'ai été en situation d'urgence sur une exploitation et cœtera … C'est vrai dans les situations nouvelles, on sait pas bien, mais dans la situation actuelle… » (L1).* L'adaptation fait partie du potentiel de l'individu, ce qu'il serait capable de faire si les circonstances étaient favorables, si on lui demandait de le faire, si cela était possible. Ces compétences s'inscrivent alors dans l'ordre des possibles et renvoie au « pouvoir agir » du candidat, plus qu'au « savoir agir » (Toupin, 1995). Des jurés s'interrogent alors sur la prise en compte ou non que ces possibles dans l'évaluation des compétences des candidats. La prise d'initiatives, on la trouve aussi dans les capacités d'anticipation des candidats : *« … il a toujours sur le tracteur un marteau, de la ficelle, donc c'est quelqu'un qui ne va pas partir comme ça à l'aveuglette, et qui du coup se rendra compte qu'il a besoin d'un marteau, il va revenir jusqu'à la campagne, le chercher, (…). C'est quelqu'un qui déjà anticipe les problèmes q'il peut y avoir… » (Jury N).* On constate donc que les ressources relationnelles, les capacités d'adaptation, les

[10] En utilisant le terme de ressources, nous nous reportons alors à la définition de compétences proposée par Le Boterf : « La compétence est une construction, le résultat d'une combinaison pertinente de plusieurs ressources. Ces ressources relèvent d'un double équipement dans lequel la personne peut puiser pour construire ses compétences : l'équipement incorporé de la personne (connaissances ; savoir-faire, expérience, motivation, valeurs, …) et l'équipement de l'environnement (réseaux relationnels, instruments, banque de données, …). Ce double équipement constitue un double patrimoine plus ou moins développé et utilisé » (2002)

[11] Cette remarque est particulièrement valable pour le jury L s'adressant à des agents de service de remplacement.

prises d'initiatives sont fortement valorisées lors des commentaires, mais des interrogations subsistent quant à la prise en compte ou non du potentiel des candidats, de son « pouvoir agir », lors des bilans.

4) Les logiques et les différents « rapports à » qui sous tendent les postures évaluatives des jurés.

Dans l'analyse des activités énonciatives des membres de jurys, lors des bilans et de la consultation des dossiers, nous distinguerons trois logiques que nous avons croisées avec les différents « rapport à » qui proviennent des résultats des différentes étapes du processus de validation:

— **Une logique d'identification et de reproduction**

L'évaluation du candidat est sous tendue par une logique de l'identification et de la reproduction. La culture de métier prend ici toute sa place. Le juré a alors tendance à attendre des réponses qui soient conformes à ce qu'il fait, ou à ce qu'il aurait fait dans cette situation là. Il cherche à reconnaître dans le discours ou les attitudes du candidat des façons de faire, des façons de dire, ou des façons d'être (« *les mains d'un bosseur* » ou « *des mains de caviste* ») et d'être avec les autres, qui correspondent pour lui à la valeur du métier. Cette identité de métier, qui les rassemble, s'appuie sur des valeurs partagées, décrites précédemment ; elle opère par identifications. Le sentiment d'appartenance à un même groupe social justifie ainsi la modalité d'échange et les critères d'évaluation deviennent alors plutôt des critères de reconnaissance.

Certains membres de jurys sont également porteurs d'un idéal culturel de maîtrise du métier qui repose sur l'habitus professionnel. La transmission s'est opérée d'homme à homme, ou de maître à apprenti, comme dans le compagnonnage. Pour Zarca, à propos de l'identité artisanale, l'identification opère de deux façons, une identification mimétique « le geste appris est un geste regardé » et par l'intimité, qui se construit à partir de rapports de confiance et de coopérations nécessaires (1988). Lors de entretiens, celle-ci se manifeste par des jeux de connivence et de complicité entre les membre de jury et des candidats. Les jurés ont tendance à donner des conseils ou à faire des commentaires interprétatifs. Les modalités énonciatives sont plus centrées sur la description de l'activité que sur son analyse, dans des situations précises de travail.[12] Ils les encouragent à dire le faire, plus qu'à dire ce qu'ils ont appris du faire (jury N, à partir des distinctions faites précédemment.

— **Une logique d'expertise**

Cette logique se manifeste dans l'analyse des dossiers par les jurés et dans l'analyse des entretiens avec les candidats. Lors des entretiens, les membres des jurys sont dans une démarche d'explicitation des ressources des candidats, à partir de

[12] Ce qu'on a appelé la « posture impliquée » dans la première partie de notre travail.

modalités énonciatives à dominante explicative et réflexive, centrées sur l'analyse de l'activité et la problématisation de la situation (en leur soumettant des situations réelles ou hypothétiques)[13]. La prise de conscience et la décentration sont nécessaires dans une démarche d'analyse et de réflexivité. Il faut qu'ils prennent conscience de leur propre fonctionnement, du fait qu'autrui peut fonctionner différemment qu'eux. « V*ous réagissez parce que vous avez une autre pratique et là je crois qu'on l'évalue sur sa pratique à lui* » dira l'un d'eux à ses collègues (L1). Les candidats sont conduits ainsi à dire le faire, mais aussi à dire ce qu'ils ont appris du faire. Pour Le Boterf, la réflexivité engage le sujet dans une démarche de conceptualisation qui suppose un cadre d'analyse (typologie des pratiques, concept de compétences, démarches types,….). Cette démarche a été réalisée par les membres du jury qui ont participé à la conception du référentiel de certification (jury L). C'est un travail indispensable pour que les membres des jurys aient une lecture des activités et des situations présentées par les candidats. L'écart entre les capacités de réflexivité des membres des jurys qui ont participé au référentiel et les autres est, en ce sens, tout à fait significatif. La formation des membres des jurys apparaît alors indispensable.

— **Une logique de centration sur le parcours**

Cette logique est sans doute plus difficile à identifier. On peut la retrouver de façon transversale dans les deux autres logiques définies précédemment. Elle est davantage centrée sur la personne, son parcours, sa trajectoire. Celle-ci est sollicitée, dans les entretiens, à partir de questions qui portent sur la formation des candidats (l'un des candidats n'a jamais été scolarisé par ex.), son expérience professionnelle, ses ruptures de parcours (un candidat a été engagé dans l'armée avant de devenir ouvrier agricole), son orientation professionnelle (vers l'étranger). Cette logique désigne également les centres d'intérêt du candidat (en rapport avec son métier), son avis (sur l'avenir du métier dans la région), ses relations avec la famille de l'exploitant. Elle renvoie à un questionnement de la part des jurés qui sollicite l'implication des candidats, et qui déborde parfois sur l'évocation des conditions de travail et le développement de la profession. Le candidat est alors appréhendé dans sa globalité. Ce qui est demandé aux candidats, c'est de faire la preuve de ses compétences, mais aussi de dire ce qui les a motivées, le contexte dans lequel elles se sont développées, leurs effets éventuels en terme de perspectives de développement professionnel pour celui-ci. Les jurés sortent alors d'une certaine neutralité pour interpeller le candidat sur des choix personnels, des opinions, des événements significatifs de sa vie, dans une tentative qui est, pour eux, sans doute, de mieux en saisir le sens. Ils demandent parfois aux candidats de s'autoévaluer. A partir d'événements de vie ou d'événements professionnels qu'ils jugent significatifs, ils cherchent à saisir le sens d'une logique de parcours, à en comprendre les fondements, plutôt qu'à dégager une liste de compétences.

Le tableau ci-dessous tente de synthétiser les résultats de notre recherche en mettant en évidence les composantes des postures évaluatives des membres de jurys.

[13] L'objectif est d'amener la personne à décrire avec précision son activité depuis le diagnostic du problème jusqu'à l'évaluation de la solution adoptée et à formaliser les savoirs théoriques, pratiques, implicites, pertinents ou non qui ont guidé son action. Ce que nous avons appelé dans la première partie du rapport « la posture d'explicitation » empruntée à la méthode des entretiens d'explicitation (Vermesch 1994, 1997).

Tableau 2 :les logiques et les différents « rapports à » qui sous tendent les postures évaluatives des membres de jurys

Postures évaluatives	Rapport au candidat	Rapport aux savoirs	Rapport à l'expérience	Rapport à l'activité langagière
Logique d'identification et de reproduction	Identification à la personne du candidat Dominante culturelle	Reconnaissance d'une culture de métier	Approche descriptive, centrée sur la tâche et l'activité	Dire/ écrire le faire
Logique d'expertise	Prise de distance par rapport au candidat Dominante cognitive	Mobilisation des ressources relationnelles et cognitives du candidat	Approche descriptive et réflexive, centrée sur l'analyse de l'activité et de la situation (recours à des situations hypothétiques)	Dire/écrire sur le faire Dire/écrire ce qu'il a appris du faire
Logique de centration sur le parcours	Prise en compte de différents aspects (social, affectif, expérientiel, cognitif) Approche globale de la personne	Mobilisation des ressources relationnelles et cognitives du candidat Valorisation de la dimension expérientielle	Approche descriptive et réflexive, centrée sur l'analyse de l'activité et de la situation (recours à des situations hypothétiques)	Dire/écrire le faire Dire/écrire sur le faire Dire/écrire ce qu'il a appris du faire

Conclusion

Les résultats de notre recherche montrent que l'activité d'évaluation ne va pas de soi pour les professionnels des jurys de VAE dans le secteur agricole. Si les textes officiels qui existent sur ce dispositif soulèvent déjà beaucoup de questions (Feutrie 2004), son appropriation par les professionnels implique aussi des enjeux identitaires et institutionnels importants. Que recouvre la fonction d'évaluateur dans ce dispositif ? Y a t-il des compétences et des savoirs spécifiques qu'un membre de jury doit développer pour remplir son rôle d'évaluateur ? Quelles conditions faut-il réunir

pour lui permettre de passer du statut de l'acteur au statut de l'évaluateur ? Nous essaierons, à partir de ces questions, d'identifier plusieurs points de discussion :

Le premier porte sur le sens donné à l'évaluation par les membres de jurys. Comme l'écrit Clot : « Du coup, qu'est ce qu'on évalue ? L'expérience antérieure, la transformation de l'expérience, la possibilité de développement du sujet dans cette expérience là » (2002, p.65). Dans la démarche de V.A.E., s'agit t-il de décrire les activités menées par les candidats, d'identifier et d'analyser la singularité de l'action, d'identifier « au cœur de l'action, la conceptualisation »[14], dans une intention formative de développement personnel et professionnel des candidats ? Il nous semble important de faire participer les membres du jury à ce débat et de les aider à considérer le travail d'évaluation du jury non pas de façon statique, comme une comparaison entre le référentiel préexistant et les compétences déclaratives des candidats, mais d'une façon dynamique, comme un processus de sollicitation de la réflexion du candidat à partir de son expérience.

Importance également d'emmener les membres du jury à conceptualiser leur propre expérience professionnelle en participant à la construction du référentiel de compétences, en échangeant entre eux sur leurs activités de travail et en développant des capacités réflexives. En effet, seuls les membres de jurys qui ont participé à la conception du référentiel se sont inscrits dans une logique d'expertise. Les autres en ont été exclus et ils se sont reportés à la connaissance qu'ils avaient du métier, en tant que salarié et employeur, et les critères à partir desquels identifier le prototype du « bon » professionnel. Le partage de ces items doit se faire dans le cadre d'une action commune entre les différents membres de jurys, en analysant leur représentation du métier, les compétences mises en œuvre et la nature de l'activité qui les attend.

Vient ensuite la question de la communication. Avoir une attitude d'écoute et de compréhension face au candidat et s'interroger sur la signification des mots qu'ils utilisent lors des entretiens. « Les mots constituent ainsi le support (spatial, graphique et sémantique) de l'évaluation ; en chacun d'eux se croisent les significations, indépendantes des sujets, que rassemble une culture, et le sens que chacun peut leur attribuer » Astier (2004). Chacun des acteurs est « soumis à l'épreuve du sens des mots d'autrui », dans le cas qui nous intéresse, ceux du candidat, ceux du dossier, ceux des membres du jury. Ce croisement de sens repose sur les représentations de chacun, leurs attentes, leurs jugements, leurs façons d'appréhender la situation nouvelle qui est la leur, leur rapport au savoir tel qu'on l'a analysé précédemment, etc. La sensibilisation du jury face à ces variations du langage peut influencer sa méthode de questionnement qui, par ailleurs, est une technique à développer :

- Techniques de relances, de reformulations « en écho » et « clarifiantes », relances de précision et d'approfondissement, différenciations entre les questions de « comment » et de « pourquoi ».
- Rôle de la description dans l'analyse des compétences : de quelle description s'agit-il ? Comment aider le candidat à l'effectuer et à l'approfondir ? Quelle

[14] Il s'agit de la reprise du titre d'une contribution de G. Vergnaud (1996) à un ouvrage collectif résumant l'un des postulats fondateurs de la didactique professionnelle.

place donner à d'autres modalités énonciatives qui viennent compléter celle de la description, comme les demandes de diagnostic et d'évaluation ?

- Choix des situations de travail à décrire. Comment faire la part entre les situations réelles de travail et les situations hypothétiques qui peuvent être proposées comme une sorte de « résolution de problème » ? Pourquoi les questions qui portent sur des situations précises sont-elles plus révélatrices de la singularité de l'action que les questions qui portent sur un modèle d'action générale ?
- Sollicitation et mobilisation des savoirs des candidats à partir des situations à décrire, de natures différentes (scolaires, procédurales, pratiques, …).

Références

ASTIER P. (2004), « La validation des acquis de l'expérience, épreuves de commensurabilité », *Education Permanente*, n°158/2004-1, 25-37.

AUBRET J. & GILBERT P. (1994), *Reconnaissance et validation des acquis*, Paris, PUF, Que sais-je?

BUCHETON D. (2003), « Le journal de bord: un écrit de travail en formation » dans *Le mémoire professionnel des enseignants observatoire des pratiques et levier pour la formation*, sous la direct. de J. Crinon, Paris : L'Harmattan.

CHERQUI HOUOT Isabelle (2001), *Validation des acquis de l'expérience et universités quel avenir ?* Paris, L'Harmattan.

CLOT Y. (1999), *La fonction psychologique du travail*, Paris, PUF.

CLOT Y. (2002), La validation des acquis professionnels, l'analyse psychologique du travail et les problèmes de l'expérience, in Creq/Desco, *La construction de la certification : quelles méthodes, pour quels usages ?:* journées d'études Creq/Desco, Paris, 13 et 14 septembre 2001, Marseille, Cereq, 59-68.

FEUTRIE M. (2004), « Une autre évaluation, une autre validation pour l'expérience », dans *Education Permanente,* n°158, 1, .99-114.

Le BOTERF G. (2002), *L'ingénierie et l'évaluation des compétences*, Paris, Ed. d'Organisation.

MORISSE M. (à paraître 2007), Propos sur la mise en mots de savoirs formels et informels dans l'écriture de dossier de VAE, *Actes du 3ème colloque mondial de l'auto-formation*, Université Cadi Ayyad, Marrakech.

MORISSE M.(2006), L'écriture réflexive est-elle formative ? in Cros F. (ed.), Ecrire sur sa apratique pour développer des compétences professionnelles, Paris, L'Harmattan.

MORISSE M., PAGONI M., (2006), *Analyse de l'activité de jurys en V.A.E. dans le secteur agricole : enjeux de construction d'une « posture évaluative »,* FAFSEA, Université de Lille 3, laboratoire Théodile.

VERGNAUD G. (1996), « La théorie des champs conceptuels», in *Brun J. (éds), Didactique des mathématiques,* Neuchâtel, Délachaux et Niestlé, 197-274.

VERMERSCH P. (1994), *L'entretien d'explicitation en formation initiale et en formation continue,* Paris, ESF.

VERMERSCH P., MAUREL M. dir (1997), *Pratiques de l'entretien d'explicitation,* Paris, ESF.

TOUPIN L. (1995), *De la formation au métier,* Paris : ESF.

SYMPOSIUM 4 : PROBLÈMES MÉTROLOGIQUES ET VALIDATION D'OUTILS D'ÉVALUATION

Stéphane Germain, Centre Hospitalier Affilié universitaire de Québec

Christina St-Onge, université Laval

Pierre Valois, Université Laval

Marie-Hélène Hébert, Université Laval

Eric Frenette, Université du Québec à Rimouski

Résumé

Divers types de problèmes métrologiques sont associés à la construction et à la validation d'outils d'évaluation en éducation, de même qu'à l'interprétation des scores individuels qui en découle. Le présent symposium à justement pour objectif de traiter de quelques-uns de ces problèmes et d'y apporter des solutions ou du moins des pistes de recherche.

Ce symposium s'inscrit dans le continuum d'une réflexion engagée depuis plusieurs années sur le rôle des théories de la mesure dans le processus de validation des outils d'évaluation. Dans la continuité de ces travaux de recherche, nombre de chercheurs ont été amené à réfléchir sur les façons d'estimer les erreurs de mesure et, partant, d'en restreindre les effets. Pour ce faire certains d'entre eux ont eu recours à des modèles paramétriques, d'autres à des modèles non paramétriques.

Dans ce symposium, les six présentateurs abordent le problème de validité sous différents angles paramétriques et non paramétriques. La première communication traite du problème de la fidélité d'outils d'évaluation des apprentissages dans une approche par compétence. Au moyen d'exemples traduisant la réalité québécoise, la communication montrera l'intérêt de la théorie de la généralisabilité pour assurer la fiabilité des mesures de rendement scolaire obtenues. À cet effet, les principaux fondements théoriques et pratiques propres à une étude de généralisabilité seront abordés. On verra comment s'assurer si les résultats observés sont accidentels, ou s'ils seraient répétables à d'autres correcteurs, d'autres moments, d'autres instruments, etc.

Les deuxième et troisième (étude de type Monte Carlo) communications portent respectivement sur le concept de biais, notamment sur les procédures utilisées afin de vérifier les différentes formes de biais et sur la robustesse des tests d'ajustement des modèles de réponse à l'item eu égard à différents types de biais.

La quatrième communication consiste en une étude Monte Carlo et vise à éprouver l'effet de la procédure d'estimation de CCI utilisée sur les taux de détection de trois

indices dérivés de la TRI en fonction de la distribution des données et du nombre de sujets. Cette présentation mettra aussi l’emphase sur l’importance d’évaluer l’ajustement des données au modèle de la TRI utilisé.

Enfin, la cinquième communication traitera de quelques modèles et méthodes d'estimation des CCI paramétriques et non paramétriques, puis de la théorie de la régularisation de Tikhonov et de son application possible au problème de l'estimation des CCI. Finalement, une librairie de routines numériques pour mettre en pratique différentes méthodes d’estimation sera présentée, ainsi qu'un module R et un module Excel convivial faisant usage de cette librairie.

Enfin, la dernière communication porte sur le développement d’un module Excel d’analyse d’item convivial et facilement accessible aux enseignants des différents ordres d’enseignement. Un exemple d’application sera présenté en mettant notamment l’accent sur les courbes caractéristiques d’items estimées à l’aide de la régression par noyau.

http://www.jemacs.uni.lu/ 2007 volume 1, No 1

ESTIMATIONS DES COURBES CARACTÉRISTIQUES D'ITEMS PAR MAXIMUM DE VRAISEMBLANCE MARGINALE PÉNALISÉE

Stéphane Germain, Centre Hospitalier Affilié universitaire de Québec
Pierre Valois, Belkacem Abdous, Université Laval,

Résumé :

Divers types de problèmes métrologiques sont associés à la construction et à la validation d'outils d'évaluation en éducation, de même qu'à l'interprétation des scores individuels qui en découle. Le présent symposium à justement pour objectif de traiter de quelques-uns de ces problèmes et d'y apporter des solutions ou du moins des pistes de recherche.

Ce symposium s'inscrit dans le continuum d'une réflexion engagée depuis plusieurs années sur le rôle des théories de la mesure dans le processus de validation des outils d'évaluation. Dans la continuité de ces travaux de recherche, nombre de chercheurs ont été amené à réfléchir sur les façons d'estimer les erreurs de mesure et, partant, d'en restreindre les effets. Pour ce faire certains d'entre eux ont eu recours à des modèles paramétriques, d'autres à des modèles non paramétriques.

Dans ce symposium, les six présentateurs abordent le problème de validité sous différents angles paramétriques et non paramétriques. La première communication traite du problème de la fidélité d'outils d'évaluation des apprentissages dans une approche par compétence. Au moyen d'exemples traduisant la réalité québécoise, la communication montrera l'intérêt de la théorie de la généralisabilité pour assurer la fiabilité des mesures de rendement scolaire obtenues. À cet effet, les principaux fondements théoriques et pratiques propres à une étude de généralisabilité seront abordés. On verra comment s'assurer si les résultats observés sont accidentels, ou s'ils seraient répétables à d'autres correcteurs, d'autres moments, d'autres instruments, etc.

Les deuxième et troisième (étude de type Monte Carlo) communications portent respectivement sur le concept de biais, notamment sur les procédures utilisées afin de vérifier les différentes formes de biais et sur la robustesse des tests d'ajustement des modèles de réponse à l'item eu égard à différents types de biais.

La quatrième communication consiste en une étude Monte Carlo et vise à éprouver l'effet de la procédure d'estimation de CCI utilisée sur les taux de détection de trois indices dérivés de la TRI en fonction de la distribution des données et du nombre de sujets. Cette présentation mettra aussi l'emphase sur l'importance d'évaluer l'ajustement des données au modèle de la TRI utilisé.

Enfin, la cinquième communication traitera de quelques modèles et méthodes d'estimation des CCI paramétriques et non paramétriques, puis de la théorie de la régularisation de Tikhonov et de son application possible au problème de l'estimation des CCI. Finalement, une librairie de routines numériques pour mettre en pratique

différentes méthodes d'estimation sera présentée, ainsi qu'un module R et un module Excel convivial faisant usage de cette librairie.

Enfin, la dernière communication porte sur le développement d'un module Excel d'analyse d'item convivial et facilement accessible aux enseignants des différents ordres d'enseignement. Un exemple d'application sera présenté en mettant notamment l'accent sur les courbes caractéristiques d'items estimées à l'aide de la régression par noyau.

Introduction

Selon la théorie de réponse aux items (TRI), les observations indépendantes, sont des vecteurs aléatoires de réponses à des items. La loi de probabilité de ces vecteurs est supposée dépendre d'une variable aléatoire latente de telle façon que conditionnellement à cette variable latente, les items sont indépendants. Donc, la loi de probabilité des observations est complètement caractérisée par la loi de la variable latente et les lois des items conditionnelles à la variable latente.

Dans la présente communication, nous nous intéressons au cas où les items sont binaires (0 ou 1) et la variable latente, appelée l'habileté, est unidimensionnelle et distribuée selon la loi normale standard. La loi de chaque item conditionnelle à l'habileté est donc une loi de Bernoulli, dont le paramètre est une fonction de l'habileté. Ces fonctions sont appelées les courbes caractéristiques d'items (CCI).

Nous commencerons par présenter quelques modèles et méthodes d'estimation des CCI paramétriques et non paramétriques, puis nous présenterons la théorie de la régularisation de Tikhonov et nous montrerons comment nous l'avons appliquée au problème de l'estimation des CCI.

Finalement, nous présenterons une librairie de routines numériques pour mettre en pratique les méthodes présentées, ainsi qu'un module R et un module Excel convivial faisant usage de cette librairie.

Les estimateurs paramétriques

En TRI paramétrique, les CCI sont supposées suivre un certain modèle avec un nombre fini de paramètres inconnus et possiblement différents pour chaque item. À titre d'illustration, nous utiliserons le modèle logistique à deux paramètres (2PLM). Ces paramètres sont souvent appelés discrimination et difficulté.

a) Le maximum de vraisemblance jointe

La méthode du maximum de vraisemblance jointe (JMLE) consiste à maximiser la vraisemblance en fonction des paramètres d'items et des habiletés. Malheureusement, dans cette situation, la maximisation ne peut pas être séparée ni par item ni par sujet.

Cette maximisation étant donc en 2n+N dimensions dans le cas du 2PLM, elle dépasse la capacité de calcul de la plupart des ordinateurs.

Afin de contourner cette difficulté, Birnbaum dans Lord et Novick (1968) a proposé une procédure itérative en deux étapes. Au préalable, une estimation des habiletés à partir des scores brutes est obtenue en appliquant l'inverse de la fonction de répartition normale aux rangs des scores. Par la suite, les paramètres d'items sont estimés par maximum de vraisemblance en remplaçant les habiletés par leurs estimations. Enfin, les habiletés sont à leur tour estimées par maximum de vraisemblance en remplaçant les paramètres d'items par leurs estimations. Les deux dernières étapes sont itérées jusqu'à convergence.

En plus du caractère ad hoc de la méthode itérative utilisée, les habiletés doivent être centrées et réduites à chaque itération, et l'estimation par JMLE souffre aussi de l'absence de preuve théorique de sa convergence (voir Baker & Kim, 2004) pour un nombre fini d'items. La difficulté à démontrer la convergence en probabilité de cet estimateur, provient de l'estimation des habiletés nécessaire à cette méthode. En effet, lorsque le nombre d'observations augmente le nombre d'habiletés à estimer augmente aussi, mais le nombre d'items utilisés pour estimer ces habiletés reste fixe.

b) Le maximum de vraisemblance marginale

Afin de pallier aux lacunes du JMLE, Bock et Lieberman (1970) ont développé une procédure basée sur le maximum de vraisemblance marginale (MMLE) pour estimer seulement les paramètres d'items en l'absence des habiletés. La vraisemblance marginale est le la probabilité de notre ensemble d'observations en fonction des paramètres d'items. Notons que contrairement à la vraisemblance jointe, les habiletés ne sont pas libres, mais sont intégrées. C'est-à-dire que la vraisemblance marginale donne la probabilité de nos observations sans connaissance des habiletés. Le MMLE a l'avantage de converger en probabilité avec un nombre fini d'items.

Un estimateur non paramétrique

a) La régression par noyau

Si nous supposons les habiletés connues, l'estimation des CCI se résume à un problème de régressions. En effet, par définition les CCI sont des espérances conditionnelles à l'habileté. Ramsay (1991) a donc appliqué l'estimateur à noyau de Nadaraya-Watson (voir Nadaraya 1964) au problème de l'estimation des CCI.

Tout comme pour l'estimation paramétrique par JMLE, une estimation préalable des habiletés est obtenue par les rangs des scores, puis il est possible d'estimer les

habiletés de nouveau par maximum de vraisemblance avec les CCI estimées, et cette procédure peut-être itérée jusqu'à convergence. Étant donné que cette méthode produit des estimations des habiletés, elle souffre du même problème de convergence que le JMLE lorsque le nombre d'items est fixe.

Objectif

Tel que le MMLE a pallié aux limites du JMLE, nous tenterons de pallier à certaines des limites de la régression par noyau. Considérons les probabilités d'observer chacun des vecteurs aléatoires possibles. Bien qu'il soit facile de calculer ces probabilités à partir des CCI, l'opération inverse est difficile. En fait, la solution de l'opération inverse n'est même pas unique. Notons de plus, que nous n'observons pas ces probabilités, mais seulement leurs estimations à partir des observations.

Une approche classique pour résoudre ce type de problème mal posé, est la régularisation de Tikhonov (Tikhonov & Arsenin, 1977). La solution par régularisation de Tikhonov est définie comme la solution qui minimise une distance entre la solution est les observations pénalisées par une mesure de la régularité de la solution. La quantité de pénalisation appliquée dépend d'un paramètre de lissage. Cette méthode nous assure un juste milieu entre une solution exacte, mais irrégulière, et une solution très régulière, mais trop éloignée des observations. De plus, cette méthode bénéficie d'une démonstration de sa convergence en probabilité due à Vapnik et Stefanyuk (1979).

Nous avons choisi une distance et une mesure de régularité afin d'obtenir un critère semblable au MMLE, mais pénalisé (PMMLE). De plus cet estimateur a la propriété de généraliser le modèle 2PLM. En effet, plus la quantité de pénalisation est grande, plus les CCI estimées sont près du 2PLM.

La convergence en probabilité de cet estimateur sera démontrée et appuyée par des simulations. Nous présenterons des logiciels qui permettent l'application de cet estimateur.

Références

Baker, F. B. & Kim, S.-H. (2004). Item response theory : Parameter estimation techniques (Second edition). Dekker Media, New York.

Bock, R. D. & Lieberman, M. (1970). Fiting a response model for n dichotomously scored items. Psychometrika, 35, 179–197.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Addison-Wesley, Reading. With contributions by A. Birnbaum.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika, 56, 611–630.

Tikhonov, A. & Arsenin, V. (1977). Solutions of ill-posed problems. Wiley, New York.

Vapnik, V. N. & Stefanyuk, A. R. (1979). Nonparametric method for reconstructing probability densities. Automat. Remote Control, 39, 1127-1140.

EFFET DE LA PROCÉDURE D'ESTIMATION DE CCI SUR LES TAUX DE DÉTECTION DE L'INDICE D'AJUSTEMENT DES SCORES INDIVIDUELS, l_Z : UNE ÉTUDE MONTE CARLO.

Christina St-Onge, Pierre Valois, Belkacem Abdous, Université Laval

Stéphane Germain, Centre Hospitalier Affilié universitaire de Québec

Mots-clés : indices d'ajustement des scores individuels, vraisemblance, patrons de réponses

Résumé:

Une catégorie d'indices statistiques est utilisée pour évaluer la vraisemblance ou l'invraisemblance des patrons de réponses obtenus à un test ou à un questionnaire, il s'agit des indices d'ajustement des scores individuels (Drasgow, Levine, & McGlaughlin, 1987; Drasgow, Levine & Zickar, 1996; Levine & Drasgow, 1982; Levine & Rubin, 1979; Meijer, 1996, 1998; Meijer & Sijtsma, 1995). Ces indices peuvent être regroupés en deux familles : les indices heuristiques et les indices dérivés de la TRI. Les chercheurs souscrivent toutefois à l'idée voulant que les indices standardisés d'ajustement des scores individuels dérivés de la TRI sont supérieurs aux indices heuristiques même si les indices d'ajustement des scores individuels dérivés de la TRI ont quelques petites lacunes, notamment leur précision qui peut être influencée par le choix du modèle d'estimation de CCI.

Une étude Monte Carlo basée sur un devis expérimental complètement croisé (2×2×5×7) a permis d'étudier l'effet que peut avoir la procédure d'estimation de CCI sur la précision de l'indice d'ajustement des scores individuels, l_Z, en fonction de la distribution des données, de la taille d'échantillon, et du taux d'aberrance, et ce, en contrôlant pour le nombre d'items et le type d'aberrance.

Les résultats observés révèlent que, dans l'ensemble, l'utilisation du modèle logistique à deux paramètres pour estimer les CCI semble maximiser la précision de l'indice d'ajustement des scores individuels l_Z. En d'autres mots, dans les cas où l'ajustement des données aux différents modèles n'aurait pas été vérifié, le choix du modèle logistique à deux paramètres s'avérerait le plus judicieux.

Introduction

En contexte d'évaluation, plusieurs événements non planifiés peuvent avoir lieu. Par exemple, certains individus peuvent tricher ou deviner les réponses à certaines questions (Bracey & Rudner, 1992), manquer de motivation à la fin du test, devoir se familiariser avec un nouveau format de questions, manquer de temps s'ils sont trop méticuleux (Wright, 1977). De telles situations peuvent avoir d'importantes répercussions sur la vraisemblance des patrons de réponses de ces individus, voire même sur la validité de leur score.

Lors d'une évaluation on s'attend à ce que les résultats obtenus aux épreuves d'évaluation donnent une estimation de l'habileté qui soit représentative du niveau d'habileté réelle de l'individu (Smith, 1986). Différents modèles de mesure peuvent être utilisés afin de vérifier la fidélité des épreuves utilisées en contexte d'évaluation. Les modèles les plus connus sont la Théorie Classique des Tests (TCT), la théorie de la généralisabilité ainsi que la Théorie des Réponses aux Items (TRI). Par ailleurs, il existe plusieurs procédés de validation, dont celui proposé par Messick (1989, 1995a et b), qui permettent l'évaluation de la validité des scores obtenus à partir de tests ou de questionnaires. Ces procédés permettent l'étude de la fidélité et de la validité pour un groupe donné et non pour une personne en particulier. Cependant, les analyses statistiques traditionnelles ne permettent pas l'étude de la vraisemblance des patrons de réponses des individus (Harnisch & Tatsuoka, 1983; Meijer & van Krimpen-Stoop, 2001; Schmitt, Cortina & Whitney, 1993).

Une catégorie d'indices statistiques fut développée pour évaluer la vraisemblance ou l'invraisemblance des patrons de réponses obtenus à un test ou à un questionnaire, il s'agit des indices d'ajustement des scores individuels (Drasgow, Levine, & McGlaughlin, 1987; Drasgow, Levine & Zickar, 1996; Levine & Drasgow, 1982; Levine & Rubin, 1979; Meijer, 1996, 1998; Meijer & Sijtsma, 1995). Les indices d'ajustement des scores individuels peuvent être regroupés en deux familles : les indices heuristiques et les indices dérivés de la TRI. Les indices heuristiques d'ajustement des scores individuels (*group based person-fit statistics)* reposent sur la prémisse qu'un patron de réponses est adéquat lorsque les items faciles sont réussis et que les items difficiles sont échoués (Harnisch & Linn, 1981; Levine & Rubin, 1979). Une recension des écrits permet de déceler deux problèmes associés à ces indices. Dans un premier temps, Harnisch et Linn (1981) ont observé des relations linéaires et non linéaires entre la valeur des indices d'ajustement des scores individuels et le score brut des individus. Deuxièmement, il semble que la distribution des indices heuristiques soit inconnue. Il est alors impossible, selon Meijer et Sijtsma (2001), de déterminer la probabilité statistique qu'un patron de réponses soit vraisemblable ou invraisemblable.

Plusieurs chercheurs dont Drasgow, Levine et McGlaughlin (1991), Drasgow, Levine et Williams (1985), Levine et Drasgow (1983), Molenaar et Hoijtink (1990, 1996), Smith (1985, 1986), ainsi que Tatsuoka et Linn (1983) ont tenté de palier aux limites des indices heuristiques. Pour ce faire, ils ont élaboré les indices d'ajustement des scores individuels dérivés de la TRI *(IRT based person-fit statistics)*. Selon la TRI, la probabilité de chaque réponse donnée est déterminée à l'aide du niveau d'habileté du candidat et des Courbes Caractéristiques d'Items (CCI). Par exemple, pour le modèle logistique à un paramètre, si le candidat réussit les items ayant un niveau de difficulté supérieur à son niveau d'habileté et ne réussit pas les items ayant un niveau de difficulté inférieur à son niveau d'habileté, on dira du patron de réponses qu'il ne s'ajuste pas au modèle théorique.

Drasgow *et al.* (1985), Tatsuoka et Linn (1983), et Smith (1985, 1986) ont observé que les premiers indices d'ajustement des scores individuels dérivés de la TRI étaient corrélés avec le niveau d'habileté, ce qui s'avère problématique. Dans les circonstances, de nouveaux indices ont été développés, soit les indices standardisés d'ajustement des scores individuels - toujours dérivés de la TRI - lesquels permettent de corriger ce problème corrélationnel (Drasgow, Levine, & Williams, 1985;

Tatsuoka & Linn, 1983; Smith, 1985, 1986). Cependant, certains problèmes persistent dont notamment leur précision qui peut être influencée par le choix du modèle d'estimation de CCI.

Les chercheurs souscrivent toutefois à l'idée voulant que les indices standardisés d'ajustement des scores individuels dérivés de la TRI sont supérieurs aux indices heuristiques (voir Drasgow, Levine & McLaughlin, 1987; Harnisch & Tatsuoka, 1983; Li & Olejnik, 1997; Meijer & Sijtsma, 2001). Les indices d'ajustement des scores individuels de la TRI sont souvent standardisés, ce qui a l'avantage de les rendrent indépendants du niveau d'habileté. Enfin, ils semblent offrir de meilleurs taux de détection que les indices heuristiques (Drasgow, Levine & McLaughlin, 1987; Harnisch & Tatsuoka, 1983; Li & Olejnik, 1997).

Problématique

Certaines des limites des indices d'ajustements des scores individuels semblent être associées aux modèles de la TRI et aux méthodes d'estimation des CCI utilisés. Pensons notamment aux répercussions potentielles liées au choix du modèle de la TRI. Si le modèle utilisé pour prédire la réponse que devrait avoir un individu ne s'ajuste pas correctement aux données, on peut supposer que ces prédictions ne seront pas très précises (Bertrand & Blais, 2004), ce qui a été observé pour certains indices d'ajustement des scores individuels dérivés de la TRI (voir Cole, 1993 ; Drasgow, 1982 ; Noonan, Boss & Gessaroli, 1992 ; Li & Olejnik, 1997). Ces chercheurs ont observé qu'un plus grand nombre de patrons de réponses était faussement identifié comme étant invraisemblable lorsque le modèle utilisé pour estimer les CCI ne correspondait pas au modèle utilisé pour simuler les données.

À notre connaissance, bien que fondamental, cet aspect n'a pas été étudié rigoureusement à ce jour. En d'autres mots, nous ne connaissons pas l'effet que peut avoir l'utilisation d'un modèle logistique pour estimer les CCI, lorsque celui-ci suppose une distribution des données différente que celle observée, sur les taux de détection des indices d'ajustement des scores individuels dérivés de la TRI.

Objectif de l'étude

Nous proposons d'étudier, dans le cadre d'une étude Monte Carlo, l'effet que peut avoir la procédure d'estimation de CCI sur la précision de l'indice d'ajustement des scores individuels, l_z, en fonction de la distribution des données, de la taille d'échantillon et du taux d'aberrance, et ce en contrôlant pour le nombre d'items et le type d'aberrance. Cette étude nous semble pertinente, car elle nous permettra d'évaluer l'effet de l'ajustement des données au modèle de la TRI utilisé lors de la détection de patrons de réponses invraisemblables à l'aide d'un indice d'ajustement des scores individuels dérivés de la TRI.

Méthodologie

Simulation des données

Pour chaque cellule de ce devis expérimental complètement croisé ($2\times2\times5\times7$; deux tailles d'échantillon, deux taux d'aberrance, cinq méthodes d'estimation de CCI et sept distributions de données), 100 répétitions furent simulées.

Les Vecteurs de Réponses Vraisemblables (VRV).

Les Vecteurs de Réponses Vraisemblables (VRV) ont été simulés selon la méthodologie proposée par Levine et Rubin (1979). Pour chaque sujet *i*, un θ fut sélectionné au hasard dans une distribution normale (0,1) de θ. Ensuite, les scores d'items ont été créés en simulant, pour chaque item *g*, une valeur uniforme entre 0 et 1. Cette valeur était alors comparée à la probabilité de réussir l'item étant donné le niveau d'habileté θ qui avait préalablement été calculé sur la base de paramètres d'items pour le test. Ainsi, si la valeur générée était inférieure à la probabilité de réussir l'item étant donné θ, « 1 » était attribué comme score à l'item ($X_g=1$). À l'inverse, si la valeur générée était supérieure à la probabilité de réussir l'item étant donné θ alors « 0 » était attribué comme score à l'item ($X_g=0$). Cette procédure a été répétée pour chaque sujet du groupe pour ainsi former une matrice de scores.

Les distributions de données.

Afin d'évaluer l'influence que peut avoir l'interaction entre la distribution des données et le modèle d'estimation de CCI, plusieurs distributions de données furent simulées. Pour ce faire, plusieurs modèles logistiques et paramètres d'items ont été éprouvés, pour un total de sept distributions différentes (voir le Tableau 1). Pour chacune de ces distributions, les paramètres de difficulté étaient générés aléatoirement à partir d'une distribution uniforme [-2.5, 2.5].

Tableau 1- Distribution des données

Distribution	Modèle	Étendue de la discrimination	Étendue de la pseudo-chance
1PLM	1PLM	a = 1	
2PLM; disc. élevée	2PLM	a ~ U[1.0, 1.5]	
2PLM; disc. faible	2PLM	a ~ U[.5, 1.0]	
3PLM; disc. élevée & chance faible	3PLM	a ~ U[1.0, 1.5]	c ~ U[0.15, 0.25]
3PLM; disc. élevée & chance faible	3PLM	a ~ U[1.0, 1.5]	c ~ U[0.0, 0.15]
3PLM; disc. faible & chance élevée	3PLM	a ~ U[.5, 1.0]	c ~ U[0.15, 0.25]
3PLM; disc. faible & chance élevée	3PLM	a ~ U[.5, 1.0]	c ~ U[0.0, 0.15]

Les tailles d'échantillon

Cole (1993) ainsi que Drasgow (1982) ont observé que la taille d'échantillon avait peu d'effet sur les taux de détection d'indices d'ajustement des scores individuels. Cependant, ils avaient utilisé des groupes de plus de 500 sujets. Désirant

étudier l'effet que peut avoir, indirectement, un petit échantillon sur la précision d'un indice d'ajustement des scores individuels dérivé de la TRI, nous avons choisi de simuler des groupes de 50 et 500 observations dans le cadre de cette étude. Quoiqu'il n'est pas recommandé d'utiliser les modèles paramétriques avec aussi peu que 50 sujets, nous avons jugé essentiel d'inclure un petit groupe dans notre simulation afin d'en apprendre davantage sur l'effet que peut avoir cette variable sur la précision de l'indice d'ajustement des scores individuels, l_Z.

Les méthodes d'estimation de CCI

Nous avons retenu cinq procédures d'estimation de CCI différentes, soit la régression Nadaraya-Watson (voir Ramsay, 1991), la régression logistique locale (voir Cleveland & Loader, 199), les modèles logistiques à deux et trois paramètres estimés par maximum marginal de vraisemblance (voir Baker & Kim, 2004). Les paramètres d'items utilisés pour simuler les données furent utilisés comme cinquième méthode pour tracer les CCI. L'objet de cette étude ne traitant pas essentiellement de ces méthodes d'estimation, le lecteur curieux est avisé de consulter les textes mis en références.

Le nombre d'items

Les études antérieures démontrent que les taux de détection des indices d'ajustement des scores individuels augmentaient avec le nombre d'items (Emons, Glas, Meijer, & Sijtsma, 2003; Emons, Sijtsma & Meijer, 2004; Karabatsos, 2003; Li & Olejnik, 1997; Meijer, 1994a, 1994b; 1996; Nering & Meijer, 1998; Reise & Due, 1991; Rogers & Hattie, 1987; Rudner, 1983). Dans les circonstances, nous avons décidé de fixer le nombre d'items à 40, ce qui est représentatif des tests utilisés en évaluation des apprentissages.

Les types et les taux d'aberrance

Le type d'aberrance généré consistait en des scores faussement élevés. La méthodologie de Levine et Rubin (1979) reprise et modifiée par Drasgow (1982) a été utilisée pour simuler ce type d'aberrance. Premièrement, des patrons de réponses associés à un $\theta > 0$ furent choisis au hasard. Des items étaient ensuite choisis aléatoirement. Toutefois, une modification fut apportée à la procédure de Drasgow : seulement les items échoués étaient choisis aléatoirement pour ensuite leur attribuer la bonne réponse. Les patrons de réponses aberrants ont été simulés avec des taux d'aberrance de 20 % et 40 %.

Le taux d'aberrants

Le pourcentage de patrons de réponses aberrants dans chaque groupe simulé a été fixé à 5 %. Afin de refléter la réalité autant que possible, les fonctions de réponses aux items seront estimées à partir du groupe total, soit à partir des patrons de réponses normaux et ceux aberrants. Les travaux de Levine et Drasgow (1982) sont importants à ce sujet, puisqu'ils ont démontré qu'un indice d'ajustement des scores individuels dérivés de la TRI peut être résistant à l'erreur d'estimation des paramètres lorsqu'il y a peu de patrons de réponses aberrants.

Analyses Statistiques

Les courbes "Receiver Operating Characteristic (ROC)"

Les courbes *Receiver Operating Characteristic (ROC)* permettent de présenter non seulement le pourcentage de vrais positifs correctement identifiés, mais aussi le pourcentage de faux positifs, vrais négatifs et faux négatifs observés. De plus, l'aire sous la courbe est souvent utilisée en tant qu'indice de la précision d'une courbe ROC (voir Hanley, 1989). Ainsi, ces valeurs seront utilisées pour présenter la précision de l_Z pour chaque croisement des niveaux de facteurs retenus.

Résultats et discussion

Le degré de précision de l_Z, c'est à dire l'aire sous les courbes (AUC) ROC, pour chaque croisement des facteurs retenus se trouvent dans le Tableau 2. On peut y observer que la précision de cet indice d'ajustement des scores individuels dérivés de la TRI semble être peu influencée par la taille de l'échantillon. Par exemple, pour 20 % d'aberrance, une distribution *2PLM-h* et des CCI estimés avec le 2PLM, la précision de l_Z est de .91 et de.92 pour les groupes de 50 et 500 sujets, respectivement. Ces résultats convergent avec ceux de Cole (1993) et Drasgow (1982) qui ont observé que la taille d'échantillon avait peu d'effets sur les taux de détection des indices d'ajustement des scores individuels.

Tableau 2- La précision (AUC) de l_Z pour tous les croisements de facteurs

	20					40				
n	LF	TG	2P	3P	Par.	LF	TG	2P	3P	Par.
					1PLM					
50	.68	.64	.87	.85	.85	.48	.44	.78	.81	.86
500	.67	.83	.87	.88	.88	.51	.78	.84	.83	.86
					2PLM-h					
50	.71	.75	.91	.88	.92	.54	.51	.82	.86	.89
500	.73	.84	.92	.92	.93	.61	.81	.89	.90	.90
					2PLM-l					
50	.62	.6	.77	.78	.80	.34	.36	.66	.7	.75
500	.62	.79	.80	.80	.81	.42	.71	.74	.72	.77
					3PLM-hh					
50	.65	.63	.84	.84	.87	.35	.36	.74	.76	.80
500	.64	.77	.86	.87	.88	.40	.7	.75	.76	.80
					3PLM-hl					
50	.71	.68	.87	.88	.94	.47	.47	.83	.87	.88
500	.7	.82	.90	.90	.92	.52	.77	.87	.85	.88
					3PLM-lh					
50	.49	.52	.73	.65	.82	.20	.24	.51	.47	.61
500	.54	.72	.73	.73	.77	.28	.63	.57	.58	.6
					3PLM-ll					
50	.58	.55	.74	.78	.80	.30	.29	.63	.69	.73
500	.59	.76	.77	.79	.79	.37	.68	.7	.71	.7

L'effet du taux d'aberrance, sur la précision de l_Z, n'est pas aussi faible que celui de la taille d'échantillon. Il est possible d'observer une nette diminution de la précision lorsqu'on augmente le taux d'aberrance. Notons, par exemple, que la précision de l_Z est de .91 dans la condition suivante : la distribution suit une 2PLM-

h, 50 sujets composent le groupe, les CCI sont estimées avec le 2PLM et le taux d'aberrance est fixé à 20 %. Cette précision diminue à .82 lorsque le taux d'aberrance est fixé à 40 %. Ces résultats contredisent ceux de Drasgow, Levine et McLaughlin (1987), Emons, Sijtsma et Meijer (2004), Karabatsos (2003), Levine et Rubin (1979), Nering et Meijer (1998), Raîche et Blais (2003) ainsi que Rudner (1983) qui avaient observé que les taux de détection d'indices d'ajustement des scores individuels augmentaient lorsque le taux d'aberrance augmentait. Cependant, les résultats observés convergent avec ceux de Emons, Sijtsma et Meijer (2004), ainsi que Glas et Meijer (2003). Ce résultat suggère qu'il pourrait exister un taux d'aberrance optimal. Ce taux d'aberrance optimal maximiserait la précision de l'indice d'ajustement des scores individuels, ce qui signifie qu'un taux d'aberrance plus faible ou plus élevé serait associé à une diminution de la précision de l'indice. Il est clair que d'autres études sont nécessaires pour étudier, plus en détail, ce phénomène.

Tel qu'il a été observé par d'autres chercheurs (Cole, 1993; Emons, Sijtsma & Meijer, 2004; Meijer, 1994; Meijer, Molenaar & Sijtsma, 1994), la discrimination des items a eu un effet considérable sur la précision de l_Z. Il est possible d'observer que la précision de l_Z augmentait lorsque les items étaient plus discriminants, ce qui était prévisible. Lorsque les items sont discriminants, on peut supposer que les patrons de réponses correspondant à des niveaux d'habiletés plus élevés diffèrent plus des patrons de réponses représentant des niveaux d'habiletés plus faibles que lorsque les items ne sont pas discriminants. Ainsi, lorsque la réponse à certains items discriminants est modifiée pour simuler un type donné d'aberrance, il peut être supposé que le patron de réponses différait plus des autres patrons de réponses que si les réponses des items peu discriminants étaient modifiées pour simuler un type donné d'aberrance.

Notons que l'effet les méthodes non paramétriques d'estimations de CCI fut testé, car nous pensions que ces méthodes pourraient avoir un meilleur ajustement aux données avec de petits groupes (voir Ramsay, 1991; Santor & Ramsay, 1998). Selon les résultats observés, il apparaît que la précision de l_Z était plus élevée lorsque les CCI étaient estimées à partir de modèles paramétriques, et ce, même dans le cadre du petit échantillon de 50 sujets.

De plus, les méthodes non paramétriques furent choisies, car les modèles paramétriques peuvent être trop restrictifs dans certaines occasions (Santor & Ramsay, 1998) et que l'utilisation d'un modèle qui n'est pas adéquatement ajustée aux données pourrait faussement faire augmenter les taux de détection d'indices d'ajustement des scores individuels dérivés de la TRI (voir Cole, 1993; Drasgow, 1982; Noonan, Boss & Gessaroli, 1992; Li & Olejnik, 1997). Étant donné que les méthodes non paramétriques n'imposent pas un modèle aux données, il y avait lieu de croire que leur utilisation pourrait faire augmenter la précision d'indices d'ajustement des scores individuels dérivés de la TRI. Au contraire, il semble que la précision de l'indice l_Z ait été influencée par la présence de vecteurs de réponses aberrants dans l'échantillon utilisé pour estimer les CCI, ce qui peut rendre moins précis l'évaluation de la vraisemblance des patrons de réponses.

Lorsque nous comparons la précision de l_Z dans le cas où des courbes estimées sont utilisées et non les vrais paramètres d'items, on peut observer (Figure 1)

que dans la plupart des situations, l_Z étaient plus précis lorsqu'il était utilisé avec les CCI estimées par le 2PLM. Cette observation a également été faite pour quelques distributions simulées avec un modèle à trois paramètres. On peut donc supposer que l'effet de pseudo-chance peut diminuer l'identification de patrons de réponses aberrants. Ainsi donc, l'utilisation du 2PLM pour estimer les CCI pourrait, dans ce cas, maximiser la précision de l_Z .

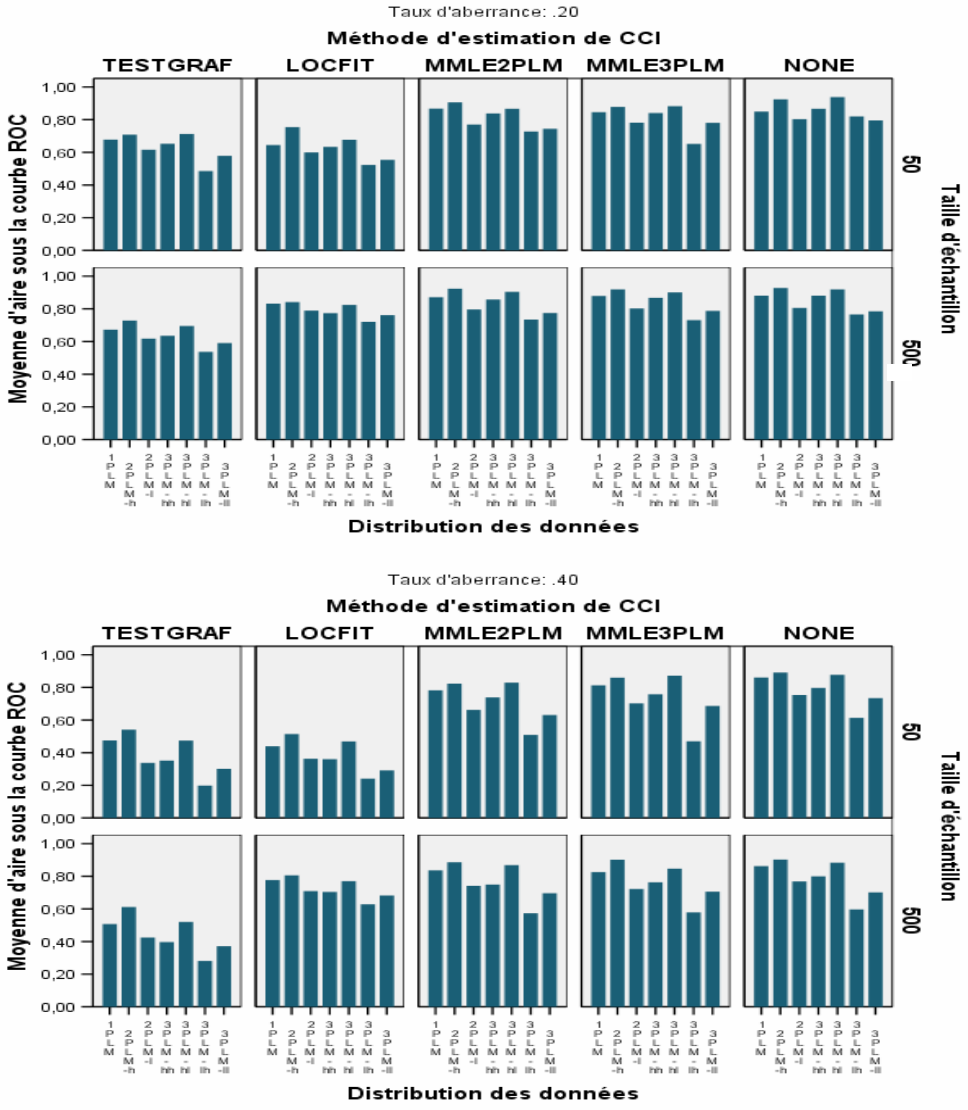


Figure 1 : Représentation graphique de la moyenne d'aire sous les courbes ROC de l_Z

L'examen des histogrammes de la Figure 2 révèle qu'en général, l_Z est plus précis lorsqu'il est utilisé avec des CCI estimées avec le modèle qui s'ajuste aux données. Par exemple, on peut remarquer que la précision de l_Z est de 91 % lorsque la distribution est 2PLM-h, il y a 50 sujets, 20 % et les CCI sont estimées avec le 2PLM tandis que la précision diminue à 88 % lorsque les CCI sont estimées avec le 3PLM. Cependant, nous pouvons remarquer que lorsque les items sont peu discriminants et que le facteur de pseudo-chance est élevée (la distribution 3PLM-lh), la précision de l_Z est de 65 % lorsque les CCI sont estimées avec le 3PLM, mais qu'elle augmente à 73 % lorsque les CCI sont estimées avec le 2PLM. De plus, lorsque les données suivent une distribution 3PLM-hh, l_Z a une précision de 88 % lorsqu'il est utilisé avec des CCI estimées par le 3PLM tandis que sa précision est de 87 % lorsque les CCI sont estimées par le 2PLM.

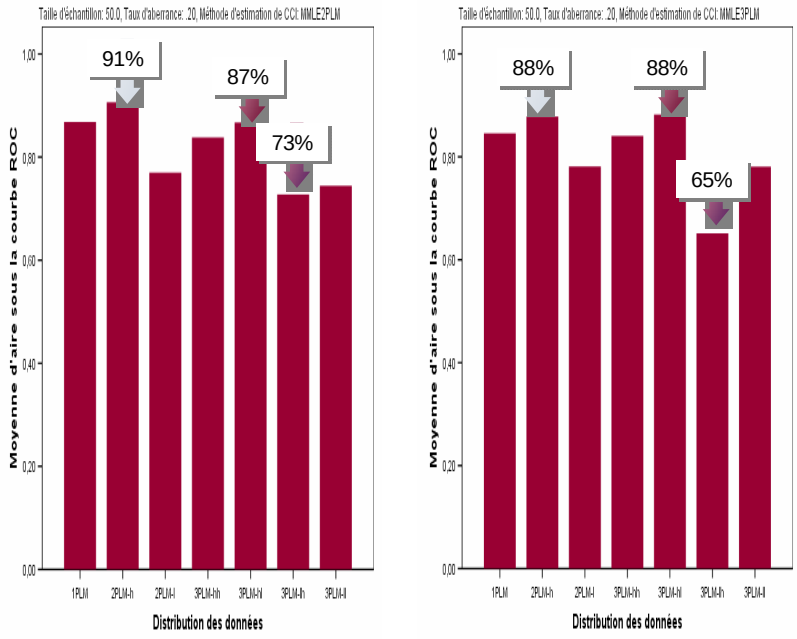


Figure 2 : La précision de l_Z pour les sept distributions, un groupe de 50 sujets et 20% d'aberrance lorsque les CCI sont estimées avec le 2PLM et le 3PLM.

Somme toute, les résultats observés révèlent que, dans l'ensemble, l'utilisation du modèle logistique à deux paramètres pour estimer les CCI semble

maximiser la précision de l'indice d'ajustement des scores individuels l_Z . En d'autres mots, dans les cas où l'ajustement des données aux différents modèles n'aurait pas été fait, le choix du modèle logistique à deux paramètres s'avérerait le plus judicieux.

Références

Baker, F.B., Kim, S.-H. (2004). *Item Response Theory*. Marcel Dekker, Inc: New York, USA.

Bertrand, R. & Blais, J.-G. (2004). *Modèles de mesure : L'apport de la théorie de réponse aux items*. Presses de l'Université du Québec : Ste-Foy, Canada.

Cleveland, W.S., Loader, C. (1996). Smoothing by local regression: Principles and methods. In W. Haerdle and M. G. Schimek, (Eds.), *Statistical Theory and Computational Aspects of Smoothing*, 10-49. Springer, New York.

Cole, G. (1993). *A Monte Carlo Study of the effects of four factors on the effectiveness of the* l_Z *and ECI4z appropriateness indices.* Unpublished doctoral dissertation. University of Ottawa, Ontario, Canada.

Drasgow, F. (1982). Choice of test models for appropriateness measurement. *Applied Psychological Measurement, 6,* 297-308.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59-79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171-191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasurement individuals. *Applied Measurement in Education, 9,* 47-64.

Emons, W. H. M., Glas, C. A. W., Meijer, R. R., & Sijtsma, K. (2003). Person Fit in order-restricted latent class. *Applied Psychological Measurement, 27,* 459-478.

Hanley, J.A. (1989). Receiver Operating Characteristic (ROC) methodology". *Critical Reviews in Diagnostic Radiology, 29.* 307-335.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns : Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133-146.

Harnisch, D. L., & Tatsuoka, K. K. (1983). *A comparison of appropriateness indices based on item response theory.* In R. K. Hambleton (ed.) Applications of item response theory (pp.104-122) Vancouver, Canada: Kluwer.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology, 35,* 42-56.

Levine, M. V., & Drasgow, F. (1983). *Appropriateness measurement: Validating studies and variable ability models.* In D. J. Weiss (ed.) iNew horizons in testing: Latent trait test theory and computerized adaptive testing (pp.109-113) New York: Academic Press.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of educational statistics, 4,* 269-290.

Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21,* 215-231.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan Publishing Company.

Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14,* 5-8.

Messick, S. (1995b). Validity of psychological assessment : Validation of inferences person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Meijer, R. R. (1996). The influence of the presence of deviant score patterns on the power of a person-fit statistic. *Applied Psychological Measurement, 20,* 141-154.

Meijer, R. R. (1998). Consistency of test behaviour and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology, 71,* 147-160.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new development. *Applied Measurement in Education, 8,* 261-272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology Review: Evaluation Person-Fit. *Applied Psychological Measurement, 25,* 107-135.

Meijer, R. R., & van Krimpen-Stoop, E. M. L. A. (2001). *Person fit across subgroups: An achievement testing example.* In A. Boomsma, M.A. van Duijn, & T.A.B. Snijders (Eds.)., Essays on item response theory, 377-390. New-York: Springer-Verlag.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55,* 75-106.

Molenaar, I. W., & Hoijtink, H. (1996). Person-fit and the Rasch model with an application to knowledge of logical quantors. *Applied Measurement in Education, 9,* 27-45.

Noonan, B. W., Boss, M. W., & Gessaroli, M. E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indices. *Applied Psychological Measurement, 16,* 345-352.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Schmitt, N. S., Cortina, J. M., & Whithney, D. J. (1993). Appropriateness fit and criterion related validiy. *Applied Psychological Measurement, 17,* 143-150.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 46,* 433-444.

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement, 46,* 359-372.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7,* 81-96.

Wright, B.D. (1977) Solving measurement problems with the Rasch model. *Journal of Educational Measurment, 14,* 97-115.

DÉVELOPPEMENT D'UN MODULE EXCEL D'ANALYSE D'ITEMS

Pierre Valois, Christina St-Onge, Université Laval
Stéphane Germain, Centre Hospitalier Affilié universitaire de Québec

Résumé :

Divers types de problèmes métrologiques sont associés à la construction et à la validation d'outils d'évaluation en éducation, de même qu'à l'interprétation des scores individuels qui en découle. Le présent symposium à justement pour objectif de traiter de quelques-uns de ces problèmes et d'y apporter des solutions ou du moins des pistes de recherche.

Ce symposium s'inscrit dans le continuum d'une réflexion engagée depuis plusieurs années sur le rôle des théories de la mesure dans le processus de validation des outils d'évaluation. Dans la continuité de ces travaux de recherche, nombre de chercheurs ont été amené à réfléchir sur les façons d'estimer les erreurs de mesure et, partant, d'en restreindre les effets. Pour ce faire certains d'entre eux ont eu recours à des modèles paramétriques, d'autres à des modèles non paramétriques.

Dans ce symposium, les six présentateurs abordent le problème de validité sous différents angles paramétriques et non paramétriques. La première communication traite du problème de la fidélité d'outils d'évaluation des apprentissages dans une approche par compétence. Au moyen d'exemples traduisant la réalité québécoise, la communication montrera l'intérêt de la théorie de la généralisabilité pour assurer la fiabilité des mesures de rendement scolaire obtenues. À cet effet, les principaux fondements théoriques et pratiques propres à une étude de généralisabilité seront abordés. On verra comment s'assurer si les résultats observés sont accidentels, ou s'ils seraient répétables à d'autres correcteurs, d'autres moments, d'autres instruments, etc.

Les deuxième et troisième (étude de type Monte Carlo) communications portent respectivement sur le concept de biais, notamment sur les procédures utilisées afin de vérifier les différentes formes de biais et sur la robustesse des tests d'ajustement des modèles de réponse à l'item eu égard à différents types de biais.

La quatrième communication consiste en une étude Monte Carlo et vise à éprouver l'effet de la procédure d'estimation de CCI utilisée sur les taux de détection de trois indices dérivés de la TRI en fonction de la distribution des données et du nombre de sujets. Cette présentation mettra aussi l'emphase sur l'importance d'évaluer l'ajustement des données au modèle de la TRI utilisé.

Enfin, la cinquième communication traitera de quelques modèles et méthodes d'estimation des CCI paramétriques et non paramétriques, puis de la théorie de la régularisation de Tikhonov et de son application possible au problème de l'estimation des CCI. Finalement, une librairie de routines numériques pour mettre en pratique

différentes méthodes d'estimation sera présentée, ainsi qu'un module R et un module Excel convivial faisant usage de cette librairie.

Enfin, la dernière communication porte sur le développement d'un module Excel d'analyse d'item convivial et facilement accessible aux enseignants des différents ordres d'enseignement. Un exemple d'application sera présenté en mettant notamment l'accent sur les courbes caractéristiques d'items estimées à l'aide de la régression par noyau.

Problématique

Bon nombre de chercheurs souscrivent à l'idée que l'analyse d'items est une étape essentielle pour apprécier les qualités métrologiques de l'ensemble d'un outil d'évaluation et des items le constituant (Auger, 2000). Ce type d'analyse nous permet de juger de la qualité métrologique des résultats obtenus lors d'une évaluation.

Or, dans les différents milieux d'enseignement, il arrive trop souvent que l'analyse d'items ne soit pas effectuée et que des décisions soient prises à partir de scores dont nous ne connaissons pas les qualités métrologiques. Pourquoi en est-il ainsi ? Une réponse possible est sans doute l'ampleur de la tâche qu'on associe avec cette analyse. En effet, les enseignants doivent connaître et avoir accès à des logiciels spécialisés en statistiques ou avoir suffisamment de connaissances pour programmer ces calculs à l'aide de logiciels plus courant. De plus, il faut, par la suite, que les enseignants interprètent les résultats obtenus qui ne sont pas toujours présentés dans une forme très conviviale.

Comment pourrions-nous faire pour inciter les professionnels, responsables d'évaluation, à procéder à une analyse d'items ? Est-ce qu'un logiciel convivial pourrait les aider à mieux réaliser une analyse d'items ?

Objectif

L'objectif de la présente communication consiste à proposer un outil informatique convivial pouvant inciter les enseignants a effectué l'étape d'analyse d'items pour mieux juger les qualités métrologiques de leurs outils d'évaluation.

Les critères retenus lors du développement de l'outil informatique étaient :

1) la facilité d'utilisation (convivialité),
2) la possibilité de faire une analyse d'items selon la théorie classique des tests et/ou une analyse selon un modèle de la théorie de réponses aux items,
3) l'intégration dans un tableur connu tel qu'EXCEL,
4) l'émission d'un rapport simple et facile à interpréter, et
5) aucun frais pour l'utilisateur.

Module Excel : analyse d'items

Étant donné les objectifs de facilité d'utilisation et de moindre coût d'utilisation, l'outil informatique a été développé comme un module Excel. Une fois installé, l'utilisateur peut avoir accès au module par l'entremise d'une option ajouté au menu « outils », et ce, à partir de n'importe quelle feuille de calcul Excel. Cette intégration dans Excel permet aux utilisateurs familiers avec ce logiciel ou tout autre tableur électronique semblable, d'être à l'aise rapidement avec l'outil et évite aussi d'avoir à réinventer un format de fichier et une interface graphique. Par exemple si les données à analyser sont déjà dans une feuille de calcul Excel, l'utilisateur doit tout simplement sélectionner la matrice des réponses et cliquer sur le bouton « courbes caractéristiques d'items », à titre d'exemple..

Une fois installé, le module crée deux options dans le menu « outils » d'Excel, soit celles d'analyse d'items et de courbes caractéristiques d'items. Parmi les statistiques données par l'option analyse d'items, on retrouve des statistiques descriptives, telles que la moyenne, l'écart type, la médiane, l'indice interquartile et le alpha de Cronbach, ainsi que une analyse d'items selon la théorie classique des tests qui comprend la moyenne à l'item, la difficulté, la discrimination et les scores minimum et maximum. L'option des courbes caractéristiques, quant à elle, produit les courbes caractéristiques de chaque item, estimé à l'aide de la régression par noyau (voir Ramsay 1991, 2000).

Mise à l'essai du module

Phase 1 : Test par des spécialistes en mesure et évaluation dans le cadre du cours « Évaluation des apprentissages au secondaire »

Nous avons testé ce module Excel dans le cadre du cours « Évaluation des apprentissages au secondaire », dispensé par Internet lors de la session d'hiver 2006 à 145 étudiants. Les résultats que nous présenterons sont ceux obtenus dans le cadre du premier travail qui fut évalué à partir de huit critères.

Phase 2 : Test par des chercheurs habitués avec ce genre d'analyse

Ultérieurement, nous demanderons à quelques chercheurs d'utiliser ce logiciel pour faire les analyses d'items qu'ils auraient habituellement faites avec un logiciel tel que

SPSS. Nous recueillerons leurs commentaires afin d'améliorer le module Excel et le guide d'utilisation pour faciliter son utilisation.

Phase 3 : Test avec des utilisateurs potentiels non habitués avec ce genre de logiciel et ce genre d'analyse

Avant de rendre ce module Excel disponible gratuitement sur le web, nous procéderons à une dernière étape de vérification. Nous solliciterons l'aide d'enseignants du primaire et du secondaire pour mettre à l'épreuve ce module Excel. Nous leur remettrons le guide de l'utilisateur pour les aider à installer et utiliser le module Excel. Nous recueillerons leurs commentaires et impression afin de procéder à l'amélioration dudit module.

Conclusion

Jusqu'à présent le module Excel s'est avéré très convivial et utile pour l'analyse d'items. Les résultats fournis par les différentes options du module ont été d'une très grande utilité. Prochainement, nous développerons une option qui permettra d'obtenir un profil d'apprentissage pour chacun des étudiants.

FIABILITE D'OUTILS D'EVALUATION DES APPRENTISSAGES DANS UNE APPROCHE PAR COMPETENCES : L'APPORT DE LA THEORIE DE LA GENERALISABILITE

Marie-Hélène Hébert, Université Laval

Virginie Duclos, Université Laval

Mots-clés : théorie de la généralisabilité, évaluation des compétences, réforme québécoise de l'éducation

Résumé :

Avec l'arrivée du *Programme de formation de l'école québécoise* (Gouvernement du Québec, 2001), axé sur le développement de compétences, il n'y a pas que l'organisation scolaire et la pédagogie qui ont été bousculées. Les pratiques évaluatives aussi. Étroitement liées à l'apprentissage, celles-ci ont dû être ajustées à cette réforme du système d'éducation québécois et, par conséquent, au renouveau pédagogique qui s'installe dans les établissements scolaires de la province. Selon certains chercheurs, il n'y a pas si longtemps encore, par l'entremise d'items à choix multiples notamment, les enseignants québécois ne se contentaient bien souvent que de donner un score à un élève. Question d'arithmétique et d'objectivité. Mais aujourd'hui, dans cet univers de productions complexes, le jugement des enseignants et son train de subjectivité prend beaucoup d'importance (Scallon, 2004). Or, ce jugement dit d'« experts », il faudra le justifier : les différents acteurs du système scolaire québécois demanderont des comptes, d'où la nécessité de contrôles de qualité. Contrôles de qualité fort justifiés, puisqu'il semblerait que les résultats obtenus avec ces nouvelles formes d'évaluation ne soient pas des plus fiables (Parkes, 2001). Vu ce manque supposé de fiabilité, il est tout à fait légitime de se questionner : une observation conduite un jour donné, avec un tel outil et un tel correcteur, mène-t-elle à un résultat différent de celui qu'on obtiendrait à un tout autre moment, avec un tout autre évaluateur et un tout autre instrument ? Au moyen d'un exemple traduisant la réalité québécoise, la communication montrera l'intérêt de la théorie de la généralisabilité pour assurer la fiabilité des mesures de rendement scolaire obtenues. À cet effet, les principaux fondements théoriques et pratiques propres à une étude de généralisabilité seront abordés. On verra comment s'assurer si les résultats observés sont accidentels, ou s'ils seraient répétables à d'autres correcteurs, d'autres instruments, etc.

Problématique

Avec l'arrivée du nouveau *Programme de formation de l'école québécoise* (Gouvernement du Québec, 2001), axé sur le développement de compétences, il n'y a pas que l'organisation scolaire et la pédagogie qui ont été bousculées. Les pratiques évaluatives aussi. Étroitement liées à l'apprentissage, celles-ci ont dû être ajustées à

cette réforme du système d'éducation québécois et, par conséquent, au renouveau pédagogique qui s'installe peu à peu dans les établissements scolaires de la province.

Selon certains chercheurs, il n'y a pas si longtemps encore, par l'entremise d'items à choix multiples notamment, les instituteurs québécois ne se contentaient bien souvent que de donner un score à un élève. Score obtenu en comptant, évidemment, le nombre d'items réussis. Mais aujourd'hui, dans cet univers de productions complexes, le jugement des enseignants et son train de subjectivité prend beaucoup d'importance (Scallon, 2004). Or, ce jugement dit d'« experts », il faudra tôt ou tard le justifier : les différents acteurs (élèves, parents, administrateurs) du système scolaire québécois demanderont des comptes, d'où la nécessité de contrôles de qualité.

Contrôles de qualité fort justifiés, puisqu'il semblerait que les résultats obtenus avec ces nouvelles formes d'évaluation ne soient pas des plus fiables (Linn, 1994 ; Parkes, 2001). Vu ce manque supposé de fiabilité, il est tout à fait légitime de se questionner : une observation conduite un jour donné, avec un tel outil et un tel correcteur, mène-t-elle à un résultat différent de celui qu'on obtiendrait à un tout autre moment, avec un tout autre évaluateur et un tout autre instrument (interrogation inspirée d'un passage du livre de Cardinet & Tourneur, 1985, p.7) ? Pas forcément. Les juges sont vraisemblablement de sévérité différente, les tâches de difficulté différente et les occasions d'évaluer, si nombreuses. Mais est-ce à la limite de l'acceptable ? Est-ce aussi dans la marge de fiabilité conventionnellement acceptée ?

Effectivement, n'est-il pas véridique d'affirmer que certains enseignants démontrent, par instants, de la tolérance envers certains élèves ? Malgré l'utilisation d'échelles uniformes et descriptives, l'effet de halo s'en trouverait par moments omniprésent. Que dire à présent du degré de sévérité propre à un correcteur ? Si certains se plaisent dans la sévérité, d'autres, en échange, baignent dans l'indulgence. Au chapitre des tâches maintenant, certaines situations seraient ou bien corsées ou bien aisées pour tout étudiant, alors que d'autres ne le seraient que pour certains élèves. Puis, n'oublions guère le moment de l'évaluation. Un élève grippé réussira-t-il aussi bien un lundi matin qu'un vendredi après-midi, en meilleure forme ? Un correcteur sera-t-il de même sévérité face à une calligraphie médiocre ? À 8 h qu'à 22 h la même journée ?

En somme, la réussite d'une tâche, voire la compétence qu'elle prétend inférer, serait-elle indépendante du correcteur, de la tâche et du moment où l'évaluation fut pratiquée ? Ou, au contraire, en serait-elle tributaire ? À la lumière de ce qui vient d'être énoncé, sanctionner la capacité d'un élève au moyen d'un nombre restreint de situations d'évaluation serait donc une utopie, à moins que l'on en vienne à prouver que la tâche, le correcteur et l'occasion choisis n'importent.

Théorie de la généralisabilité

Pour y arriver ou du moins tenter d'y arriver, on se tournera vers un outil communément appelé *la théorie de la généralisabilité* (Bain & Pini, 1996 ; Cardinet & Tourneur, 1985). Un outil statistique dont l'utilité peut être ainsi résumée : ne pouvant connaître le véritable potentiel d'un élève qu'en lui présentant toutes les

tâches imaginables du domaine d'appartenance de la compétence à inférer pour ensuite les voir annoter par tous les correcteurs admissibles ou par un juge dit infaillible, ce qui est en pratique illogique, on devra plutôt se contenter de déterminer, par le biais d'une étude de généralisabilité, si le résultat moyen observé, avec un nombre de tâches et de correcteurs limité, s'en trouve généralisable à tous les correcteurs et tâches concevables. Dans l'affirmative, on pourra dès lors certifier que le résultat moyen observé, moyennant quelques tâches et juges sélectionnés, traduit le véritable potentiel de l'élève.

Pour savoir si effectivement les résultats observés au moyen de quelques tâches et de quelques correcteurs échantillonnés s'en trouvent généralisables à tous les correcteurs et à toutes les tâches imaginables, un coefficient de généralisabilité sera alors calculé. À chaque étude, selon l'intention de la mesure, un choix d'indice devra néanmoins s'imposer : le coefficient de généralisabilité dite « relative » s'emploiera pour hiérarchiser ou comparer entre eux les différents objets de mesure (p. ex. : qui est le meilleur élève de la classe ?), tandis que le coefficient de généralisabilité dite « absolue » s'utilisera pour comparer les scores observés à un standard quelconque (p. ex. : qui passe le seuil de 60 % ?) (Bain & Pini, 1996).

Objectif de la recherche

Au moyen d'un exemple traduisant la réalité québécoise, la communication montre l'intérêt de la théorie de la généralisabilité pour assurer la fiabilité des mesures de rendement scolaire obtenues. À cet effet, les principaux fondements théoriques et pratiques propres à une étude de généralisabilité sont abordés. On verra comment s'assurer, au moyen du logiciel spécialisé EduG (version 1.3), si les résultats observés sont accidentels, ou s'ils seraient répétables à d'autres correcteurs, d'autres instruments, etc.

Un exemple fictif

À titre d'exemple, supposons une classe de 32 élèves de 6e année, trois correcteurs issus d'une même équipe-école et quatre situations-problèmes à accomplir visant à sanctionner leur capacité à *résoudre une situation-problème mathématique*[1]. Situations-problèmes fictives annotées par les trois correcteurs fictifs au moyen de la grille d'appréciation suivante, fidèle au nouveau *Programme de formation de l'école québécoise* pour l'éducation préscolaire et l'enseignement primaire (p.127) :

Tableau 1 - Grille d'appréciation de la compétence *Résoudre une situation-problème mathématique* inspirée du *Programme de formation de l'école québécoise.*

Critères	Très satisfaisant	Satisfaisant	Insatisfaisant	Nettement insatisfaisant
Production d'une solution correcte : démarche et résultat	3	2	1	0
Explicitation (orale ou écrite) des éléments pertinents de la solution	3	2	1	0
Explicitation adéquate (orale ou écrite) de la validation de la solution	3	2	1	0

[1] Compétence tirée du *Programme de formation de l'école québécoise* (Gouvernement du Québec, 2001, p.126).

Une simple multiplication (32 élèves x 3 correcteurs x 4 tâches) nous permet dès lors de constater qu'on se retrouve avec 384 résultats globaux à traiter (exprimés en pourcentages) pour mettre en évidence les différentes moyennes observées entre les élèves, les tâches et les correcteurs, comme en témoigne le tableau tronqué 2, dont l'original, de loin trop volumineux pour être présenté dans sa totalité, comporte pas moins de 32 colonnes :

Tableau 2 - Résultats fictifs et moyennes fictives (M), exprimés en pourcentages, de 32 élèves fictifs (E) à quatre tâches fictives (T), corrigées par trois correcteurs fictifs (C).

	E_1				E_2				..	E_{31}				E_{32}			
	T1	T2	T3	T4	T1	T2	T3	T4		T1	T2	T3	T4	T1	T2	T3	T4
C_1	100	89	100	89	99	88	99	99		78	89	88	88	88	88	99	88
C_2	100	100	99	99	88	99	100	99		99	88	88	88	99	88	88	88
C_3	100	100	100	99	89	99	99	99		78	67	88	88	89	78	67	67
M	95,42				88,08					78,92				78,92			

Suite à l'examen des résultats de ce tableau, nous sommes à même de nous questionner : les valeurs moyennes fictives observées (M) reflètent-elles le véritable potentiel de chaque élève à *résoudre une situation-problème mathématique* ? Ou, au contraire, sont-elles simplement le reflet de fluctuations aléatoires introduites par la démarche de mesure elle-même, à savoir les quatre tâches échantillonnées et les trois correcteurs ciblés (interrogation inspirée d'un passage du livre de Cardinet & Tourneur, 1985, p.7) ? Chaque élève n'étant exposé qu'à un échantillon de quelques tâches et sa performance n'étant évaluée que par trois correcteurs, il y a fort à parier que ces 384 résultats, dont sont à peine reproduites ci-haut 48 entrées, ne sont pas exempts d'erreur de mesure. Les juges sont vraisemblablement de sévérité différente et les tâches de difficulté différente. Mais est-ce à la limite de l'acceptable ? Dans la marge de généralisabilité conventionnellement acceptée ? Un ensemble de 384 résultats étant de beaucoup trop volumineux pour visualiser à l'œil nu l'importance des fluctuations d'échantillonnage affectant vraisemblablement les valeurs moyennes observées, tels une interaction Élèves x Tâches indue ou encore un effet Correcteurs x Tâches plus élevé que ce qui peut être toléré, le recours à la théorie de la généralisabilité, et à ses quatre phases caractéristiques, s'avère ici nécessité.

Phase d'observation

La phase d'observation se veut la genèse d'une étude de généralisabilité. En prenant conscience des conditions et des contextes propres au dispositif de mesure, elle requiert l'identification des facettes, le dénombrement des niveaux observés de celles-ci et la précision des relations qu'elles entretiennent entre elles.

Dans l'exemple mentionné au début de cette sous-section, trois sources de variation ou facettes peuvent être identifiées : les Élèves (E), les Correcteurs (C) et les

Tâches (T). L'élève E_1 et ses 31 camarades ayant été soumis à quatre situations visant à inférer leur capacité à *résoudre une situation-problème mathématique*, situations ayant été annotées par la correctrice C_3 et deux autres collègues, les niveaux observés des facettes Elèves, Tâches et Correcteurs se chiffrent respectivement à 32, 4 et 3. Ces informations sont résumées au tableau 3 suivant :

Tableau 3 - Dispositif d'évaluation : plan d'observation.

Facettes	Niveaux observés	Nombre de niveaux
E	E_1 à E_{32}	32
T	T_1, T_2, T_3, T_4	4
C	C_1, C_2, C_3	3

Au regard du dispositif de mesure nous préoccupant, les quatre situations d'évaluation ayant été corrigées par tous les correcteurs, les trois correcteurs ayant annoté tous les élèves et les 32 élèves ayant pris part à toutes les situations, les données exploitées s'en trouvent structurées suivant un plan d'observation à trois facettes croisées Correcteurs x Élèves x Tâches, lequel se représente visuellement par le diagramme d'Euler-Venn suivant :

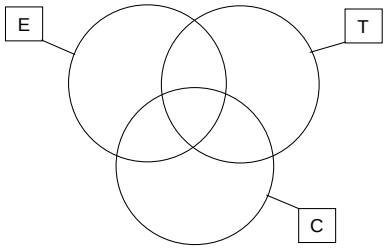


Figure 1 - Illustration du plan d'observation à trois facettes croisées du dispositif d'évaluation CET.

Les facettes identifiées, les niveaux observés dénombrés et les interrelations énumérées, le plan d'observation de l'exemple fictif ne saurait toutefois être complet sans une énumération exhaustive des sept sources de variation ou d'erreur pouvant, de près ou de loin, assombrir la fiabilité d'un tel dispositif d'évaluation : Élèves, Tâches, Correcteurs, Élèves x Tâches, Élèves x Correcteurs, Tâches x Correcteurs et un résidu, Élèves x Tâches x Correcteurs, confondu avec les autres sources d'erreur non incluses dans le dispositif d'évaluation (écoles, classes, occasions, etc.). Le diagramme d'Euler-Venn ci-après en fait foi :

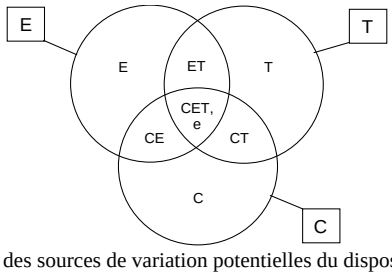


Figure 2 - Illustration des sources de variation potentielles du dispositif d'évaluation CET.

Les sources de variation maintenant énumérées, il ne reste plus qu'à dériver les carrés moyens, préludes à leur quantification, pour affirmer complétée la phase d'observation propre à une étude de généralisabilité. Conséquemment à ce qui avait été annoncé, les étapes dites de calculs ont été ignorées, quoique assurées par un logiciel spécialisé. En appliquant aux données fictives partiellement compilées au tableau 2 le logiciel des études de généralisabilité EduG, nous obtenons, au terme de la phase d'observation, le tableau 4 ci-dessous :

Tableau 4 - Sources de variation et calcul des carrés moyens du dispositif d'évaluation CET.

Plan d'observation			
Sources de variation	Sommes des carrés	Degrés de liberté	Carrés moyens
C	174,1302	2	87,0651
E	14488,8333	31	480,2849
T	532,7917	3	177,5972
CE	401,3698	62	6,4737
CT	221,0990	6	36,8498
ET	12567,5417	93	135,1349
CET	344,0677	186	1,8498

Phase d'estimation

Au sortir de la phase d'observation, la seconde phase, celle d'estimation, demande à ce que soit précisée la façon dont les niveaux observés des facettes considérées ont été échantillonnés (Mokonzi, 2003). Suivant le mode d'échantillonnage privilégié, trois types de facettes sont alors plausibles : facette fixée, facette aléatoire finie et facette aléatoire infinie (Bain & Pini, 1996).

Pour l'exemple à l'origine de ce texte, en supposant que les 32 élèves aient été choisis parmi des milliers d'élèves disponibles, que les quatre situations d'évaluation aient été puisées à même des milliers de tâches admissibles et que les trois correcteurs chargés de les annoter aient été ciblés parmi des centaines d'individus aptes à les apprécier, force est d'admettre que les facettes ont été échantillonnées au hasard plutôt que considérées dans leur totalité. Elles sont dites d'une part aléatoires. D'autre part, puisqu'elles ont été constituées à partir de populations théoriquement illimitées, l'adjectif « infini » leur est attribué : ce sont des facettes aléatoires infinies. Le tableau 5, en plus d'y préciser facettes et niveaux observés des facettes, relate ces informations :

Tableau 5 - Dispositif d'évaluation : plans d'observation et d'estimation.

Facettes	Niveaux observés	Niveaux admissibles	Type
E	32	Infini	Aléatoire infinie
T	4	Infini	Aléatoire infinie
C	3	Infini	Aléatoire infinie

À ce stade, sur la base des informations fournies aux phases d'observation et d'estimation, une analyse de la variance estime l'influence, voire l'importance, des

diverses sources de variation précédemment identifiées (Bain & Pini, 1996). La phase d'estimation tirant son nom de l'estimation des composantes de variance aléatoires, celles-ci sont dérivées des carrés moyens calculés à l'issue du plan d'observation. Dans la perspective de l'analyse de la variance suit le calcul des composantes de variance mixtes, puis celui des espérances de variance mixtes. En appliquant aux données partiellement compilées au tableau 2 le logiciel EduG, nous obtenons, au sortir de la phase d'estimation, le tableau 6 suivant, quasiment semblable, quoique plus détaillé que le tableau 4 précédent :

Tableau 6 - Analyse de la variance du dispositif d'évaluation CET.

Plan d'observation				Plan d'estimation				
Sources de variation	Sommes des carrés	Degrés de liberté	Carrés moyens	Comp. aléatoires	Comp. mixtes	Espér. variance mixtes	Erreurs types	%
C	174,1302	2	87,0651	0,3562	0,3562	0,3562	0,5021	0
E	14488,8333	31	480,2849	28,3772	28,3772	28,3772	9,9882	37
T	532,7917	3	177,5972	0,0777	0,0777	0,0777	1,2031	0
CE	401,3698	62	6,4737	1,1560	1,1560	1,1560	0,2900	1
CT	221,0990	6	36,8498	1,0938	1,0938	1,0938	0,5758	1
ET	12567,5417	93	135,1349	44,4283	44,4283	44,4283	6,5361	57
CET	344,0677	186	1,8498	1,8498	1,8498	1,8498	0,1908	2

Phase de mesure

Au sortir de la phase d'estimation, on est à même d'entreprendre la phase de mesure, où commence, au dire des auteurs consultés, l'application de la « véritable » théorie de la généralisabilité. L'analyse de la variance étant maintenant une étape franchie, entrent ici en jeu des concepts fort essentiels à une étude de généralisabilité, notamment les notions de variance vraie, de variance d'erreur et de coefficients de généralisabilité.

Alors que l'analyse de la variance traitait toutes les facettes « sur un même pied » (Cardinet & Tourneur, 1985), le plan de mesure demande à ce qu'un « distinguo » soit introduit entre elles : dans le jargon de la généralisabilité, l'appellation « facette de différenciation » est employée pour qualifier toute facette cible de la mesure (objet d'étude), alors que le vocable « facette d'instrumentation » est réservé à toute facette instrument de la mesure (condition d'observation) (Mokonzi, 2003). Dans l'exemple ici analysé, l'intention de la mesure étant de déterminer si les valeurs moyennes présentées au tableau 2 témoignent du véritable potentiel des élèves *à résoudre une situation-problème mathématique* plutôt que de fluctuations aléatoires introduites par

la démarche de mesure elle-même, la facette Élèves s'en trouve cataloguée « facette de différenciation ». Ces valeurs moyennes ayant été observées au moyen de quelques tâches et de quelques correcteurs, les facettes Tâches et Correcteurs s'en trouvent, quant à elles, étiquetées « facette d'instrumentation ».

Des suites de ce *distinguo* et du calcul préalable des espérances de variance mixtes sont calculées ce qu'on nomme dans la perspective de l'analyse de la généralisabilité les variances vraie et d'erreur. Les espérances de variance mixtes sont, pour ce faire, réparties en deux catégories : les unes, exprimant les différences réelles entre les objets de la mesure par rapport à la caractéristique mesurée, correspondent à la variance dite vraie ; les autres, exprimant les différences entre les diverses conditions d'observation, correspondent, après s'être vu appliquer une pondération liée au nombre de niveaux observés et au mode d'échantillonnage des facettes considérées, à la variance dite d'erreur (Cardinet & Tourneur, 1985).

Cela étant, l'étude de généralisabilité aboutit à l'étape décisive où l'analyse permet le calcul de deux indices de généralisabilité, le coefficient de généralisabilité relative[2] et le coefficient de généralisabilité absolue[3], indiquant la précision avec laquelle on peut généraliser, d'un résultat obtenu dans des conditions particulières, à la valeur théorique recherchée (Cardinet & Tourneur, 1985), à savoir le véritable potentiel de l'élève. En plus de réorganiser les espérances de variance mixtes du tableau 6 en fonction du plan de mesure privilégié, le tableau 7 synthétise, pour la situation fictive nous intéressant, les résultats de l'analyse des paramètres de généralisabilité obtenus au moyen du logiciel EduG. Une fois de plus, les formules théoriques et le détail des étapes dites de calculs ont été ignorés.

Tableau 7 - Analyse de la généralisabilité du plan de mesure E/CT.

Sources de variation	Variance vraie	Sources de variation	Variance d'erreur relative	Variance d'erreur absolue	%
		C		0,1187	0
E	28,3772				
		T		0,0194	0
		CE	0,3853	0,3853	3
		CT		0,0911	0
		ET	11,1071	11,1071	93
		CET	0,1542	0,1542	1
Totaux	28,3772		11,6466	11,8759	
Coefficient de généralisabilité relative 0,7090 ; absolue 0,7050 (calculs non montrés ici)					

Au regard du tableau 7, plusieurs statistiques intéressantes semblent s'en dégager. D'une part, si on veut déterminer quels élèves se situent en deçà du seuil critique de 60 %, le coefficient de généralisabilité absolue est à examiner. D'autre part, s'il s'agissait plutôt de sélectionner les quatre meilleurs élèves pour former une équipe toutes étoiles en vue d'un concours régional de résolution de problèmes

[2] Obtenu à partir de la variance d'erreur relative.

[3] Issu de la variance d'erreur absolue.

mathématiques, le coefficient de généralisabilité relative est à prioriser. Un coefficient de généralisabilité supérieur à 0,80 étant souhaité, le dispositif à trois facettes croisées CET serait ni approprié à la sélection des élèves doués (0,7090), ni à l'identification des élèves en voie d'échouer (0,7050).

Somme toute, une attitude critique s'imposait : les valeurs moyennes observées sont davantage le reflet de fluctuations aléatoires, introduites par la démarche de mesure elle-même, à savoir tâches et correcteurs échantillonnés, que du véritable potentiel des élèves à *résoudre une situation-problème mathématique*. Chaque élève n'étant exposé qu'à un échantillon de quelques tâches et sa performance n'étant appréciée que par quelques correcteurs, les 384 résultats fictifs partiellement rapportés au tableau 2 ne seraient pas exempts d'erreur de mesure.

À cet effet, se fiant à l'importance relative en pourcentages rapportée à la dernière colonne du même tableau 7, la composante de variance concourant le plus à l'erreur de mesure dans l'appréciation de la performance mathématique des élèves serait associée à l'interaction statistique Élèves x Tâches (ET). Effectivement, sa valeur (11,1071) représente pas moins de 95 % de la variance d'erreur relative (ou 95 % du total de la colonne de gauche) et 93 % de la variance d'erreur absolue (ou 93 % du total de la colonne de droite). Les élèves réagiraient différemment aux différentes tâches proposées, et ce, indépendamment de leur capacité à résoudre une situation-problème mathématique et du niveau de difficulté des situations-problèmes auxquelles on les a confrontés.

Phase d'optimisation

Les valeurs de généralisabilité relative et absolue trouvées n'attestant pas d'une bonne fiabilité du dispositif d'évaluation pour l'objectif visé, à savoir inférer la capacité des élèves *à résoudre une situation-problème mathématique*, la phase d'optimisation vise à améliorer, voire à optimiser, l'efficience du dispositif d'évaluation proposé (Bain & Pini, 1996 ; Cardinet & Tourneur, 1985). L'objectif de l'exemplification n'étant au départ que d'illustrer si les valeurs moyennes observées au tableau 2 reflétaient ou non le véritable potentiel d'un élève à *résoudre une situation-problème mathématique*, voilà pourquoi il a été choisi de se limiter à une présentation très sommaire de cette quatrième et dernière phase propre à une étude de généralisabilité[4].

En gros, le logiciel spécialisé EduG pourrait être mis de nouveau à contribution en reprenant tous les calculs après avoir augmenté le nombre de niveaux observés des facettes d'instrumentation (Tâches et Correcteurs). Ainsi, il se pourrait qu'en doublant le nombre de tâches administrées ou en triplant le nombre de correcteurs mandatés, les coefficients de généralisabilité atteignent des valeurs acceptables pour prendre de bonnes décisions. L'approche demeure théorique et il revient aux « décideurs » d'apprécier les coûts que peut entraîner cette optimisation du dispositif d'évaluation : correcteurs à engager, tâches à créer et à administrer, etc.

[4] Pour le détail des pistes d'optimisation, on se rapportera à Cardinet et Tourneur (1985).

Conclusion

À ce stade, une étude de généralisabilité peut être dite achevée. Étude fictive il va sans dire, mais qui avait pour but de montrer l'intérêt de la théorie de la généralisabilité pour assurer la fiabilité des mesures de rendement scolaire obtenues dans le cadre de la mise en œuvre du nouveau *Programme de formation de l'école québécoise*.

Bien que les tenants d'une approche par compétences aient souvent tendance à privilégier la validité au détriment de la fiabilité, il n'en demeure pas moins que des résultats fiables, conformes au véritable potentiel de l'élève, ne soient pas à négliger. Comme en témoigne la *Politique d'évaluation des apprentissages* (Gouvernement du Québec, 2003) nouvellement rafraîchie au goût du jour :

> La rigueur se traduit par une évaluation soucieuse d'exactitude et de précision. Elle est soutenue par une démarche formelle ou informelle, selon les situations. L'utilisation d'une instrumentation de qualité pour la collecte des données sur l'apprentissage et pour leur interprétation contribue à assurer la fidélité de l'évaluation. Il est essentiel que les informations recueillies soient pertinentes et suffisantes si l'on veut se prononcer sur les apprentissages des élèves. Une évaluation rigoureuse doit conduire à poser les jugements les plus justes possibles afin de prendre des décisions et de mener des actions qui vont servir à faire progresser l'élève, à l'orienter dans son cheminement scolaire et à reconnaître officiellement ses apprentissages. (Gouvernement du Québec, 2003, p.11)

Et si d'un seul coup, par manque de justice, d'équité, d'égalité, de cohérence, de rigueur et de transparence, nos diplômes perdaient toute crédibilité ?

Références

Bain, D., & Pini, G. (1996). *La généralisabilité : mode d'emploi*. Genève : Centre de recherches psychopédagogiques.

Cardinet, J., & Tourneur, Y. (1985). *Assurer la mesure*. Berne : Peter Lang.

EduG (Version 1.3) [Logiciel informatique]. (2002). Lévis : EDUCAN.

Gouvernement du Québec (2001). *Programme de formation de l'école québécoise. Éducation préscolaire. Enseignement primaire*. Québec : ministère de l'Éducation.

Gouvernement du Québec (2003). *Politique d'évaluation des apprentissages*. Québec : ministère de l'Éducation.

Linn, R. L. (1994). Performance assessment: policy promises and technical measurement standards. *Educational researcher*, *23*, 4-14.

Mokonzi, G. B. (2003). Le modèle de la généralisabilité : une théorie de la mesure en éducation. *Mesure et évaluation en éducation, 26*, 5-18.

Parkes, J. (2001). The role of transfer in the variability of performance assessment scores. *Educational assessment, 7*, 143-163.

Scallon, G. (2004). L'évaluation des compétences et l'importance du jugement. *Pédagogie collégiale, 18*, 14-20.

http://www.jemacs.uni.lu/ 2007 volume 1, No 1

ADAPTATION OU TRADUCTION DE TEST : ÉTUDE DES DIFFÉRENTES FORMES DE BIAIS.

Eric Frenette, Université du Québec à Rimouski
Richard Bertrand, Université Laval

Résumé :

Divers types de problèmes métrologiques sont associés à la construction et à la validation d'outils d'évaluation en éducation, de même qu'à l'interprétation des scores individuels qui en découle. Le présent symposium à justement pour objectif de traiter de quelques-uns de ces problèmes et d'y apporter des solutions ou du moins des pistes de recherche.

Ce symposium s'inscrit dans le continuum d'une réflexion engagée depuis plusieurs années sur le rôle des théories de la mesure dans le processus de validation des outils d'évaluation. Dans la continuité de ces travaux de recherche, nombre de chercheurs ont été amené à réfléchir sur les façons d'estimer les erreurs de mesure et, partant, d'en restreindre les effets. Pour ce faire certains d'entre eux ont eu recours à des modèles paramétriques, d'autres à des modèles non paramétriques.

Dans ce symposium, les six présentateurs abordent le problème de validité sous différents angles paramétriques et non paramétriques. La première communication traite du problème de la fidélité d'outils d'évaluation des apprentissages dans une approche par compétence. Au moyen d'exemples traduisant la réalité québécoise, la communication montrera l'intérêt de la théorie de la généralisabilité pour assurer la fiabilité des mesures de rendement scolaire obtenues. À cet effet, les principaux fondements théoriques et pratiques propres à une étude de généralisabilité seront abordés. On verra comment s'assurer si les résultats observés sont accidentels, ou s'ils seraient répétables à d'autres correcteurs, d'autres moments, d'autres instruments, etc.

Les deuxième et troisième (étude de type Monte Carlo) communications portent respectivement sur le concept de biais, notamment sur les procédures utilisées afin de vérifier les différentes formes de biais et sur la robustesse des tests d'ajustement des modèles de réponse à l'item eu égard à différents types de biais.

La quatrième communication consiste en une étude Monte Carlo et vise à éprouver l'effet de la procédure d'estimation de CCI utilisée sur les taux de détection de trois indices dérivés de la TRI en fonction de la distribution des données et du nombre de sujets. Cette présentation mettra aussi l'emphase sur l'importance d'évaluer l'ajustement des données au modèle de la TRI utilisé.

Enfin, la cinquième communication traitera de quelques modèles et méthodes d'estimation des CCI paramétriques et non paramétriques, puis de la théorie de la régularisation de Tikhonov et de son application possible au problème de l'estimation des CCI. Finalement, une librairie de routines numériques pour mettre en pratique

différentes méthodes d'estimation sera présentée, ainsi qu'un module R et un module Excel convivial faisant usage de cette librairie.

Enfin, la dernière communication porte sur le développement d'un module Excel d'analyse d'item convivial et facilement accessible aux enseignants des différents ordres d'enseignement. Un exemple d'application sera présenté en mettant notamment l'accent sur les courbes caractéristiques d'items estimées à l'aide de la régression par noyau.

Introduction

Au cours des dernières décennies, les évaluations internationales en éducation ont connu un essor considérable. Entre autres, les programmes *International Assessment of Educational Progress* (Lapointe, Mead, Askew, 1992), *Programme International pour le Suivi des Acquis Scolaires* (OCDE, 2000) et *Trends International Mathematics and Science Study* (Beaton *et al.*, 1996) évaluent les compétences scolaires des élèves de différents pays, en plus de recueillir de l'information sur un ensemble de données sociodémographiques. Parmi les grands enjeux des programmes d'évaluation internationale, il y a celui de fournir des comparaisons justes et équitables de la réussite des élèves, et ce, en dépit des différences culturelles des pays participants. Une grande importance doit donc être accordée à l'équivalence entre les adaptations ou traductions de tests selon les langues ou les cultures, car il en va de la justesse des comparaisons entre les pays (Ercikan, 1998; Hambleton, 1993; O'Leary, 2002; Wainer, 1994). À cette fin, van de Vijer & Leung (1997) ont développé une classification à trois niveaux pour identifier les différentes sources potentielles de biais : biais de concept, biais de méthode et biais d'item.

Cadre théorique

Selon Bertrand *et al.* (2001), le biais de concept fait référence au cas où la traduction ou l'adaptation des instruments se traduirait par un glissement du sens du concept prétendument mesuré entre les différentes adaptations du test pour chacun des pays. Selon ces auteurs, le biais de méthode est présent lorsque la réponse de l'élève dépend, de façon générale, de la procédure d'administration de l'instrument, du format des questions ou encore de la façon stéréotypée de répondre de l'élève. Enfin,

le biais d'item entre les différentes adaptations d'un même test survient lorsque deux étudiants d'habileté égale n'ont pas la même probabilité de réussite à un item simplement en raison du fait qu'ils appartiennent à des groupes distincts (Angoff, 1993; Bertrand *et al.*, 2001).

Objectif de la présentation

Plusieurs procédures ont été proposées pour détecter le biais de concept (Bertrand *et al.*, 2001; Gierl, 2000; Van de vijer & Leung, 1997) et le biais de méthode (Hambleton, 2001), mais la plupart des procédures statistiques ont été développées pour contrôler le biais d'item (Mantel & Heanszel, 1959; Roussos, Schniphe, & Pashby, 1999). L'objectif de cette communication est de présenter les différentes méthodes utilisées afin de détecter ces formes de biais. Les conditions préliminaires à la vérification de ces sources de biais seront aussi présentées.

Biais de concept

Afin d'évaluer le biais de concept, il est important de s'assurer que les différentes adaptations d'un test possèdent la même structure factorielle. L'évaluation du biais de concept consiste à faire ressortir la structure factorielle du test selon les différents groupes et comparer les saturations obtenues par les groupes. Plus les facteurs comparés sont semblables, plus il y a équivalence factorielle. Bien que les procédures identifiant la structure factorielle les plus utilisées soient l'analyse factorielle confirmatoire (van de Vijver & Leung, 1997) et l'analyse factorielle exploratoire (Bertrand *et al.*, 2001), plusieurs autres procédures ont été proposées : DIMTEST, TESTFACT et NOHARM. Sur le plan de l'interprétation absolue, deux indices proposés par Van de Vijver et Leung (1997) permettent de comparer les deux structures factorielles: (1) le coefficient de corrélation et (2) le coefficient d'identité. Trois indices à interprétation relative peuvent être utilisés afin de comparer les structures factorielles : (1) le pourcentage d'items communs aux deux structures pour chacun des facteurs, (2) la distance euclidienne entre les saturations des items de chacun des facteurs des structures comparées et (3) la racine carrée moyenne.

Biais de méthode

Lors d'évaluation internationale, tant le contenu des indicateurs que le processus d'administration des instruments d'évaluation doivent être standardisés afin de permettre des comparaisons internationales valides. Le processus de standardisation permet d'atténuer le plus possible un type de biais de méthode. Une autre approche utilisée afin d'estimer le biais de méthode consiste à évaluer des concepts qui font référence explicitement ou implicitement à la motivation liée à la procédure d'administration, à la motivation que les élèves pouvaient avoir à répondre aux questions du test, ou encore à la désirabilité sociale. Par la suite, afin de détecter le biais de méthode, on peut avoir recours à l'analyse de variance et à l'analyse de covariance. L'analyse de variance est utilisée afin de vérifier si le degré associé à chacun des concepts est différent selon les régions administratives ou pour des groupes présentant des caractéristiques différentes (langue, sexe). Quant à l'analyse de covariance, elle a pour but de déterminer s'il existe des différences significatives entre certains pays quant au rendement dans une matière, et ce, en contrôlant l'effet de certaines variables nuisibles (p. ex. la désirabilité sociale) d'un pays à l'autre. Par ailleurs, le niveau de biais de méthode est estimé en comparant la différence (en valeur absolue) des moyennes ajustées et des moyennes non ajustées de chacun des pays. Plus la valeur de cet écart sera élevée (et négative), plus l'ajustement des scores moyens par la différence au niveau du concept sera grand et donc plus le biais de méthode sera considéré important.

Biais d'item

Avant de vérifier le biais d'item entre différents pays, il doit être vérifié en fonction des caractéristiques d'un pays (sexe, langue, régions administratives). Le biais d'item affecte grandement la propriété d'invariance et, par conséquent, rend plus difficile les comparaisons entre les pays. Dans le but d'identifier le biais d'item, plusieurs procédures ont été proposées : la régression logistique (Camilli & Shepard, 1994; Clauser & Mazor, 1998), la méthode de Mantel-Haenszel (Holland & Thayer, 1986), l'indice NCDIF non compensatoire (Raju, van der Linden, & Feer, 1995), l'approche par comparaison de modèle (Thissen, Steinberg, & Wainer, 1993), la méthode basée

sur l'aire (indice UPD; Shepard, Camilli, & Williams, 1984) et TESTGRAF (Ramsay, 1998). Une analyse de ces différentes méthodes sera discutée.

Références

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In p. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.3-23). Hillsdale, NJ: Lawrence Erlbaum.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics assessments in the middle school years : IEA's Third International Mathematics and Science Study (TIMSS)*. Chesnut Hill, MA : Boston College.

Bertrand, R., Boiteau, N., Gauthier, N., Compain, C., Frenette, É., Laprise, A., Léger-Bourgoin, N., & Jeanrie, C. (2001). *La gestion des biais de concept, des biais de méthode et des biais d'item dans le contexte des enquêtes du Programme des indicateurs de rendement scolaire (PIRS)*. Document interne. Québec : Université Laval.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement : Issues and practices*, Spring , 31-44.

Ercikan, K. (1999). *Translation DIF in TIMSS*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Montréal.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*, 4, 280-296.

Hambleton, R.K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of psychological assessment, 9*, 57-68.

Hambleton, R.K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17,* 164-172.

Holland, P. W., & Thayer, D.T. (1986). *Differential item functioning and the Mantel-Haenszel procedure*. Technical Report. Princeton, NJ: Educational Testing Service.

Lapointe, A. E., Mead, N. A., & Askew, J. A. (1992). *Learning mathematics*. Princeton, NJ: The International Assessment of Educational Progress, Educational Testing Service.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

OCDE (2000). *Connaissances et compétences : des atouts pour la vie : Premiers résultats du PISA 2000*. Rapport international. Ottawa, Canada : Statistiques Canada.

O'Leary, M. (2002). Stability of country rankings across item formats in the TIMSS. *Educational Measurement: Issues and Practice, 21*, 27-38.

Raju, N.S, van der Linden, W.J., & Fleer, P.F. (1995). IRT-based internal measure of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 4, 353-368.

Ramsay, J. O. (1998). *TESTGRAF : A program for the graphical analysis of multiple choice test and questionnaire data.* Unpublished manuscript. Montreal, Canada : McGill University.

Roussos, L.A., Schnipke, D.L., & Pashley, P.J. (1999). A generalized formula for the Mantel-Haenszel differential item functionning parameter. *Journal of Educational and Behavioral Statistics, 24, 3*, 292-322.

Shepard, L.A., Camilli, G., & Williams, D.M. (1984), Accounting for statistical artifacts in items bias research. *Journal of Educational Statistics*, 9, 93-128.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland and H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

van de Vijver, F.J.R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

Wainer, H. (1994). Problèmes de mesure. *Mesure et évaluation en éducation, 17,* 2, 115-146.

SYMPOSIUM 5: BILANS D'INNOVATIONS EN MATIERE D'EVALUATION : ANALYSE CRITIQUE ET PERSPECTIVES

Dominique Lafontaine, Département éducation et formation, Université de Liège

Alex Blanchet, Unité de recherche pour le pilotage des systèmes pédagogiques, Lausanne

Louise Bélair, Université du Québec à Trois-Rivières

Bruno Suchaut, Institut de recherche sur l'éducation, C.N.R.S, Dijon

Anne Soussi, Service de la recherche en éducation, Genève

Martine Wirthner, Institut de recherche et de documentation pédagogique, Neuchâtel

Ladislas Ntamakiliro, Unité de recherche pour le pilotage des systèmes pédagogiques, Lausanne

Patricia Gilliéron Giroud, Unité de recherche pour le pilotage des systèmes pédagogiques, Lausanne

Walter Tessaro, Faculté de psychologie et des sciences de l'éducation, Genève

---

Résumé

Les pratiques d'évaluation scolaire se sont considérablement modifiées depuis les années 80 dans la perspective de ce qu'il est convenu d'appeler la nouvelle culture d'évaluation scolaire. Le sens et la mise en œuvre de ces changements ont été régulièrement questionnés dans le cadre de l'ADMEE comme en témoigne notamment l'ouvrage collectif édité par Figari et Achouche (2001). Le présent symposium s'inscrit dans le prolongement de ce questionnement, en analysant plus particulièrement les pratiques d'évaluation.

Le sysmposium est organisé en deux parties comprenant chacune quatre communications. La première regroupe les contributions concernant les dispositifs nationaux ou régionaux d'évaluation scolaires. Trois communications abordent les innovations assez ambitieuses en matière d'évaluation scolaire en Suisse romande : Martine Wirthner se situe au niveau de la coordination romande des innovations entreprises dans les différents cantons, Alex Blanchet se penche sur les innovations réalisées dans le canton de Vaud, Anne Soussi traite des innovations introduites dans le canton de Genève. Dans la quatrième communication, Dominique Lafontaine analyse les adaptations successives du dispositif d'évaluation externe en Belgique francophone.

La deuxième partie du symposium cible plus particulièrement deux pratiques d'évaluation émergentes. D'abord, l'évaluation externe des élèves, à travers la communication de Bruno Suchaut, concernant les évaluations nationales en France

puis celle de Ladislas Ntamakiliro et Walther Tessaro, portant sur les épreuves communes en Suisse romande. Ensuite, le dossier d'évaluation, dont le bilan et les perspectives de mise en œuvre en Suisse romande et au Québec sont analysés par Patricia Gilliéron Giroud et Louise Bélair.

Figari, G. & Achouche, M. (Eds). (2001). *L'activité évaluative réintérrogée*. Bruxelles : De Boeck.

LE DOSSIER DE L'ÉLÈVE (PORTFOLIO) DANS L'APPROCHE PAR COMPÉTENCE AU PRIMAIRE AU QUÉBEC ; QUELS APPORTS ? QUELLES PERSPECTIVES ? QUELLES LIMITES ?

Louise M. Bélair, UQTR

Mots-clés : dossier progressif, approche par compétences, évaluation-régulation-bilan

Résumé :
L'implantation d'un programme de formation à l'école primaire basé sur une approche par compétences tant disciplinaires que transversales par le Ministère de l'éducation, des loisirs et des sports (MELS) en est à sa huitième année. Au fil du temps, divers documents ministériels proposant des manières innovantes de concevoir l'évaluation formative et sommative des élèves, notamment par le biais des dossiers progressifs[1] ont été produits.

Un modèle d'analyse de l'innovation à partir des représentations des acteurs face au phénomène innovant imposé, inspiré par les travaux de Rogers et de Hall et Hord, a été mis en place par Leclerc (2005). Notre recherche consiste à vérifier les apports de l'implantation d'un dossier progressif pour l'évaluation des élèves, tant du point de vue de la régulation proactive, rétroactive et interactive que du bilan certificatif à effectuer. Ce choix méthodologique permet de mettre l'accent sur ce qui a favorisé leur adhésion personnelle et volontaire à ce changement majeur, de cerner les atouts qui ont facilité cette mise en œuvre, tout en identifiant les difficultés rencontrées. Un questionnaire a été distribué auprès des personnes intéressées afin de connaître leur niveau d'implication dans l'utilisation et la mise en œuvre d'un dossier avec leurs élèves. Nous avons sélectionné cinq enseignants démontrant le plus haut niveau d'adhésion afin de pouvoir analyser leurs propres représentations face à cet outil d'évaluation et à ses effets. Des entrevues sont en cours de réalisation auprès d'enseignants dits adoptants et innovants du primaire de la région de Trois-Rivières et de la Mauricie.

Les résultats présentés feront alors état de leurs préoccupations et de leurs représentations tout en insistant sur les forces à mettre en place pour faciliter une adhésion plus grande auprès du personnel enseignant au Québec. Nous tenterons alors de répondre à la question : Quels sont les enjeux et les implications de ces innovations pour la pratique ?

Problématique

L'implantation d'un programme de formation à l'école primaire basé sur une approche par compétences tant disciplinaires que transversales par le Ministère de l'éducation, des loisirs et des sports (MELS) en est à sa huitième année. Au fil du

1 Aussi appelés portfolios

temps, divers documents ministériels proposant des manières innovantes de concevoir l'évaluation formative et sommative des élèves, notamment par le biais des dossiers progressifs (portfolios) ont été produits. Dans un document administratif publié en 2002, le MELS a également substitué les termes « formatif » et « sommatif » par ceux de régulation (en faisant référence aux écrits de Allal en1993) et de bilan-synthèse. Ces orientations passablement récentes, d'un même qu'un changement de terminologie ont certes entraîné de l'inconfort au sein du corps enseignant notamment en raison du fait que plusieurs d'entre eux se sentent peu pou prou formés voire outillés pour entreprendre ce virage. De tels constats ne sont pas uniques à la situation québécoise puisqu'ils reflètent les tendances observées dans divers établissements en quête d'innovations pédagogiques (Gauthier et Mellouki, 2006 ; Gather-Thurler, 1996 et 1998). Toutefois, un certain nombre d'enseignants « adoptants »[2] sont entrés de plain-pied dans ce processus et ont tenté l'expérimentation de ces nouveaux dispositifs d'évaluation des compétences. (Bélair, 1999). Notre étude vise donc à répertorier quelques pratiques entourant l'usage des nouveaux outils d'évaluation par des enseignants oeuvrant au primaire.

Question de recherche :

Comment le dossier (portfolio) a-t-il été mis en œuvre, utilisé et finalement adopté par des enseignants innovants et convaincus de la réforme actuelle au Québec ?

Cadre de référence

•*Les divers types de portfolio*

Afin de mieux prendre en compte les différentes utilisations des portfolios dans le contexte de la réforme québécoise, il était nécessaire au préalable de dresser un portrait des multiples usages et des diverses fonctions qui orientent les appellations variées en vigueur. 4 types d'outils qui prévalent actuellement dans le milieu scolaire et qui traînent dans leur sillage certaines confusions dans leur mise en œuvre sont relevés.

Portfolio professionnel (Richard Desjardins)

Ce portfolio surtout utilisé dans le secteur professionnel, est élaboré pour que l'utilisateur puisse démontrer sa personnalité professionnelle en tenant compte de ses savoirs, savoirs-faire et savoirs-être (Desjardins, 2002). L'étudiant cumule ainsi des productions et effectue une synthèse, en vue de présenter l'ensemble de ses acquis et expériences à un éventuel employeur.

2 Ce terme issu de la recherche de Rogers (1995) désigne des personnes qui présentent un degré d'adhésion élevé envers une innovation.

Le processus d'évaluation entériné par ce type de portfolio cible principalement l'autorégulation par l'utilisateur puisque ce dernier s'apparente davantage à un dossier de présentation plutôt qu'à un outil permettant l'évaluation d'apprentissages.

Comme ce dossier vise une mise en valeur lors d'une entrevue éventuelle pour un emploi, il est conçu essentiellement comme une présentation pouvant laisser place à une certaine forme d'autoévaluation. En milieu scolaire, ce type de portfolio est peu utile puisqu'il n'implique aucune rétroaction de la part de l'enseignant et que conséquemment il y a absence d'interaction en vue d'une progression dans les apprentissages.

Portefeuille de compétences (Aubret 2002, Meyer 1995, Runtz-Christan, 2001)

Ce dossier personnel s'apparentant au dossier de réussite (Tardif, 2006) comporte un ensemble de pièces servant de preuves des compétences sociales et professionnelles, développées ou en voie de l'être par le titulaire, de même que la démarche suivie pour sa réalisation. Ce dossier qui se présente comme une mémoire des expériences en vue d'un emploi est utilisé entre autres pour la validation des acquis d'expériences. Il est également parfois repris dans les stages en formation professionnelle.

Quoique ce portefeuille de compétences ressemble au dossier professionnel de Desjardins (2002), il ajoute la dimension de preuves à l'appui face à ce qui est avancé par son titulaire et en ce sens, il met davantage l'accent sur l'aspect autoévaluatif. On note cependant que le formateur est toujours absent du processus d'élaboration de ce dossier.

Encore ici, on se trouve en présence d'un outil peu pertinent pour l'apprentissage des élèves puisqu'il ne permet pas la prise en compte par l'enseignant de la progression dans les apprentissages.

Dossier d'apprentissage (Doyon et Juneau, 1991 ; Goupil, 2000 ; Forgette-Giroux et Simon, 1998)

Ce recueil continu et systématique qui témoigne du progrès de l'élève se retrouve surtout en milieu scolaire. L'élève est alors partie prenante de l'évaluation du niveau de maîtrise de ses compétences.

L'évaluation globale du dossier face aux compétences visées s'effectue ici à partir de la collection de travaux de l'élève ainsi que des traces pertinentes de ses réalisations qui font foi du développement des compétences. Les travaux insérés au dossier sont ainsi sélectionnés en fonction de critères établis par l'enseignant et par l'élève (MEQ). À cet égard, le processus entourant la conception du dossier exige des moments de coévaluation par les pairs, souvent sous forme de jury, et d'autorégulation de la part de l'élève. L'enseignant porte alors un regard, autant sur l'évolution de l'élève dans le développement de ses compétences que sur le bilan et la synthèse qu'il en effectue à la fin du processus. Des critères tels la pertinence, la cohérence, la profondeur, la précision, etc permettent de juger la qualité du dossier dans son ensemble.

Ce dossier a été élaboré à l'origine pour permettre à l'élève de comprendre son évolution dans ses apprentissages. Au fil du temps, son usage a été quelque peu modifié pour permettre l'établissement par l'enseignant ou par l'élève, d'un bilan de ses acquis, mais il permet difficilement de juger du niveau de compétence des élèves. Forgette-Giroux et Simon (1998) relèvent d'ailleurs qu'il est encore difficile de définir les critères précédemment mentionnés en tenant compte de la complexité des compétences. Toutefois, la grande difficulté de cet outil réside dans le fait qu'il est souvent confondu avec le journal de bord ou avec un simple archivages de preuves à conserver pour justifier la compétence (Tardif, 2006).

Processfolio ou dossier progressif (Barth, 2004 ; Bélair, 1999)

Inscrit dans la lignée du dossier d'apprentissage, le dossier progressif s'y distingue toutefois par le fait que l'élève juge de l'évolution de son niveau de développement de compétences ainsi que par le fait que l'enseignant est médiateur et consultant dans ce processus. Ce dossier progressif exige ainsi maints aller-retours entre l'enseignant et l'élève, lesquels prennent souvent la forme d'entretiens.

La mise en place de ce dossier exige un travail d'interrelations autant entre élèves qu'entre ceux-ci et l'enseignant. Bien que les différents moments de régulations interactives permettent de statuer face à la progression dans le développement des compétences par l'élève, le processus entourant son évaluation est particulièrement alourdi (Lebel et Bélair,2004).

Ce dossier diffère du dossier d'apprentissage puisque la sélection des preuves ne s'effectue pas de la même manière. Le dossier d'apprentissage préconise la sélection des meilleurs travaux pour illustrer le niveau de compétence tandis que le dossier progressif comprend également des brouillons pour démontrer l'évolution et la progression de l'élève. Sa grande difficulté réside toutefois dans son utilisation, laquelle est lourde et souvent peu adaptée à des publics d'élèves du primaire notamment au plan de l'établissement de critères susceptibles de rendre compte de *l'évolution du développement des compétences.*

À la lecture de ces différents outils, on est appelés à constater la présence de différentes appellations dont les significations et les enjeux diffèrent. Toutefois, au-delà de cette variété, certains invariants se manifestant tout de même, et il semble important de les mettre en exergue pour une meilleure compréhension de l'utilisation de ce type d'outils dans le milieu scolaire.

Des invariants dans la recension d'écrits

Les recherches et parutions portant sur le dossier ou le portfolio situent ce type d'outil dans une perspective socioconstructiviste (Behrens, 1998 ; Bélair, 1999, 2000 ; Jorro, 1996 ; Shepard, 2000 ;Tardif, 2006 etc) mettant ainsi en parallèle le caractère social de l'outil en raison des multiples communications qu'il exige et le principe de la construction des savoirs qu'il suppose.

La recension des écrits rapporte également que tous les dossiers doivent inclure non seulement des productions, mais aussi des réflexions explicatives sur les compétences ou les apprentissages développés (Mottier et Allal, 2004 ; Bélair, 1999 ; Bourassa, 1996 ; Desjardins, 1999 ; Lafortune et Deaudelin, 2001, etc). Quoique certains types de dossiers, tels le portefeuille de compétences par exemple, misent surtout sur les productions, ils incluent souvent des notes explicatives ou argumentatives susceptibles de convaincre un éventuel employeur.

Plusieurs auteurs s'entendent aussi au niveau des diverses retombées de ces outils chez les élèves. Huba et Freed (2000) mentionnent que ces dossiers améliorent la perception de soi, Selon Allal, Riedweg et Riedweg (1999) ces outils permettent de développer l'estime de soi et favorisent les régulations métacognitives chez les élèves. Pour leur part, Gronlund et Cameron (2004) parlent du développement chez les élèves, de la communication et de la collaboration. Tardif (2006) ajoute la dimension de la prise en compte individuelle des différences en soulignant toutefois les difficultés soulevées par ce processus. Ce même auteur fait également état de deux types de dossiers : le dossier de réussites «qui vise à rendre compte du niveau de développement des compétences au terme d'une période de formation» et le dossier de progression «qui vise à documenter la construction progressive de la trajectoire de développement de chaque compétence au cours d'une période de formation» (p. 288).

Une définition générique du dossier permettant d'englober les quatre types d'outils présentés, tout en conservant les spécificités propres à chacun d'eux est énoncée par St-James (1998) qui présente le dossier comme un outil d'autoévaluation et de coévaluation dynamique facilitant le dialogue, la négociation et la réflexion sur des thèmes, des objectifs ou des compétences abordées. Ce type d'outils permet

également à l'étudiant de fournir les preuves factuelles de son niveau de compétences et de justifier la qualité de ses performances, tout en lui permettant de «prendre-parti» et de se positionner face aux compétences visées (Lebel et Bélair, 2004). Enfin, cet outil favorise la discussion et permet donc une meilleure prise en compte par l'élève de ses forces et faiblesses (Bélair, 1999).

Une grille de lecture contenant les différents paramètres reliés à ce type d'outils à été conçue à la suite de la revue de littérature. Cette grille a servi à constituer des repères pour interroger divers enseignants utilisateurs de dossiers et voir ainsi comment ils s'approprient ces dispositifs et selon quelles perspectives ils les utilisent.

Tableau 1 : Une grille de lecture émergente de la revue de littérature

La nomination	*La terminologie et la définition de l'outil* *L'auteur de référence ou encore la conception qui est sous-jacente*
Le type d'utilisation	*une discipline,* *interdisciplinaire* *par cycle* *par année scolaire*
Les parties de l'outil	*table des matières,* *sections obligatoires et facultatives*
Les procédures de régulation et de bilan	*autoévaluation,* *coévaluation,* *examen,* *observation,* *entretiens, etc*

La grille de correction (S'il y a lieu)	*Par compétence ou par objectifs,* *globale ou spécifique,* *processuelle ou produit* *notée, chiffrée, lettrée, par niveaux, ou aucune*
Les déterminants de l'évaluation	*Ce qui est visé par cet outil et pour quelles fins (quoi)* *Qui évalue et selon quelles modalités (qui)* *La procédure d'utilisation (comment)* *La raison d'être de cet outil (pourquoi)* *Les périodes d'utilisation (quand)*
Les avantages et les difficultés au regard :	*Du dispositif comme tel (présentation, organisation)* *Du type d'utilisation (cours ou stage)* *Des élèves (accueil, tâche)* *Du travail que cela exige de l'enseignant (dimensions formatives et sommatives)* *De l'institution (règlements, notation obligatoire, support)*
À titre de conclusion	*Expérience à refaire ?* *à revoir ?* *à éliminer ?*

Méthodologie

Un modèle d'analyse de l'innovation à partir des représentations des acteurs face au phénomène innovant imposé, inspiré par les travaux de Rogers et de Hall et Hord, a été mis en place par Leclerc (2005). Ce modèle d'analyse a permis de sélectionner des personnes démontrant un taux élevé d'adhésion face à cet outil d'évaluation. Ce choix méthodologique permet ainsi de mettre l'accent sur ce qui a favorisé l'adhésion personnelle et volontaire d'enseignants à ce changement majeur, de cerner les éléments de divers ordres qui ont facilité cette mise en œuvre, tout en identifiant les difficultés rencontrées. Des entrevues ont été réalisées en juin 2006 auprès d'enseignants réputés adoptants et innovants du primaire de la région de Trois-Rivières et de la Mauricie. Ces six enseignants, dont 5 femmes et un homme oeuvrent à tous les paliers d'enseignement préscolaire et primaire. Ils ont entre 12 et 30 ans d'expérience et utilisent le portfolio surtout depuis 2 à 5 ans quoique deux d'entre eux l'utilisent depuis 12 ans, donc avant la réforme actuelle.

Résultats

Les résultats qui suivent font état de leurs préoccupations et de leurs représentations face à l'utilisation du dossier. De même, des éléments entourant les enjeux et les implications de ces innovations pour la pratique étant ressortis au fil des discussions, ils seront également présentés.

Quelques constats préliminaires

Le tableau suivant présente la synthèse des entrevues menées auprès des participants.

Éléments	Compilation des réponses des 6 personnes interrogées.
La nomination	La plupart parlent de dossier d'apprentissage, 2 de portfolios. Aucun auteur de référence, tous soulignent le document officiel ministériel
Le type d'utilisation	*5 utilisent un dossier disciplinaire, 1 mélange à la fois les dimensions disciplinaires et transversales.*

	Les dossiers disciplinaires renferment plusieurs disciplines, mais elles sont juxtaposées et non intégrées.
Les parties de l'outil	*Tous utilisent un dossier ayant des sections pré-déterminées afin de faciliter le travail de lecture.*
Les procédures de régulation et de bilan	*L'idée du dossier est accompagnée la plupart du temps de la dimension autoévaluative ou encore d'une coévaluation entre l'enseignant et l'élève.* *Ils disent tous se servir du dossier pour «se donner une idée» au moment des bulletins.* *Par contre, le dossier sert principalement à l'élève lorsqu'il le présente à ses parents au moment de la remise des bulletins. Cela semble être la fonction principale du dossier pour les tous.*
La grille de correction (S'il y a lieu)	*Aucun n'a discuté spécifiquement du lien entre les compétences et le dossier. Ils ont un regard global sur son ensemble, sans grille descriptive de niveaux de compétence.* *La moitié s'en servent comme une collection de produits réalisés et choisis par l'élève, l'autre moitié veut tenir compte de la progression.*

	Aucune note globale n'est associée au dossier.
Les déterminants : Quoi Qui	*Pour les 6 personnes, le dossier sert à colliger les productions des élèves et à permettre aux élèves de présenter l'état de ses travaux à ses parents. Il donne une idée sur certaines compétences d'ordre disciplinaire.*
Comment Pourquoi	*Le dossier n'est pas évalué en tant que tel. «Ils sont bien assez évalués comme cela!»(R3). L'évaluation est faite sous forme de régulation par l'élève (autoévaluation) et son enseignant (nommé coévaluation ici).*
Quand	*L'utilisation diffère entre les participants. 1 personne parle de cahier de production à faire signer aux parents. 1 autre parle de fichier de réalisation accompagné de signets pour la production que j'ai aimé, qui m'a posé le plus défi, qui représente une faiblesse, etc.* *Pour 4 d'entre eux, le dossier sert simplement d'indicateur au moment d'insérer la note au bulletin. Par contre il permet selon 1 participant, de fournir beaucoup plus de renseignements qu'un simple B au bulletin. 1 autre négocie avec l'élève les résultats à insérer pour les productions. Cela est si riche que la plupart veut poursuivre l'expérience.* *Certains l'utilisent de manière ponctuelle au mois ou à la semaine, d'autres l'utilisent selon les besoins de chaque élève.*

Les avantages Les difficultés	*Tous les participants voient un lien direct entre le dossier et la réforme amorcée au primaire au Québec. Par contre, 2 d'entre eux utilisaient déjà un dossier avant la venue de cette réforme.* *Le principal avantage pour 1 personne, c'est de permettre de suivre les élèves qui n'ont pas le même rythme et de s'y adapter. Tous affirment que cela rend service aux élèves, ils voient où ils en sont et surtout ils sont fiers de présenter leur travail à leurs parents. Ils disent que les élèves sont fiers, que c'est la joie sans stress, qu'isl ont hâte de présenter leur dossier à leurs parents.* *Un autre avantage concerne les parents. Tous disent que ceux-ci ont une meilleure vision des apprentissages de leurs enfants et font même pression pour que les classes subséquentes utilisent cet outil.* *4 trouvent par contre que cela représente un surplus de travail, mais qui vaut la peine pour les élèves.* *Finalement, ils affirment tous que les réactions parfois négatives de collègues sont souvent dues au manque de formation et d'information, au manque de soutien et surtout à la lourdeur de la tâche.*
Expérience à refaire	*Tous veulent poursuivre l'expérience qu'ils croient positives ; ils en sont convaincus. Par contre, ils affirment que si des collègues abandonnent, c'est souvent à cause de leurs convictions personnelles, mais aussi à cause de la lourdeur du système de*

	gestion d'un tel outil, de la manipulation ou encore du manque de ressources. Aussi, la peur de ne pas attirer les parents aux rencontres avec les élèves. *Il faut définitivement conserver cet outil, mais peut-être revoir sa gestion et fournir plus d'aide.*

À la lecture des résultats des entrevues et à l'instar de la recherche de Eyssautier-Bavay (2004), on constate d'emblée qu'il semble bien exister une confusion d'ensemble, voire des usages contradictoires, puisque indépendamment du type de dossier auxquels les utilisateurs réfèrent, les caractéristiques propres à ces dossiers sont occultées.

Premier constat: les dossiers sont constitués en fonction des disciplines et non à partir des compétences. On ne peut donc pas les associer au dossier professionnel ou au portefeuille de compétences puisque les travaux ou productions ne sont pas consignés par compétences, mais bien par matières. Tout au plus on peut qualifier ce type d'outils de journaux de bord.

Second constat : ces dossiers ne sont pas spécifiquement évalués, mais servent plutôt d'appoint. Les participants parlent d'autoévaluation et de coévaluation enseignant-élève, sans faire mention des régulations continuelles. On peut donc difficilement qualifier ces outils de dossiers progressifs, puisque l'aspect fondamental qui les caractérise n'est pas invoqué. A priori, cela semble s'apparenter davantage à un dossier d'apprentissage, mais le fait qu'ils ne confèrent pas de statut formel aux évaluations et que celles-ci ne sont pas systématiquement fondées sur les compétences, ne permet pas de situer cet usage sous la bannière du dossier d'apprentissage non plus. À cet égard, aucun système impliquant un jury ou une coévaluation par les pairs n'est mentionné par les participants**.** Également ces dossiers ne servent pas à l'évaluation certificative puisqu'ils ne sont pas conçus pour permettre à l'enseignant d'effectuer un bilan des apprentissages en fonction des compétences. On peut donc conclure d'emblée que ces outils n'ont pas une fonction d'évaluation définie, tout au plus, on peut parler d'une fonction d'autorégulation pour les élèves.

Un troisième constat s'illustre dans les propos des participants: ce qu'ils nomment des dossiers, ne sont en fait que des outils de présentation des travaux effectués par les élèves à l'intention des parents. Il ne s'agit donc pas de dossiers permettant de faire état du niveau de développement des compétences, mais plutôt d'une collection de productions témoignant du travail réalisé au cours du semestre.

Finalement, des réponses formulées dans le cadre des entretiens, lesquelles portaient sur l'utilisation proprement dite de ces outils révèlent que, bien que tous

désirent poursuivre l'expérience et trouvent le dossier pertinent, des problèmes reliés à sa faisabilité, surtout dans la lourdeur reliée à la gestion de l'outil, à sa réalisation et au mode de présentation de l'outil par les élèves sont soulevés. Certains croient même que cette lourdeur est une des causes principales de la non-adhésion de leurs collègues à cette forme de consignation des travaux.

Conclusion

La plupart des participants rencontrés envisagent d'abord et avant tout ce type d'outil comme un élément motivateur pour les élèves, et mentionnent ne pas l'utiliser formellement dans un contexte d'évaluation. Cette situation qui n'est pas transférable, mais bien propre au milieu ciblé par l'étude, nous interpelle tout de même sur la «vocation» du portfolio en classe. De fait, à l'examen des résultats, on est appelé à penser que même si les participants utilisent ces types d'outils, il semble bien qu'ils le font dans la lignée de ce que Doyon et Juneau en 1991, nommaient le dossier d'apprentissage, sans avoir apporté de réelles modifications aux usages qui prévalaient à l'époque et incidemment, sans avoir inséré les compétences qui sont désormais au cœur du programme de formation de l'école québécoise. Des enseignants d'autres régions seront ultérieurement interrogés sur cette même problématique qui mérite d'être examinée.

Références.

Allal, L. & Mottier Lopez, L. (2005). L'évaluation formative : revue de publications en langue française. In *L'évaluation formative, pour un meilleur apprentissage dans les classes secondaires*, 265-290. Paris: Publications OCDE

Allal, L., Riedweg, B. & Riedweg, B. (1999). Le portfolio, un outil permettant de développer des régulations métacognitives et une meilleure estime de soi ? http://adoc38.free.fr/portfolio.htm

Allal, L. (1993). Régulations métacognitives et évaluation formative. document inédit. Université de Genève.

Aubret, J. (2002). *Le portefeuille de compétences : le portefeuille des acquis de formations et d'expérience*. Paris : ARIFOR

Barth, B.-M. (2004), "Constructivisme : quelle place pour l'évaluation ? (l'exemple du process-folio en formation universitaire)", *Revue des HEP de Suisse romande et du Tessin*, 1, 119-124.

Bataille, M. (1996). Modalités d'implication des acteurs dans les processus d'innovation. In F. Cros & G. Adamczewski (dir.). *L'innovation en éducation et en formation* (pp. 119-128). Bruxelles : De Boeck.

Behrens, M. (1998). Le portfolio dans la dynamique de l'entretien. tripartite. Mesure et évaluation en éducation Vol. 20 n°3, 5-31.

Bélair, L.M. (1999). *L'évaluation dans l'école*. Paris : ESF.

Bélair, L.M. (2000). Savoir construire la relation éducative. In M. Théberge (dir.). *Former à la profession enseignante*, 23-52. Montréal : Éditions Logiques.

Bélair, L.M. (2002).

Bourassa, M. (1996). L'évaluation-rencontre – repenser la culture de l'évaluation pédagogique, *Mesure et Évaluation en Éducation*, 19(2), 23-32.

Desjardins, F. (1999). La classe de 2025 : autonomie intellectuelle et interaction mondiale, In L.M. Bélair (dir.) *L'éducation en 2025*, 149-173. Ottawa : CFORP.

Desjardins, R. (2002). *Le portfolio de développement professionnel continu*. Montréal : De la Chenelière.

Doyon, C. & Juneau, R. (1991). *Faire participer l'élève à l'évaluation de ses apprentissages*, Montréal : Beauchemin.

Eyssautier-Bavay, C. (2004). *Le portfolio en éducation : concept et usages*, Texte réalisé dans le cadre d'une thèse, Université Joseph Fourier Grenoble 1.

Forgette-Giroux, R. & Simon, M. (1998). L'application du dossier d'apprentissage à l'université, *Mesure et Évaluation en Éducation*, 20(3), 85-103.

Gather Thurler, M. (1996). Innovation et coopération entre enseignants: liens et limites. In M. Bonami et M. Garant (dir.) *Systèmes scolaires et pilotages de l'innovation : émergence et implantation du changement*, 145-168. Bruxelles : De Boeck.

GatherThurler, M. (1998). Savoirs d'action, savoirs d'innovation. In G. Pelletier (dir.) *Former des dirigeants de l'éducation. Apprentissage dans l'action*, 101-131. Bruxelles : De Boeck.

Gauthier, C. & Mellouki, M. (2006). La recherche en formation à l'enseignement : état des lieux. Dans Gauthier, C. & Mellouki, M. (dir*). La formation des enseignants au Québec à la croisée des chemins*, 3-40. Québec : PUL.

Goupil, G. (2000). *Comment constituer mon portfolio et l'afficher sur la toile*, http://www.Edumedia.risq.qc.ca/songe.htm.

Gouvernement du Québec (2002). *L'évaluation des apprentissages au préscolaire et au primaire. Cadre de référence*. Québec : MELS.

Gronlund, N.E. & Cameron, I.J. (2004). *Assessment of Student Achievement*. Toronto : Pearson Education Canada.

Huba, M.E. & Freed, J.E. (2000). *Learnong-centered Assessment on College Compuses. Shifting the Focus from teaching to Learnong*. Boston : Allyn & Bacon.

Jorro, A, (1996). Pour une culture plurielle de l'évaluation, entre usages et archétypes, *Mesure et Évaluation en Éducation*, 19(2), 5-22.

Lafortune, L. & Deaudelin, C. (2001). *Accompagnement socioconstructiviste*, Montréal : PUQ.

Lebel, C. et Bélair, L.M. (2004). Évaluer pour trans-former, évaluer pour qualifier. *Évaluer pour mieux rendre compte*, actes du colloque AQPC (107-116). Montréal : AQPC.

Leclerc, M. (2005). *Les représentations des enseignants en regard de l'intégration des TIC dans une école élémentaire*, Thèse de doctorat, Université d'Ottawa.

Meyer N. (1995). Portefeuille de compétences à la validation des acquis In C. Hansen (dir.) *Éducation permanente*, (pp. 65-71) supplément éducation nationale, 123.

Mottier Lopez, L. & Allal, L. (2004). Le portfolio : pratiques et perspectives. Educateur (numéro spécial 04), 46-51.

Rodgers, E.M. (1995). *Diffusion on Innovation*, New-York : The Free Press.

Runtz-Christan, E. (2001), *Portefeuille de compétences : Moyen d'évaluation d'un cours universitaire*, Université de Fribourg, Département des sciences de l'éducation, Formation des enseignantes et des enseignants, Document inédit.

Savoie-Zajc, L. (1993). *Les modèles de changement planifié en éducation*,Montréal : Les éditions Logiques.

Shepard, L. (2000), The role of assessment in a learning culture, *Educational Researcher*, 29 (7), 4-14.

St-James (1998) ST. JAMES, S. (1998), “Le portfolio : un levier pour l’amélioration de l’enseignement”, *Mesure et évaluation en éducation*, 21, 2, 65-84.

Tardif, J. (2006). *L’évaluation des compétences ; documenter le parcours de développement*. Montréal : De la Chenelière-éducation.

LE DOSSIER D'EVALUATION OU PORTFOLIO EN SUISSE ROMANDE : VISEES INSTITUTIONNELLES, MISES EN ŒUVRE ET PERSPECTIVES D'AVENIR

Patricia Gilliéron Giroud, Unité de recherche pour le pilotage des systèmes pédagogiques (URSP)

Mots-clés : dossier d'évaluation ou portfolio, évaluation des apprentissages, .

Résumé :
Le dossier d'évaluation, appelé également dossier d'apprentissage ou portfolio, a été introduit dans les classes de Suisse romande dans les années 90. La mise en œuvre de ce support s'est réalisée dans le cadre de réformes scolaires plus larges. Associé à l'amélioration des pratiques professionnelles et au développement de l'évaluation formative, le dossier est porteur d'attentes pédagogiques fortes, parfois extrêmes. Après plusieurs années d'utilisation dans les classes primaires et/ou secondaires, le support peut donner lieu, aujourd'hui, à un bilan sur son évolution dans le terrain. Ce bilan se fera sur la base d'entretiens menés auprès de responsables cantonaux ou de prises d'information, effectuées auprès d'enseignants, au cours de la mise en application de ce support et ceci dans trois cantons romands : Neuchâtel, Genève et Vaud. Sur la base de quelques définitions, l'article qui va suivre évoquera le glissement de descriptions théoriques du portfolio vers des descriptions plus concrètes, à caractère institutionnel, destinées aux enseignants. Ce glissement, comme nous le verrons, est porteur des impératifs scolaires en matière de communication aux familles et de justification des décisions prises au moment de la promotion ou de la certification des élèves. Ils viennent se surajouter à la mission formative et réflexive du dossier. Quelques difficultés d'application du dossier évoquées par les enseignants seront ensuite évoquées. Elles relèveront notamment la confusion entraînée par un dossier d'évaluation chargé de multiples missions et attentes institutionnelles. L'article se terminera par une série de constats, suivis de propositions, suggérant d'une part de tenir compte des bilans réalisés en Suisse romande et ailleurs sur l'utilisation du dossier ou du portfolio et, d'autre part, de repenser fondamentalement le rôle attribué à ce support dans les classes de l'école obligatoire .

Introduction

Au milieu des années 90, le *dossier d'évaluation*, appelé aussi en milieu scolaire *dossier de l'élève, dossier d'apprentissage* ou *portfolio*, a été investi par l'enseignement obligatoire de différents cantons de Suisse romande. Utilisé depuis de nombreuses années dans le monde artistique, le dossier est d'abord une manière de présenter un ou des travaux personnels. Transposé dans le monde économique et social, son rôle se précise : il valorise et valide les acquis professionnels nécessaires à l'obtention d'un emploi. En milieu scolaire, le dossier est multifonctionnel (formatif, informatif, certificatif). Il vise en particulier à introduire diverses innovations

pédagogiques favorisant l'amélioration des pratiques d'enseignement et d'évaluation des apprentissages.

Dans un premier temps, le texte présenté ici rappelle quelques intentions théoriques qui ont présidé à l'introduction du dossier à l'école, puis expose les formulations choisies par l'institution scolaire, au moment de leur mise en œuvre dans les classes primaires et/ou secondaires de trois cantons romands : Genève, Neuchâtel et Vaud. Après plusieurs années de mise en application, des informations prises auprès de responsables scolaires ou d'enseignants permettront, dans un second temps, de faire un bref bilan critique et de s'interroger sur les perspectives d'avenir de l'utilisation du dossier dans l'enseignement obligatoire. Les constats permettront, plus largement, quelques réflexions sur le déroulement des réformes scolaires récentes.

La nature du bilan qui découle des observations est fortement positionnée par le point de vue (pris ici comme lieu de discours) de la chercheure. Les missions de l'URSP impliquent, en effet, un questionnement des visées institutionnelles à travers l'analyse de leur mise en œuvre dans le terrain, instaurant ainsi un va-et-vient entre les intentions déclarées par le système et la recevabilité – voire la faisabilité – de celles-ci par les professionnels. Les missions d'aide au pilotage ancrent les recherches dans une perspective évaluative, utile à la régulation des prises de décisions institutionnelles. Elles déterminent ainsi la formulation des constats et des propositions présentés dans la conclusion.

1.Fondements théoriques du dossier

Les écrits théoriques qui concernent le dossier sont nombreux. La littérature évoquant ce support s'inscrit dans des thématiques diverses telles que la psychologie cognitive, la pratique réflexive, l'évaluation formative, la validation des acquis, l'utilisation des technologies de l'information et les outils certificatifs. Il n'est dès lors pas surprenant que les noms donnés à ce support varient autant : *dossier d'évaluation, dossier d'apprentissage, dossier de l'élève, dossier de présentation, dossier de progression, dossier de réussites, portfolio, passeport de compétences, journal de formation, journal de bord, etc.* Les significations et, corollairement, les missions attribuées au dossier, diffèrent d'un écrit à l'autre. Explorer la littérature se rapportant au dossier ou au portfolio devient dès lors une aventure riche et périlleuse (Weiss, 2000).

Les premiers développements du portfolio en milieu scolaire sont américains et ceci dès le début des années 90 (voir notamment Paulson & Paulson, 1990). Les chercheurs européens et québécois se sont plus tardivement intéressés à ce support et de manière inégale, selon les degrés de la scolarité. L'enseignement primaire a été le premier terrain de mise en œuvre (Derycke, 1998), les applications du dossier au secondaire étant plus rares. Si les descriptions du dossier varient, nous pouvons nous arrêter sur une définition américaine du portfolio, fréquemment reprise dans les recherches des Sciences de l'éducation :

Le portfolio est un assemblage finalisé des travaux de l'élève qui démontrent ses efforts, ses progrès et ses acquisitions dans un ou plusieurs domaines. Il implique la participation de l'élève à la sélection des contenus, à la définition des critères de sélection et d'appréciation des travaux, ainsi que des manifestations de réflexion personnelle de la part de l'élève (Paulson, Paulson & Meyer, 1991 ; traduit par Allal et al., 1998).

Dans cette définition, il est question :

- d'assemblage finalisé soit d'un recueil cohérent de travaux visant un but ; le terme *finalisé* s'applique aussi à une situation d'apprentissage complexe et se réfère ainsi à une situation-problème, une démarche de projet, une recherche ou une situation de communication ;
- de visibilisation des efforts et des progrès de l'élève dans les apprentissages ;
- d'acquisitions librement mises en valeur dans un ou plusieurs domaines ;
- de participation active de l'élève, et ceci sous deux aspects : le choix du contenu du portfolio et les différentes caractéristiques des documents ;
- d'évaluation des travaux par l'élève soit de co-évaluation ou d'autoévaluation ;
- de manifestations de réflexions personnelles faisant appel à des activités métacognitives.

Le portfolio vise ici essentiellement une mise en valeur et une « exploitation » des apprentissages au bénéfice de l'élève. Cette définition évoque différentes facettes du dossier qui seront reprises par l'institution scolaire.

2. Définitions du portfolio en lien avec sa mise en œuvre dans les classes

Les définitions théoriques, souvent exprimées de manière complexe (paradoxalement par souci de précision et de clarté), sont généralement reformulées par les systèmes scolaires. Les questions de compréhension et d'appropriation par les praticiens obligent à faire des choix et nécessitent un vocabulaire plus commun et plus accessible. A la suite de leurs travaux sur le portfolio, deux chercheures canadiennes proches des milieux scolaires, vont proposer une définition plus pratique qui sera, par la suite, largement reprise par les écrits sur la mise en application de ce support (Simon & Forgette-Giroux, 1994). Cette définition est la suivante :

Le portfolio est un recueil cumulatif et continu d'indicateurs du cheminement de l'élève (étudiant) dans ses apprentissages, sélectionnés et commentés par l'élève (étudiant) et l'enseignant ou l'enseignant à des fins d'évaluation.

Elle introduit le mot *indicateurs*, un terme emprunté au monde économique, et implique non seulement l'élève mais également l'enseignant dans la sélection des documents et dans les commentaires sur ceux-ci. Le mot *dossier d'évaluation*, qui apparaît dans le titre de l'ouvrage, a été créé par ces deux chercheures. Dans les pays francophones, leur document est considéré comme un texte fondateur en matière de portfolio.

Dès le milieu des années 90, l'intérêt pour le dossier d'évaluation est en plein essor. Les tentatives de mises en œuvre se multiplient et les premiers ajustements ont lieu. En 1998, Forgette-Giroux et Simon proposent une autre définition du portfolio. Il s'agit…

… *d'un recueil cumulatif et systématique d'une variété de données qui témoignent du progrès de l'élève par rapport à la maîtrise d'une compétence jugée à partir d'une échelle descriptive*.

Cette nouvelle définition introduit la notion de progrès ainsi qu'une seconde notion, devenue depuis incontournable dans les réflexions sur le portfolio : celle de compétence et de son évaluation.

La même année, Blouin (1998) propose une définition proche des intentions du programme d'enseignement primaire du Québec. Il définit le portfolio comme étant…

… *un ensemble de travaux significatifs, choisi par l'élève, en collaboration avec son enseignant, illustrant ses progrès et ses différents savoirs dans un ou plusieurs domaines, s'échelonnant dans le temps (au moins une année scolaire), ayant des normes de performance clairement définies, gardant des traces des réflexions de l'élève et de son auto-évaluation, contenant des commentaires de l'enseignant favorisant la communication élève-enseignant, pairs et parents* (Blouin, 1998).

Le programme québécois introduit un aspect du portfolio n'apparaissant pas encore dans les définitions évoquées plus haut : la communication entre les élèves et les enseignants, entre enseignants et entre ceux-ci et les parents. Cette dernière définition inclut toutes les facettes du dossier d'évaluation ou du portfolio utilisé en milieu scolaire. Elle révèle que l'école reformule les définitions théoriques en tentant de les simplifier, mais aussi de les adapter en fonction de ses impératifs et de ses besoins internes. C'est multifonctionnel et en pleine évolution que le dossier d'évaluation va être introduit dans les pratiques de classe de Suisse romande.

3. Introduction du dossier d'évaluation ou du portfolio dans les classes de Suisse romande

La Suisse, comprenant 26 cantons et demi-cantons autonomes en manière de politique scolaire, s'est lancée dans des réformes structurelles et/ou pédagogiques selon des modalités et des calendriers différents. La partie francophone de la Suisse, constituée de 7 cantons ou parties de cantons (Berne, Fribourg, Genève, Jura, Neuchâtel, Valais et Vaud), n'échappe pas à cette diversité. L'introduction du dossier d'évaluation dans les classes illustre bien les différents choix cantonaux en matière de calendrier et de modalités de mise en application. Nous nous limiterons ici à évoquer le cas de trois cantons : Neuchâtel, Genève et Vaud.

3.1 Le portfolio dans le canton de Neuchâtel

Le canton de Neuchâtel a mené des réflexions sur l'évaluation du travail des élèves dès le milieu des années 60 (GRETEL, 2004)[1]. Il faudra toutefois attendre 1984 pour qu'une trentaine de classes expérimente de nouvelles formes d'évaluation au primaire. Une large consultation de tous les milieux intéressés aboutit, en 1990, à un arrêté du Conseil d'Etat instaurant des codes-lettres de la 1re à la 3e année primaire, en remplacement des mentions et des notes. Cette décision sera étendue aux degrés 4 et 5 dès 1997.

Au secondaire, la réforme de l'évaluation est plus tardive. Un rapport, déposé en 2002, instaure la note de synthèse, le code chiffré, la suppression des moyennes et l'amélioration de la communication avec les parents. Le projet prévoit une introduction progressive dans les classes dès 2005.

Dans ce canton, au primaire comme au secondaire, les réflexions sur le portfolio ne sont pas très vives. Les travaux s'orientent davantage vers la forme et le contenu de documents accompagnant le bulletin scolaire : les *journaux hebdomadaires* ou le *carnet de communication* (Service de l'enseignement obligatoire, 2004). L'évaluation formative est favorisée sans référence explicite à l'utilisation d'un support ou d'un recueil d'informations. La communication avec les parents se réalise particulièrement au moment de l'évaluation certificative, au travers d'entretiens, de discussions sur les épreuves de connaissances scolaires et de rencontres de réseaux impliquant les différents intervenants.

La mise en place très progressive des nouvelles modalités d'appréciation du travail des élèves dans le canton de Neuchâtel a évité les discussions passionnées et parfois virulentes qui ont eu lieu dans d'autres cantons. Le récent rapport du *Groupe de pilotage Evaluation* relève toutefois que les attentes des parents sont aujourd'hui plus importantes et qu'il s'agit de donner un retour d'information clair et pertinent sur les apprentissages effectués, les savoirs acquis et les compétences développées (Op. cit. p. 5). Les réflexions actuelles portent sur l'évaluation des élèves en grandes difficultés et l'introduction, pour ceux-ci, d'un recueil d'informations de type portfolio.

3.1.1 Bilan de la mise en application du portfolio dans le canton de Neuchâtel

Comme nous l'avons vu précédemment, le portfolio ou le dossier d'évaluation n'a pas été introduit en tant que tel au primaire comme au secondaire et n'a donc pas donné lieu à un bilan. Par contre, un suivi de la forme donnée aux trois bulletins d'informations, ainsi qu'une analyse de la distribution des codes de fin d'année, a été effectué par le Service de l'enseignement obligatoire (2004). L'un des constats de

[1] La plupart des informations sur les réformes scolaires des cantons de Neuchâtel, Genève et Vaud sont tirées d'un document rédigé par le Groupe de référence pour l'évaluation du travail des élèves (GRETEL), publié par la Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin (CIIP) en 2004. Les indications sur le portfolio ou le dossier d'évaluation proviennent, quant à elles, d'entretiens avec des responsables scolaires ou de prises d'informations plus formelles auprès des enseignants.

l'étude relève la grande diversité des pratiques d'évaluation et ceci tant à l'intérieur d'un même collège que dans la même classe, selon les périodes de l'année.

3.2 Le dossier d'évaluation dans le canton de Genève

L'école primaire genevoise a introduit, dès 1994, un livret scolaire en 1re et 2e année, centré sur la progression des apprentissages. Conjointement à l'introduction de ce livret, les enseignants ont développé des dossiers d'évaluation, destinés prioritairement à informer les parents. En parallèle, dès 1995 (avec la mise en chantier de la rénovation de l'école primaire), quinze écoles planchent sur divers outils de transmission de l'évaluation sans notes aux familles. La rénovation vise notamment à renforcer la communication entre l'école et les parents au moyen de plusieurs rencontres annuelles prenant la forme d'entretiens individuels, de réunions collectives, de permanences ou de *soirées portfolio*. Dans certaines écoles, le dossier prend le nom de *portfolio* car il contient, en plus des travaux, des réflexions de l'élève sur ses apprentissages.

Aucune information sur l'emploi de ce support ne figure dans la loi scolaire ou son règlement d'application. Par contre, des indications apparaissent dans le programme. L'institution scolaire demande aux enseignants de fournir un dossier d'évaluation comme complément au livret. Ce dossier peut prendre différentes formes mais doit, minimalement :

- illustrer la progression des apprentissages de l'élève (en lien avec les objectifs d'apprentissage du programme),
- se présenter sous la forme d'un recueil de divers types de travaux,
- servir de support aux entretiens avec les parents.

L'enseignant ou l'équipe d'enseignants décide de la forme du dossier (recueil de bilans ou portfolio), du nombre de disciplines figurant dans le recueil et du degré d'implication de l'élève dans l'évaluation.

Un article paru dans la *Revue française de pédagogie* (Weiss, 2000) résume le rôle du dossier ou du portfolio en lien avec les fonctions de l'évaluation dans l'enseignement primaire. Nous reprenons ci-après uniquement le contenu ainsi que les destinataires prioritaires du support:

Tableau 1 : Fonctions de l'évaluation et utilisation du dossier d'évaluation ou portfolio dans le canton de Genève

Fonctions de l'évaluation	**Formative**	**Informative**	**Certificative**
Destinataires prioritaires	L'élève et l'enseignant	Les parents et l'élève	L'institution, l'élève et les parents
Outils/Moyens	Un *dossier d'évaluation* de type *portfolio* (impliquant, entre autres, des démarches d'autoévaluation montrant la progression de l'élève dans ses apprentissages)	Le *dossier d'évaluation* sous forme de recueil d'évaluations ou de *portfolio*	Le bilan certificatif inscrit dans le livret scolaire repose notamment sur des éléments tirés du *dossier d'évaluation* dont le but est de mettre en exergue les acquisitions des élèves

Dans ce tableau synthétique, les termes *dossier d'évaluation* et *portfolio* se confondent. Les destinataires ainsi que les missions du support se révèlent multiples.

3.2.1 Bilan de la mise en application du portfolio dans le canton de Genève

La mise en œuvre du dossier d'évaluation ou du portfolio dans les classes n'a pas fait l'objet d'un bilan systématique auprès des enseignants du primaire. Une étude récente (Soussi et al., 2006), sur les pratiques d'évaluation, indique que le dossier d'évaluation est davantage pratiqué au primaire qu'au secondaire, et en particulier dans les établissements de la catégorie socioprofessionnelle la moins favorisée. De nombreux articles ont paru sur la mise en œuvre du dossier ou du portfolio, soit dans des revues pédagogiques ou scientifiques, soit dans des Actes de colloques. Les informations qui vont suivre sont tirées, d'une part, d'entretiens avec des responsables scolaires de l'enseignement primaire et, d'autre part, de quelques articles choisis.

Comme nous l'avons vu précédemment, le portfolio genevois veut être à la fois recueil de documents et démarche d'apprentissage. Après de nombreux ajustements, il vise toutefois aujourd'hui une fonction prioritaire d'accompagnement, d'information aux familles lors de la transmission semestrielle du livret scolaire. Il s'agit d'un classeur conséquent, assez identique pour chaque élève, répertoriant toutes les

disciplines scolaires. Il contient des documents authentiques[2] d'apprentissage, des instruments d'autoévaluation et des tests d'évaluation (Wegmuller, 2005). Mais cette fonction informative du portfolio ne semble pas satisfaire totalement les enseignants. Ils souhaiteraient que cet instrument soit d'abord utile à l'élève, à son implication dans ses apprentissages et à ses évaluations. Ils aimeraient donc prioriser les démarches d'évaluation formative.

Les difficultés liées à la gestion des fonctions informative et formative du portfolio sont souvent évoquées. Malgré le fait que l'institution scolaire genevoise ait laissé une certaine souplesse dans l'application du dossier, la demande de transmission d'informations complémentaires accompagnant le carnet a orienté le dossier vers une fonction quasi essentiellement informative, à destination des parents. La formation des enseignants se heurte aux questions que les professionnels se posent, qui sont le plus souvent celles des liens entre le dossier et le livret scolaire, de l'utilité de ce support en regard des notes, des différences entre portfolio et dossier d'évaluation, du ou des destinataires et du temps scolaire pris *pour faire parler les élèves de leurs apprentissages* (Hohl, 2005). Les enseignants s'interrogent, par ailleurs, sur l'implication des élèves et sur les effets des régulations qui leur sont proposées.

3. 3 Le dossier d'évaluation dans le canton de Vaud

Les travaux qui ont préparé la mise en œuvre de la réforme scolaire EVM (Ecole vaudoise en mutation) ont débuté en 1994. Ils ont permis de coordonner les réflexions de nombreux groupes de travail qui « planchaient » depuis plusieurs années sur diverses thématiques. Le projet, mis en forme dans un exposé des motifs paru en 1996 et accepté en votation populaire en décembre de la même année (suite à un référendum), sera introduit progressivement dans les classes, dès la rentrée scolaire 97-98. Les finalités de la réforme scolaire s'articulent autour de la différenciation pédagogique, de l'évaluation formative, de la lutte contre l'échec scolaire et de la redéfinition des filières du secondaire inférieur. Des appréciations puis des codes chiffrés, dès le second cycle primaire (dès la 4e année), remplacent les anciennes notes (échelle de 10 et demi-points donnant lieu à des moyennes au dixième).

Parmi les innovations introduites dans le canton de Vaud, le dossier d'évaluation prend, dès le départ, une place centrale. Au cœur des réflexions, il figure dans la loi scolaire, ainsi que dans le règlement d'application. Il subira plusieurs changements de missions dans le Cadre général de l'évaluation (2000, 2005), sorte de vade-mecum des enseignants. Ces derniers sont tenus d'établir un dossier d'évaluation pour chaque élève, dès son entrée dans la scolarité obligatoire. Le dossier l'accompagne jusqu'à la fin de sa scolarité et favorise notamment le dialogue entre les maîtres, l'élève et ses

[2] Un document issu d'une évaluation authentique provient, d'une part, d'une situation plus proche des conditions d'utilisation des savoirs et savoir-faire dans la vie extrascolaire et, d'autre part, d'une situation d'enseignement/apprentissage mise en œuvre au fil du degré ou du cycle. L'évaluation authentique ne se limite donc pas à une prise d'informations sous forme de test, fiche, interrogation orale et autres contrôles spécifiquement élaborés à des fins d'évaluation (Mottier Lopez & Allal, 2004).

parents. Le terme *portfolio* n'est pas utilisé. Les informations qui vont suivre sont tirées de ces divers documents officiels.

Tableau 2 : Fonctions de l'évaluation et utilisation du dossier d'évaluation dans le canton de Vaud

Fonctions de l'évaluation	**Formative**	**Informative**	**Certificative**
Destinataires prioritaires	L'élève et l'enseignant	Les parents et l'élève	L'institution, les parents et l'élève
Missions	– Communiquer et suivre l'élève – Faciliter le dialogue et indiquer la progression – Communiquer en vue de la certification		
Contenu	Les travaux de l'élève, les travaux significatifs, les informations sur la capacité à s'adapter à la nouveauté, les bilans périodiques, la décision de mise en niveaux, les épreuves cantonales de référence, le carnet périodique et les documents certificatifs		

Le dossier d'évaluation est surtout un outil de communication. Ses destinataires, ses missions et son contenu sont multiples.

3.3.1 Bilan de la mise en application du dossier d'évaluation dans le canton de Vaud

Un suivi de la réforme scolaire a été réalisé par l'URSP[3]. Les informations qui vont suivre sont tirées de diverses publications (voir notamment Bétrix & Martin, 1999 ; Bétrix, 2000 ; Blanchet, 2001), parues dans la revue *Deux points : « Ouvrez les guillemets*. Elles relèvent, par cycle ou par degré, les points importants évoqués par les enseignants[4] des établissements explorateurs (entre 12 et 15 établissements comprenant plusieurs bâtiments scolaires). Une étude établissant un bilan de la mise en application du dossier d'évaluation dans le canton de Vaud est à paraître (Gilliéron Giroud, 2006). Voici quelques informations recueillies.

Point de vue des enseignants du premier cycle primaire (1re-2e année)

– L'introduction du dossier est reconnue comme plutôt positive ou très positive par plus de la moitié des établissements explorateurs du premier cycle primaire.

[3] Unité de recherche pour le pilotage des systèmes pédagogiques (URSP), Rte du Signal 11, 1014 Lausanne. www0.dfj.vd.ch/ursp

[4] Nom utilisé au masculin dans le texte, par souci de lisibilité ; il est générique et inclut le féminin.

Le dossier est avant tout employé dans une perspective informative et quelquefois formative,

- Le dossier d'évaluation reste le plus souvent en classe ; mis à disposition des parents, il est pourtant rare que ceux-ci viennent le consulter. Ce sont les travaux et les bilans amenés régulièrement à la maison qui constituent le lien entre l'école et les parents,

- Au niveau des difficultés rencontrées, le dossier est perçu comme une charge supplémentaire qui prend beaucoup de temps. Le choix des documents et la rédaction de commentaires représentent un très gros travail,

- Il est difficile de faire la différence entre portfolio formatif et dossier informatif, voire sommatif. Le dossier fait double emploi avec le carnet d'informations et ne joue pas son rôle formatif. La participation active des élèves est difficile ; de nombreux enseignants estiment que les élèves sont trop jeunes pour profiter réellement de la réflexion sur les travaux et les progrès effectués ou à effectuer. Dans les classes qui ont plusieurs enseignants, la cohérence des pratiques n'est pas évidente,

- Certains enseignants estiment que la formation à l'utilisation de ce support fait défaut.

Point de vue des enseignants du deuxième cycle primaire (3e-4e année)

- Moins de la moitié des enseignants des établissements explorateurs du deuxième cycle primaire reconnaissent l'introduction du dossier comme plutôt positive ou très positive. L'aspect certificatif du dossier semble légèrement privilégié par rapport à l'aspect formatif,

- 4 établissements sur 15 estiment que le dossier permet aux élèves de constater leurs progrès et leurs lacunes et donc de les aider dans leurs apprentissages ; un établissement trouve le dossier utile au pilotage de la différenciation,

- Au niveau des difficultés rencontrées, le dossier est trop gourmand en temps. Il est difficile à gérer. Le mélange entre les différentes fonctions de l'évaluation rend son application peu évidente.

Point de vue des enseignants du cycle de transition (5e-6e année)

- Au cycle de transition, le dossier est clairement utilisé (il s'agit d'une demande institutionnelle) comme aide à l'orientation des élèves dans les trois voies subséquentes,

- Certains établissements essaient de promouvoir la participation de l'élève (faire le point sur sa manière d'apprendre ou sur ses apprentissages) sur la base des

bilans d'apprentissage, alors que d'autres les classent sans commentaires, à l'image du carnet périodique,

- La difficulté la plus importante concerne la gestion du dossier : choisir et faire choisir les travaux, penser aux remédiations, écrire des commentaires, avoir des relations avec les autres enseignants, faire signer les travaux, vérifier s'ils ne sont pas perdus, s'ils sont bien classés, etc. Pour certains enseignants, le dossier fait double emploi, c'est une simple mode.

Point de vue des enseignants dans les degrés 7 à 9

- La plupart des enseignants des établissements explorateurs 7-9 conservent les travaux mais pas forcément dans le dossier d'évaluation,

- Dans la moitié des établissements, le choix des travaux est partagé, mais l'enseignant impose toutefois la conservation de certains documents et l'élève complète à sa guise ; le choix est ressenti comme difficile, voire subjectif : *il vaut mieux finalement tout garder,*

- La majorité des établissements ont un avis négatif concernant l'introduction d'un dossier d'évaluation. La gestion de ce support est difficile, voire impossible. La quantité de travail et le temps investis sont importants, pour une faible utilisation ou « pour rien »,

- L'un des problèmes majeurs est la perte des documents par les élèves, notamment pour les élèves peu structurés, et le coût du contrôle qui permettrait de l'éviter.

En conclusion : quelques constats et propositions

Les constats et les propositions découlant des diverses observations exposées ici sont nombreux. En voici quelques-uns.

Premier constat

L'introduction du dossier d'évaluation ou du portfolio dans les classes était porteuse de grands espoirs, alimentés notamment par les réflexions des Sciences de l'éducation. Visant prioritairement l'amélioration des pratiques d'évaluation, le support s'est chargé de missions informative (combler le vide laissé par l'abandon des notes et des moyennes, auprès des familles) et certificative (justification des décisions prises par l'enseignant) nécessaires au fonctionnement du système scolaire. Les définitions choisies par l'école se caractérisent ainsi par une multiplication des missions, des rôles, des contenus et des destinataires du dossier, entraînant une importante confusion chez les enseignants. La mise en application de certaines

innovations scolaires se heurte à la fois aux besoins et aux impératifs extérieurs qui régissent l'école (sociaux, politiques et économiques) et au fonctionnement complexe de la classe, comportant des dimensions particulières, n'existant nulle part ailleurs et qui, de surcroît, sont encore peu connues.

Propositions

Ces réflexions invitent à reconsidérer :

- l'introduction, à l'école, de concepts ou notions encore « en chantier » en pédagogie (mais aussi en psychologie, en sociologie ou en économie),
- la généralisation d'innovations n'ayant pas bénéficié, au préalable, d'une solide exploration, tenant compte des impératifs liés au contexte de la classe.

Des expérimentations, à petite échelle, avec des professionnels volontaires chevronnés, sont néanmoins à valoriser; porteuses de réflexions nouvelles et permettant l'émergence de pratiques pertinentes, elles pourraient constituer la base d'explorations plus larges. La diffusion et le partage de leurs résultats sont alors essentiels.

Deuxième constat

Des innovations comme l'emploi du portfolio, l'abandon des notes et des moyennes, l'évaluation formative ou l'évaluation des compétences, visent à améliorer l'école et contiennent une large part d'idéal, constitutive du principe d'éducabilité régissant toute action de formation. Les difficultés rencontrées, lors de la mise en œuvre d'une innovation « prometteuse », ne découragent pas tous les enseignants de réussir son introduction. Un tiers des enseignants seraient novateurs et réceptifs. Les différentes manières d'aborder les réformes scolaires ont pour effet majeur une croissance de l'hétérogénéité des pratiques enseignantes. Cet écart pourrait, à terme, mettre en danger la cohésion du système d'enseignement. L'introduction de réformes scolaires touchant aux pratiques enseignantes, et donc aux représentations individuelles et à l'identité professionnelle, met l'institution scolaire face à des choix stratégiques nouveaux.

Propositions

La nécessité de faire des choix complexes (dans les contenus innovants aussi bien que dans les modalités d'introduction des réformes dans les classes) suggère la nécessité d'un rapprochement fort entre les différentes instances qui se préoccupent du champ scolaire (universités, centres de recherche pédagogique, Hautes écoles pédagogiques et Départements). Une orientation concertée, à la fois plus cohérente et plus valide sur les changements et les innovations, semble actuellement tout à fait indispensable.

La prégnance de la formation (initiale et continue) des enseignants est forte. Les contenu de la formation ne peuvent porter sur des thématiques nouvelles floues, instables et parfois difficilement applicables.

Troisième constat

Les récentes réformes scolaires montrent toute la difficulté à reconnaître la grande part d'intuition inhérente au métier d'enseignant. Les traces concrètes de la progression des apprentissages des élèves, souhaitées par le portfolio, illustrent bien ce propos. Cette demande, simple en apparence, est en réalité extrêmement difficile à réaliser ; les traces des apprentissages sont, en effet, extrêmement difficiles à comparer. La perception des progrès des apprentissages est l'un des aspects de cette intuition des professionnels, documentée à la fois par des travaux sommatifs et par des observations moins formelles sur des contenus d'apprentissage généralement très divers. Depuis quelques années, le besoin de justifier les décisions scolaires auprès des parents a non seulement orienté les missions du dossier d'évaluation ou du portfolio vers la communication et l'information aux familles, mais a créé un climat peu propice à l'introduction d'innovations faisant place à l'intuition et au jugement professionnel.

Propositions

La nécessité d'attester des progrès réalisés par l'élève dans ses apprentissages peut se visibiliser au travers d'épreuves externes à la classe. Ces dernières sont actuellement en pleine expansion dans l'enseignement obligatoire. Les évaluations externes pourraient avoir notamment comme effet de réduire le besoin de justification des choix ou des décisions prises quotidiennement par les enseignants.

Quatrième constat

Les études sur la mise en application du dossier d'évaluation ou du portfolio, réalisées ces cinq dernières années, relèvent de concert les nombreuses difficultés rencontrées par les enseignants ; elles mettent également en évidence plusieurs incertitudes théoriques. La pratique, toute récente, de l'évaluation des compétences, les inconnues qui entourent les processus d'apprentissage des élèves (rendant souvent peu évident le choix de régulations efficaces), le questionnement émergent et complexe sur les démarches ou les objets médiateurs du discours de l'apprenant, les incertitudes entourant la notion de progression, etc., sont autant de questionnements complexes en cours. Au cœur de ces champs de réflexions, et malgré les nombreuses tentatives de mise en œuvre, le dossier ou le portfolio n'a pas, aujourd'hui, de mission évaluative claire bien établie.

Propositions

Le moment semble propice au répertoire et à la synthèse des différents constats de recherche sur le dossier (que nous percevons convergents) et à une redéfinition de sa mission dans le cadre de la scolarité obligatoire.

Les besoins actuels en matière de communication avec les familles orientent le dossier vers une mission d'information, principalement au moment de la transmission des livrets scolaires. Un portfolio de type « dossier de présentation », proche du support utilisé dans le domaine des arts, laisserait toute latitude à d'éventuelles activités d'autoévaluation. La constitution et l'emploi d'un portfolio utile aux élèves rencontrant des difficultés scolaires est une piste à explorer avec tout le recul actuel sur la mise en œuvre du dossier d'évaluation dans les classes.

Références

Allal, L. & al. (1998). Le portfolio dans la dynamique de l'entretien tripartite. *Mesure et évaluation en éducation*, 20 (3), 5-31.

Bétrix, D. & Martin, D. (1999). Le dossier d'évaluation au premier cycle primaire. *Deux points : « Ouvrez les guillemets,* 7, 21-23.

Bétrix, D. (2000). Le dossier d'évaluation au 2e cycle primaire. *Deux points : « Ouvrez les guillemets*, 15, 6-9.

Blanchet, A. (2001). Le dossier d'évaluation trois ans après. *Deux points : « Ouvrez les guillemets*, 15, 10-14.

Blouin, H. (1998). Repenser des pratiques évaluatives: proposition d'une démarche de renouvellement pédagogique centrée sur l'utilisation du portfolio. Essai présenté à la Faculté des études supérieures de l'Université Laval pour l'obtention du grade de maîtrise ès arts.

Derycke, M. (1998). Suivi pédagogique : grille critériée et dossier d'apprentissage. *Mesures et évaluation en éducation*, 20 (3), 33-53.

Evaluation du travail des élèves dans la scolarité obligatoire vaudoise. *Cadre général*. (2000 et 2005). Lausanne : DFJ-SENEPS puis DGEO.

Forgette-Giroux, R. & Simon, M. (1998). L'application du dossier d'apprentissage à

l'université. Mesure et évaluation en éducation, 20 (3), 85-103.

Gilliéron Giroud, P. (à paraître en 2006). *Le dossier d'évaluation dans la scolarité obligatoire : entre idéal et réalité.* Lausanne : URSP.

GRETEL (2004). Lignes directrices pour un système d'évaluation du travail des élèves dans la scolarité obligatoire. Neuchâtel : CIIP.

Hohl, J-M. (2005). Le portfolio : un outil d'évaluation à la fois au service des apprentissages et support de la communication évaluative aux familles ? *Actes du colloque de l'ADMEE 05*. Reims.

Mottier Lopez, L. & Allal, L. (2004) Le portfolio : pratiques et perspectives. *Revue romande Educateur*, no spécial 04, 46-51.

Paulson, F.L. & Paulson, P.R. (1990). *How do portfolios measure up ?: a cognitive model fort assessing portfolios*. (S.I) : (s.n).

Service de l'enseignement obligatoire - Groupe de pilotage Evaluation (2004). Recom-mandations relatives à la rédaction des bulletins d'informations et des carnets scolaires. Neuchâtel.

Simon, M. & Forgette-Giroux, R. (1994). Vers une utilisation rationnelle du dossier d'apprentissage. *Mesure et évaluation en éducation*, 16 (3 et 4), 27-40.

Soussi, A. & all. (2005). *Pratiques d'évaluation : ce qu'en disent les enseignants (à l'école obligatoire et dans l'enseignement postobliatoire général).* Genève : SRED.

Wegmuller, E. (2005). Le portfolio : entre ses intentions et sa praticabilité. *Actes du colloque de l'ADMEE 06*. Reims.

Weiss, J. (2000). Le portfolio, instrument de légitimation et de formation. *Revue Française de Pédagogie*, no 132, 11-22.

PRATIQUES D'EVALUATION EXTERNE EN SUISSE ROMANDE :

APPORTS ET LIMITES

Ladislas Ntamakiliro, URSP Lausanne
Walther Tessaro, Université de Genève

Mots-clés : pratiques d'évaluations, évaluation externe, suisse romande

Résumé :

Dans les différents cantons de Suisse romande, l'évaluation externe des élèves a connu un développement important ces dernières années. Autrefois limitée à la fin de l'école primaire ou de l'école obligatoire, elle tend à se généraliser à d'autres moments du cursus scolaire, en particulier à la fin des cycles d'apprentissage. Conçue pour compléter l'évaluation interne des élèves par les enseignants, elle est également destinée au pilotage du système d'enseignement.

Une analyse de cette pratique d'évaluation est esquissée en référence aux standards de la nouvelle culture de l'évaluation scolaire. Différentes questions soulevées par les épreuves externes sont abordées. Les différentes fonctions de ces épreuves sont-elles cohérentes ? Les cadres de référence utilisés pour la notation des élèves sont-ils appropriés ? L'exploitation des résultats en vue du pilotage du système d'enseignement est-elle pertinente ?

Introduction

En Suisse romande, les pratiques d'évaluation externe, plus couramment appelées épreuves communes ou épreuves de référence, se sont considérablement développées ces dernières années, en lien avec les réformes des systèmes d'évaluation engagées depuis le début des années 90. Auparavant, elles étaient principalement passées à la fin de l'école primaire (6e)ou de l'école obligatoire (9e). Depuis, elles sont de plus en plus administrées aux différents degrés de l'école obligatoire, surtout à la fin des cycles d'apprentissage. Une autre caractéristique des évaluations externes des élèves émerge : elles sont de plus en plus envisagées comme un moyen de pilotage du système d'enseignement.

Ces pratiques d'évaluation scolaire un peu particulières ont déjà fait l'objet d'écrits portant principalement sur leur fonctions, leur compatibilité avec les évaluations internes des enseignants, leur perception par les enseignants, leur validité et leur fiabilité (Davaud et Cardinet, 1995 ; Lafontaine, Ntamakiliro, Tessaro et Weiss, 2003, Ntamakiliro, 2003 ; Tessaro, 2005). Notre communication au 19e colloque de

l'admee, dans le cadre du symposium « Bilan d'innovation en matière d'évaluation scolaire » s'inscrit dans le prolongement de ces travaux. Il s'agit d'analyser les différentes modalités d'évaluation externe, leurs fonctions, leur évolution récente, leur apport et leurs limites.

Différentes modalités d'évaluation externe

Les termes *évaluation externe* et *épreuve externe* sont utilisés pour désigner toutes les pratiques d'évaluation centralisées à l'échelle nationale ou régionale. Au sens strict, ces termes sont inappropriés, comme le relèvent Monseur et Demeuse (2005). Dans la plupart des cas, en effet, ces épreuves sont administrées et corrigées par les enseignants. Au sens large, cependant, ces termes sont cependant les mieux indiqués pour caractériser toutes sortes de pratiques d'évaluation des élèves dont le seul dénominateur commun est d'être partiellement ou entièrement élaborées à l'extérieur de la classe, à l'opposé des évaluations internes assurées par les enseignants.

En Suisse romande, d'un canton à l'autre voire à l'intérieur d'un même canton, des termes différents sont utilisés : épreuve de référence, épreuve cantonale de référence, épreuve cantonale, épreuve commune, épreuve commune cantonale. Cette situation peut être définie par une tour de Babel terminologique, dans le sens que Van Der Marren (1996) donne à cette expression. Si concepteurs et responsables cantonaux pensent œuvrer à la même tâche, les objectifs et l'utilisation de ces épreuves sont cependant fort différents.

Un classement des pratiques d'évaluation externes devient dès lors la première étape de notre essai d'analyse. Deux critères s'avèrent pertinents à cet effet: la standardisation[1] et l'analyse statistique des résultats.

Tableau 1 : trois modalités d'évaluation externe

1.	Epreuves externes non standardisées
2.	Epreuves externes standardisées
3.	Epreuves externes standardisées avec statistiques des résultats

1 Selon De Landsheere (1979), le critère de standardisation est satisfait si tous les élèves d'une même catégorie sont soumis à une même épreuve et si les conditions d'administration et de correction sont uniformisées.

La construction des épreuves externes -standardisées ou non- est un travail d'équipes d'experts en didactiques de la branche, de spécialistes en évaluation, d'enseignants et d'inspecteurs d'écoles ou de responsables pédagogiques. En principe, ces épreuves sont adaptées aux objectifs pédagogiques définis dans le plan d'études. Elles sont validées suivant les procédures reconnues, notamment par l'analyse des items et le calcul de l'accord interjuges basés sur les données d'un prétest effectué auprès d'un échantillon restreint. Souvent, des tableaux de spécification permettent de clarifier les liens entre les objectifs d'apprentissage et les items de l'épreuve. La validité de contenu des épreuves est toutefois limitée du fait des nombreuses contraintes qui les caractérisent : la durée, l'outil papier-crayon, le moment de l'évaluation, le degré d'autonomie des élèves, etc. Par ailleurs, le niveau de connaissances et de compétences attendu, plus largement les standards de performances, ne sont pas assez précisés dans les plans d'études, ce qui ne facilite pas la construction d'épreuves sommatives valides.

Les épreuves externes standardisées sont passées le même jour, suivant les mêmes protocoles et par tous les élèves d'une même volée. Les critères de correction, plus ou moins précis selon les épreuves, sont généralement fournis aux enseignants chargés de les appliquer strictement. Cette standardisation est toutefois restreinte dans la mesure où ce sont les mêmes enseignants qui font passer les épreuves et qui corrigent les copies de leurs élèves.

Les résultats aux épreuves standardisées sont communiqués par une note chiffrée ou une appréciation qualitative. En théorie, l'échelle utilisée est critériée ou normative. En pratique, les cadres de référence sont souvent confondus. Un subtil dosage de points de vue critériés et normatifs est contraint par les considérations inhérentes à cette évaluation certificative (Chevallard, 1986; Roegiers, 2004). Si le cadre de référence est critériel, les seuils de réussite fixés a priori ne sont souvent adoptés qu'a posteriori, en tenant compte de la distribution des résultats du prétest ou du test proprement dit. Si le cadre de référence est normatif, le barème, particulièrement le seuil inférieur de la note de réussite, peut être ajusté en fonction de la difficulté de l'épreuve.

Dans certains cas, les résultats aux épreuves communes sont récoltés et soumis aux analyses statistiques pertinentes pour le pilotage du système d'enseignement. Les analyses statistiques sont communiquées aux différents acteurs concernés, principalement les enseignants, les directeurs d'établissements et les autorités scolaires (Tessaro, 2005). Le choix des analyses statistiques pertinentes pour le pilotage de l'enseignement est très débattu, d'autant plus lorsque les épreuves sont élaborées dans une logique de l'apprenant (Weiss, 2003). Se pose notamment la question de savoir si les résultats aux épreuves communes sont bien indiqués pour l'évaluation des établissements et des enseignants.

Situation de l'évaluation externe des élèves de l'école obligatoire en Suisse romande

La Suisse est une confédération de cantons ayant chacun son propre système d'enseignement. Les sept cantons entièrement ou partiellement francophones qui nous intéressent ici ont ainsi chacun leur système d'enseignement. Ceci explique en partie pourquoi les pratiques d'évaluation externe diffèrent d'un canton à un autre.

Tableau 2 : modalités d'évaluation externe en suisse romande en 2005-2006

	Berne	Fribourg	Genève	Jura	Neuchâtel	Valais	Vaud
1re							
2e			L1, M, E		**L1**		L1
3e							
4e		L2, M	L1, L2, M		**L1**	L1, M	L1, M
5e	**L1, L2, M**					**S**	
6e	**L1, L2, M**	**L1, M**	L1, L2, M	L1, L2, M	L1, L2, M	L1, L2, M	L1, M
7e			L1,L2,M,Lt,S,A				
8e			L1,L2,M,Lt,S,A			L1, L2, M	
9e			L1,L2,M,Lt,S,A		L1, L2, M	L1, L2, M	

L1 : langue 1 (français)
L2 : langue 2 (allemand)
M: mathématiques
Lt: latin
A : anglais
S : sciences
E : Ecriture-graphisme

Ce tableau reflète un état de la situation pour l'année 2005-2006. Le nombre et les disciplines varient d'une année à l'autre, souvent au sein du même canton. A l'école primaire, ce sont en grande majorité les disciplines Français et Mathématiques qui sont évaluées. L'Allemand est évalué chaque année à Genève dans cet ordre d'enseignement, alors qu'il l'est de façon épisodique dans les autres cantons à l'image d'autres branches comme Sciences, Géographie ou Histoire. Dans le secondaire inférieur, cette discipline apparaît cependant systématiquement lorsque des épreuves externes sont organisées.

Les fonctions des épreuves externes

Pour caractériser les fonctions des différentes modalités d'évaluation externe, nous nous sommes basés sur les informations fournies par l'enquête de l'IRDP (Landry, 2005) concernant les pratiques d'évaluation dans les sept cantons francophones. Dans son enquête sur les pratiques cantonales d'évaluation, l'IRDP s'intéresse notamment aux finalités des épreuves externes. Les services de l'enseignement des différents Départements de l'instruction publique y expriment eux-mêmes les visées et les fonctions des épreuves externes dont ils sont responsables.

Tableau 3: Finalités des évaluations externes, d'après enquête de l'IRDP (Landry, 2005)

Epreuves non standardisées	- Harmonisation des exigences - Aider l'enseignant à ajuster son enseignement - Aider l'enseignant à évaluer et noter le travail des élèves - Préciser les attentes de l'institution - Offrir des balises externes à la classe - Donner des outils d'évaluation aux enseignants - Evaluation formative
Epreuves standardisées	- Orientation - Sélection - Certification
Epreuves standardisées et analyses statistiques	- Complément à l'évaluation certificative - Régulation institutionnelle de l'enseignement - Harmoniser les exigences de l'enseignement dans le canton - Information sur le degré de maîtrise des objectifs du programme - Aide à l'orientation - Certification

La plupart des épreuves externes, standardisées ou non, sont conçues avant tout pour gérer le cursus de l'élève. Elles sont élaborées dans une perspective formative, certificative ou pronostique. Cependant, les épreuves de ce type ont des retombées qui dépassent la fonction première annoncée. Elles contribuent aussi à la construction collective des attentes du système dans chaque discipline d'enseignement et exercent ainsi un effet global de régulation sur les pratiques didactiques et pédagogiques et sur les relations famille-école. Les formes choisies pour analyser les résultats peuvent renforcer la qualité de l'information fournie aux différents acteurs et favoriser des actions de régulation. L'exploitation de ces résultats peut se faire potentiellement au niveau de *la classe*, de *l'établissement scolaire*, *des instances de formation* et *du système scolaire* (Tessaro, 2005).

Au niveau de la classe, les informations véhiculées par l'épreuve permettent une adaptation de l'enseignement pour la volée suivante, dans une régulation formative de deuxième ordre. Les choix effectués par les concepteurs d'épreuves sont des indicateurs très importants pour l'enseignant. Si ces choix sont explicités dans les

documents d'accompagnement, l'effet produit sur les pratiques de classe est inévitable.

Sur le plan d'une équipe pédagogique, les résultats d'une épreuve externe sont des informations importantes qui permettent de réfléchir au suivi collégial des élèves. La réflexion commune évite d'attribuer les résultats aux seuls titulaires du degré concerné et responsabilise chaque enseignant. Pour les autorités scolaires, l'analyse des résultats contribue à mieux identifier les forces et les faiblesses du système et à prendre les mesures d'ajustement qui s'imposent.

Les épreuves standardisées constituent un enjeu important pour les élèves (Bishop, 1995). Les résultats ont le plus souvent de réelles conséquences sur leur cursus. Ils sont parfois inscrits dans le livret de l'élève et contribuent d'une façon ou d'une autre aux décisions de promotion à la fin d'un cycle ou d'un degré.

Evolution récente des pratiques dans les cantons de Genève et de Vaud

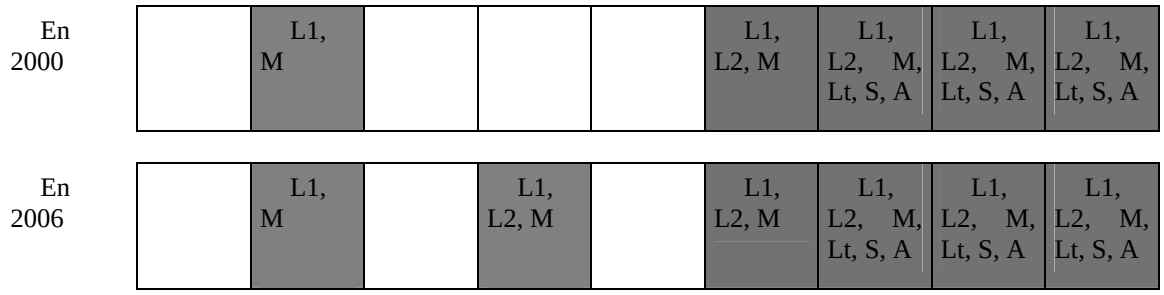
L'évaluation externe prend une place de plus en plus importante dans le systèmes d'évaluation comme outil d'évaluation formative, certificative et pronostique des élèves et comme moyen de régulation du système d'enseignement. L'évolution récente des pratiques d'évaluation externe dans cantons de Genève et de Vaud illustre ce phénomène.

Dans le canton de Genève, si l'évaluation externe en 6e, 7e, 8e et 9e s'inscrit dans une longue tradition, la mise en place d'épreuves en fin de 2e et en fin de 4e est, quant à elle, beaucoup plus récente. L'introduction de ces épreuves est intimement liée à la réforme du système d'enseignement et d'évaluation mise en œuvre depuis 1994. Cette réforme a consisté notamment à passer d'une organisation de l'enseignement primaire en degrés à une organisation en deux cycles de quatre ans : le cycle élémentaire (de – 2 à 2) et le cycle moyen (de 3 à 6). En 2000, les épreuves externes ont été mises en place aussi en fin de 2e, pour soutenir l'évaluation certificative prévue à la fin du cycle élémentaire.

En 2006, la rénovation du système d'enseignement primaire a fait l'objet d'ajustements. Le cycle moyen de quatre ans, jugé trop long, a été scindé en deux cycles moyens, I et II. En conséquence, de nouvelles épreuves externes ont été introduites en fin de 4e dans le but, là aussi, de consolider l'évaluation certificative.

Tableau 4 : Evolution récente des pratiques d'évaluation externe dans le canton de Genève

	1re	2e	3e	4e	5e	6e	7e	8e	9e
Jusqu'en 1999						L1, L2, M	L1, L2, M, Lt, S, A	L1, L2, M, Lt, S, A	L1, L2, M, Lt, S, A



Dans le canton de Vaud, la pratique d'épreuves externes non standardisées en 6e et parfois à d'autres degrés était de tradition jusqu'en 2001. En 2002, en plein débat politique sur la réforme scolaire mise en place à partir de 1997, les épreuves externes standardisées avec analyse statistique des résultats ont été introduites en 5e et en 6e principalement pour assurer plus de validité et d'équité au processus d'orientation des élèves à l'entrée en 7e. En 2005, le système d'évaluation scolaire est entièrement revu. Le nouveau cadre général de l'évaluation prévoit des épreuves externes standardisées à différents degrés de l'école obligatoire. En 2005, celles-ci sont assurées à la fin de chaque cycle de deux ans.

Tableau 5 : Evolution récente des pratiques d'évaluation externe dans le canton de Vaud

	1re	2e	3e	4e	5e	6e	7e	8e	9e
En 2001						**L1, M**			
En 2003					L1, M	L1, M			
En 2006		L1		L1, M		L1, M			

Les épreuves externes comme régulateur de l'enseignement

Grâce aux épreuves externes et à l'analyse statistique des résultats des élèves, les responsables scolaires souhaitent des indices relativement fiables du niveau des élèves et de ses variations. De cette façon, l'évaluation des enseignants et des établissements, ne serait-ce qu'à titre informel, est envisageable.

L'impact des épreuves externes sur l'efficacité et l'équité du système d'enseignement est relativement bien documenté (Bishop, 1995 ; Crahay, 1996 ; Monseur et Demeuse, 2005). Analysant les pratiques d'évaluation externe aux Etats-Unis, McDonnell (1994) attribue cet impact à la force de persuasion des enseignants et de régulation de leur travail que ces épreuves comportent, ainsi qu'aux standards de contenu et de performance qu'elles véhiculent. Darling-Hammond (1994) est cependant sceptique et se demande si une régulation top-down de la qualité de l'enseignement est possible et souhaitable, tout en mettant en garde contre deux écueils :

- le risque de s'illusionner si l'on n'est pas critique vis-à-vis des résultats des élèves aux épreuves externes ;
- le risque de contrarier et démotiver les enseignants et responsables locaux qui se préoccupent d'atteindre le meilleur niveau possible dans le contexte de leur établissement.

Dans le canton de Vaud, les réactions contre l'utilisation d'épreuves externes standardisées aux fins du pilotage top-down de l'enseignement n'ont pas manqué, au moment de mettre en place les moyens d'assurer une évaluation externe standardisée en 2002. Elles venaient d'enseignants, de directeurs d'établissements parfois même de ceux qui étaient chargés d'élaborer ces épreuves. S'il fallait absolument développer l'évaluation externe des élèves, il suffisait de multiplier les épreuves non standardisés dans toutes les disciplines et d'en améliorer la qualité. Dans le canton de Genève, également, l'importance croissante des épreuves externes dans le système d'évaluation à l'école primaire fait parfois réagir certains enseignants qui craignent une possible utilisation des résultats afin d'évaluer les enseignants et les établissements.

Références

Bishop, (1995) The impact of curriculum-based external examination on school priorities and student learning. *International Journal of Educational Research*, 23 (8), 653-752.

Crahay, M. (1996). *Peut-on lutter contre l'échec scolaire ?* Bruxelles : De Boeck.

Davaud, C. & Cardinet, J. (1995). La problématique des épreuves communes. In D, Lavault (Ed.), *Les pratiques d'évaluation en éducation.* Montréal : Editions de l'ADMEE.

Darling-Hammond, L. (1994). National standards and assessments : will they improve education ? *American Journal of Education*, 102, 478-508.

De Landsheere, G. (1979). *Dictionnaire de l'évaluation et de la recherche en éducation*. Paris : Presses Universitaires de France.

Lafontaine, D., Ntamakiliro, L., Tessaro, W. & Weiss, J. (2003). Epreuves externes : évaluation de la qualité des apprentissages ou des curricula. *Actes du 15è colloque international de l'ADMEE-Europe et congrès annuel de la SSRE*, Université de Lausanne, septembre 2002.

Landry, F. (2005). *Pratiques cantonales concernant l'organisation d'épreuves, d'examens, de tests ainsi que l'obtention de certificats ou diplômes. Année scolaire 2005-2006*. Neuchâtel : IRDP.

McDonnell, L. M. (1994). Assessment policy as persuasion and regulation. *American Journal of Education*, 102, 394-420.

Monseur, C. & Demeuse, M. (2005). Les évaluations externes permettent-elles une régulation efficace ? In M. Demeuse et C. Monseur (Ed.), *Vers une école juste et efficace*. Bruxelles : De Boeck.

Ntamakiliro, L. (2003). La place des épreuves communes standardisées dans le système d'évaluation des élèves. *Deux points : « ouvrez les guillemets »*, 19, 29-31.

Tessaro (2005). Epreuves externes: quelle utilité pour les enseignants ? *Résonances, 7*, 10-11.

Van Der Maren, J.- M. (1996). *Méthodes de recherche pour l'éducation*. Bruxelles: De Boeck.

LES ÉVALUATIONS NATIONALES EN FRANCE : OUTILS PÉDAGOGIQUES ET OUTILS DE PILOTAGE, INTERÊTS, LIMITES ET NOUVELLES PERSPECTIVES.

Bruno Suchaut, Université de Bourgogne et Irédu C.N.R.S., Dijon, France

Résumé :

Depuis 1989, le système éducatif français s'est doté d'outils d'évaluation des élèves à différents niveaux de la scolarité primaire et secondaire[1]. Ces évaluations de masse, distribuées et utilisées à l'échelon national ont, d'une part contribué de manière progressive à la diffusion de la culture de l'évaluation à l'ensemble de la communauté éducative et, d'autre part, permis aux enseignants d'approfondir leur réflexion pédagogique (Joutard, Thélot, 1999 ; M.E.N., 2000). L'objectif initial de ces évaluations nationales était de «*permettre l'observation des compétences et apprécier les réussites et les difficultés éventuelles de chaque élève considéré individuellement, à un moment précis de la scolarité. Elles fournissent aux enseignants des repères exploitables quant aux types d'erreurs fréquemment produites par les élèves au cours de leur apprentissage…* »[2]. Même si la dimension diagnostique reste aujourd'hui prioritaire, ces évaluations peuvent être utilisées comme des instruments de pilotage car elles rendent objectivement compte du niveau d'acquisition des élèves à plusieurs niveaux : national, régional, départemental, établissement scolaire.

Pour le chercheur en éducation, les évaluations présentent également un intérêt dans le sens où celles-ci balaient un vaste ensemble de compétences et de connaissances alors que les tests couramment mobilisés dans les recherches françaises sont beaucoup plus restrictifs sur ce plan. Ce texte expose une analyse originale des épreuves des évaluations nationales de CE2 de français et de mathématiques sur la base d'une recherche récente qui étudie les résultats des évaluations de CE2 (de septembre 1999) et de sixième (de septembre 2002) d'une même cohorte d'élèves (Morlaix, Suchaut, 2006). Nous essaierons de montrer en quoi les outils d'évaluation livrent des informations capitales sur les apprentissages réalisés par les élèves, à

[1] Les évaluations nationales mis en place par la D.E.P. (Direction de l'Evaluation et de la Prospective) ont concerné à l'origine trois niveaux de la scolarité : CE2 (3ème année de l'enseignement élémentaire), sixième (1ère année de l'enseignement secondaire) et seconde (5ème année de l'enseignement secondaire).

[2] Note de service 2002-105 du 30.04.2002, B.O. N°19 du 9 mai 2002.

condition de ne pas se limiter pas à la seule approche institutionnelle, notamment à la classification des compétences établies par les concepteurs des épreuves.

Introduction

Nous rappellerons dans un premier temps les précautions à prendre dans l'utilisation des résultats des évaluations nationales, puis, dans un second temps, nous exposerons une méthodologie originale qui permet de mieux appréhender la structure des acquisitions des élèves. Nous énoncerons pour conclure les perspectives nouvelles qu'offre cette nouvelle approche dans différents domaines.

L'utilisation des résultats des évaluations nationales : précautions et limites

Les exercices proposés dans les épreuves d'évaluation reflètent les programmes scolaires officiels en vigueur et constituent à ce titre une référence commune pour l'ensemble des enseignants et des élèves[3]. Sur la base des simple taux de réussite aux items, il est possible d'avoir une vision du niveau d'acquisition des élèves à un moment donné de leur scolarité et d'identifier les lacunes de chacun en vue d'y remédier au cours de l'année scolaire. Outre cette perspective diagnostique et pédagogique, les résultats des évaluations fournissent des indicateurs objectifs de la réussite des élèves, au sein de la classe, de l'école, mais aussi à des niveaux supérieurs. Les comparaisons spatiales et temporelles peuvent ainsi servir d'éléments utiles au pilotage d'une circonscription primaire ou d'un département. Le Ministère insiste toutefois sur l'utilisation inadaptée qui pourrait être faite des résultats des évaluations, ceux-ci ne pouvant, en aucun cas, être considérés comme des normes à atteindre, mais seulement comme un diagnostic, en début de cycle[4], des réussites et des difficultés des élèves. Il est également clairement énoncé dans le discours officiel que les épreuves d'évaluation ne rendent que partiellement compte des compétences et des connaissances des élèves puisque la passation est collective et limitée à des exercices écrits. La prudence est aussi conseillée quant à l'interprétation de

[3] Différents personnels de l'Education nationale participent à la conception et à la sélection des exercices : enseignants, inspecteurs de l'Education nationale, Inspecteurs généraux…

[4] L'enseignement primaire français est organisé en 3 cycles. Le cycle I (cycle des apprentissages premiers) concerne l'école maternelle ; le cycle II (cycle des apprentissages fondamentaux) regroupe la grande section de maternelle, le cours préparatoire et le CE1 ; le cycle III (cycle des approfondissements) intègre le CE2, le CM1 et le CM2.

l'évolution au cours du temps des taux de réussite, les épreuves n'étant pas la plupart du temps identiques d'une année sur l'autre.

L'institution éducative annonce donc nettement les limites à une utilisation des évaluations nationales à des fins de pilotage par les résultats. Il est vrai que les taux de réussite en eux-mêmes n'ont pas de signification bien précise et il pourrait en effet être hasardeux de les interpréter de manière directe. A titre d'illustration, le graphique suivant présente les pourcentages de réussite en français à l'entrée en CE2 entre 1996 et 2005. On voit très nettement que les résultats varient sensiblement au cours de ces dix années, la moyenne des scores globaux oscille en effet entre 60,5% (en 2001) et 73,5% (en 2004) avec un chiffre moyen de 67,3%. Des écarts sensibles apparaissent aussi pour un même domaine ; par exemple, les résultats en production d'écrits chutent fortement entre 1996 et 1997, alors que pour la même période les scores dans la dimension des outils de la langue connaissent une augmentation[5]. Les différences constatées d'une année sur l'autre renvoient principalement à la composition des épreuves, les chiffres étant proches quant les exercices sont identiques et des écarts marqués apparaissent quand les épreuves sont totalement renouvelées. Le poids variable accordé à chaque domaine d'acquisition et le degré de difficulté des exercices composant ces domaines expliquent aussi pourquoi, selon les années, c'est telle ou telle dimension des acquis qui semble la moins (ou la mieux) maîtrisée. Dans tous les cas, les allures des courbes du graphique témoignent du caractère fluctuant des taux de réussite et les variations temporelles constatées ne renvoient pas nécessairement à une baisse ou une hausse du niveau des élèves.

[5] Pour l'année 2005, le domaine habituellement nommé « outils de la langue » est scindé en deux champs « reconnaissance des mots » et « écriture et orthographe ». Nous avons reporté sur le graphique les scores correspondant à l'ensemble des items des deux champs.

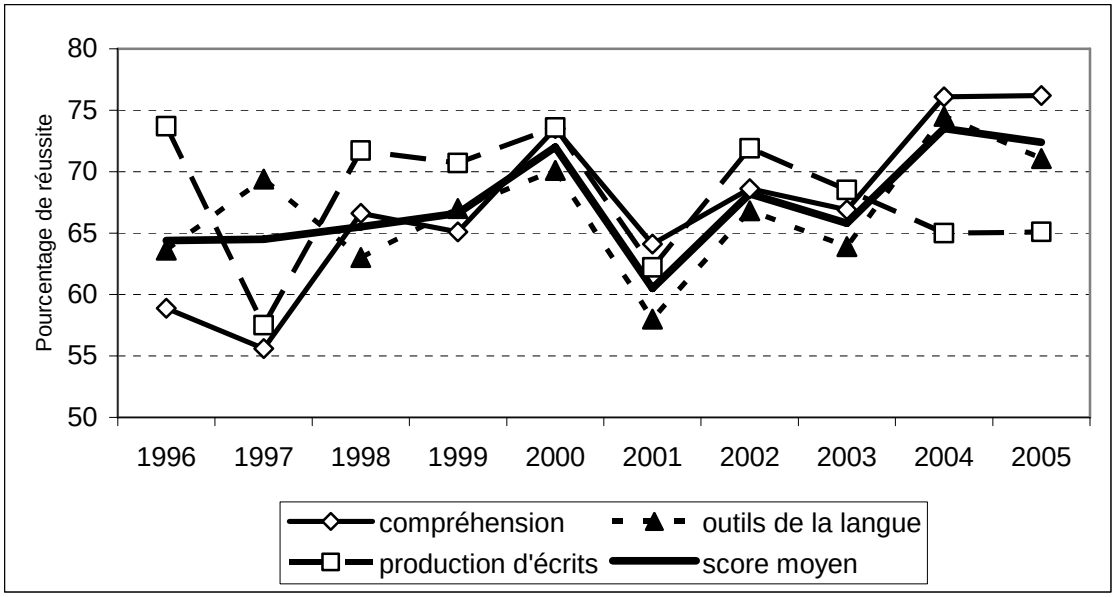


Figure1 : Evolution temporelle des résultats aux évaluations nationales de français au CE2

On peut également s'interroger sur l'échelle de mesure utilisée qui varie dans l'absolu entre 0% et 100% de réussite ; cette échelle ne fait pas référence à des critères de mesure précis et ne peut être considérée comme une mesure étalonnée du niveau des élèves. Les concepteurs des épreuves s'attendent-ils à ce qu'un élève qui aurait assimilé la moitié du programme du cycle II obtienne 50% de réussite au test ? La réponse est incertaine dans la mesure où un nombre infime d'élèves réussit la totalité des exercices présents dans les cahiers d'évaluation, ce qui ne veut évidemment pas dire qu'aucun élève français ne maîtrise correctement totalement le programme du cycle II. Ces limites ne signifient pas que les résultats des évaluations nationales ne peuvent pas être utilisés à des fins de comparaisons spatiales et temporelles, mais ces comparaisons impliquent l'usage d'une mesure relative et non absolue du niveau des élèves ; ceci est d'ailleurs aisé à mettre en œuvre par les standardisation des scores dans une échelle arbitraire.

Au-delà de ce premier problème lié à l'interprétation des taux de réussite, on peut s'interroger plus en profondeur sur la mesure des acquisitions des élèves. Les évaluations nationales sont composées d'items regroupés dans des exercices, chaque exercice étant censé évaluer la maîtrise d'une seule compétence. Les exercices

appartiennent à des champs (ou domaines) différents, eux-mêmes appartenant à une discipline (mathématiques ou français). A titre d'exemple, les épreuves CE2 de 1999 comportaient 91 items en français et 80 items en mathématiques ; le tableau 1 fournit la répartition de ce nombre d'items ainsi que les compétences qui y sont associées dans les différents domaines.

Tableau 1 : Nombre d'items et de compétences dans les différents champs
Evaluations de CE2 de 1999

		Nombre de compétences	Nombre d'items
Français	Compréhension	6	41
	Outils de la langue	7	40
	Production d'écrits	2	10
Mathématiques	Travaux géométriques	8	17
	Mesure	8	22
	Travaux numériques	6	32
	Résolution de problèmes	5	9

On observe un double déséquilibre dans cette répartition. En premier lieu, on note que les différents champs de l'évaluation ne sont pas représentés avec la même intensité dans les épreuves. Une quarantaine d'items évaluent la compréhension et les outils de la langue alors que seulement 9 items mesurent la résolution de problèmes et 10 items ont trait à la production d'écrits. Cette disproportion pourrait se justifier par la place plus ou moins importante accordée dans les programmes scolaires à ces différents domaines d'acquisition, mais aucun document émanent du Ministère indique que les épreuves ont été élaborées en choisissant le nombre d'items selon une technique de stratification des contenus d'enseignement (Laveault, Grégoire, 1997).

En second lieu, on relève un fort déséquilibre entre le nombre d'items et le nombre de compétences. Ainsi, les 6 compétences de compréhension sont mesurées par 41 items (soit en moyenne un peu moins de 7 items pour une compétence) alors que les 5 compétences en résolution de problèmes sont évaluées par seulement 9 items (soit en moyenne moins de 2 items pour une compétence). L'échelle utilisée pour apprécier la réussite des élèves est donc très variable d'une compétence à l'autre selon le nombre d'items retenus. Pour prendre des exemples extrêmes se rapportant toujours à l'évaluation CE2 de 1999, la compétence «*comprendre un texte…*» est évaluée dans

12 items alors que la compétence «*construire un figure simple sur un quadrillage…*» ne concerne qu'un seul item. Selon le cas, la graduation de la réussite est très fluctuante, allant de la simple dichotomie « échec / réussite » (score de 0 ou 1) à une échelle en 12 scores. Certaines compétences sont donc mesurées dans des situations plus nombreuses et plus variées que d'autres et on peut douter de la précision de la mesure pour les compétences qui sont évaluées à travers un très petit nombre d'items. L'institution s'interroge elle aussi sur la pertinence des protocoles ne faisant intervenir que très peu d'items pour mesurer une compétence (M.E.N., 2000). Ces remarques nous interrogent sur la validité des épreuves au sens psychométrique du terme et le constat de la maîtrise d'un bon nombre de compétences sera davantage le fait d'une inférence à partir de situations d'évaluation forcément limitées. Cela pose d'autant plus problème s'il s'agit de compétences qui se manifestent principalement dans des situations complexes (Gérard, 2005).

La question du seuil de réussite des compétences se pose également. Il serait possible de fixer un seuil de réussite commun à toutes les compétences en adoptant les principes en usage (75% est généralement le seuil retenu), mais les dissymétries entre les différentes échelles rendent cette solution trop imparfaite, notamment pour les compétences qui comportent très peu d'items. Une solution consisterait alors à avoir recours à des juges ou à des experts (des professionnels de terrain compétents dans le domaine) qui définiraient eux-mêmes et individuellement les critères d'acceptation de la performance. Plusieurs techniques statistiques peuvent ensuite être mobilisées pour exploiter les différents jugements et définir un score de réussite correspond à chaque compétence (Laveault, Grégoire, 1997). Cette méthode, même si elle présente des intérêts évidents, sans doute du fait de sa lourdeur de mise en œuvre, n'a pas été retenue par les concepteurs des évaluations nationales. Cette question du seuil de réussite reste donc entière quand on souhaite savoir si une compétence est maîtrisée ou non par un élève.

Indépendamment des points relatifs à la mesure, les évaluations nationales peuvent donner lieu à des interrogations en ce qui concerne la définition même des compétences évaluées. Ces compétences sont définies par les commissions chargées de l'élaboration des épreuves et leur nombre peut varier d'une année sur l'autre. Les évaluations CE2 auxquelles nous nous référons dans cette recherche comportent 15 compétences en français et 27 en mathématiques (tableau 2). Les résultats des évaluations nationales sont exploités généralement à deux niveaux dans deux perspectives différentes. Le premier niveau, utilisé dans une perspective de pilotage, concerne une analyse globale des scores totaux, par discipline ou par champ. Le second niveau, utilisé dans une optique pédagogique, se centre sur la réussite ou

l'échec aux compétences. Ces exploitations classiques des évaluations nationales fournissent une image incomplète des apprentissages effectivement réalisés par les élèves, soit en se situant à un niveau trop global qui ne permet pas d'identifier quelles dimensions des apprentissages sont les plus centraux, soit en dissociant les compétences les unes des autres sans examiner les liens qu'elles entretiennent entre elles. Quand on adopte au contraire une démarche basée sur la recherche des liens entre les multiples facettes des acquisitions des élèves, on soulève un certain nombre d'interrogations sur la pertinence de la définition institutionnelle des compétences et de leur rattachement à des domaines d'acquisition[6].

[6] Les analyses qui vont suivre sont basées sur l'étude d'un échantillon de 671 élèves pour lesquels nous disposons des résultats détaillés aux évaluations de CE2 (1999) ainsi que d'informations socio-démographiques et scolaires.

Tableau 2 : Liste des compétences des évaluations CE2 de 1999

Compétences de français
Savoir utiliser une bibliothèque ou une BCD, repérer et identifier des ouvrages appartenant à des domaines textuels différents.
Distinguer des textes différents (récits, documents,…) en utilisant des indices extérieurs au texte.
Comprendre et savoir appliquer les consignes courantes du travail scolaire.
Comprendre un texte et montrer qu'on l'a compris.
Reconstituer la chronologie des évènements dans des textes de statuts variés
Se représenter les relations spatiales et les lieux évoqués dans des textes de statuts variés.
Savoir produire un écrit bref appartenant à un type de texte défini et dans une situation de communication déterminée en s'appuyant sur une consigne, une sollicitation, des documents de référence.
Savoir produire, de manière plus autonome, un texte cohérent dans le cadre d'une situation de communication déterminée.
Repérer les usages typographiques courants et se situer dans l'espace graphique d'une page imprimée.
Reconnaître les mots écrits.
Mieux comprendre les mots d'un texte.
Copier un mot, une phrase, un texte, en respectant les exigences de présentation et en écrivant lisiblement.
Ecrire sous la dictée des mots courants, de petites phrases ou de petits textes.
Transformer un texte en appliquant des règles simples.
Identifier certains aspects d'un texte.
Compétences de mathématiques
Se repérer et se déplacer dans un quadrillage.
Utiliser les instruments de dessin pour achever un tracé.
Construire une figure simple sur un quadrillage en utilisant des propriétés de la figure.
Compléter par pliage (symétrie) une figure dessinée sur un quadrillage.
Associer une figure à une description.
Compléter un plan à partir de consignes.
Se repérer dans l'espace.
Tracer une figure à partir de consignes.
Se repérer dans la journée
Mesurer ou tracer un segment de longueur donnée.
Ranger des longueurs.
Associer une unité usuelle à une grandeur
Utiliser le calendrier.
Comparer des distances.
Résoudre un problème faisant intervenir une grandeur.
Choisir l'unité la mieux adaptée à un mesurage.
Effectuer des additions, posées, en ligne ou à poser.
Calculer des produits et des différences (calcul exact ou approché)
Calculer mentalement (calcul exact ou approché)
Transcrire en lettres des nombres écrits en chiffres et inversement.
Ranger des nombres
Comparer des nombres donnés sous formes diverses.
Lire et/ou remplir un tableau à double entrée.
Exploiter un document « brut »
Résoudre un problème à une opération.
Résoudre une situation de partage ou de groupement.
Effectuer un choix et en formuler la justification.

Une première remarque concerne la distinction entre les deux disciplines. Si en français, toutes les compétences sont effectivement reliées positivement les unes aux autres (les coefficients de corrélation sont tous significatifs au seuil de 1%), on relève une plus bien grande hétérogénéité des situations en mathématiques puisque certaines compétences n'entretiennent aucune liaison statistique entre elles. Par exemple, les compétences «*associer une unité usuelle à une grandeur*» et « *lire et/ou remplir un tableau à double entrée* » ne sont pas corrélées (r = - 0,01, non significatif). On note également que les liens statistiques entre certaines compétences issues de disciplines différentes peuvent être plus intenses que ceux existants entre des compétences originaires d'une même discipline. Par exemple, *« calculer mentalement »* (mathématiques) et « é*crire sous la dictée des mots courants, de petites phrases ou de petits texte »* affichent un coefficient de corrélation élevé (r = +0,41, significatif à .01). En outre, les corrélations les plus fortes ne concernent pas obligatoirement des compétences appartenant à un même champ. Ainsi, en français les compétences : *« comprendre un texte et montrer qu'on l'a compris»* et «*identifier certains aspects d'un texte* » (corrélation de +0,45) appartiennent respectivement aux champs de la compréhension et des outils de la langue. En mathématiques les compétences «*calculer mentalement* » et « *résoudre un problème à une opération* » (corrélation de +0,50) se rattachent respectivement aux travaux numériques et aux problèmes numériques. La distinction entre les différents domaines d'apprentissage ne va donc pas de soi et laisse la place à un certain arbitraire quand on observe les relations statistiques entre les scores obtenus par les élèves aux différentes compétences.

Une seconde remarque, plus fondamentale, s'interroge sur la pertinence de l'identification de certaines compétences[7]. On observe globalement que les items appartenant à un même exercice, donc dans la grande majorité des items étant censés mesurer la même compétence, ne sont pas plus corrélés entre eux que les items appartenant à des exercices différents. Certains exercices présentent même des corrélations nulles entre les items les composant, c'est le cas pour l'exercice 7 de français qui évalue la compétence « *comprendre et savoir appliquer les consignes courantes du travail scolaire* » à l'aide de 5 items (tableau 3).

[7] Les analyses portent sur les corrélations entre l'ensemble des items présents dans les épreuves.

Tableau 3 : Corrélations entre les items de l'exercice 7 de français
Evaluations de CE2 de 1999

	Item 28	Item 29	Item 30	Item 31
Item 29	+0,02 n.s.			
Item 30	+0,16***	+0,00 n.s.		
Item 31	-0,01 n.s.	+0,04 n.s.	+0,00 n.s.	
Item 32	+0,17***	+0,09**	+0,14***	+0,03 n.s.

n.s. : non significatif, ** : significatif au seuil de 5%, *** : significatif au seuil de 1%

Il n'existe ainsi aucune relation statistique entre les items 28 et 29, 28 et 31, 29 et 30, 29 et 31, 30 et 31, 31 et 32 alors que les items 30 et 32, 30 et 28, 28 et 32, 29 et 32 sont positivement corrélés. Cette situation peut s'expliquer quand on se reporte au contenu des différents items proposés dans cet exercice. Si ces tâches mesurent bien toutes une aptitude ayant trait à un comportement scolaire, en l'occurrence «*savoir appliquer des consignes*», elles sont en revanche d'une nature différente d'un item à l'autre. Les items 28, 30 et 32 mobilisent une activité de discrimination visuelle semblable (identifier une ou deux lettres dans un mot simple) qui explique la corrélation relevée entre ces trois items. La réussite de l'item 29 nécessite que l'élève sache se repérer dans l'espace (il s'agit de mettre une croix entre un carré et au-dessus d'une ligne) ; cette activité est donc assez éloignée de la précédente ce qui peut expliquer l'absence de corrélation entre l'item 29 et les items 28 et 30.

De façon complémentaire, les analyses suggèrent que certains items proposés dans les épreuves pourraient s'associer différemment pour mesurer la maîtrise des compétences. Par exemple, la compétence visée par l'exercice 24 de mathématiques «*exploiter un document brut* » et mesurée par les items (73 et 74) pourraient aussi bien être intégrée dans la mesure d'une autre compétence, c'est en tout cas ce que nous préconise notre approche empirique. Ainsi, l'analyse des corrélations entre items indique une corrélation positive et significative entre l'item 73 et 35 (r = + 0,10, significatif au seuil de 1%), ce dernier item mesurant à lui seul la compétence *«résoudre un problème faisant intervenir une grandeur* », compétence intuitivement proche des habilités requises pour répondre à la question de l'item 73 qui implique une comparaison de grandeurs numériques.

Nous mettons ici l'accent sur la difficulté à saisir concrètement la notion de compétence dans le domaine de l'éducation et surtout à la définir de façon univoque.

Il est toutefois possible aujourd'hui de retenir une définition relativement consensuelle de la compétence, celle-ci pouvant être considérée comme une aptitude à la mise en œuvre d'un ensemble organisé de savoirs, de savoirs faire d'attitudes permettant d'accomplir un certain nombre de tâches (Crahay, 2006). Une compétence intègre une mise en relation de connaissances, qu'elle s'applique à une famille de situations et qu'elle est orientée vers une finalité (Gillet, 1991). Plus généralement, la notion de compétence renvoie à un ensemble intégré de connaissances susceptibles d'être mobilisées pour accomplir des tâches (Crahay, 2006). En ce qui nous concerne, nous nous situons à la marge du débat théorique relatif à cette question en nous contenant de faire émerger les diverses compétences mise en œuvre par les élèves à partir de situations d'évaluation. Le terme de compétence intègre ici dans sa définition les différentes capacités, aptitudes, connaissances et attitudes qui mobilisent chez les individus des ressources cognitives pour réaliser une tâche.

Le choix de situations d'évaluation des compétences relève d'un certain arbitraire auquel les épreuves nationales n'échappent pas. Les remarques précédentes montrent bien que les apprentissages des élèves s'inscrivent dans une logique d'interdépendance dont l'approche disciplinaire ne peut rendre compte, mais aussi qu'une définition a priori des compétences, n'est pas adéquate. Pour progresser dans ce domaine et permettre une identification empirique des compétences, nous avons adopté une méthode basée sur l'analyse des corrélations entre les résultats aux items. Ce seront les corrélations les plus fortes entre items qui nous permettrons d'identifier des compétences en regroupant les items qui mesurent des aptitudes proches chez les élèves. Notre démarche s'inspire globalement des procédures utilisées pour les validations empiriques a posteriori des épreuves d'évaluation et plus précisément les validations empiriques internes (De Ketele, Gérard, 2005).

L'identification des compétences selon une approche empirique

Cette approche au niveau des items permet tout d'abord d'utiliser une échelle de mesure commune puisque tous les items présentent le même barème de cotation : 0 pour une réponse erronée, 1 pour la réponse attendue. Le second avantage, découlant nécessairement du précédent, est que la question du seuil de réussite ne se pose plus et que celui-ci est défini objectivement par la réussite ou l'échec, sans possibilité de situations intermédiaires[8]. L'analyse des relations entre items passe en premier lieu

[8] On mentionnera que cette dichotomie (échec / réussite) n'est pas adaptée à une interprétation diagnostique des évaluations nationales et que les barèmes de cotation de certains items comportent à l'origine d'autres paliers qui permettent une analyse des erreurs des élèves.

par le calcul de corrélations. Compte tenu du nombre d'items (171 items pour le CE2 en 1999), la matrice de corrélations produit 14535 coefficients. L'examen des corrélations permet d'identifier une grande variété des situations, certains exercices pouvant être perçus comme très cohérents dans la mesure où les items qui les composent tendent à mesurer la même chose (corrélations fortes) alors que d'autres exercices sont composés d'items beaucoup plus indépendants les uns des autres (corrélations faibles ou nulles). Cela confirme l'observation précédente selon laquelle certaines compétences seraient mesurées de façon imparfaite.

Une difficulté liée à notre approche tient à la sélection des corrélations inter-items en vue de l'étude approfondie des relations statistiques. Le nombre très élevé de corrélations nous contraint à effectuer des choix et nous avons sélectionné uniquement les relations les plus fortes parmi l'ensemble. Le seuil sur lequel nous nous basons pour effectuer ce choix revêt forcément un caractère arbitraire. Nous avons décidé de ne retenir que les corrélations supérieures à +0,20 : elles sont au nombre de 317 ; ce seuil est en grande partie déterminé par les possibilités d'analyse qui tiennent d'une part aux capacités de traitement du logiciel (il existe une contrainte d'arbitrage entre le nombre de corrélations et le nombre d'observations) et d'autre part aux possibilités d'interprétation des résultats ; celles-ci seraient peu opérationnelles si tous les liens entre items, notamment les plus faibles étaient examinés en détail. On signalera que les 317 corrélations représentent la grande majorité des items (141 items sur les 171).

Afin d'organiser l'analyse des relations entre items, une phase préparatoire est nécessaire pour étudier individuellement chaque corrélation et dresser ainsi une cartographie de l'ensemble des situations. Le principe de cette étape préalable est d'identifier des blocs de relations au sein desquels on retrouve le plus souvent les mêmes items. Cette procédure est systématique puisque pour chaque item, on identifie tous les autres items qui lui sont associés dans les corrélations. Au terme de cette phase, on aboutit à un ensemble de 29 regroupements comportant chacun un nombre d'items varié. La phase suivante consiste à étudier chacun de ces regroupements de corrélations entre items et d'isoler les compétences à l'aide d'analyses en variables latentes.

L'objectif est ici d'analyser les liaisons statistiques entre items de façon à mettre à jour des variables latentes pouvant être interprétées comme des compétences, des aptitudes ou des capacités mobilisées par les élèves dans les évaluations nationales. Les modèles ont été estimés avec le logiciel LISREL[9] ; ils permettent de rendre

[9] LISREL :LInear Structurel RELationship.

explicite une dimension latente des phénomènes étudiés en postulant l'existence de variables inobservées (ou inobservables) qui rendent compte des relations ente variables observées, en l'occurrence ici les résultats aux items. L'adoption de cette démarche contraint à la formulation d'un modèle théorique postulé « a priori ». Elle vise en fait *« à vérifier la validité d'une théorie causale préalablement formulée en testant l'ajustement d'un modèle mathématique à des données observées»* (Aish-Van Vaerenbergh, 1997). Le logiciel LISREL permet d'estimer des modèles de mesure en établissant des relations entre les variables latentes (compétences supposées) et leurs indicateurs (les items). Pour illustrer cette démarche, nous présentons les estimations d'un modèle de mesure faisant intervenir un regroupement de 8 items de français.

Sur ces 8 items, 5 items (items 62, 63, 64, 66, 67) appartiennent au même exercice et sont censés mesurer la même compétence «*écrire sous la dictée des mots courants, de petites phrases ou de petits textes*»[10]. La réussite à la dictée dépend de plusieurs éléments : de compétences orthographiques, de la capacité à se remémorer la graphie des mots préalablement observés et, dans une certaine mesure, d'une capacité d'attention. L'item 60 est supposé rendre compte de la compétence « *copier un mot, une phrase, un texte…* » mais cet item n'évalue qu'un aspect de la production des élèves : le respect de la ponctuation (majuscules et points). L'item 12 fait partie d'un exercice dans lequel on demande aux élèves de repérer des types d'écrits à partir d'un extrait de livres différents ; la compétence visée est *«distinguer des textes différents en utilisant des indices extérieurs au texte »*. Enfin, l'item 16 vise la compétence : « *comprendre un texte et montrer qu'on l'a compris* ». Un texte est proposé aux élèves et ceux-ci doivent répondre à des questions concernant ce texte.

Une première étape a consisté à définir un variable latente qui détermine les résultats des élèves à tous ces items. Les statistiques fournies avec ce premier modèle de mesure indiquent que celui-ci peut-être amélioré et le logiciel LISREL suggère qu'une nouvelle variable latente peut être introduite dans l'analyse en isolant les items 64 et 66. Un second modèle a donc été estimé en définissant deux variables latentes La première variable « *comp1* » est déterminée par les items 12, 16, 60, 62, 63, la seconde « *comp2* » par les items 64 et 66. Le graphique suivant présente le diagramme causal associé au modèle de mesure tel que le produit le logiciel. Il est

[10] Dans cet exercice, il est demandé aux élèves d'écrire la phrase suivante, dictée par l'enseignant : «*pendant la récréation, les garçons et les filles jouent aux billes.* ». Cette phrase a été auparavant écrite au tableau, les élèves étant invités à mémoriser l'orthographe des mots. La phrase a été ensuite effacée et un exercice différent a été proposé aux élèves ; ce n'est qu'après cet exercice que la phrase a fait l'objet d'une dictée.

d'usage de représenter le modèle de mesure par un diagramme dans lequel les variables latentes sont symbolisées par des ellipses et les variables observées (ou indicateurs de ces variables latentes) sont matérialisées par des rectangles. Les flèches en traits pleins matérialisent l'intensité des relations qui lient chacun des indicateurs (items) à la variable latente (compétence), un coefficient de régression (et son degré de significativité) pour chaque indicateur fournit une indication sur la validité du modèle[11] (Morlaix, 2002).

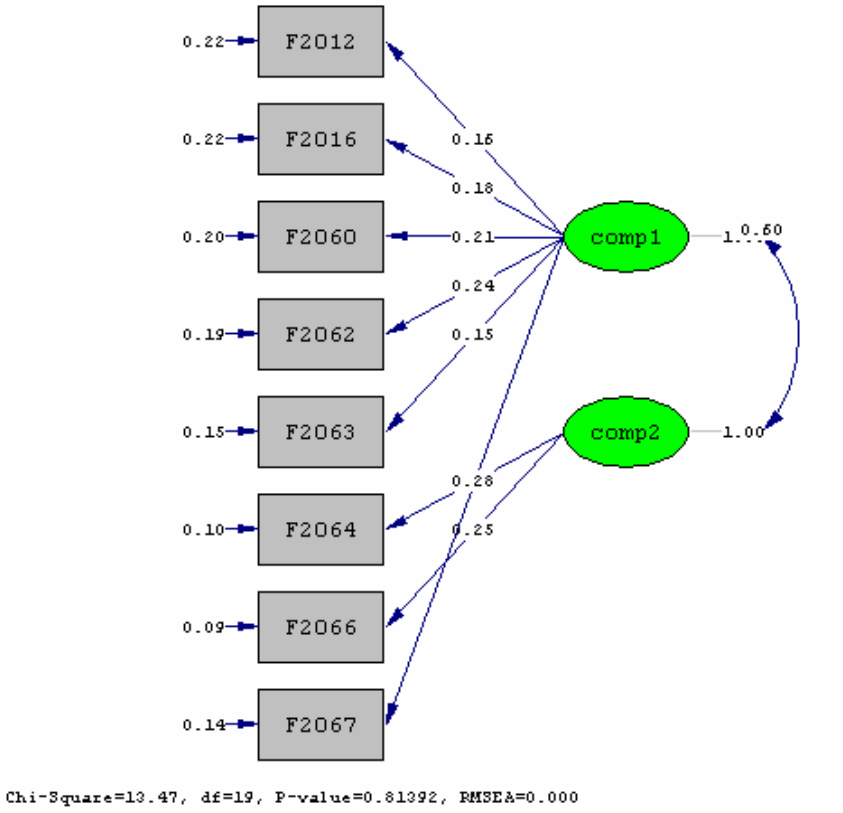


Figure 2 : Modèle de mesure établi à partir d'un regroupement de 8 items
Evaluations CE2, 1999

[11] Le logiciel LISREL fournit des informations sur la qualité du modèle grâce à plusieurs indicateurs statistiques : χ^2 (khi carré) mesure la distance séparant les deux matrices de corrélations, celle théorique et celle portant sur les corrélations estimées, la π value (P value) mesure la probabilité d'obtenir la valeur du χ^2 correcte, l'indice RMSEA (Root Mean Square Error of Approximation) se rapporte à la moyenne des résidus du modèle et à leur significativité.

On voit apparaître sur le schéma la corrélation entre les deux variables latentes (coefficient de +0,60), cela témoigne de la proximité des deux compétences. Toutes les relations matérialisées sur le schéma sont significatives. Il semble donc bien que les items 64 et 66 mesurent une compétence spécifique qui rende compte des aptitudes des élèves à orthographier correctement des noms écrits au pluriel (« *garçons* » et «*filles* »). Les 6 autres items (12, 16, 60, 62, 63 et 67) mesureraient pour leur part une compétence différente. Il est difficile d'interpréter rigoureusement ces relations sans avoir recours à des éléments théoriques. En effet, les résultats des élèves aux items de ce premier regroupement semblent dépendre de plusieurs aptitudes : mémorielles, attentionnelles, orthographiques…. En tout cas, nous sommes sans doute loin de la seule compétence orthographique formulée dans les évaluations nationales. Les modèles de mesure estimés sur la totalité des regroupements d'items ont ainsi donné lieu à l'identification de 63 variables latentes de composition variée (Morlaix, Suchaut, 2006) qui forment une recomposition des compétences de l'évaluation CE2. Parmi les 63 variables mises à jour, 27 d'entre elles (soit 43%) correspondent, souvent de façon partielle, à des regroupements d'items déjà présents dans les épreuves nationales. Le recouvrement avec les compétences des épreuves est très imparfait puisque seules 5 variables correspondent exactement à des compétences figurant dans les évaluations (tableau 2).

L'étape suivante a consisté à identifier des ensembles de compétences plus consistants sachant que certaines variables latentes peuvent, du fait de la méthode employée, être très proches les unes des autres[12]. Nous avons donc conduit de nouvelles analyses qui s'appuient sur la matrice de corrélations entre les différentes variables latentes. Les coefficients de corrélation présentent des valeurs allant de 0 à +0,80 (ce qui est un chiffre très élevé). Pour isoler les compétences majeures parmi l'ensemble des variables, nous avons sélectionné les corrélations les plus importantes (supérieures à +0,70). Cela permet de dégager trois grands groupes de compétences qui sont représentées sur le graphique suivant (les chiffres correspondent aux numéros associés aux variables latentes).

[12] En effet, les modèles de mesure ont été estimés séparément sur chacun des regroupements d'items.

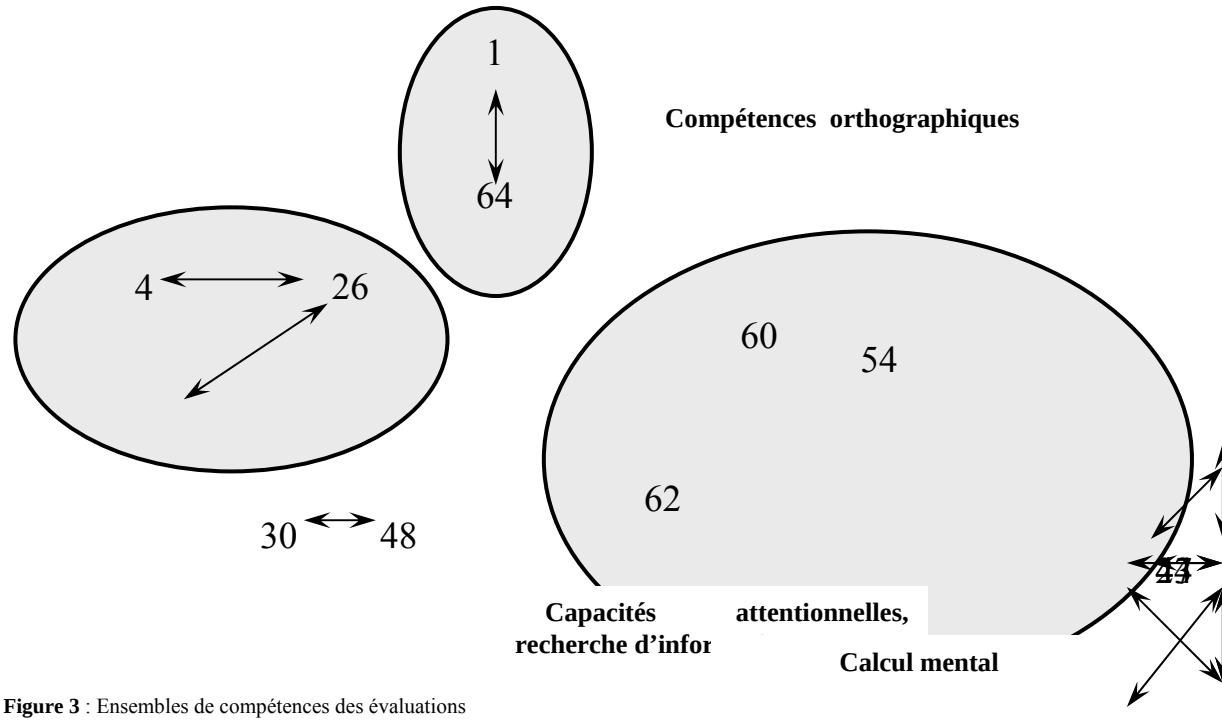


Figure 3 : Ensembles de compétences des évaluations CE2 (1999)

Un premier ensemble est composé des variables « *comp1* » et « *comp64* ». Il s'agit ici clairement de compétences orthographiques qui sont mesurées avec ces deux variables. En effet, la presque totalité des items rendant compte de ces variables latentes concernent deux exercices de dictée. Le deuxième ensemble regroupe les variables « *comp4* », « *comp26* » et « *comp62* ». Cet ensemble d'apparence disparate prend sens quand on l'examine sous l'angle de la psychologie cognitive[13]. Autant les items de français que ceux de mathématiques présents dans ce regroupement évaluent la capacité que peuvent avoir les élèves à rechercher de l'information plus ou moins complexe à partir de supports divers (texte, mots, calendrier, emploi du temps, plan, énoncé de problème). Ce sont donc les capacités attentionnelles des élèves qui sont mises à contribution pour la maîtrise de cette compétence globale. Le troisième ensemble établi sur les corrélations les plus fortes, regroupe 7 variables latentes. Le

[13] Des collègues psychologues du L.EAD (Université de Bourgogne), Pierre Barrouillet et Valérie Camos, ont été associés à cette partie de la recherche, ils ont notamment contribué à donner une signification à certaines variables latentes identifiées par les modèles LISREL.

lien commun entre ces variables est également clair puisque les items de calcul mental interviennent systématiquement pour chacune d'entre elles et 21 fois au total dans ce regroupement.

Les acquisitions des élèves à l'entrée au CE2 s'organisent principalement autour de ces trois compétences qui ne sont pas de même nature. Si l'acquisition des compétences orthographiques dépend principalement d'un enseignement systématique, les deux autres compétences sont davantage associées à des processus plus complexes qui interviennent de façon transversale dans de nombreuses situations d'apprentissage. On devrait s'attendre à ce que ces compétences majeures contribuent fortement à l'explication des différences de réussite entre élèves à l'entrée au CE2. Pour vérifier cela, une régression « pas à pas » a été estimée avec comme variable dépendante le score global de CE2 et comme variables explicatives l'ensemble des compétences mises à jour précédemment. On constate alors que les trois compétences les plus prédictives (« *comp48* », « *comp4* », « *comp64* ») appartiennent chacune à un des groupes identifiés auparavant. Ces trois compétences expliquent à elles seules 82% de la variance du score global, elles sont bien au centre des acquisitions des élèves à l'entrée au CE2. Les habiletés en calcul mental, les capacités attentionnelles (et à trouver rapidement des informations dans des supports variés), la maîtrise de l'orthographe structurent ainsi fortement les résultats des élèves au début du cycle III.

Nouvelles perspectives d'utilisation des évaluations nationales

Nous insisterons pour conclure sur les pistes qu'ouvre notre démarche d'analyse. Nous avons mis l'accent dans ce texte sur le fait que la définition des compétences des évaluations nationales française était contestable dans la mesure où un même exercice peut mobiliser des compétences parfois nombreuses et diverses alors que la catégorisation « officielle » se limite à l'identification d'une seule compétence pour cet exercice. Notre méthode, basée sur une analyse statistique des relations entre les résultats obtenus aux items, permet au contraire d'associer des situations d'évaluation mobilisées dans des exercices distincts et de faire émerger des compétences capables de mieux cerner la structure des apprentissages des élèves.

Parmi la soixantaine de compétences identifiées, certaines ont un statut particulier dans le sens où elles rendent compte, à elles seules des différences d'acquisition des élèves à l'entrée au cycle III. Ces compétences concernent trois domaines :

l'orthographe, l'attention et la recherche d'information, le calcul mental[14]. On rejoint ici un constat réalisé par la D.E.P. sur la base du panel 1997[15] selon lequel l'épreuve d'attention partagée administrée à l'entrée au CP s'est révélée fortement corrélée avec la réussite générale à l'évaluation des CE2, deux années plus tard ; « l*a capacité d'attention pourrait donc bien constituer un facteur majeur de la réussite* » (M.E.N., 2000). Un premier intérêt à notre travail est donc de pouvoir proposer des épreuves d'évaluation réduites, concentrées sur les items mesurant les compétences dégagées avec notre approche. Cela pourrait être très utile pour identifier les élèves en difficulté à l'entrée en CE2 ; cette identification pouvant être réalisée au niveau de la classe dans une perspective diagnostique, au niveau de l'école dans une perspective de pilotage.

Notre recherche a aussi permis d'approfondir les relations entre les compétences en mettant en évidence une structure hiérarchique (Morlaix, Suchaut, 2006), certaines compétences étant indispensables à l'acquisition d'autres compétences. Sans entrer ici dans le détail de cette structure, signalons que la compétence de CE2 qui se trouve au sommet de la hiérarchie correspond à la maîtrise de la soustraction qui apparaît être un processus nécessitant de la part des élèves diverses capacités et habiletés préalables. La compétence se situant au plus bas niveau de la structure hiérarchique concerne des exercices qui visent la compréhension de consignes simples (alors que les évaluations nationales ont répertorié ces exercices dans le domaine de la mesure). En prenant en compte la structure hiérarchique d'apprentissage effective des élèves, c'est-à-dire en respectant le fait, trivial à présent, que certains apprentissages ne peuvent se réaliser que si d'autres déjà maîtrisés (Bloom, 1979), les résultats évoqués précédemment débouchent naturellement sur un questionnement lié à la planification des activités d'enseignement, et plus largement au contenu des programmes scolaires. Il est aussi essentiel de s'interroger aussi sur le temps que l'on doit consacré à l'acquisition des différentes compétences, sachant que le temps alloué est très lié aux progrès des élèves et que celui-ci s'inscrit, pour l'enseignant, dans une logique d'arbitrage entre des activités diverses et concurrentielles (Morlaix, 2000 ; Suchaut, 1996).

Les évaluations nationales ont sans aucun doute contribué à faire évoluer positivement le système éducatif français en dotant l'institution d'outils pertinents utilisables à différents échelons. Ces outils, malgré leurs qualités initiales, demandent

[14] . Ces deux derniers éléments pouvant intervenir dans des situations d'apprentissage très variées qui dépassent largement l'approche disciplinaire.

[15] Elèves entrés au CP en 1997.

une certaine prudence quand on souhaite interpréter correctement les résultats des élèves, surtout dans une perspective d'évaluation diagnostique. Plus de 15 ans après leur mise en place, on peut souhaiter une évolution significative au niveau de la définition des compétences censées être évaluées à travers les situations proposées dans les épreuves. Les évaluations nationales sont dans un sens une bonne illustration de la vision simplificatrice et réductionniste que peut avoir de nos jours la notion de compétence dans le domaine de l'éducation. Peut-être que notre approche empirique contribuera modestement à éclairer « *la caverne d'Ali Baba conceptuelle*» (Crahay, 2006) que représente la notion de compétence, c'est en tous cas ce que l'on a essayé de faire apparaître dans notre démonstration. Il n'en reste pas moins que des recherches complémentaires doivent être conduites pour tenter de mieux cerner ce que les situations d'évaluation mesurent réellement chez les élèves, quels concepts sont manipulés, quels processus d'apprentissage sont mobilisés.

Pour conclure, cette recherche a montré que les évaluations nationales pouvaient constituer une base empirique d'analyse très riche, à partir du moment où l'on ne se contente pas des regroupements d'items déjà proposés dans les documents d'accompagnement. L'utilisation d'analyses statistiques en variables latentes a permis d'identifier des compétences et des habiletés différentes de celles visées par les évaluations diagnostiques qui peuvent donner lieu à une réflexion nouvelle de la part des acteurs. La hiérarchisation des compétences peut quant à elle fournir des indications quant à la définition des curricula, en priorisant certaines compétences considérées comme centrales et se développant en amont du cycle III.

Références

Aish-Van Vaerenbergh, AM.(1997), « Modèles statistiques et inférences causales : analyse de structures de covariances avec LISREL ». In *Faut-il chercher aux causes une raison ? L'explication causale en sciences humaines.* Aish-Van Vaerenbergh, AM. et al. Librairie philosophique Vrin, 106-130.

Bloom, B.S. (1979) *Caractéristiques individuelles et apprentissages scolaires*, Bruxelles, Labor, Paris, Nathan.

Crahay M. (2006), Dangers, incertitudes et incomplétude de la logique de la compétence en éducation. *Revue française de pédagogie*, N°154, 97-110.

De Ketele J.M., Gerard F.M. (2005), La validation des épreuves d'évaluation selon l'approche par les compétences, *Mesure et Évaluation en Éducation*, À paraître.

Gerard F.M. (2005), *L'évaluation des compétences à travers des situations complexes*. Actes du Colloque de l'Admee-Europe, IUFM Champagne-Ardenne, Reims, 24-26 octobre 2005.

Gillet P. [éd.] (1991), *Construire la formation : outils pour les enseignants et les formateurs*. Paris, ESF.

Laveault D., Grégoire J. (1997), *Introduction aux théories des tests en sciences humaines*. De Boeck Université.336 pages

M.E.N. / Direction de l'Enseignement Scolaire (2000). *L'exploitation de l'évaluation nationale en CE2 : la lecture. Programme national de pilotage : actes du séminaire national*. C.R.D.P. de l'Académie de Versailles.

Morlaix S. (2000). Rechercher une meilleure répartition du temps scolaire en primaire pour favoriser la réussite au collège. *Revue française de pédagogie.* N°130, 121-131

Morlaix S. (2002). Intérêts et apports de l'analyse des variables latentes pour le chercheur en sciences sociales : exemple d'application à l'économie de l'éducation. *Orientation scolaire et professionnelle*, vol 31., N°1, 117-138.

Morlaix S., Suchaut B. (2006). *Evolution et structure des compétences des élèves à l'école élémentaire et au collège. Une analyse empirique des évaluations nationales*. Rapport pour l'UNSA-Education. 196 p. et annexes. Dijon, IREDU.

Suchaut B. (1996), « La gestion du temps à l'école maternelle et primaire : diversité des pratiques et effets sur les acquisitions des élèves ». *L'année de la recherche en sciences de l'éducation*, 123-153.

NOUVELLES PRATIQUES D'EVALUATION : LA SITUATION DANS LE CANTON DE GENEVE

Anne Soussi, Service de la recherche en éducation, Genève (Suisse)
anne.soussi@etat.ge.ch.

Mots-clés: évaluation, innovation, pratiques

Résumé:

Dans le canton de Genève comme ailleurs, des innovations sont régulièrement expérimentées mais pas forcément généralisées. Nous allons présenter les principales indications fournies par l'Institution concernant l'évaluation et montrer comment les enseignants de l'école obligatoire et du postobligatoire (secondaire II) s'en emparent et comment leurs pratiques évoluent à l'école obligatoire et au secondaire II. La recherche dont il sera question se base sur plusieurs types d'information: une analyse documentaire des textes officiels donnant aux enseignants des indications sur l'évaluation (règlements, plans d'études, etc.), des entretiens et une enquête par questionnaire.

Enfin, pour conclure, une brève présentation des nouveaux prolongements dans le domaine de l'évaluation à Genève sera effectuée.

Les trois niveaux d'enseignement fonctionnent de manière très différenciée et les innovations ne sont pas forcément de même nature. Ainsi, à l'école primaire dès la troisième année, il existe deux systèmes d'évaluation, avec ou sans notes, en lien avec l'implication dans des projets d'école (selon la rénovation de l'enseignement primaire). En plus de l'évaluation sommative/certificative et formative, l'évaluation informative est recommandée sur la base de dossiers d'évaluation, de portfolios et d'entretiens avec les parents.

Au secondaire I, l'évaluation sommative/certificative, formative et diagnostique sont préconisées. L'évaluation sommative est censée donner lieu à des notes (livret scolaire) deux, trois ou quatre fois par an selon les établissements.

A l'école obligatoire, des évaluations externes sous la forme d'épreuves communes ont lieu à des différents moments-clé de la scolarité sous une forme sommative ou diagnostique (en particulier au secondaire I).

Enfin, au secondaire II, règne une certaine autonomie et peu d'indications sont données à l'ensemble des enseignants en matière d'évaluation. Au moment de notre analyse documentaire, les évaluations externes étaient absentes, ce qui a évolué depuis.

Les pratiques et représentations des enseignants sont de ce fait assez différentes d'un ordre à l'autre même si certains grands principes, comme les finalités de l'évaluation, font l'objet d'un certain consensus.

Nouvelles pratiques d'évaluation : la situation dans le canton de Genève

Dans le canton de Genève comme ailleurs, des innovations sont régulièrement expérimentées mais pas forcément généralisées. Dans un premier temps, nous présenterons les principales indications fournies par l'Institution concernant l'évaluation puis montrerons comment les enseignants de l'école obligatoire et du postobligatoire (secondaire II) s'en emparent. La recherche dont il sera question se base sur plusieurs types d'information: une analyse documentaire des textes officiels donnant aux enseignants des indications sur l'évaluation (règlements, plans d'études, etc.), des entretiens et une enquête par questionnaire.

Enfin, pour conclure, une brève présentation des nouveaux prolongements dans le domaine de l'évaluation à Genève sera effectuée.

Un système d'enseignement et d'évaluation complexe

Au moment où nous avons analysé la situation, elle se présentait de manière particulièrement complexe, notamment à l'école primaire. Depuis 1994, l'école primaire connaît une réforme qui se généralise au fur et à mesure, impliquant de plus en plus d'établissements scolaires. Au premier cycle qui dure jusqu'à la fin de la deuxième année (élèves de 8 ans), l'évaluation n'est pas chiffrée, se basant sur des appréciations et des commentaires. Dans le deuxième cycle, (jusqu'à la fin du primaire: élèves de 12 ans), deux situations coexistent en fonction de l'engagement des écoles dans des projets liés à la rénovation de l'enseignement primaire. Les écoles en projet pratiquent une évaluation sans notes avec des commentaires et des appréciations comme dans les plus petits degrés. Dans les autres écoles, les notes apparaissent dès la troisième année comme dans la suite du cursus.

Dossiers d'évaluation, portfolios constituent la base du processus et des entretiens avec les parents sont prévus au moins trois fois par an afin de les informer.

Au secondaire I (élèves de 12 à 15 ans), la situation n'est pas beaucoup plus simple avec deux types de structures en parallèle. La plus répandue (16 établissements sur 19) comporte des regroupements différenciés (2 ou 3 selon les établissements) où les élèves sont répartis sur la base de leurs résultats scolaires. Dans le second système, les élèves fréquentent, au 7e degré, des classes hétérogènes puis sont répartis dans des niveaux en mathématiques et allemand.

Au moment de notre état de situation, les établissements jouissaient d'une certaine autonomie du point de vue de la périodicité de l'évaluation (certificative) à raison de deux, trois ou quatre fois par an.

Enfin, dans le niveau d'enseignement postobligatoire (secondaire II), trois types de formation ou d'orientation sont offerts aux élèves: une première qui mène à un diplôme de type baccaléauréat général, une autre qui mène à un diplôme de culture générale, et l'enseignement professionnel, en école ou en alternance école-entreprise, qui peut mener à un certificat d'aptitudes fédéral (entre le CAP et le BTS) ou à une maturité professionnelle (de type baccalauréat). Nous ne nous sommes intéressés qu'aux deux premiers programmes qui regroupent environ la moitié des élèves, le troisième étant régi par des règlements fédéraux et par profession.

L'enseignement postobligatoire se caractérise par une grande autonomie des établissements et des enseignants, avec peu de directives communes. Les deux programmes de formation sont assez différents: l'un relativement traditionnel avec un accent très fort sur l'évaluation sommative/certificative et l'autre plus innovant. Les trois formes d'évaluation sont présentes dans les textes (sommative/certificative, formative et diagnostique). De manière générale, à ce niveau d'enseignement, l'évaluation externe est quasi absente et l'évaluation sommative/certificative très présente.

De manière générale, à l'école obligatoire, l'évaluation externe vient compléter l'évaluation à plusieurs moments clés du cursus dans différentes matières (français, maths et allemand au primaire, auxquels se rajoutent d'autres matières telles que biologie, anglais, etc.). Cette évaluation est en pleine expansion.

Plusieurs formes d'évaluation sont recommandées aux enseignants tout au long du cursus scolaire: l'évaluation sommative/certificative et formative pour les trois niveaux d'enseignement avec en plus l'évaluation informative au primaire (Allal & Wegmuller, 2004; Weiss, 2003) ainsi que l'évaluation diagnostique (de Ketele, 1993) au secondaire I surtout.

En vue d'harmoniser les principes d'évaluation tout au long de la scolarité, un cadre de référence a été mis sur pied mais n'a pas été officialisé étant donné un accueil mitigé et variable des enseignants selon les moments du cursus scolaire.

Les principales indications concernant l'évaluation

L'école primaire comporte trois principales formes d'évaluation: formative, sommative/certificative et informative. L'évaluation formative met l'accent sur l'élève à qui elle permet de se situer. Un autre élément important est souligné; celui de la progression permettant à l'enseignant de piloter le passage d'une séquence à l'autre. L'évaluation sommative/certificative a avant tout une fonction de bilan, pour attester du degré d'atteinte des objectifs. Quant à l'informative, développée dans le contexte genevois et romand (cf. Allal et Wegmüller, 2004; Weiss, 2003; Gretel, 2004) , distinguée en tant que telle essentiellement au primaire, sa fonction principale est d'informer les parents sur les apprentissages de leur enfant. Par ailleurs, depuis de nombreuses années, l'évaluation se pratique non seulement de manière interne, c'est-à-

dire élaborée et gérée par l'enseignant mais de façon externe au moyen d'épreuves communes à tous les élèves d'un même degré. Au moment de nos analyses, elles existaient en fin de 2ème et de 6ème année dans un certain nombre de branches scolaires telles que le français et les mathématiques pour les deux degrés concernés ainsi qu'en écriture-graphisme (2ème) et en allemand (6ème).

Au secondaire I, trois formes existaient: formative et sommative/certificative ainsi que l'évaluation diagnostique. L'évaluation formative met surtout l'accent sur le diagnostic et la remédiation tandis que l'évaluation sommative/certificative a à la fois une fonction de bilan mais aussi d'orientation : elle sert à valider les acquis et compétences des élèves en vue de leur orientation. L'évaluation diagnostique n'est pas définie en tant que telle mais il est précisé que les évaluations communes de début d'année ont une valeur diagnostique. On peut la rapprocher de la définition de de Ketele: une évaluation qui s'effectue avant la séquence d'apprentissage et qui a pour but de diagnostiquer et pronostiquer. L'évaluation externe était en pleine expansion au moment de notre état de situation et présente dans de nombreuses branches à fin de la 9ème année (fin de la scolarité obligatoire). Parallèlement, en début de 8ème année, on trouvait une forme d'évaluation diagnostique commune.

Enfin, après la scolarité obligatoire, les différents règlements et textes de cadrage analysés ne donnent pas de définitions explicites des différents types d'évaluation qu'ils mentionnent.. On peut toutefois noter un essai d'harmonisation à l'Ecole de culture générale, programme de formation ne menant pas à un diplôme de type baccalauréat et qui accueille des élèves qui ne se destinent pas à une formation professionnelle. En effet, dans ce type d'école, face à une autre population d'élèves, on essaie d'innover. De nouveaux plans d'étude ont été élaborés et comportent pour chaque branche scolaire, une rubrique concernant l'évaluation. Au Collège (qui correspond au lycée), les rares directives se font au niveau local (par établissement ou par matière scolaire).

Les différents documents analysés mettaient en évidence un accent différent notamment au niveau de la définition des types d'évaluation et de la fréquence de ces dernières. Par contre, les objets (compétences, connaissances) et les instruments étaient assez peu différenciés si l'on excepte, certains outils d'évaluation davantage préconisés pour l'évaluation formative et à l'école primaire, tel que le dossier d'évaluation.

Quand on observe de manière plus précise les indications dans les documents analysés, on peut relever les éléments suivants:

- l'évaluation informative (présente uniquement à l'école primaire) se caractérise par l'utilisation du livret scolaire, du dossier d'évaluation, du portfolio et de la pratique de l'entretien. Elle porte sur le travail scolaire et le comportement et a lieu semestriellement.
- L'évaluation diagnostique, existant surtout au secondaire I, se présente sous la forme d'une évaluation externe, écrite et plutôt non chiffrée. Elle se situe en début de formation.

- L'évaluation sommative/certificative se décline différemment selon le moment du cursus scolaire. A l'école primaire, elle est prévue en fin de cycle. Elle peut être chiffrée ou non selon le type d'école (école dite en projet de rénovation ou non). Dans le secondaire I ou II, elle peut avoir lieu à la fin d'une période (trimestre ou semestre), d'une année voire d'un cycle. Elle est toujours notée. Dans la scolarité obligatoire, elle peut être interne (élaborée et gérée par l'enseignant) ou externe. Dans le secondaire II, elle est seulement interne.
- L'évaluation formative au primaire se présente notamment sous la forme de portfolio et de dossier d'évaluation. Dans le secondaire, on utilisera plutôt des questionnaires et des grilles d'autoévaluation. De manière générale, à l'école obligatoire, ce type d'évaluation est censée être non chiffrée.

Les pratiques des enseignants ou ce qu'ils en disent

Les enseignants des trois niveaux d'enseignement (primaire, secondaire I et II) ont été interrogés sur leurs pratiques d'évaluation au moyen d'un questionnaire. Nous allons maintenant illustrer par quelques exemples de ce qu'ils font des indications fournies par l'institution, notamment du point de vue des innovations.

Les finalités et la conception de l'évaluation

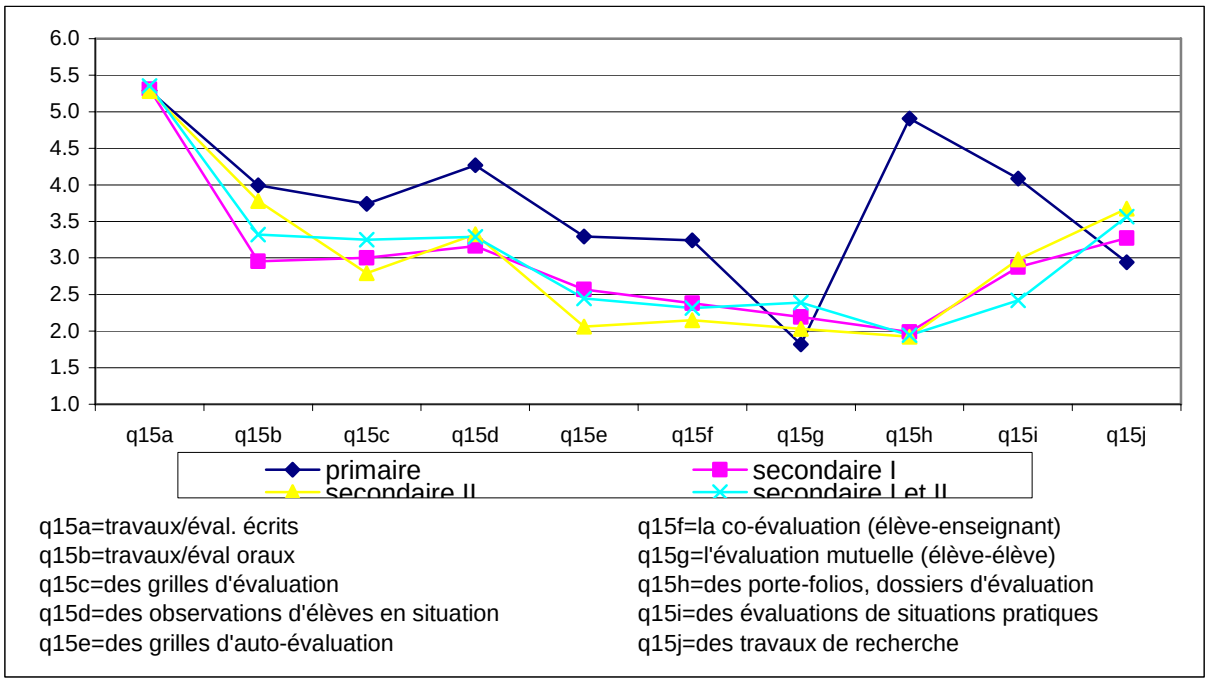
De manière, les enseignants des trois niveaux d'enseignement considèrent que l'évaluation a surtout pour finalités de situer les élèves dans un processus d'apprentissage et leur donner des repères, d'orienter et de réguler mais pas à classer les élèves. Les finalités principales semblent donc prendre en compte aussi bien l'évaluation formative que la sommative/certificative. La fonction d'information et notamment des parents est aussi une finalité importante, de manière encore plus marquée au primaire qu'au secondaire.

On peut toutefois différencier un peu les enseignants selon leur provenance (primaire ou secondaire). Ainsi, au primaire, l'évaluation considérée comme une sanction ou servant de carotte pour motiver les élèves est fortement rejetée. On peut sans doute relier cela à l'accent fort mis au primaire sur la progression de l'élève et l'absence de notes dans une partie des écoles.

Si l'évaluation est considérée par tous les répondants comme devant être intégrée au processus d'enseignement, le rôle de l'enseignant est toutefois variable selon les moments du cursus. Ainsi, les enseignants de l'école obligatoire rejettent l'idée d'un enseignant seul responsable de l'évaluation et ne s'intéressant qu'aux résultats de l'évaluation, conception s'apparentant à celle d'un instructeur-contrôleur selon Jorro (2000).

Les instruments utilisés

Quel que soit le niveau d'enseignement, les travaux écrits sont décrits comme étant les plus utilisés comme on peut le constater dans le graphique suivant. On retrouve le même type de tendance que notamment dans l'enquête française réalisée auprès d'enseignants du collège (secondaire I) (Braxmeyer, 2005).

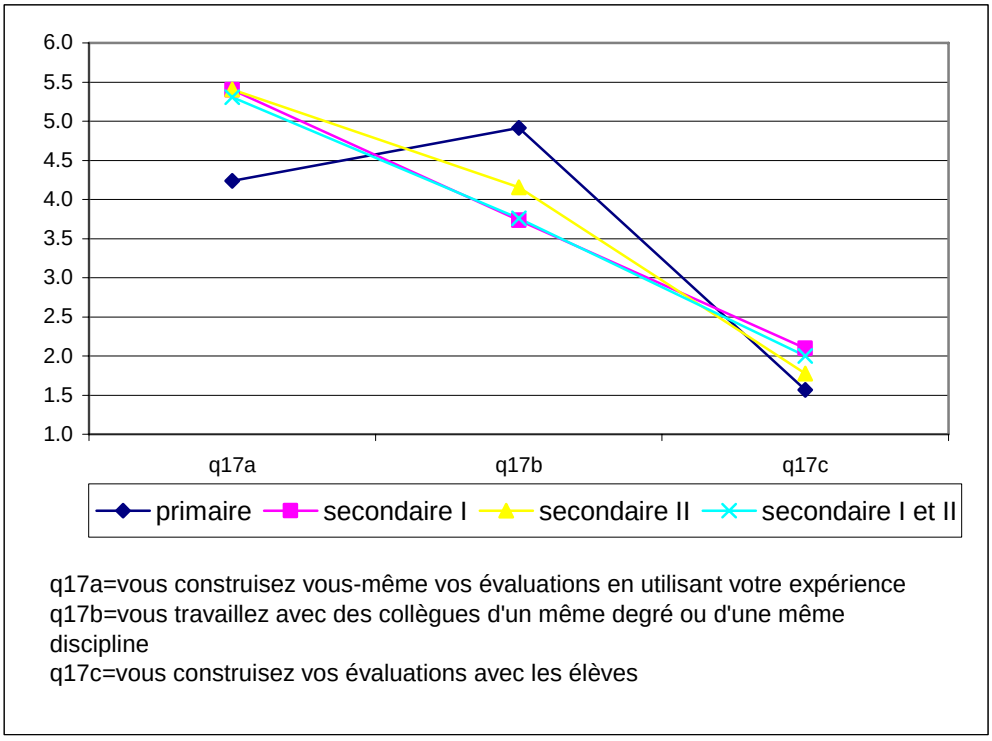


Graphique 1- Les instruments utilisés dans les trois ordres d'enseignement[1]

Toutefois, à l'école primaire, on observe une assez grande diversité d'instruments (comme dans les indications données par les règlement): grilles d'évaluation, portfolios et dossiers d'évaluation ainsi qu'observations de situations pratiques. L'utilisation des portfolios et dossiers d'évaluation qui accompagnent l'évaluation informative donne lieu à des écarts particulièrement importants entre les pratiques de l'école primaire et celle de l'école secondaire (I ou II) puisqu'elle est importante à l'école primaire et peu pratiquée dans la suite de la scolarité (en conformité avec les indications institutionnelles).

[1] Pour la plupart des questions, les enseignants sont censés se positionner sur une échelle de 6 points en fonction du degré d'accord ou de la fréquence. Les modalités vont de 1 (*jamais* ou *pas du tout d'accord*) à 6 (*très souvent* ou *tout à fait d'accord*), la moyenne se situant à 3.5. Par exemple, si en moyenne, on obtient la valeur de 2.5, cela signifie que l'affirmation est plutôt rejetée ou qu'il s'agit d'une pratique peu courante. Si par contre, elle se situe autour de 4 points, il y a un certain accord ou il s'agit d'une pratique fréquente.
La 4ème catégorie de répondants (secondaire I et II) concerne les personnes qui enseignent à la fois dans le secondaire I et dans le secondaire II. Leurs tendances de réponse sont le plus souvent intermédiaires entre les deux niveaux d'enseignement, un peu plus proche du secondaire I.

Un autre aspect différenciateur, lié à la fois aux innovations du primaire et à la structure, réside dans la collaboration entre enseignants. Ainsi, les travaux d'évaluation sont élaborés le plus souvent sur la base des plans d'études, de fiches ou d'exemples provenant de manuels ou encore d'anciennes épreuves de l'enseignant voire de collègues. Cette élaboration s'effectue plutôt individuellement au secondaire et davantage en collaboration au primaire comme on peut l'observer dans le graphique suivant.



Graphique 2-Manière de construire les évaluations

Représentations des deux principaux types d'évaluation

Les représentations des deux principales formes d'évaluation (sommative/certificative et formative) données par les enseignants sont relativement proches des indications fournies par les règlements respectifs et diffèrent donc d'un niveau d'enseignement à l'autre à l'exception de quelques caractéristiques. Ainsi, pour la majorité des enseignants interrogés, l'évaluation formative se caractérise par sa fonction de diagnostic et de remédiation, ainsi que par le moment où elle intervient, c'est-à-dire au cours d'une phase d'enseignement. L'évaluation sommative/certificatve a quant à elle une fonction de sélection et d'orientation, et intervient à quels moments de l'année, de préférence à la fin d'une phase d'enseignement ou d'une année voire

d'un cycle. Les deux évaluations sont annotées et donnent des repères sur les acquis et les lacunes des élèves.

Par contre, certaines caractéristiques sont, d'après les enseignants du secondaire, communes aux deux types d'évaluation: mesurer la progression et réguler l'enseignement, tandis que pour ceux du primaire, ces caractéristiques sont spécifiques à l'évaluation formative. On observe donc une différenciation plus marquée des deux types d'évaluation dans les déclarations des enseignants primaires que chez les enseignants du secondaire. Peut-être faut-il attribuer cela à la structure: dans un cas, des enseignants généralistes avec une plus grande liberté pour la gestion du temps (entre autres) et de l'autre, des personnes enseignant des matières scolaires spécifiques pour lesquelles parfois la dotation horaire hebdomadaire est réduite.

Par ailleurs, quand il s'agit du type de travaux ou des contenus, les deux évaluations ne semblent pas se différencier aux dires des enseignants: ce sont les travaux écrits qui priment comme dans beaucoup de contextes scolaires ainsi que les connaissances et compétences.

Modification de pratiques

Nous avons questionné les enseignants à propos des changements dans leur pratique au cours des années. Pour la majorité d'entre eux, les réponses vont dans le sens d'un changement de pratiques. Ces changements concernent surtout l'utilisation de critères plus précis et le fait de communiquer aux élèves les critères d'évaluation. Certains éléments sont davantage marqués au primaire, tels que le changement en fonction des réformes et l'introduction de l'évaluation formative, ou au secondaire comme l'abaissement du seuil d'exigences.

Différenciation de l'évaluation

Si l'on observe un certain consensus sur le fait que l'évaluation dépend du moment du cursus scolaire et de la discipline, les enseignants du secondaire se différencient de leurs collègues quand on parle d'une ***même évaluation pour tous par souci d'équité***. Ils sont peu favorables à cette idée. On peut supposer que la question de l'hétérogénéité au sein d'une classe semble plus difficile à gérer qu'au primaire, par exemple. Par ailleurs, une des innovations introduite à ce niveau dans les années 90 prévoyait les mêmes objectifs pour tous malgré une structure organisée principalement en filières différenciées. Il est intéressant de souligner que les enseignants français du Collège, niveau comparable au secondaire I genevois, mentionnent comme attente principale par rapport à l'évaluation la prise en compte de l'hétérogénéité des élèves en diversifiant l'évaluation.

Opinion sur l'évaluation externe

Malgré des pratiques très différentes selon le niveau d'enseignement concernant les évaluations externes (pratique bien ancrée au primaire, plus récente au secondaire I et plutôt absente au secondaire II), les avis sont très proches. Les enseignants sont d'accord avec l'idée que les évaluations externes permettent de réguler les pratiques et de générer des idées et rejettent les aspects négatifs proposés (par exemple l'enseignant connaît ses élèves et est seul capable de les évaluer, ou bien l'évaluation externe ne tient pas compte de ce qui se fait en classe, ou encore a pour fonction

d'évaluer les enseignants). On observe toutefois quelques différences entre les niveaux d'enseignement: l'idée comme quoi l'évaluation externe ne tiendrait pas compte de ce qui se fait en classe est moins fortement rejetée par les enseignants du secondaire.

Profils d'enseignants

Afin de synthétiser l'information, nous avons essayé de caractériser les enseignants en fonction de leurs réponses aux différentes questions et avons ainsi défini des profils d'enseignants au moyen de l'analyse de correspondance:

- Le premier groupe d'enseignants se caractérise par une certaine diversification des outils d'évaluation (et notamment l'utilisation de portfolios et de dossiers d'évaluation ou encore des balises), une certaine collaboration lors de la construction d'épreuves, pas d'information donnée aux élèves avant une épreuve pour ne pas les stresser, un impact de l'évaluation sur la suite de l'enseignement et la réutilisation de la formation continue. Les besoins relevés en matière d'évaluation sont les suivants: de la formation continue et davantage de collaboration entre collègues mais aussi des consignes claires ainsi que des outils d'évaluation. Ils ont une opinion relativement positive de l'évaluation externe, génératrice d'idées nouvelles, permettant une certaine objectivité et donnant des repères précis. Ils estiment que la nécessité de mettre des notes est à l'origine d'une certaine incompatibilité entre leur conception de l'enseignement apprentissage et leurs pratiques d'évaluation que l'on pourrait qualifier de formatives. Cette vision de l'enseignement est plutôt socioconstructiviste (c'est l'élève qui construit ses connaissances avec l'aide de l'enseignant). On trouve dans ce groupe davantage d'enseignants primaires.

- Le deuxième groupe estime que l'évaluation a pour finalité d'obliger les élèves à travailler, mettre des notes ou des appréciations, qu'il est nécessaire de multiplier les contrôles pour faire progresser les élèves, que l'évaluation est de la responsabilité du maître qui construit seul ses évaluations et qui établit les barèmes a posteriori. Pour eux, l'évaluation ne peut pas être la même tout au long de la scolarité car les enjeux sont différents et la manière d'évaluer dépend de la discipline d'enseignement.
  Les caractéristiques se rapprochent davantage de l'évaluation sommative, la vision de l'enseignement est plus transmissive (c'est l'enseignant qui transmet les connaissances) que dans le premier groupe. On y trouve plus d'enseignants du secondaire, particulièrement du secondaire II.

- Enfin, le troisième groupe a une position un peu intermédiaire: l'évaluation a pour finalités de vérifier les acquis, de voir où en est l'élève, le situer voire remédier à ses difficultés. Elle doit s'adapter au niveau de la classe, on utilise les directives officielles, la formation continue est réutilisée. Les élèves n'ont pas de documents à leur disposition pendant les épreuves. En matière de modifications apportées au fil des années, ils disent mieux intégrer l'évaluation dans leur enseignement, préciser aux élèves les critères d'évaluation ou encore n'avoir rien changé. Ce dernier groupe se caractérise notamment par l'expérience récente des enseignants. Il s'agit surtout des

enseignants voltigeant entre le secondaire I et II. Leur conception de l'enseignement est comme celle du premier groupe, proche d'une approche socioconstructiviste.

L'évaluation selon la discipline enseignée

Selon la matière enseignée, on observe aussi un certain nombre de différences dans les pratiques d'évaluation. Les tendances sont très proches de celles relevées dans l'enquête française (Braxmeyer, 2005).

De manière générale, si l'ensemble des enseignants du secondaire I et II (l'école primaire n'a pas été prise en compte dans cette analyse en fonction des disciplines) sont d'accord avec l'idée qu'on n'évalue pas de la même manière dans toutes les disciplines, les réponses au questionnaire ne donnent pas lieu à de grandes différences quand on les analyse en fonction des domaines disciplinaires. Seules quelques dimensions font exception: les *documents laissés à la disposition des élèves lors des évaluations*, les *outils d'évaluation* ou encore la *différenciation de l'évaluation*. Ainsi, les enseignants de *français* et des *sciences de l'homme* (histoire, géographie, éducation citoyenne, etc.) laissent des documents à disposition pendant les épreuves contrairement à leurs collègues d'autres domaines, tels que les ***langues II, les mathématiques*** ou les ***sciences de la nature*** (biologie, physique, chimie, observation scientifique, etc.).

Concernant les outils d'évaluation, les enseignants ***d'art et d'éducation physique*** se distinguent de leurs collègues (comme dans d'autres enquêtes telles celle réalisée au Collège en France) qui privilégient les travaux écrits, sans doute par la nature de la discipline, par l'utilisation de dossiers d'évaluation, d'évaluation de situations pratiques ou encore de la co-évaluation (élève/enseignant). De même, d'autres éléments sont pris en compte tels que la participation ou l'implication des élèves en cours. Le barème est établi a posteriori sur la base des résultats obtenus.

Il faut toutefois souligner que les différences de pratiques entre domaines de disciplines sont moins importantes que celles observées entre les niveaux d'enseignement à l'exception du domaine ***art et éducation physique*** où l'on privilégie tout autant la différenciation et l'adaptation en fonction du niveau de la classe ou des élèves au secondaire II que dans la première partie du secondaire.

Les représentations de l'évaluation formative et sommative/certificative sont dans l'ensemble assez semblables d'une discipline à l'autre, à quelques nuances près. Ainsi, de nouveau, les enseignants ***d'art et d'éducation physique*** se distinguent de leurs collègues. L'évaluation formative, à elle seule, sert à réguler l'enseignement et intervient régulièrement, vision plus proche des enseignants du primaire tandis que l'évaluation sommative/certificative seulement a une fonction de différenciation des élèves. On peut sans doute supposer que les finalités de ces disciplines ne sont pas tout à fait les mêmes que dans d'autres qui ont davantage pour fonction la sélection des élèves.

La situation actuelle

Actuellement, l'évaluation a encore connu des évolutions et des modifications auront certainement lieu prochainement. A l'école obligatoire, une certaine

harmonisation a été apportée dès la rentrée 2005: au primaire, l'évaluation avec notes a été généralisée dès la troisième année (et les appréciations traduites en notes) et au secondaire I , la fréquence des évaluations a été fixée à trois fois par an dans tous les établissements. Par ailleurs, l'évaluation externe a été encore développée à l'école obligatoire (3 fois à l'école primaire et deux au secondaire I) et harmonisée sous la forme d'évaluation-bilan seulement. L'évaluation diagnostique externe a été supprimée au profit d'une évaluation à la fin de la 7ème année (première année du secondaire I).

Enfin, après l'école obligatoire, les changements sont moins importants au niveau de l'évaluation si ce n'est l'introduction d'évaluations communes dans le deuxième type de formation (culture générale). Relevons que c'est surtout dans ce deuxième type de formation accueillant un autre type d'élèves que l'on cherche davantage à innover que dans la formation plus classique où l'évaluation se gère surtout au niveau de l'école.

De manière générale, si les pratiques évoluent un peu, leur changement est un processus lent (Gather Thurler, 1993) et les représentations changent souvent avant les pratiques. Ainsi, si les enseignants ont des représentations assez différentes de l'évaluation sommative/certificative et de la formative, dans la pratique, il est assez difficile de les différencier du point de vue du type de contenu ou des outils utilisés (Crahay, 2003; Perrenoud; 1998). Des contraintes structurelles et liées à la pratique quotidienne (Bélair, 1993; Van Nieuwenhoven et Jonnaert, 1993) peuvent sans doute expliquer ces écarts entre représentations et pratiques.

Références bibliographiques

Allal, L. Wegmuller, E. (2004). Finalités et fonctions de l'évaluation. In *L'Educateur*, numéro spécial, 4-7.

Bélair, L.M. (1993). L'acte d'évaluer les apprentissages: problématique de la formation. In. R. Hivon (dir.) *L'évaluation des apprentissages*. Montréal: Editions Logiques.

Braxmeyer, N., Guillaume, J.-C., Lairez, Th. Lévy, J.-F. (2005). *Les pratiques d'évaluation des enseignants en collège.* Paris: Les Dossiers, Ministère de l'Education nationale de l'Enseignement supérieur et de la Recherche, Direction de l'évaluation et de la prospective.

Crahay, M. (2003, 2ème édition). *Peut-on lutter contre l'échec scolaire*? Bruxelles: De Boeck.

De Ketel, J.-M. (1993). L'évaluation conjuguée en paradigmes. In *Revue Française de pédagogie*, no 103, 59-80.

Gather Thurler, M. (1993). Amener les enseignants vers une construction active du changement. Pour une nouvelle conception de la gestion de l'innovation. *Education et recherche*, 15, 2, 218-235.

GRETEL (2004). *Lignes directrices pour un système d'évaluation du travail des élèves dans la scolarité obligatoire.* Neuchâtel: Conférence intercantonale de l'Instruction publique de la Suisse romande et du Tessin.

Jorro, A. (2000). *L'enseignant et l'évaluation. Des gestes évaluatifs en question*. Bruxelles: de Boeck.

Perrenoud, Ph. (1998). *L'évaluation des élèves. De la fabrication de l'excellence à la régulation des apprentissages. Entre deux logiques*. Bruxelles: De Boeck.

Weiss. J. (2003). Une évaluation informative pour la régulation des apprentissages et des formations. In Laurier, M.D. (collectif dirigé par). *Evaluation et communication*. Outremont (Québec).

Van Nieuwenhoven, C., Jonnaert, Ph. (1993). Une approche des représentations des enseignants du primaire à propos de l'évaluation. In *Mesure et évaluation en éducation, 16*, 3 et 4, 41-79.

BILAN DES REFORMES DE L'EVALUATION DU TRAVAIL DES ELEVES EN SUISSE ROMANDE

Martine Wirthner, IRDP - Neuchâtel

Mots-clés : évaluation du travail des élèves – Suisse romande – historique – bilan

Résumé : Depuis une trentaine d'années, de nombreux travaux en évaluation se sont développés en Suisse romande. L'accent a été mis très rapidement sur l'évaluation du travail des élèves, au service de l'apprentissage. Publications théoriques et propositions ont fleuri, puis, des réformes ont été entreprises pour les concrétiser. Peu à peu, les différentes fonctions de l'évaluation ont été définies et ont donné lieu à des outils diversifiés à l'intention des enseignant-e-s, mais aussi des parents et des autorités scolaires.

Dans la présentation envisagée pour ce symposium, il s'agira de comprendre l'évolution de ces travaux depuis les années 70 jusqu'à aujourd'hui, c'est-à-dire comment on a passé de projets et réalisations fortement centrés sur l'évaluation formative, en lien avec l'émergence d'une théorie de l'apprentissage constructiviste et socio-constructiviste, de l'enseignement par objectifs et de l'essor des didactiques, à des projets et réalisations axés prioritairement sur l'évaluation des performances des élèves, en lien avec l'émergence des bilans de compétences et du pilotage du système éducatif. Dans ce sens, seront mis en relation les réformes scolaires proprement dites touchant au travail en classe, les travaux de recherche et les prises de position politiques, avec les décalages qui apparaissent immanquablement entre eux.

Ce bilan ouvrira les questions concernant les perspectives pour ces prochaines années, mais aussi les autres présentations prévues dans le cadre du symposium, concernant plus particulièrement des situations locales de Suisse romande. Il s'agira encore de se demander dans quelle mesure le parcours de l'évaluation dessiné pour la Suisse romande est le reflet d'un cheminement plus général, touchant l'ensemble des systèmes éducatifs du monde francophone, voire du monde occidental.

Introduction

En guise d'introduction, je voudrais rappeler les critiques qui sont nées dès les années 1930 à l'égard de l'évaluation scolaire, plus particulièrement à propos des notes. Cardinet (1991) montre en effet que le point de départ de la recherche sur l'évaluation scolaire se situe dans ces années, avec plusieurs études critiques qui ont mis en évidence le manque de fiabilité des notes au baccalauréat français. La comparaison des notes de plusieurs jurys faisait apparaître de grandes différences d'appréciation, conduisant Piéron (1963) – qui lui-même a effectué des expériences en décimologie – à affirmer que, finalement, « la réussite au baccalauréat dépend plus de l'examinateur que de l'examiné » ! De même, il a été montré que pour une même discipline, dans un même degré scolaire, les moyennes de classe à des tests

pouvaient varier très fortement, alors même que la distribution des notes présentait pour chaque classe à peu près le même profil (Reuchlin et Bacher, 1969, cités par Cardinet). D'autres critiques ont fleuri depuis, par exemple pour déplorer le fait d'utiliser les notes à des fins diverses et de leur attribuer plusieurs fonctions évaluatives à la fois (sélective, régulatrice et certificative) (Cardinet, 1991, 1995). Parallèlement, dès les années 1960, les travaux américains de Skinner sur l'enseignement programmé allaient modifier profondément les conceptions traditionnelles de l'évaluation en montrant comment définir précisément les étapes nécessaires pour arriver au savoir, à l'objectif fixé, et comment mesurer son degré d'atteinte par l'élève ou l'étudiant-e. Bloom et ses collaborateurs ont systématisé ces conceptions en distinguant évaluation sommative et évaluation formative.

En Suisse romande, l'ouvrage que publient Allal, Cardinet et Perrenoud en 1979 rend compte de cette évolution critique, et notamment de l'émergence de l'évaluation formative, prenant sa source dans l'analyse des systèmes et axée sur la régulation de l'apprentissage de l'élève.

Soulignons encore que, selon Cardinet (1991), les nouvelles conceptions de l'évaluation qui se dessinent alors correspondent aux nouvelles exigences de la société. En effet, à une société qui avait essentiellement le souci de donner une instruction soignée à ses élites, c'est-à-dire à la minorité appelée à tenir les leviers du pouvoir politique ou industriel, succède une société (d'après la seconde guerre mondiale) soucieuse de former le plus possible de citoyens capables de prendre des initiatives dans une économie libérale en plein essor. D'ailleurs, dans ce contexte, la concurrence devient forte sur les marchés. Il s'agit donc que la majorité réussisse sa formation scolaire. L'évaluation scolaire apparaît ainsi comme étroitement liée à l'évolution sociale, ce qui se révèle à nouveau vrai aujourd'hui, comme nous le verrons plus loin.

Les années 1980 ou le déploiement de nouvelles conceptions de l'évaluation. Prédominance de l'évaluation formative

En Suisse, la source d'importants changements dans le domaine de l'évaluation scolaire, en particulier du travail de l'élève, se situe dans un projet de recherche d'envergure appelé SIPRI (Examen de la situation de l'école primaire). Ce projet a duré de 1980 à 1986, sous l'égide de la CDIP, c'est-à-dire de la Conférence suisse des chefs de Départements de l'Instruction publique. Il portait sur l'ensemble de la situation de l'école primaire, incluant une étude sur les fonctions et les formes de l'évaluation scolaire. Des écoles dites de contact, expérimentales, ont offert un terrain d'essai des nouvelles propositions élaborées dans le cadre de ce projet. Des groupes de travail se sont également mis à la tâche, composés de chercheurs, d'enseignant-e-s, d'autorités scolaires et de parents. Dans le rapport final (1986), une synthèse des résultats a été présentée, sous forme de thèses. Un des chapitres de ce rapport concernait l'évaluation, intitulé « Appréciation du travail des élèves ». Il réunissait les positions majeures prises à cette époque en matière d'évaluation. En priorité, les auteurs ont insisté sur la fonctionnalité de l'évaluation et ont mis en évidence les fonctions suivantes :

L'évaluation peut être un moyen de diagnostic didactique et de planification pour l'enseignant-e ; il est important que ce dernier distingue, dans ses pratiques, les différentes fonctions que l'évaluation peut avoir, c'est-à-dire :

- une aide à l'apprentissage de l'élève ;
- un bilan scolaire ;
- un instrument pronostic pour l'orientation des élèves ;
- un moyen d'information.

A partir de cette distinction, le rapport SIPRI définit également les évaluations de la manière suivante :

- l'évaluation formative, sert à l'amélioration, à l'orientation et au contrôle du processus d'apprentissage, du comportement de l'élève et de l'enseignant-e dans la perspective de la maîtrise des objectifs d'apprentissage ;
- l'évaluation sommative, sert à faire un bilan de la somme des savoirs et des savoir-faire acquis à un moment donné d'une période de formation et permet d'établir si les objectifs sont atteints et, donc, si l'enseignement a porté ses fruits. Le rapport ne fait toutefois pas de distinction claire entre évaluation sommative et certificative ;
- l'évaluation pronostique, orientée vers l'avenir, sert à recommander une orientation à l'élève, ceci à divers moments de sa scolarité.

Les thèses du rapport mettaient encore en avant la valeur de l'autoévaluation, dans le but d'accorder davantage d'autonomie aux élèves et une meilleure compréhension de leur travail et de ses finalités. Elles encourageaient enfin l'information aux parents, dans le but de leur faire part des progrès et des difficultés de leur(s) enfant(s).

Ces thèses sur l'évaluation supposaient le développement, en parallèle, d'une pédagogie par objectifs et de la différenciation de l'enseignement. Elles prônaient également l'éloignement des temps de promotion et une formation des enseignant-e-s à l'évaluation, en particulier pour favoriser la mise en place d'une évaluation formative dans l'enseignement/apprentissage.

A travers SIPRI, on se rend compte de l'importance qu'a eue la recherche pour mettre à la lumière à la fois des questions, des projets, des expériences et des recommandations touchant à de nouvelles conceptions de l'évaluation. Les thèses qu'un projet comme SIPRI a soutenues appelaient de profonds changements dans l'école. Si des changements ont effectivement eu lieu dans cette décennie à propos des plans d'études et à l'intérieur de disciplines telles que les mathématiques et le français langue première, à la fin des années 1980 il était encore trop tôt pour qu'ils se concrétisent dans le domaine de l'évaluation. A cette époque, le Centre suisse pour la coordination de la recherche en éducation (CSRE) d'Aarau créait un réseau suisse « Evaluation », tandis qu'en Suisse romande naissait, à l'initiative de l'IRDP, un groupe intercantonal réunissant enseignant-e-s et chercheur-e-s, qui ont, au fil de leurs travaux, constitué un ensemble de connaissances sur les conditions théoriques et pratiques de l'évaluation formative.

On remarquera encore la lente progression, en Suisse en tout cas, des concepts et des réalisations concernant l'évaluation, en particulier l'évaluation formative, depuis

les premiers travaux critiques portant sur l'évaluation jusqu'aux propositions concrètes ; il aura fallu en effet près d'un demi-siècle !

Par ailleurs, on se rend bien compte que la distinction des fonctions de l'évaluation, la mise en avant de l'évaluation formative et de l'autoévaluation au service de l'enseignement/apprentissage, ont des conséquences sur la manière d'envisager celui-ci, surtout le déroulement des apprentissages des élèves. Ainsi, dans la mesure où l'évaluation formative doit informer le mieux possible l'élève de ses difficultés, ses capacités, mais aussi de la pertinence de ses démarches, d'une part l'interaction entre enseignant-e-s et élèves prend une place centrale, d'autre part l'autonomie de l'élève par rapport à son travail est considérée comme une capacité importante à développer. L'enseignement devient de plus en plus l'organisation par l'enseignant-e des activités scolaires susceptibles de donner aux élèves les moyens d'améliorer leur travail en classe, du temps pour apprendre, les ressources pour la gestion autonome de leurs apprentissages. Il y a donc incontestablement un lien fort entre didactique et évaluation, ce qui fait dire d'ailleurs à Cardinet que l'évaluation est moins un problème de mesure que de dialogue. Jean Cardinet, il faut le dire ici, spécialiste de la mesure et des statistiques, a joué un rôle incontestable dans la réflexion sur l'évaluation scolaire en Suisse romande. Ayant développé un regard critique vis-à-vis d'une vision de la mesure à l'école, en particulier la note pour répondre à toutes les fonctions de l'évaluation, il a construit une position argumentée autour de l'évaluation formative, soucieux de défendre l'apprentissage de tout élève, sans l'entrave précoce de la note et de la sélection. De son côté, Bain (1987 ; Bain & Schneuwly, 1993) a ardemment défendu ce lien et a fait des propositions pour illustrer des manières de l'élaborer. Ces conceptions de l'évaluation font finalement confiance à la construction des savoirs par les élèves eux-mêmes, pour peu que les conditions favorables pour apprendre soient assurées par l'enseignement. L'apprentissage est l'élément dominant, mettant ainsi « l'élève au centre » de l'enseignement/apprentissage comme le déplorent bien des détracteurs actuels de l'école ; nous reviendrons plus loin sur cette critique.

Les années 1990 ou le temps des réalisations

Pour que des réalisations voient le jour dans l'école en Suisse romande, il a fallu que les responsables politiques manifestent leur intention de changement et prennent les rênes des réformes projetées. Pour asseoir leurs décisions en matière d'évaluation, ils commencent, en 1993, par mandater une commission romande, sous la responsabilité scientifique de l'IRDP en la personne de J. Weiss, en demandant « de proposer à l'intention des autorités cantonales des voies et moyens pour faire progresser les approches et pratiques de l'évaluation formative dans les divers degrés de la scolarité obligatoire, compte tenu des structures, objectifs et missions actuels desdits degrés » (Weiss, 2004)). Pour les membres de la commission – représentants des enseignant-e-s, chercheur-e-s, autorités scolaires cantonales – cette mission a constitué un défi intéressant. Il s'agissait désormais de faire des propositions concrètes dans un domaine précis, l'évaluation formative, encore peu exploré. Dans

cette perspective, la commission a voulu partir de réalisations concrètes et a mis au jour des pratiques existantes dans les cantons. De plus, en choisissant de ne pas se centrer uniquement sur l'évaluation formative, elle a envisagé l'évaluation sous ses différentes facettes pour en définir une articulation. Finalement, en 1996, la commission dépose son rapport, intitulé « Vers une conception cohérente de l'évaluation pour la scolarité obligatoire en Suisse romande et au Tessin. Une évaluation pour apprendre et pour choisir. Voies et moyens » (Weiss, Réd.). Neuf propositions sont développées, à l'intention des mandataires (les chefs des Départements de l'instruction publique de Suisse romande et du Tessin – la partie italophone de la Suisse) (voir annexe). Des exemples de pratiques observées sont également fournis, pour appuyer les propositions faites.

En précisant, là encore, les fonctions de l'évaluation, les auteurs du rapport ont clarifié le rôle des différents partenaires de l'école ; ainsi, le travail essentiel des enseignant-e-s est d'organiser l'enseignement/apprentissage dans leur classe ; dans ce sens, ils sont amenés à apprécier régulièrement la situation scolaire, à examiner les travaux de leurs élèves et à les informer des conséquences pour l'apprentissage. Dans cette perspective, il ressort que l'évaluation formative est première, se faisant au jour le jour, sous des formes le plus souvent souples et peu formalisées. Des évaluations sommatives sont également possibles, au terme d'une phase d'apprentissage par exemple. Elles sont dites provisoires, utiles pour réorienter les apprentissages. Un autre rôle est attribué aux enseignant-e-s, celui d'informer les parents de l'état du travail de leur(s) enfant(s). Le rapport octroie aux parents la responsabilité du choix du parcours de leur(s) enfant(s), en collaboration avec les enseignant-e-s. Il s'agit là d'une nouveauté qui occasionnera par la suite des débats animés, notamment chez bon nombre d'enseignant-e-s, désireux de conserver la responsabilité de l'orientation/sélection de leurs élèves. Les autorités scolaires, de leur côté, ont également un rôle à jouer par rapport à l'évaluation ; il est en effet recommandé qu'elles fournissent aux enseignant-e-s des repères leur permettant de situer leurs élèves, leur classe, par rapport aux autres. Grâce à ces repères, les enseignant-e-s ont la possibilité de réguler plus aisément leur enseignement.

En donnant à l'évaluation formative une place privilégiée, en limitant les certifications, en déchargeant les enseignant-e-s- de la tâche d'orientation/sélection, il est certain que les auteurs du rapport de 1996 espéraient que les structures scolaires de plusieurs cantons, encore largement et précocement sélectives à l'époque, se modifient au profit d'une orientation véritable et souple qui ne serait pas réalisée trop tôt dans la scolarité des élèves.

Enfin, les auteurs du rapport reconnaissent la nécessité de former les enseignant-e-s aux nouvelles modalités d'évaluation, de favoriser les projets d'établissement touchant à ce domaine, et à les accompagner par des personnes-ressources. La recherche doit également être mise à contribution, pour le suivi de nouveaux projets ou pour approfondir des questions encore en suspens.

Avant de prendre les décisions relatives à ces différentes recommandations, les autorités scolaires de Suisse romande ont soutenu l'organisation d'un grand colloque

axée principalement sur l'évaluation formative et sa mise en pratique dans les classes. Le but principal de ce colloque était en effet de proposer les moyens concrets de réalisation de programmes cantonaux d'évaluation. Les Actes de cette rencontre ont fourni des pistes de travail. Dès 1997, des réformes sont entreprises dans bon nombre de cantons de Suisse romande. Les documents issus des travaux de la commission romande et du colloque d'évaluation sont souvent à la source des nouvelles réalisations qui se dessinent dans les classes. La note scolaire, en particulier, est remise en question voire supprimée, au primaire tout spécialement. De nouvelles pratiques d'évaluation voient le jour, l'information aux parents se développe, l'autoévaluation également. Selon Vögeli-Mantovani (2000), là où les formes et les processus d'évaluation sont intégrés dans une conception générale prenant en considération, voire dépassant, la contradiction entre formation et sélection, et se traduisant en de multiples modalités (nouvelle structure d'orientation, élargissement des évaluations, bulletins d'apprentissage sans notes, dossiers d'élèves, etc. par exemple), la mise en place de nouvelles pratiques d'évaluation s'avère plus efficace que si les nouveautés sont introduites isolément, avec l'unique visée de changer le mode d'évaluation.

Un bilan mitigé et les prémisses d'une ère nouvelle de l'évaluation

Quelque dix ans plus tard, que pouvons-nous dire de ces réformes ? Ce temps des réformes a pour le moins engendré de nombreuses réactions, souvent passionnelles, sur les nouveaux modes d'évaluation, opposant adeptes et récalcitrants. La note scolaire, en particulier, a fait couler beaucoup d'encre, certains nostalgiques la réclamant à cor et à cris dans sa forme antérieure. C'est dans ce climat conflictuel – malgré plusieurs études cantonales réalisées dans le but d'une régulation du système d'évaluation, comme par exemple dans le canton de Berne (Merkelbach & Riesen, 2000) – qu'une nouvelle commission a été créée, en 2003, pour « …étudier et proposer, dans l'esprit de la coordination romande, des lignes directrices pour un système d'évaluation du travail des élèves dans la scolarité obligatoire… » (Article 1 du mandat rappelé dans Weiss, 2004). Alors que, dans les cantons, on fait face aux problèmes qui ont émergé à la suite de ces réformes, à des degrés divers, les autorités politiques expriment leur souci de trouver un consensus pour, notamment, affermir leur position face à la critique. Une fois encore, J. Weiss est mandaté pour la rédaction du rapport rendant compte des travaux de cette commission. A la lecture du mandat, apparaît immédiatement le fait que l'évaluation formative n'est plus prédominante, mais intégrée dans un système plus vaste dont les articulations sont précisément à redéfinir, à la fois en tenant compte de la diversité des situations cantonales et en trouvant un accord entre toutes et tous.

Un bilan en demi-teinte

Le rapport fait mention des critiques principales qui ont cours dans les milieux scolaires de Suisse romande et parle d'un bilan global en demi-teinte. Ainsi, deux cantons n'ont finalement pas entrepris de réforme, l'un – le Valais – à la suite d'un rejet en votation populaire, l'autre – Fribourg – à la suite des critiques et incertitudes résultant de la phase de consultation du projet envisagé. Dans ce canton, il a même été décidé un moratoire à l'extension de toute nouvelle forme d'évaluation au-delà de la 4e primaire.

A l'inverse, certains cantons ont rencontré peu de résistances, alors même qu'ils ont entrepris des réformes de structures de leur système scolaire (Berne, Jura et Neuchâtel). Mais il faut dire qu'à Berne, des régulations importantes ont été réalisées par rapport au projet de départ concernant les modalités d'évaluation ; dans le Jura, aucune réforme n'a été accomplie en ce qui concerne l'évaluation, et dans le canton de Neuchâtel, la réforme touchant à l'évaluation n'a été effective que dans l'école primaire.

Ce sont dans les cantons de Genève et de Vaud, que les critiques ont été les plus virulentes, surtout à propos de la note. A Genève, malgré une enquête montrant un taux élevé de satisfaction de la part des parents, par exemple pour les notes assorties de commentaires et les dossiers d'élève (Nidegger, 2003), une initiative populaire a été déposée, mettant en cause le système des notes et le découpage de la scolarité obligatoire en cycles d'apprentissage de quatre ans (deux ans dans les autres cantons). Une initiative a également été déposée dans le canton de Vaud, réclamant des notes dès le primaire, alors même que, là aussi, une enquête apportait des informations plus nuancées sur l'opinion des parents d'élèves (Abbet, 2002). De manière générale, ce sont les faiblesses des réformes entreprises qui ont été montées en épingle.

Quelles sont, plus précisément, les principales critiques adressées à ces réformes ?

- L'évaluation formative a été le plus souvent mal comprise ; par exemple, elle a été assimilée à tout ce qui n'est pas la note.
- Les procédures proposées ou imposées aux enseignant-e-s se sont avérées pour beaucoup très lourdes à mettre en place dans la classe ; ces derniers se sont plaints fréquemment du temps passé à observer, commenter, corriger le travail de leurs élèves.
- Pour de nombreux parents, les nouvelles modalités d'évaluation sont apparues peu compréhensibles ; en perdant la note, ils avaient l'impression de perdre un repère familier capable de les renseigner sur le travail de leur(s) enfant(s). Paradoxalement, l'absence de la note a été souvent considérée comme laissant place à davantage d'arbitrarité et d'iniquité dans l'évaluation ; en outre, elle a pu être également ressentie comme un manque d'information sur l'avenir scolaire de l'enfant que bien des parents voient en relation avec les résultats des autres élèves de la classe[1].

[1] Si la sélection en filières est réalisée sous forme de répartition des élèves selon un pourcentage calculé à l'avance par filière, on peut comprendre la demande des parents : pour entrer dans une « bonne » filière, il vaut mieux, pour l'enfant, se trouver dans les meilleurs de la classe.

- L'absence de référence à la moyenne de la classe au profit d'une référence aux objectifs et aux critères d'évaluation a dérouté plus d'un parent.
- La note est vue par beaucoup comme élément de motivation à travailler pour l'élève ; en la supprimant au profit de la motivation interne (le plaisir ou la volonté à apprendre), il y a leurre et les élèves manquent en fait d'encouragements externes.
- Une évaluation par objectifs est finalement exigeante, car elle amène à la réussite dans toutes les disciplines ; il n'est plus possible à l'élève de compenser une mauvaise note dans une branche par une bonne dans une autre. Les échecs risquent donc d'être plus nombreux. Pour les éviter, des systèmes, parfois bien compliqués, voire « tordus » ont été mis en place, dans le secondaire en particulier.
- Aussi bien des enseignant-e-s que des parents se sont plaints d'un manque d'information officielle, de lacunes dans la communication à propos de l'évaluation de la part des autorités scolaires et politiques. Ils ont pu alors se sentir seuls porteurs des réformes en cours et, pour les enseignant-e-s, peu soutenus par leur hiérarchie.
- A vouloir trop parfaitement et précisément évaluer le travail de l'élève, à le référencer aux objectifs à atteindre, aux critères énoncés, une dimension subjective, voire fantasmatique ou mythique, finalement présente dans l'évaluation traditionnelle, a eu tendance à disparaître, ce qu'ont déploré certains.

Il serait faux, toutefois, de ne souligner que les critiques qui ont émergé à la suite des réformes conduites en Suisse romande ; des points positifs méritent eux aussi d'être mentionnés. Par exemple, les réformes de l'évaluation entreprises au premier cycle, jusqu'en 2e année primaire, sont acquises partout, et même jusqu'à la fin de l'école primaire, pour autant que des décisions d'orientation ne soient pas à prendre.

Bien des enseignant-e-s ont apprécié les nouvelles modalités d'évaluation présentées et les ont faites leurs : l'évaluation formative, critériée, le jugement professionnel, l'appréciation des situations d'élèves avec les pairs, l'entretien avec les parents. Des projets d'établissement ont vu le jour et se sont déroulés avec succès.

Même si certains souhaitent le rétablissement des notes, une majorité de parents disent leur satisfaction et soutiennent les pratiques des enseignant-e-s.

Les années 1990 ont ouvert une période d'instabilité, non seulement dans les pratiques de l'évaluation à l'école, mais pour tout ce qui touche à l'école. Le rapport de la commission romande mandatée en 2003 (cf. le rapport de J. Weiss, 2004) atteste d'un phénomène qui ne cesse de croître depuis une vingtaine d'années (selon Hutmacher, 2004) : l'essor de l'évaluation formelle (méthodique et structurée), par exemple les épreuves de référence à l'intention des enseignant-e-s, les certifications, les épreuves à large échelle, internationales (PISA, TIMSS, etc.) permettant le pilotage du système scolaire.

Vers une évaluation formelle

Un tel essor ne naît pas *ex nihilo*, mais s'accompagne d'autres phénomènes sociaux qui renforcent les raisons de son émergence. Les mondes politique, économique, du travail ont développé, dans les pays anglo-saxons d'abord et surtout, des modes d'évaluation basés sur la réflexivité et le projet. Il s'agit avant tout de se doter d'instruments et de méthodes à même d'améliorer l'efficacité des activités propres à ces différents domaines. Ainsi s'est développée une branche d'activité évaluative spécialisée et professionnalisée, issue plus particulièrement de l'économie et recourant aux méthodes et aux résultats des sciences sociales.

Comment cela se manifeste-t-il dans le monde de l'école ?

Hutmacher souligne la transformation des politiques éducatives selon deux axes :

La bascule du pilotage par les inputs vers le pilotage centré sur les outputs d'une part, la bascule de la priorité donnée à l'enseigner vers la priorité donnée à l'apprendre d'autre part.

Jusqu'ici, la priorité a été donnée aux entrées du système : la formation des enseignant-e-s, l'élaboration de plans d'études, de moyens d'enseignement, la mise à disposition de salles avec ordinateurs dans les établissements, etc. Améliorer les conditions d'enseignement était vu comme susceptible d'améliorer les conditions de l'apprentissage des élèves. Depuis la deuxième moitié des années 1980, l'accent s'est déplacé sur la question des résultats de cet apprentissage, en rapport avec les buts poursuivis (efficacité de l'enseignement) et en rapport avec les ressources engagées (efficience). Les milieux politiques et économiques sont désormais intéressés de savoir quelles peuvent être les conséquences des acquisitions des élèves sur la vie sociale, culturelle, et économique. Les enquêtes internationales telles que PISA sont là pour apporter des réponses à cette question.

Parallèlement se développe l'idée forte que tout le monde peut – doit – apprendre, tout le long de la vie. Nous avons vu que la promotion de l'évaluation formative en classe allait dans ce sens, favorisant avant tout l'apprentissage, si possible dans le respect des rythmes propres aux élèves. Mettre l'élève au centre du processus d'enseignement/apprentissage relève de ce nouveau credo que souligne Hutmacher. Toutefois, ce n'est plus le processus d'apprentissage lui-même qui est jugé intéressant mais ses résultats en termes d'acquisitions et de compétences.

Que voit-on se passer ? La formation de l'enseignant-e se transforme, vers un accroissement des exigences et des compétences demandées en fin de cursus. Le métier se modifie, on parle de profession, impliquant de savoir réfléchir à ce qu'on fait, de collaborer avec ses collègues, de se renouveler au fil des demandes sociales, de s'adapter au monde qui ne cesse lui-même de changer. L'établissement scolaire prend davantage d'importance, dans une perspective décentralisatrice ; des projets s'y développent, une plus forte autonomie de fonctionnement lui est accordée ; il ressemble de plus en plus à une petite entreprise. Le système bureaucratique mis en place par l'Etat semble avoir fait son temps, largement critiqué dans certains milieux

qui lui reprochent son inertie, son manque d'efficacité (échec scolaire), et de freiner, voire d'empêcher la créativité et l'innovation.

Il n'est pas ici le lieu d'une analyse de ces phénomènes, mais plutôt de constater leur répercussion sur l'évaluation scolaire aujourd'hui.

L'école de Suisse romande dans le XXIème siècle : vers quelles évaluations ?

L'école en Suisse romande n'échappe pas à ces influences socio-économiques et politiques. A l'aide de trois exemples de recherche, nous montrerons que les responsables politiques en matière d'éducation sont actuellement essentiellement soucieux de mesurer et de contrôler les compétences que les élèves construisent à l'école afin d'accroître l'efficacité du système scolaire, mais aussi de justifier l'utilisation des deniers publics.

Ainsi, l'enquête PISA, commencée en 2000, a engendré de nombreuses réactions dans la presse helvétique et dans les milieux proches de l'école. En 2000, en ce qui concerne la littératie, domaine prioritaire d'évaluation, les résultats ont été considérés comme peu satisfaisants dans le cadre d'une école suisse coûteuse et, de ce fait, généralement vue comme performante. Les réactions négatives d'un large public, relayées notamment par la presse, ont fourni aux autorités scolaires et politiques l'occasion de prendre des mesures de correction du système, comme par exemple de rendre obligatoire l'entrée à l'école à quatre ans, dans un cycle d'apprentissage allant jusqu'à la fin de la 2P, essentiellement consacré à l'entrée dans l'écrit. Dans plusieurs cantons, des responsables de l'enseignement de la lecture ont été désignés dans le but de mettre sur pied des activités de lecture diversifiées et motivantes, et pour encourager cet enseignement à tout niveau de la scolarité obligatoire.

Dans cette même perspective, un projet suisse est né (HarmoS), d'harmonisation de l'école par la construction de standards de formation dans les disciplines de la langue première, des langues étrangères, des mathématiques et des sciences naturelles (CDIP, 2004). Des groupes d'experts dans ces disciplines sont actuellement au travail pour définir des modèles et des niveaux de compétences dans leur domaine et pour construire des épreuves permettant de vérifier la pertinence de ces modèles. Il s'agit de définir quelles sont les compétences attendues (standards de base[2]) à la fin de chacun des cycles d'apprentissage précisés dans le cadre de ce projet. Il est prévu, ensuite, de créer des instruments de mesure de l'atteinte de ces standards par les élèves dans le but d'un pilotage des systèmes scolaires du pays, c'est-à-dire d'augmenter la qualité de fonctionnement des écoles et de la formation qu'elles dispensent, d'appréhender de manière objective les résultats atteints par les élèves, de rendre compte de façon systématique des effets de la formation, d'assurer l'autonomie

[2] Selon les initiateurs d'HarmoS, les standards de formation peuvent être de deux ordres : des standards de performance fondés, par domaine disciplinaire, sur un cadre de référence incluant des niveaux de compétence ; des standards déterminant certains critères en matière de contenu ou de réalisation.

des établissements, d'augmenter l'égalité des chances et d'améliorer la mobilité à l'intérieur du système. Le programme est donc pour le moins ambitieux !.

En Suisse romande, les autorités scolaires (Conférence intercantonale de l'Instruction publique de Suisse romande et du Tessin – CIIP/SR-TI) ont, dans la foulée du rapport romand de 2003, exprimé leur volonté que soient créées des épreuves de référence pour la Suisse romande. Il est intéressant de remarquer qu'à l'origine, la proposition issue du rapport de J. Weiss, insistait sur la fonction diagnostique de telles épreuves, prévues pour fournir aux enseignant-e-s des repères utiles à leur enseignement. Or, à la lecture de la mesure prise par la CIIP/SR-TI, on constate une sorte de glissement vers des épreuves de référence de type bilan, intervenant en fin de cycle. Dans ce cas, ces épreuves deviendraient plus profitables au pilotage des systèmes scolaires qu'à celui de l'enseignement proprement dit.

Ces exemples, brièvement décrits, illustrent la volonté politique de « monitorage » du système scolaire. La situation helvétique en matière scolaire, particulièrement complexe puisque l'école est de fait de la responsabilité cantonale, est actuellement prise en main à différents niveaux : au niveau national avec le projet HarmoS et l'inscription dans la Constitution fédérale d'articles précisant les modalités d'une harmonisation pour la Suisse de certains paramètres du système éducatif ; au niveau régional, romand, avec les mesures prises par la CIIP/SR-TI à propos des épreuves de référence mais aussi des profils de compétence à élaborer en fin de scolarité obligatoire ; à niveau cantonal, lieu également décisionnel et des mises en œuvre proprement dites. A ces niveaux s'ajoute le niveau international, puisque la Suisse participe à la plupart des enquêtes proposées notamment par l'OCDE.

L'évaluation scolaire… un débat qui se poursuit

Comme le soulignait déjà Cardinet et comme le développait Hutmacher, l'évaluation scolaire dépasse les murs des écoles et est intimement liée au contexte socio-économique et politique du moment. On assiste ainsi, en Suisse en tout cas, à un double clivage : celui des évaluations interne et externe, la première étant entièrement du ressort de l'enseignement et du corps enseignant, la seconde prise en main par le politique via les experts scientifiques de centres de recherche ad hoc. L'autre clivage en découle et concerne la recherche en éducation ; ce qui touche à l'enseignement/apprentissage est désormais confié aux organes de recherche intégrés dans les lieux de formation des enseignant-e-s, ce qui touche à l'évaluation externe, au pilotage des systèmes scolaires, à l'élaboration des tests, est du ressort de centres de recherche au service des instances politiques, qu'elles soient nationales, régionales ou cantonales.

La question est alors de savoir comment s'établissent les liens entre ces deux types d'évaluation, en quoi les résultats des grands surveys peuvent influencer ensuite l'enseignement/apprentissage dans les classes et apporter des éléments utiles aux

enseignant-e-s – et donc aux élèves. Comment éviter le décalage entre les exigences du système et ce qui se fait réellement dans les écoles ?

Jusqu'ici, l'IRDP – mais il n'est pas le seul centre de recherche dans ce cas – a pu allier ces deux types d'évaluation en envisageant conjointement des démarches quantitatives (questionnaires, épreuves auprès des élèves…) et qualitatives (entretiens, audits, observations dans les classes…), tissant des liens entre les résultats obtenus afin de faire des propositions utiles à tous les partenaires de l'école (la dernière recherche allant tout à fait dans ce sens est Mathéval : Antonietti, 2005 ; Tièche Christinat & Delémont, 2005 ; Delémont, 2006). Désormais, comment faire si les instances de recherche sont appelées à traiter séparément ce qui touche à l'évaluation et ce qui touche à l'enseignement/apprentissage ? Ne court-on pas le risque d'une mesure déconnectée de ce qui se fait en classe – et dès lors peu utile pour la remédiation de l'enseignement ? Ou alors d'un enseignement inféodé aux instruments de mesure, tenté de travailler les épreuves pour conduire les élèves à de bons résultats ?

A propos de l'enquête PISA, Bain, adoptant une position critique, formule, dans le document qu'il a écrit sur le lien entre PISA et la didactique (2003), des questions de fond, de cette nature : « Une dissociation entre les approches du bilan (PISA) et des analyses à visée diagnostique et formative nous met dans une situation absurde de recourir à deux éclairages différents, voire divergents, sur une même réalité : les acquis (voire les difficultés) en lecture de nos élèves. Nous aboutissons à une certaine aporie qui nous semble peu problématisée par les milieux concernés ». L'analyse qu'il effectue met en avant les conceptions dissemblables qui se font jour entre les options théoriques à la source des tests PISA et certaines approches didactiques, en particulier celle qu'il connaît bien pour l'avoir construite avec plusieurs de ses collègues (Bronckart, 1985), axée sur la lecture vue comme activité langagière opérant sur des genres de texte dans des contextes définis. Quoi qu'il en soit, la pédagogie actuelle tend généralement à donner du sens aux activités demandées aux élèves. Or, dans PISA, textes et contextes apparaissent souvent bien secondaires au vu des questions posées qui, elles, jouent un rôle central sur les résultats. Bain oppose également les approches statistiques de type psychotechnique, basée essentiellement sur les différences interindividuelles (par pays dans le cas de PISA) et de type édumétrique visant la mesure des apprentissages par rapport à des objectifs pédagogiques et en référence à des résultats attendus. Cette seconde approche permet mieux de repérer la nature des problèmes rencontrés par les élèves et de remédier à leurs difficultés. A la suite d'une étude telle que PISA, quels remèdes apporter à des maux à la fois très grossièrement décrits et dont on ne sait pas dans quelle mesure l'école en est responsable, puisque ne renvoyant pas aux objectifs de l'enseignement/apprentissage ? En effet, les tests et questionnaires passés en 2000 n'ont apporté que peu d'éléments diagnostiques permettant le choix de remèdes adaptés et spécifiques.

Les questions soulevées à travers cet exemple restent tout à fait d'actualité et appellent encore des réponses. Alors même que les derniers travaux des commissions romandes tentaient de définir des formes d'articulation entre les différentes

évaluations scolaires, force est de constater que l'évolution récente en la matière tend au contraire à les « désarticuler », que ce soit au sein de la classe, pour ceux qui prônent un retour de la note bonne à tout faire par exemple, ou au sein des institutions, lorsque les évaluateurs sont déconnectés des didacticiens ou des formateurs d'enseignant-e-s.

Références bibliographiques

Abbet, J.-P. (2002). *Les parents et l'orientation : orientation scolaire au cycle de transition : enquête lors de la généralisation.* Lausanne : URSP.

Allal, L., Cardinet, J. & Perrenoud, Ph. (1979). *L'évaluation formative dans un enseignement différencié*. Berne : P. Lang.

Antonietti, J.-Ph. (Ed.). (2005). *Evaluation des compétences en mathématiques en fin de 4e année primaire*. Neuchâtel : IRDP.

Bain, D. (1987). *L'évaluation formative fait fausse route.* Exposé présenté à la rencontre internationale sur l'évaluation, Dijon, 17-19 septembre 1986. Genève : CRPP (Centre de Recherches Psychopédagogiques).

Bain, D. (2003). Pisa et la lecture : un point de vue de didacticien. *Revue suisse des sciences de l'éducation*, 25 (1) 59-78.

Bain, D. & Schneuwly, B. (1993). Pour une évaluation formative intégrée dans la pédagogie du français: de la nécessité et de l'utilité de modèles de référence. In L. Allal, D. Bain et Ph. Perrenoud (Ed.), *L'évaluation formative et didactique du française*, 51-79. Neuchâtel : Delachaux & Niestlé.

Broi, A.-M., Soussi, A. & Wirthner, M. (2001). Les résultats des élèves en compréhension écrite. In Ch. Nidegger (Ed.), *Compétences des jeunes romands : résultats de l'enquête PISA 2002 auprès des élèves de 9e année*, 97-122. Neuchâtel : Institut de recherche et de documentation pédagogique (IRDP).

Bronckart, J.-P., Bain, D., Schneuwly, B., Davaud, Cl. & Pasquier, A. (1985). *Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse.* Neuchâtel : Delachaux & Niestlé.

Cardinet, J. (1991). *L'histoire de l'évaluation scolaire, des origines à demain.* Neuchâtel : IRDP.

Cardinet, J. (1995*). Où mènent les recherches sur la notation scolaire ?* Neuchâtel : IRDP.

CDIP (Conférence suisse des directeurs cantonaux de l'instruction publique). (2004). *HARMOS. Finalités et conception du projet. Juin 2004*. Berne : CDIP.

Delémont, M. (2006). *L'épistémologie des enseignants : quel impact sur les procédures des élèves en mathématiques ?* Neuchâtel : IRDP.

Hutmacher, W. (2004). L'évaluation des enseignants, entre régulation bureaucratique et régulation professionnelle. In L. Paquay (Dir.), *L'évaluation des enseignants* : *tensions et enjeux*, 59-86. Paris : L'Harmattan, 59-86.

Merkelbach, Ch. & Riesen, W. (2000). *Ombres et lumières au pays de l'évaluation : rapport de recherche sur les procédures et les instruments d'évaluation.* Tramelan : Office de recherche pédagogique (ORP).

Nidegger, Ch. (Dir.). (2003). *A la fin de la 6e primaire : enquête auprès de la volée 2000 des élèves de 6P : acquis et compétences des élèves, représentations des parents et des enseignants*. Genève : SRED.

Salamin, J.-P. (Ed.). (1986). *Ecole primaire suisse : une école pour apprendre : 22 thèses pour le développement de l'école primaire : projet SIPRI*. Berne : CDIP.

Tièche Christinat, Ch. & Delémont, M. (2005). *Pratiques et discours : le nouvel enseignement des mathématiques 1P-4P sous la loupe*. Neuchâtel : IRDP.

Vögeli-Mantovani, U. (2000). *Pour une évaluation plus formative et moins sélective. Le développement de l'évaluation scolaire en Suisse : version abrégée du rapport de tendance du CSRE, no 3*. Neuchâtel : IRDP.

Weiss, J. (1996). *Vers une conception cohérente de l'évaluation pour la scolarité obligatoire en Suisse romande et au Tessin : une évaluation pour apprendre et pour choisir : voies et moyens : rapport no 2*. Neuchâtel : IRDP.

Weiss, J. (2004). *Lignes directrices pour un système d'évaluation du travail des élèves dans la scolarité obligatoire.* Rapport du GRETEL. Neuchâtel: Conférence intercantonale de l'Instruction publique de la Suisse romande et du Tessin.

Weiss, J. & Wirthner, M. (Eds). (1998). Evaluation 97: pour une évaluation plus formative: actes du colloque, Yverdon-les-Bains, 20 novembre 1997. Neuchâtel : IRDP.

Annexe – Liste des propositions tirée de : « Vers une conception cohérente de l'évaluation pour la scolarité obligatoire en Suisse romande et au Tessin » (J. Weiss, rédacteur principal, 1996)

PROPOSITIONS

1. Développer l'évaluation formative et d'autoévaluation

Les élèves devraient apprendre à autoréguler leurs apprentissages et leurs comportements scolaires en fonction d'appréciations régulières, faites avec le maître; il s'agit pour eux de découvrir, au travers de ces échanges, les attentes de l'école à leur égard et les méthodes pour les satisfaire.

2. Fournir une aide à l'évaluation et aux choix

Se donner des moments pour apprécier et analyser ensemble la situation; les enseignants devraient en effet consacrer périodiquement du temps pour l'analyse de la situation scolaire et pour l'examen des dossiers de travaux, avec l'élève pour la régulation de ses apprentissages, avec ses parents, une à deux fois l'an, pour une appréciation en commun des meilleurs itinéraires scolaires à retenir. Cette activité d'accompagnement et de conseil relève de la globalité de la tâche de l'enseignant.

03. Evaluer pour choisir

Les parents devraient être régulièrement invités à examiner, avec l'enseignant, la situation scolaire généra1e de leur enfant et à tirer, avec lui, les conséquences des compétences ou insuffisances analysées, particulièrement lors du choix de l'itinéraire de formation le plus adéquat.

04. Créer des repères pour les enseignants

Il appartient à l'institution scolaire de mettre régulièrement à la disposition des enseignants et des élèves des données de référence issues d'évaluations régiona1es dont les items et les critères de réussite sont définis par une équipe d'enseignants concernés.

Une offre de formation à l'utilisation de ces repères est à prévoir.

15. Limiter l'évaluation certificative à quelques moments-clés de la scolarité

Ne certifier les compétences des élèves par un diplôme ou par des crédits qu'aux termes de grandes périodes de formation, marquant l'achèvement d'une étape ou d'une unité de formation. Ne considérer les bilans intermédiaires que comme des

évaluations sommatives provisoires utiles à la réorientation des différentes séquences successives de l'apprentissage.

26. Tendre à la réduction de l'importance de l'évaluation sélective, voire à sa suppression

Réexaminer les lois et règlements régissant la promotion et la sélection des élèves dans le sens d'une limitation du rôle de l'autorité scolaire aux cas exceptionnels de décisions excessives de la famille en regard des résultats obtenus par l'élève, décisions jugées préjudiciables au développement de l'élève.

0MISE EN OEUVRE

17. Apprendre à s'évaluer pour évaluer autrui

Appliquer dans les institutions de formation initia1e des formes d'évaluation conformes aux propositions du présent rapport, et faire de l'évaluation un objet de réflexion, d'analyse et de formation.

8. Favoriser le développement de projets systémiques d'évaluation

Appliquer et analyser dans le cadre d'un établissement ou d'un canton la mise en œuvre de nouvelles formes d'évaluation.

9. Accompagner les projets d'innovation par des dispositifs de recherche et d' évaluation

Mettre à disposition des équipes "en projet d'innovation" les ressources scientifiques des centres de recherche.

COMMENT CONCILIER LA NOUVELLE CULTURE DE L'EVALUATION ET LES CONDITIONS DE CERTIFICATION ? EVOLUTION DES PRATIQUES ET MODÈLES DE CONNAISSANCE

Alex Blanchet, Unité de recherche pour le pilotage des systèmes pédagogiques (URSP)

Mots-clés : Evaluation formative, compétences minimales, promotion.

Résumé :

Un grand nombre d'idées généreuses ont motivé les réformes qui ont été mises en place dans les années 90. Parmi elles, la volonté de donner à tous les élèves une base minimale, indispensable à la vie en société. Cette idée n'est pas nouvelle et correspond à l'une des missions essentielles de l'école, mais elle se trouvait à nouveau portée par des courants aussi différents que la pédagogie de maîtrise, l'application à l'école de concepts de management et le souhait de certains milieux de retrouver les savoirs instrumentaux traditionnels.

Si le but est ainsi fixé, encore faut-il avoir les moyens de l'atteindre sans porter préjudice à une partie des élèves, ceux qui pourraient ne pas atteindre les seuils fixés.

L'exposé montrera comment la réforme mise en place dans le canton de Vaud a tenté d'articuler les nouvelles idées d'une évaluation centrée sur l'apprentissage avec les contraintes liées à l'information des parents et à la certification. On se centrera notamment sur l'évolution des règles appliquées pour établir le bilan certificatif de l'élève au terme d'une année ou d'un cycle : promotion ordinaire, promotion extraordinaire, tolérance de résultats inférieurs aux seuils attendus, moyennes, compensation entre branches, double compensation.

Au-delà de la cohérence de ces règles pour attester de la réussite de l'enseignement, on tentera de cerner le modèle de la connaissance qu'elles véhiculent, la motivation et les stratégies qu'elles génèrent chez les élèves. Ces réflexions s'appuieront sur les données recueillies par différentes recherches menées par l'URSP pour rendre compte des avantages et des difficultés de la réforme mise en place et pour en effectuer le bilan après 7 ans (questionnaire auprès de 943 enseignants).

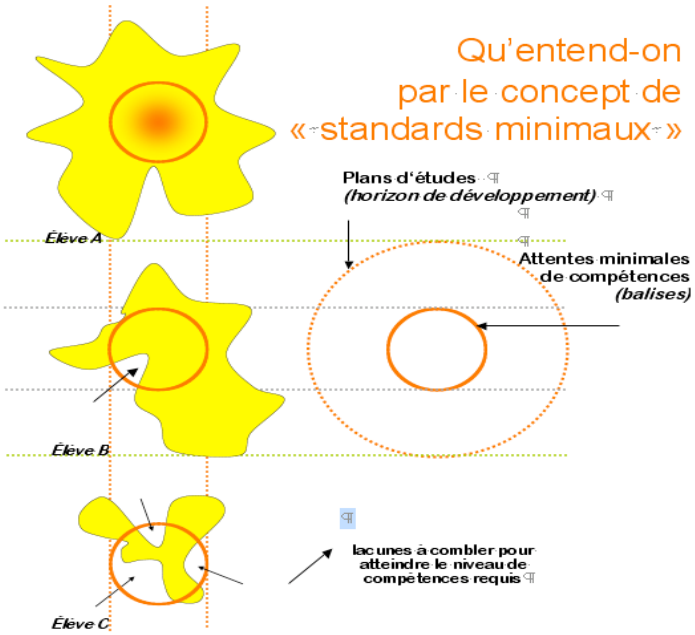
Introduction

La centration sur les objectifs minimaux

Les réformes des années 90 ont remis au premier plan l'une des missions essentielles de l'école : donner à tous les élèves les compétences minimales, indispensables à la vie en société. Cet objectif, qui est déjà celui de l'école républicaine, constitue toujours un défi, puisque nos sociétés peinent à éviter qu'une part non négligeable des élèves sortent de l'école obligatoire sans ce minimum indispensable. Les résultats de l'enquête PISA montrent qu'entre 15 et 20% environ des élèves suisses n'ont pas acquis ce bagage minimum en lecture et en mathématiques. A cet égard, la prise en charge des élèves de classes sociales défavorisées et des élèves d'origine étrangère, toujours plus nombreux, s'avère toujours difficile.

La centration sur les compétences minimales constitue actuellement une sorte de consensus par le fait qu'elle est portée par des courants très différents. La pédagogie de maîtrise, souvent associée à la pédagogie par objectifs, a proposé de découper les tâches d'apprentissage en objectifs partiels et a naturellement incité les pédagogues à choisir ceux qui étaient indispensables et ceux qui l'étaient moins. Dans une formulation récente (voir figure 1), en cours d'élaboration, on recourt à des distinctions entre compétences minimales et horizon de développement pour distinguer le cœur fondamental et les approfondissements souhaitables. Cette représentation très réductrice a pour but d'identifier des standards de connaissance minimaux utilisables pour évaluer un système scolaire.

Figure 1 – Schéma différenciant les cas d'élèves ayant atteint (A), partiellement atteint (B) ou n'ayant pas atteint (C) les objectifs minimaux (document CIIP)



Les objectifs minimaux rejoignent en cela la volonté actuelle d'appliquer à l'école des procédures d'évaluation qu'elle a longtemps refusés en invoquant la difficulté à préciser des compétences à la fois assez larges et évaluables de manière fiable.

La définition d'objectifs stratégiques ou opérationnels dont il faut ensuite évaluer l'atteinte coïncide aussi fort bien avec une vision de l'école comme une entreprise et l'utilisation de concepts de management. La définition d'objectifs communs permet d'élaborer des évaluations externes utiles au pilotage des systèmes de formation. Enfin, plus récemment, on a aussi vu surgir en Europe le souhait de certains milieux de redéfinir une fonction plus étroite de l'école, visant à redonner aux savoirs instrumentaux traditionnels la place qu'ils avaient par le passé. C'est aussi en termes d'objectifs minimaux que ces souhaits peuvent s'exprimer.

L'efficacité estimée des compétences attendues

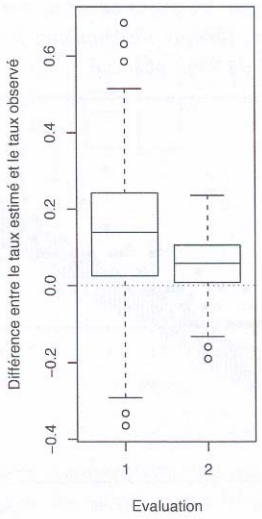
Si les objectifs d'apprentissage et les compétences qui leur sont associées sont bien acceptés, encore faut-il que ces termes recouvrent la même réalité. Lire, écrire et compter sont facilement évoqués quand il s'agit de définir les bases à acquérir à l'école obligatoire. Mais la tâche est plus difficile quand il s'agit de préciser les activités et les conditions de réalisation permettant d'estimer que ces compétences de base sont effectivement acquises. Définir une compétence minimale suppose des choix et des jugements de valeur, liés au sens que chaque expert donne à un contenu dans la construction des savoirs (de Landsheere, 1988 ; Roegiers, 2004). La Suisse a précisément entrepris d'importants travaux afin de définir des standards de performance (CDIP, 2004) susceptibles d'harmoniser les exigences de ses 26 systèmes scolaires différents. Cette démarche a nécessité la mise en place de collèges d'experts, chargés de négocier un minimum consensuel, testé ensuite auprès d'un échantillon de la population. Ce travail est certainement nécessaire à l'institution scolaire, mais il ne garantira pas, en revanche, que ce minimum soit fonctionnel pour tous les élèves, leur assurant un bon cheminement dans la construction de leurs connaissances. La raison des experts n'est pas forcément celle de l'élève.

Les adultes, qu'ils soient enseignants, formateurs ou chercheurs, semblent avoir une nette tendance à la surestimation des performances des élèves. Pour beaucoup, PISA a montré l'importance de la distance entre les performances attendues et les performances réelles des élèves. Si cette distance n'est pas plus apparente dans les évaluations habituelles, c'est que les enseignants tendent à sous-estimer la spécificité des contenus qu'ils enseignent et leur évanescence une fois l'évaluation passée. L'école sait fort bien trouver les moyens pour « oublier » une part des lacunes des élèves. Annoncer les travaux et interroger les élèves dans la suite directe d'un enseignement fait partie des pratiques habituelles, mais témoigne implicitement que le maître sait que les connaissances sont loin d'être permanentes et doivent être activées ou réactivées si l'on ne veut pas être déçu. L'écart existant entre ce que les élèves peuvent réaliser dans le prolongement direct d'un enseignement et ce qu'ils peuvent

effectivement mettre en œuvre sans cette mise en train est peut-être bien au cœur du processus d'apprentissage, comme le suggère la zone proximale de développement de Vygostky. Mais l'importance de l'écart questionne cependant la nature des compétences visées, ceci d'autant plus si elles sont qualifiées de minimales.

Dans une étude récente visant à évaluer les acquisitions des élèves après quatre années d'enseignement des mathématiques (Antonietti, 2005), il a été demandé à des experts d'évaluer les épreuves utilisées dans le test en fonction des compétences attendues, précisées dans le plan d'études. Il s'agissait d'avoir ainsi une référence basée sur l'estimation des acquis nécessaires pour réussir les tâches. Le groupe d'experts était composé de quatre enseignants, de quatre formateurs d'enseignants et de quatre chercheurs en éducation. Ces personnes ont d'abord résolu les problèmes, puis elles ont dû évaluer combien d'élèves (sur 20), possédant les compétences minimales attendues, seraient capables de répondre correctement à chacun des items (1re estimation). Après concertation entre collègues, les experts étaient informés de la proportion observée d'élèves ayant réussi chaque item. Enfin, chaque expert reprenait son estimation pour l'affiner en fonction de ces informations.

FIGURE 4.3 – *Différence entre le taux de réussite estimé par les juges et le taux de réussite observé avant et après information – évaluations 1 et 2 respectivement. Les distributions des différences sont représentées par des boîtes à moustaches.*



La figure 2 (Antonietti, 2005, p. 103) montre les taux de réussite estimés avant et après les informations en référence au taux de réussite observé (ramené à 0 dans le graphique). On voit que les avis sont très dispersés dans la première « boîte à moustaches » (la boîte représente les valeurs comprises entre le premier et le 3e quartile). Cela signifie qu'une analyse a priori ne permet pas, même quand on est expert, de définir avec sûreté un niveau de réussite pour une tâche. Après confrontation des estimations et prise de connaissance des taux de réussite, on voit que les secondes estimations sont beaucoup moins dispersées, sans pour autant se conformer aux valeurs observées. Il faut en effet souligner que cette estimation reste

nettement au-dessus des niveaux observés, puisqu'en appliquant la norme des experts, seuls 30% des élèves romands atteindraient les compétences minimales attendues. Même si l'étude laisse apparaître quelques résultats inquiétants sur les niveaux atteints en quatrième année, ils ne sont certainement pas aussi bas que l'estiment les experts. En mathématiques, la Suisse romande obtient, quelques années plus tard (9e année), des résultats honorables à l'étude PISA 2003 (Suisse : 527 points).

Cet écart important entre les attentes et les observations n'est certainement pas nouveau, mais il devient certainement plus apparent depuis que le recours à des évaluations externes s'est développé.

Objectifs minimaux et lutte contre l'échec scolaire

Le canton de Vaud est celui qui a le plus fort taux d'échec scolaire en Suisse. Le principe d'éducabilité a été au cœur de la réforme mise en œuvre dans ce canton, pour tenter d'inverser une forte tendance à utiliser le redoublement en réponse aux difficultés scolaires des élèves.

Les moyens qui ont été proposés pour mieux faire atteindre par tous les objectifs visés consistaient à offrir un meilleur environnement d'apprentissage. Il s'agit de différencier l'enseignement en fonction des individus et des difficultés rencontrées, de donner du sens à l'apprentissage grâce à la résolution de problèmes, de favoriser la construction des connaissances en tirant parti des erreurs en appliquant une évaluation formative, d'intégrer les savoirs par la confrontation à des situations complexes, de multiplier les approches grâce au travail de groupe souvent plus adapté que le cours du maître, etc. Tous ces moyens devaient, en assurant l'acquisition du minimum souhaité, diminuer de manière substantielle l'échec scolaire.

Des modifications des structures ont aussi été effectuées (cycles, filières, sections), mais nous n'en parlerons pas ici. En revanche, des modifications dans les pratiques d'évaluation ont eu pour but de remettre en question les références utilisées et d'obliger les enseignants à expliciter les objectifs visés par leur enseignement et à les communiquer lors des travaux de contrôle qui marquent le terme d'une séquence d'apprentissage. La mesure la plus importante dans ce domaine est certainement la suppression des moyennes dans chacune des branches, et la réduction de l'échelle d'évaluation à 5 positions (au niveau secondaire). Les maîtres étaient encouragés à rédiger des commentaires pour expliquer en quoi les objectifs avaient été atteints ou pas, avec aisance ou « largement atteints ».

Enquête auprès des enseignants

Les principes fondamentaux de cette réforme importante ont donné lieu à d'importants débats et à d'intenses efforts de formation. Un nouveau plan d'études a été élaboré. De nombreuses critiques ont été émises quant à la perspective idéaliste et idéologique du socio-constructivisme qui sous-tendait la réforme. Nous avons réalisé, sept ans après les débuts de cette réforme, une enquête visant à savoir quelles étaient les opinions des enseignants à propos de ses principes et leur accord avec la pratique.

(Blanchet, 2004). Un questionnaire a été envoyé à 1500 enseignants constituant un échantillon représentatif du corps enseignant. 63% ont donné leur avis sur les huit thèmes proposés. Nous ne présenterons ici que quelques résultats concernant les objectifs et leur évaluation.

La définition et la communication des objectifs d'apprentissage ont été bien accueillies par les maîtres. Plus de 85% des enseignants secondaires sont ainsi favorables à la définition d'objectifs d'enseignement/apprentissage pour chaque cycle, et 73% approuvent leur expression en termes de compétences attendues. Ils sont même 96% à préciser *aux élèves les objectifs à atteindre lors des travaux importants.* Le recours à des activités de remédiation est aussi massivement approuvé (80%), mais il se trouve pourtant une nette majorité (63%) pour estimer que *les remédiations ne permettent pas à davantage d'élèves d'atteindre les objectifs de fin de cycle ou de degré.* Sur plusieurs thèmes essentiels, on retrouve ce contraste entre des innovations pédagogiques apparemment bien intégrées et une estimation plutôt négative de leur efficacité. C'est notamment le cas pour le travail de groupe (63%) et l'autoévaluation (55.5%).

Concernant plus spécifiquement les pratiques d'évaluation, l'enquête réalisée a montré que les enseignants étaient majoritairement opposés à la suppression de la moyenne pour chaque branche (71%) et même défavorables à la suppression de la moyenne générale (moyenne des moyennes de disciplines, 57%). Ces avis tranchent avec l'adhésion déclarée aux nouvelles orientations pédagogiques centrées sur l'évaluation formative (80%), sur la résolution de problèmes (86%), la différenciation (79%), le travail de groupe (70%), etc. Cependant, pour ces maîtres, l'opposition entre ces orientations et le calcul de moyennes n'a pas lieu d'être (82%), et les notes présentent des avantages certains par rapport à la finesse de l'évaluation (86%), à la clarté des informations communiquées, à la motivation des élèves (84%) et à la protection de l'enseignant par rapport aux contestations (87%). Au cours des années suivantes, les notes et les moyennes, qui avaient été supprimées dans un premier temps, ont été réintroduites au niveau secondaire, sans moyenne entre les disciplines cependant.

La gestion des lacunes

Malgré un effort important pour différencier l'enseignement et aider les élèves dans leurs apprentissages, l'école doit constater que tous n'atteignent pas le seuil fixé dans certains domaines. On peut s'interroger sur les raisons de ces échecs : les compétences requises sont-elles mal conçues ? L'enseignement peu efficace ? Les remédiations inadéquates ? L'hétérogénéité des élèves sous-estimées ? ou bien encore les maîtres estiment-ils utile d'avoir un certain nombre d'échecs pour exercer une certaine pression favorable à l'assiduité des élèves (Perrenoud, 1988)? Probablement un peu de tout cela à la fois. Toujours est-il que l'école doit accepter des accrocs au principe d'éducabilité. Comme le dit Meirieu, on peut partager le principe d'éducabilité visant à donner à tout le minimum prévu, sans dénier que certains vont résister à tous les moyens mis en œuvre.

> L'éducabilité est d'abord le principe "logique" de toute activité éducative: si l'on ne postule pas que les êtres que l'on veut éduquer sont éducables, il vaut mieux changer de métier. C'est aussi un principe heuristique essentiel: seule la postulation de l'éducabilité de l'autre interdit à l'éducateur d'attribuer systématiquement ses échecs à des causes sur lesquelles il n'a pas pouvoir et d'engager la recherche obstinée de nouvelles médiations. (…)
>
> Mais le principe d'éducabilité peut, on le sent bien, être porteur de préoccupantes dérives: on sait ce qu'il advient quand on se donne pour objectif d'éduquer "quoi qu'il en coûte": la violence n'est pas loin, l'exclusion parfois, la "rééducation" de temps en temps. Le piège, en Education, c'est de confondre la formation d'une personne et la fabrication d'un objet, de ne pas supporter que l'autre nous échappe, se récuse... et d'abandonner le principe d'éducabilité quand l'autre ne nous paye pas des efforts que l'on a faits pour lui par sa reconnaissance, sa soumission ou sa réussite. (…) Admettre que le principe d'éducabilité soit constamment mis en échec sans, pour autant, y renoncer. Assumer la négativité de l'éducabilité, sans, pour autant, basculer dans le dépit et la suffisance, sans sombrer dans le fatalisme.

Comment gérer ces insuffisances et statuer sur leur importance pour la poursuite des études ? Que faire des échecs partiels ? Jusqu'à quand proposer des re-médiations et pour combien d'élèves ? Quand faut-il accepter l'échec de l'apprentissage et passer plus loin ? Ces questions ont souvent surgi quand les maîtres qui mettaient en œuvre les innovations pédagogiques étaient invités à exprimer les difficultés qu'ils rencontraient.

La première solution adoptée par l'école vaudoise a consisté à faire bénéficier l'élève d'une certaine marge d'erreur. Il a donc été admis qu'un élève pouvait poursuivre son cursus quand bien même il n'obtenait pas le minimum requis dans tous les domaines. Cette forme de promotion a été appelée *promotion extraordinaire sans condition* pour la différencier de la promotion ordinaire. Quand l'échec était plus important, mais ne compromettait pas nécessairement la poursuite des études, un travail supplémentaire devait être réalisé par l'élève au terme de l'année. Il ne pouvait continuer dans le degré suivant qu'à la condition que le travail soit jugé suffisant, d'où le statut de *promotion extraordinaire sous condition.*

L'expérience n'a pas été probante puisque après une année déjà, des recommandations incitaient les établissements à éviter d'utiliser les promotions conditionnelles en raison du travail supplémentaire que cela impliquait d'une part, et parce qu'il était psychologiquement et pratiquement difficile de replacer un élève dans un degré antérieur après un ou deux mois passés dans le degré (et dans la classe) supérieur(e). La re-médiation par travaux conditionnels est donc devenue très rare.

En ce qui concerne la *promotion extraordinaire sans condition*, on a rapidement constaté qu'elle concernait un très grand nombre d'élèves du niveau secondaire. Les insuffisances sectorielles touchaient 33.7% d'entre eux la première année et 43.5% la deuxième. La promotion extraordinaire a ainsi concerné autant d'élèves que la promotion ordinaire. L'extraordinaire était devenu régulier. Actuellement, le taux de promotion extraordinaire oscille entre 15 et 20% (Daeppen, 2004, 2005), mais la mesure n'est plus fiable puisque le terme *extraordinaire* a été progressivement évité pour la communication aux parents, et n'est actuellement plus mentionné dans les directives sur l'évaluation ou dans le Règlement.

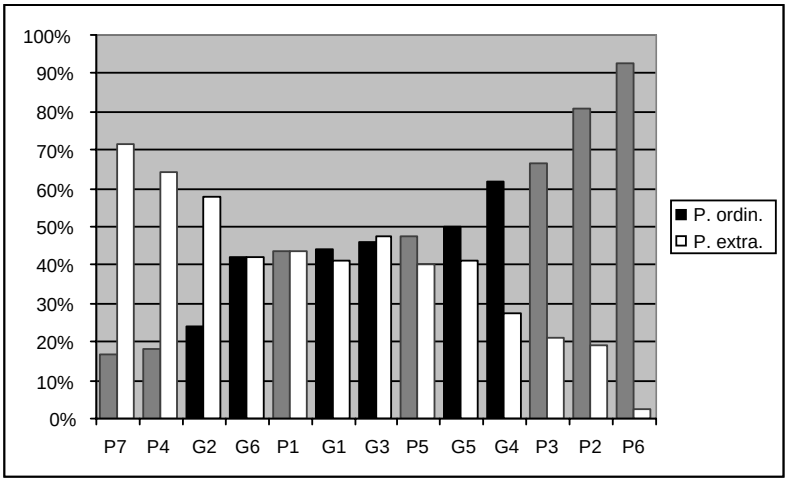


Figure 3 : Résultats de la deuxième année d'utilisation de la *promotion extraordinaire* dans les 13 établissements explorateurs de la réforme EVM (Blanchet, 2002).

La promotion extraordinaire n'a donc pas été utilisée dans le but pour lequel elle avait été conçue. Elle a en fait servi à pallier le décalage existant entre le principe selon lequel tous les enfants ou presque doivent atteindre les objectifs minimaux dans tous les domaines et la réalité scolaire actuelle qui fait que beaucoup d'élèves sont en dessous de ces niveaux dans l'une ou l'autre des disciplines.

Depuis la décision d'utiliser à nouveau les notes et les moyennes, le système scolaire a procédé à différents ajustements des points négatifs acceptables, qui ne conduisent pas à un redoublement. Ces *tolérances* ont d'abord été définies globalement, pour l'ensemble des disciplines, puis leur champ a été précisé afin d'éviter qu'une promotion puisse avoir lieu alors que tous les *points négatifs* concernent les branches les plus fortement dotées, le français et les mathématiques. Il reste donc qu'un élève peut obtenir son certificat de fin d'études avec des résultats négatifs dans ces deux branches.

Règlement : Art. 19.– Pour être promu aux 8 et 9 degrés, l'élève ne doit pas avoir plus de 3 points négatifs sur l'ensemble des disciplines, dont au maximum 2 points négatifs en français et en mathématiques.

Signification des conditions de promotion

Maintenant que les passions sont un peu retombées, on peut s'interroger sur ce que signifie ce système de promotion, dont le but déclaré est de se centrer sur les objectifs à atteindre pour chaque cycle. On a vu, grâce au nombre de promotions extraordinaires, que beaucoup d'élèves ne parviennent pas à atteindre les objectifs minimaux dans une ou plusieurs branches, et que des tolérances, ou points négatifs, ont dû être acceptés. En revanche, les points positifs ne sont pas pris en considération, ce qui équivaudrait à rétablir une forme de compensation entre les disciplines fortes et faibles et donc à une moyenne générale considérée comme incohérente. Les points

négatifs sont donc les seuls qui comptent pour la moitié des élèves environ. Cette centration sur les points négatifs n'est évidemment pas très motivante, et contraste avec la vision positive que l'école voudrait se donner avec l'évaluation formative, la différenciation des parcours d'apprentissage et l'analyse bénéfique des erreurs, par exemple.

En fonction de ces règles, on considère donc comme équivalents deux profils d'élèves assez distincts : l'un ayant deux résultats négatifs et des résultats justes satisfaisants par ailleurs, l'autre ayant aussi deux résultats négatifs, mais quelques très bons résultats dans les autres disciplines. Au moment d'établir un pronostic, les situations ne sont pourtant pas du tout semblables.

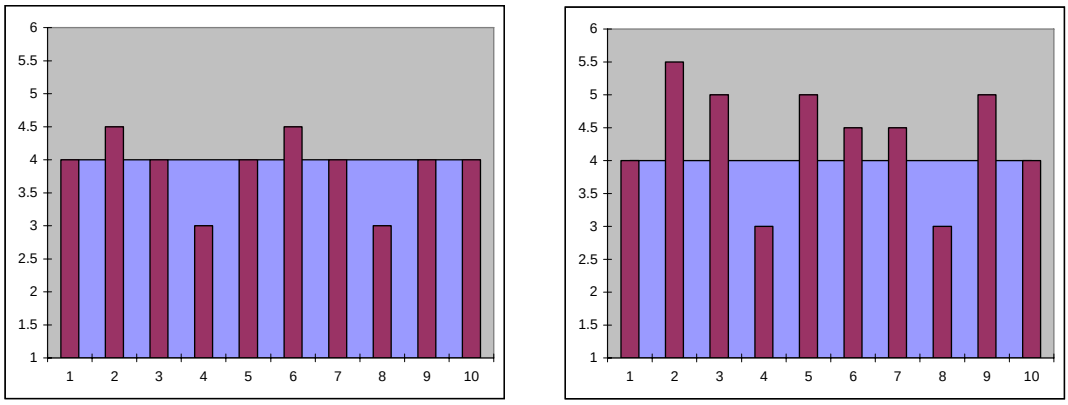


Figure 4 : représentation schématique de deux profils de résultats jugés équivalents selon le Règlement en vigueur.

Pour les élèves qui doivent gérer leur temps et leurs efforts, le développement des disciplines fortes n'a pas d'influence sur leur promotion. Ils peuvent donc les délaisser quelque peu pour se consacrer à atteindre, dans leurs branches faibles, le minimum requis. Dit d'une autre manière, leur investissement dans les branches fortes ne tient qu'à leur motivation interne, alors que le travail des branches faibles est lié à une motivation externe importante (la promotion). Dans ce contexte, le nivellement par le bas dont on tient les réformes comme responsables n'est pas tant le fait de l'école qui réduirait ses exigences, mais plutôt de l'élève qui ne se préoccupe à court terme que d'assurer ce minimum dans les disciplines où il n'est pas à l'aise. Cette crainte a été fréquemment confirmée par les maîtres que nous avons suivi au cours de la mise en œuvre de la réforme (Blanchet, 2002).

Au niveau gymnasial (lycée), la réglementation suisse a accepté la compensation habituelle entre les différentes disciplines, mais en introduisant une pondération qui la rend plus limitée. Pour chaque point négatif dans une branche, il faut en effet deux points positifs pour que la promotion puisse être acceptée (double compensation). L'introduction de cette nouvelle règle a momentanément modifié les taux d'échec des élèves, mais on constate que le système rétablit très rapidement ses taux habituels de redoublements (Stocker, 2006). Dans la mesure où ces taux sont élevés, la double

compensation n'a pas réussi à amener davantage d'élèves au minimum visé dans chaque domaine.

Dans tous les cas, on constate que l'effet sur le taux d'échec est faible. Après un temps d'adaptation, le système réajuste ses critères, et reproduit les taux d'échecs habituels. Il s'agit donc de savoir si la volonté d'assurer un niveau minimum de compétences pour tous, évidente pour donner un cadre à l'institution scolaire, est une logique qui s'applique aussi aux individus. Les différents ajustements auxquels le système a dû recourir montrent clairement qu'il ne parvient pas à endiguer la diversité des élèves. Il faut donc évaluer les avantages et les inconvénients respectifs de deux stratégies, correspondant à deux conceptions de la connaissance : assurer l'acquisition, même forcée, d'instruments de base dans tous les domaines, ou favoriser dans le plus de domaines possibles les capacités de manipulation de symboles et de raisonnement abstrait, en acceptant des secteurs au développement moins favorable.

Promotion et modèles de l'apprentissage ?

L'essentiel pour l'enseignement du niveau secondaire est de permettre à l'élève de développer la capacité de manipuler des symboles et de s'abstraire du concret. Cette capacité est essentielle dans tous les domaines complexes: activités scientifiques, économiques, et manuelles aussi, par la généralisation des machines à commande numérique. La construction de ces abstractions et leur maîtrise – que Piaget modélisait en termes de structure d'opérations formelles – peut s'effectuer dans de nombreux domaines. Que cette capacité soit exercée dans le cadre du latin, des mathématiques ou de la musique importe peu. Mais on constate que chaque individu possède un ou plusieurs domaines de prédilection pour accomplir ce passage vers l'abstrait. Il est donc essentiel que l'élève soit encouragé à développer les domaines où il excelle, par lesquels il pourra tirer le plus d'enseignement. On peut donc regretter que l'école ne tienne pas compte des résultats très positifs dans la promotion et n'encourage pas ainsi le développement des capacités les plus exigeantes.

A l'inverse, la centration sur les objectifs minimaux suppose que l'on connaisse comment se construisent les compétences des élèves, quels sont les éléments fondamentaux et à quelles performances ils correspondent. Mais on a vu combien les experts divergent dans cette estimation et ont tendance à surévaluer les compétences des élèves. En plus du fait qu'ils ne sont pas toujours d'accord entre eux sur le contenu des objectifs minimaux, ceux-ci sont décrits en termes abstraits qui laissent une large marge d'interprétation quant aux performances attendues. Les connaissances de l'élève sont généralement locales et fragiles, alors que l'enseignant souhaite qu'elles s'appliquent à toutes les situations et de manière permanente. Ce que l'élève apprend est d'abord très spécifique et dépendant du contexte. Ce n'est que petit à petit que se tissent des correspondances entre les situations d'apprentissage et que se dégage la généralité. Si les activités que nous proposons aux enfants sont fondamentales, ils ne peuvent en voir la généralité que progressivement, en fonction de cheminements très variés, probablement très personnels, sur lesquels nous n'aurons

jamais prise. Il est ainsi certainement très imprudent de croire que la progression linéaire des manuels correspond au développement des élèves.

La construction des connaissances vue comme une sorte de mécano, dont il faudrait apprendre à maîtriser chaque pièce pour que se constitue la machine d'ensemble, est donc un modèle superficiel. C'est ce que suggère le premier schéma présenté plus haut (figure 1). Plus que des pièces accumulées, les savoirs forment un réseau dont les différents nœuds se renforcent ou se contredisent, s'oublient et s'apprennent à nouveau en fonction des situations. Pour chaque savoir, l'élève doit progressivement juger du domaine d'application, du niveau d'abstraction et du degré de généralité. Tous les maîtres savent combien les enfants oublient, combien de fois il faut revenir sur une même notion, dans des contextes différents, pour que se constitue un véritable outil.

Pour favoriser cette construction, il est important de valoriser les domaines dans lesquels l'élève a une certaine aisance. Ne pas lui accorder le bénéfice de ces capacités dans le contexte de la promotion peut être un signal démotivant. Par ailleurs, ne se centrer que sur les objectifs minimaux peut générer beaucoup de frustration chez ceux qui ne parviennent pas à les atteindre, sans que l'on puisse être sûr que ces objectifs constituent des points de passage obligés dans la construction complexes des compétences visées. Dans ce contexte, le choix des conditions de certification est important et témoigne d'un modèle sous-jacent de l'apprentissage et de l'enseignement.

Références

Antonietti, P. (2005). Evaluation des compétences en mathématiques en fin de 4e année primaire. Résultats de la seconde phase de l'enquête Mathéval. Neuchâtel : IRDP.

Blanchet, A. (2000). Promotions et réorientations à l'issue de la 7e année exploratoire. Lausanne: URSP 103.

Blanchet, A. (2002). Bilan provisoire de l'exploration des degrés 7 à 9. Deux points :« ouvrez les guillemets 18, 14-21.

Blanchet, A. (2002). Promotions et réorientations à l'issue de la 8e année exploratoire. Lausanne: URSP 107.

Blanchet, A. (2004). Enseignant-e-s et responsables vaudois face aux changements de leur système scolaire. Contribution au Bilan final de la mise en oeuvre d'EVM. Lausanne : URSP.

Blanchet, A. (2005). Supprimer les notes : regard rétrospectif sur une réforme contestée. Actes du 18e Colloque de l'ADMEE. Reims : IUFM de Champagne, et publication électronique sous www0.dfj.vd.ch/ursp.

Conférence suisse des directeurs cantonaux de l'instruction publique (CDIP). (2004). HarmoS, finalités et conception du projet.

Daeppen, K. (2004). Promotions et réorientations au secondaire I : résultats de l'année 2002-2003. Lausanne : URSP 116.

Daeppen, K. (2005). Promotions et réorientations au secondaire I : résultats de l'année 2003-2004. Lausanne : URSP 122.

De Landsheere, V. (1988). Faire réussir, faire échouer. La compétence minimale et son évaluation. Paris : PUF.

Meirieu, P. Educabilité. www.meirieu.com.

Perrenoud, P. (1988) Évaluation formative : cinquième roue du char ou cheval de Troie ? Journal de l'Association pour le développement de la mesure et de l'évaluation en éducation (ADMÉE-CANADA), vol. 5, n°4, 21-28.

Roegiers, X. (2004). L'école et l'évaluation. Des situations pour évaluer les compétences des élèves. Bruxelles : De Boek..

Stocker, E. (2006). Regards sur le parcours gymnasial des jeunes vaudois. Lausanne : URSP 123.

D'UN DISPOSITIF D'ÉVALUATIONS EXTERNES CENTRÉ SUR L'ENSEIGNANT Á UN DISPOSITIF CENTRÉ SUR L'ÉTABLISSEMENT ET LE SYSTÈME
LE CAS DE LA COMMUNAUTÉ FRANÇAISE DE BELGIQUE

Dominique Lafontaine,
Département Education et Formation, Université de Liège

Mots-clés : évaluation externe, pilotage, acquis des élèves

Résumé :
Le présent article analyse l'évolution des dispositifs d'évaluation externe mis en place en Communauté française de 1994 à 2006. En 1994-95, un premier dispositif d'évaluation externe a été créé. Ce dispositif a été consolidé et étendu par le Décret instaurant la Commission de pilotage en mars 2002.

De 1994 à 2006, les évaluations externes ont eu une visée essentiellement diagnostique, avec comme destinataire principal l'enseignant dans sa classe. Durant cette période, la visée « informative » qu'auraient pu jouer les évaluations externes (à l'intention des décideurs) est apparue en Communauté française de Belgique comme relativement secondaire, où jusqu'alors aucune politique d'évaluation claire et structurée n'était en effet impulsée par les décideurs.

Avec le nouveau Décret voté fin mai 2006 par le Parlement de la Communauté française, il semble que l'on s'oriente désormais vers un dispositif davantage centré sur l'établissement, avec des visées de pilotage (fonctions informative et régulatrice) renforcées au niveau local et global. Les objectifs du futur dispositif sont en effet « d'informer les autorités et l'ensemble des acteurs sur l'évolution des acquis de cohortes d'élèves à différents moments du cursus scolaire et de permettre à chaque équipe pédagogique d'apprécier l'efficacité de son action ». Le Décret fait ici l'objet d'un examen critique afin de déterminer si les modalités d'organisation prévues pour les futures épreuves externes permettent de rencontrer les objectifs assignés à ce dispositif et l'objectif plus général d'amélioration de l'efficacité et de l'équité du système éducatif de la Communauté française, dont l'urgence est régulièrement rappelée par les campagnes PISA.

Contexte

La Communauté française de Belgique possède depuis toujours un système éducatif fortement décentralisé. On distingue trois réseaux d'enseignement : le réseau de la Communauté française de Belgique, organisé par le Ministère de l'Education; le

réseau subventionné officiel, organisé par les provinces et les communes, et le réseau subventionné libre, organisé par des instances privées, principalement catholiques.

La définition des programmes d'enseignement et des méthodes pédagogiques est de la responsabilité des réseaux. Chaque réseau est libre d'organiser des évaluations externes dans les écoles de son ressort. De nombreuses initiatives se sont ainsi multipliées au fil du temps. En revanche, dans un tel contexte, on saisit combien a pu apparaître délicate, sur le plan institutionnel, la mise en place d'un dispositif d'évaluation externe qui concerne l'ensemble des élèves de la Communauté française de Belgique tous réseaux confondus.

Première étape : création d'un dispositif léger d'évaluation externe

Un premier dispositif d'évaluation externe est créé en 1994-95. Cette création est étroitement liée à la mise en place de « socles de compétences », référentiel unique concernant tous les élèves et transcendant les programmes singuliers des réseaux. Il devient dès lors pleinement légitime, et plus aisé, sur le plan pratique, de mettre en place un dispositif destiné à vérifier dans quelle mesure les élèves atteignent les niveaux de compétences recommandés au moment attendu.

Modalités de fonctionnement

Une cellule est chargée de l'organisation des évaluations externes au sein de l'Administration générale de l'Enseignement et de la Recherche. Pour chaque opération d'évaluation externe, un groupe de travail est mis sur place. Dans ce groupe, on trouve des responsables de l'administration, des inspecteurs, des représentants des réseaux (qui sont le plus souvent des enseignants) et des chercheurs[1] recrutés sur des conventions extérieures.

Objectifs : le sens de la mesure

Dans la tête de son initiateur, ce dispositif[2] doit répondre à un double souci :

a) sortir les enseignants de leur isolement, en leur fournissant un référentiel (les résultats d'un échantillon représentatif) auquel comparer les performances de leur classe. L'outil, à visée diagnostique, est passé en début d'année;

[1] Le fait que les chercheurs soient recrutés sur contrat extérieur ne signifie pas qu'ils soient chaque année différents. Classiquement, les évaluations dans le domaine de la lecture et de l'éveil scientifique ont été confiées au Service de pédagogie expérimentale de l'Université de Liège et les évaluations dans le domaine des mathématiques à l'Université catholique de Louvain.

[2] Pour plus d'informations sur ce dispositif d'évaluation externe, on consultera Lafontaine (1999, 2001) et le site http://www.enseignement.be/prof/dossiers/eval/index.asp

b) réduire les disparités de rendement entre écoles, dont l'importance a été soulignée par les recherches internationales (IEA, Ocdé) et par le *Rapport sur l'état du système éducatif* des experts de l'Ocdé (Ocdé, 1991).

Dans les faits, le dispositif est aussi utilisé comme un « levier didactique », i.e. pour accompagner la mise en place d'une approche par compétences (test « modèle » montrant aux enseignants comment « bien évaluer », en se centrant sur les compétences essentielles, accompagnement de l'évaluation par un document de pistes didactiques exploité sur le terrain par les inspecteurs). Ainsi, si l'évaluation se veut « au service des enseignants », elle est également au service des curriculums et mise sur l'effet en retour de l'évaluation sur les pratiques (Anderson & Wall, 1996).

Les évaluations externes se sont ainsi vu assigner comme destinataire principal l'enseignant dans sa classe (en misant implicitement sur leurs capacités d'autorégulation au départ d'informations écrites). La visée « informative » que pourraient également jouer les évaluations externes (à l'intention des décideurs) apparaît en Communauté française de Belgique comme relativement secondaire dans ce premier dispositif.

Le traitement des données de l'évaluation externe pour en extraire des informations pertinentes sur l'état du système éducatif (indicateurs) ne fait pas partie des objectifs prioritaires de ce premier dispositif.

Résultats fournis aux enseignants

Chaque enseignant administre, corrige et conserve les carnets de sa propre classe. Les enseignants de l'échantillon renvoient les grilles et les questionnaires aux universités qui traitent les données. Des pourcentages moyens de réussite par item et pour l'ensemble de l'épreuve sont calculés.

Lorsque l'enseignant reçoit ces scores moyens, il peut comparer la moyenne obtenue dans sa classe avec celle obtenue par l'échantillon représentatif. Cette information permet à celui-ci de relativiser les résultats de sa classe. Des recherches (Grisay, 1984a et 1984b) ont en effet montré qu'en raison de leur isolement, les enseignants ont tendance à adapter leur enseignement et leur degré d'exigence en fonction du niveau de leur classe.

Au-delà du pourcentage moyen de réussite, des pourcentages de réussite de l'épreuve en fonction des caractéristiques de la composition des classes (pourcentage d'élèves en retard, d'allophones, d'élèves issus de milieu privilégié ou d'élèves provenant de milieux très défavorisés...) sont calculés. L'enseignant peut ainsi comparer les résultats de sa classe à ceux de classes présentant, dans leur composition socio-pédagogique, des caractéristiques approchantes.

Publicité des résultats

Le degré de publicité des résultats est strictement limité. Au niveau central, les résultats sont traités par les universités dans le respect des règles de confidentialité. Aucun palmarès, aucun classement n'est publié ni transmis aux décideurs. L'absence de garantie quant aux conditions de passation de ces tests (c'est l'enseignant qui les administre à sa propre classe) rendrait en effet toute conclusion relative à un établissement ou une classe particulière peu fiable. A un niveau local, les enseignants conservent les résultats de leur classe et rien n'empêche des autorités telles que les directeurs ou les inspecteurs de récolter les résultats des classes relevant de leur pouvoir d'intervention, en vue d'actions de micro-pilotage.

En résumé

Le dispositif d'évaluation externe mis en place en Communauté française de Belgique entre 1994 et 2002 présente plusieurs caractéristiques qui sont le reflet des singularités structurelles et organisationnelles du système éducatif :

- il n'est pas explicitement, et encore moins légalement, investi d'une mission de pilotage. Sa principale cible est l'enseignant dans sa classe, et non les décideurs en matière d'éducation;
- il est faiblement « institutionnalisé », dès lors qu'il est mis en oeuvre par un groupe de travail rassemblant, pour un temps limité, des partenaires issus de différents milieux (administration, inspection, réseaux, universités).

Etape intermédiaire : consolidation du dispositif existant en 2002

Le dispositif qui a été présenté constitue une première étape dans la mise en place des évaluations externes. En mars 2002, un texte législatif relatif au pilotage du système éducatif est adopté, qui englobe les évaluations externes.

Ce dispositif instaure la commission de pilotage, responsable des évaluations externes. A l'article 3, le Décret précise :

« La Commission (de pilotage) procède aux évaluations externes. Ces évaluations externes sont organisées au moins tous les cinq ans pour chaque cycle de l'enseignement obligatoire (…), ainsi que pour le 2e et le 3e degrés de l'enseignement secondaire. Le seul objectif de ces formes d'évaluation externe doit être l'amélioration de la qualité de l'enseignement en Communauté française. Les résultats ne peuvent être divulgués à d'autres fins que le bon accomplissement des missions reprises à l'article 3. Il est prohibé d'en faire état à des fins de publicité ou de concurrence entre établissements. »

Ce texte constitue une étape importante sur le plan symbolique, puisqu'il donne un cadre légal aux évaluations externes, cadre qui n'existait pas au préalable. Ce texte, comme on peut le constater, donne une définition aussi vaste que vague de ce que doivent être les futures évaluations externes. Seule une référence générale à

l'amélioration de la qualité est esquissée. Or, la « qualité », pour être consensuelle, n'en recouvre pas moins de multiples aspects. De quelle « qualité » peut-il s'agir ? En comparaison, la définition de ce que ne doivent pas être les évaluations externes est plus développée. La place importante occupée par cette définition « en creux » témoigne de toutes les craintes qui entourent les évaluations externes depuis leur création dans un paysage éducatif dont la concurrence entre établissements et entre réseaux constitue une composante majeure et une citadelle que les évaluations externes ne peuvent clairement venir menacer.

Dans les faits, avec le recul, il apparaît que le Décret instaurant la Commission de pilotage n'a pratiquement pas affecté les objectifs et les modalités du dispositif déjà existant. C'est la continuité qui a clairement prévalu. Tout au plus la création d'une Commission de pilotage a-t-elle instauré un lieu où doit s'effectuer formellement un retour d'information sur les résultats des évaluations externes vers les décideurs et les acteurs en matière d'éducation représentés dans cette Commission.

Des interrogations et un débat significatif

En 2003, Jean-Marc Nollet, Ministre de l'Enfance en charge de l'enseignement fondamental, sollicite l'avis de la Commission de pilotage sur deux questions relatives aux évaluations externes :

1. ***Serait-il opportun d'organiser en fin d'enseignement primaire une épreuve commune à tous les réseaux et établissements en vue de l'octroi du Certificat d'études de base sanctionnant la réussite des études primaires*** ?[3] Cette question ne concerne en rien le dispositif d'évaluations externes dont il a été question ci-avant, mais elle est d'une importance suffisante dans un système éducatif où il n'existe jusqu'alors aucune épreuve externe commune certificative pour ne pas la laisser sous silence. Le fait que cette question soit mise au débat en Commission de pilotage est le signe qu'un tabou majeur est en passe d'être levé. Chose surprenante, on n'assiste à aucune levée de bouclier notamment de la part des réseaux, et la Commission de pilotage prend une position nuancée (beaucoup de précautions sont prévues), mais unanimement en faveur d'une telle épreuve certificative commune. Ceci aboutira au Décret de juillet 2006 dont on reparlera ci-après.
2. ***Faut-il se mettre en capacité d'évaluer la « performance relative » des établissements ?*** Au travers de cette question, le Ministre demande à la Commission de pilotage de voir s'il serait opportun de repenser le dispositif d'évaluations externes afin que celui-ci soit en mesure d'apprécier la « valeur ajoutée » par le travail pédagogique et d'évaluer l'efficacité de l'école. Pour nourrir le débat autour de cette deuxième question, la Commission de pilotage a sollicité l'avis de d'experts en Sciences de l'Education, un duo, composé de

[3] Pour rappel, ce certificat, jusqu'alors, est attribué sur la base d'épreuves internes qui peuvent être du seul ressort de l'établissement ou d'épreuves externes ou semi-externes propres à un réseau, un pouvoir organisateur, une circonscription d'inspection, et qui n'ont aucun caractère obligatoire.

Vincent Dupriez-Christian Maroy et Dominique Lafontaine, qui ont remis chacun une note séparée à ce propos. Les deux notes présentent beaucoup de similitudes dans l'explication de ce que sont les indicateurs de valeur ajoutée, leur apport et leurs limites, mais diffèrent sensiblement en termes de recommandations. Alors que la note de Dupriez-Maroy penche favorablement pour le calcul de tels indicateurs, la note de Lafontaine (Lafontaine, 2005) insiste sur la nécessité d'une réflexion préalable et d'un débat en profondeur sur les objectifs d'un tel choix et le dispositif à mettre en œuvre en aval si une diffusion d'indicateurs en termes de valeur ajoutée se faisait vers les établissements. A l'issue du débat en Commission de pilotage, le Président a transmis les deux notes au Ministre et estimé « qu'il était trop tôt pour elle (la Commission) de rendre un avis tranché ». Il était prévu de reprendre le débat ultérieurement (en 2004), mais la Commission de pilotage n'est pas revenue depuis sur ce dossier.

L'émergence de ce double questionnement à la Commission de pilotage est à replacer dans le contexte des évaluations internationales – les résultats de PISA 2000 sont encore frais dans les mémoires, et de toute la réflexion qui s'en est suivi quant à l'efficacité et surtout à l'équité du système éducatif en Communauté française de Belgique (Lafontaine *et al*., 2003 ; Lafontaine & Blondin, 2004). Au travers de la mise en œuvre de telles mesures, le Ministre entend « lutter contre la dualisation du système », « assurer davantage d'équité et, partant, d'efficacité », « voir s'amplifier et se développer davantage encore les pratiques d'évaluation externe »[4]. On peut y reconnaître aussi une des marques de la montée en puissance de « l'Etat évaluateur » analysée par les sociologues de l'éducation (Dupriez, 2004).

Deuxième étape : d'un dispositif d'évaluation externe centré sur l'enseignant à un dispositif centré sur l'établissement et le système ?

Le 30 mai 2006, le Parlement de la Communauté française a adopté un Décret portant à la fois sur l'octroi du certificat d'études de base au terme de l'enseignement primaire sur la base d'une épreuve certificative commune (voir supra) et sur l'évaluation externe non certificative. Nous ne reviendrons pas ici sur l'épreuve certificative, si ce n'est pour souligner à nouveau l'importance de cette réforme, qui introduit une épreuve certificative commune à l'ensemble des réseaux et des établissements. Il s'agit là d'une première historique.

Nouvelles modalités du dispositif d'évaluations externes non certificatives

Le Décret prévoit un renforcement du dispositif existant, en instaurant des cycles triennaux d'évaluation dans différentes disciplines. La première année du cycle, il est prévu d'évaluer simultanément les élèves de 2e, de 5e primaire et de 2e secondaire en

[4] Ces citations sont extraites du canevas de l'intervention du Ministre de l'Enfance en Commission de pilotage en septembre 2003, intitulé « Pour une régulation accrue par le développement d'évaluations externes ».

lecture/production d'écrits, la deuxième année du cycle, les élèves des mêmes années d'études seront évalués en mathématiques et la troisième année du cycle, en éveil scientifique et technologique. A plus long terme, il est prévu de développer des évaluations en langues et d'autres évaluations pour les années supérieures du secondaire.

Au travers de ces cycles triennaux, « il sera veillé à apprécier l'évolution de la maîtrise des mêmes compétences aux divers moments de la scolarité visé ». En effet, les élèves évalués en 2e en lecture devraient se retrouver trois ans plus tard en 5e lorsqu'une nouvelle évaluation en lecture aura lieu. Le dispositif s'inscrit donc dans la perspective d'un suivi de cohorte, à distinguer d'un véritable suivi longitudinal, qui supposerait une identification et un suivi individuels des élèves – ceci n'est pas à l'ordre du jour.

Un seul changement organisationnel notable est à souligner : les évaluations n'auront plus lieu en début d'année scolaire, mais en cours d'année (entre janvier et mai). Ce changement à lui seul signe la fin du caractère diagnostique et formatif qui constituait le principal caractère des épreuves externes jusqu'alors.

Ceci étant dit, les épreuves continuent à être conçues par un groupe de travail composé de membres de l'administration, de chercheurs, d'inspecteurs et d'enseignants issus des différents réseaux. Les épreuves sont toujours administrées par les enseignants dans leurs classes et seront dorénavant corrigées par les enseignants d'autres écoles ou d'autres classes.

Objectifs

« Les évaluations externes non certificatives informent sur le niveau d'avancement des élèves, en poursuivant au moins un[5] des objectifs suivants :

1. Permettre à chaque équipe pédagogique d'apprécier l'efficacité de son action en établissant l'état des acquis de ses élèves par rapport aux compétences attendues et en situant les résultats de ses élèves par rapport aux résultats globaux des élèves de la Communauté française.
2. Permettre d'apprécier également l'efficacité de cette action au niveau de la zone[6].
3. Informer les autorités et l'ensemble des acteurs sur les acquis des élèves d'une année d'études ou d'un âge donné fréquentant l'ensemble des établissements d'enseignement organisé ou subventionné par la Communauté française.
4. Informer les autorités et l'ensemble des acteurs sur l'évolution des acquis de cohortes d'élèves à différents moments du cursus scolaire. »

[5] C'est nous qui soulignons cette formulation pour le moins curieuse.

[6] La zone est un découpage administrativo-géographique qui ne sert que pour l'enseignement ; la Communauté française est ainsi découpée en 10 zones. Les conseils de zones fonctionnent par réseau pour l'enseignement fondamental, alors qu'ils fonctionnent par caractère (« confessionnel – non confessionnel ») au niveau secondaire.

Comme le précédent Décret, le présent Décret rappelle avec insistance ce à quoi ne doivent pas servir les évaluations externes. Les résultats doivent rester anonymes sauf pour l'établissement concerné, l'Inspection pédagogique et les services d'animation pédagogique propres aux réseaux. Il n'est en aucune manière question de rendre publics les résultats, notamment pour les parents ; le mode de régulation « à l'anglaise » avec pression des usagers « avertis » sur les établissements est explicitement exclu par le Décret.

Le Décret insiste en outre, en matière d'exploitation des résultats, sur la mission qui incombe d'une part aux chefs d'établissement, de « procéder à un travail d'analyse et d'exploitation des résultats », et d'autre part, aux services d'inspection et d'animation pédagogique, « d'intégrer ces résultats à leurs propres investigations » et d'apporter « leur appui aux établissements scolaires dans l'analyse et l'exploitation des résultats ».

Analyse critique

Si l'on compare le dispositif instauré par ce récent Décret au dispositif antérieur, plusieurs constats se dégagent :

1. On passe d'un dispositif à visée diagnostique ou formative à un dispositif sommatif – même si le terme est absent du Décret –, qui, curieusement, se situe en cours et non en fin d'année.
2. On passe d'un dispositif qui avait pour cible principale l'enseignant dans sa classe à un dispositif davantage centré sur l'établissement ou l'équipe pédagogique d'une part, les autorités et l'ensemble des acteurs d'autre part.
3. Même si le Décret est peu explicite sur ce point, on peut penser que la forme de pilotage visée est désormais davantage orientée sur le pilotage de l'ensemble du système et des établissements que sur des formes autorégulation par l'enseignant portant essentiellement sur des aspects pédagogiques ou didactiques.

Ce triple déplacement donne à penser que l'on s'oriente, quoique de façon encore timide, vers une logique d'obligation de résultats (« rendre des comptes sur l'action éducative ») qui s'impose de nos jours dans un nombre croissant de systèmes éducatifs. La mise en avant de l'analyse et de l'exploitation des résultats au niveau de l'établissement est le signe de cet infléchissement, mais on soulignera par ailleurs que le Décret et ses étapes préparatoires ne vont pas au bout de cette logique, puisque le calcul d'indicateurs de valeur ajoutée[7] n'est pas prévu dans le Décret et que ce débat n'est pas revenu sous les feux de l'actualité. De même, les épreuves continuent à être administrées par les enseignants à leur propre classe, ce qui peut jeter un doute sur la fiabilité des résultats obtenus s'il s'agissait de mesurer vraiment l'efficacité de l'action pédagogique dans une logique d'obligation de résultats. Sans mettre aucunement en cause la bonne volonté des enseignants, on peut craindre, si ces évaluations externes en venaient à soulever de réels enjeux dans les établissements,

[7] Question posée par le Ministre de l'Education à la Commission de pilotage en 2004, mais entre-temps, la composition du Gouvernement a changé, et les portefeuilles ministériels ont aussi changé de main.

que ceux-ci ne conservent pas tout le recul nécessaire pour administrer les épreuves en toute impartialité – puisqu'ils seraient dès lors juge et partie. D'un autre côté, le dispositif prévu ne s'inscrit pas non plus de façon résolue dans une perspective de pilotage différente qui inciterait les équipes pédagogiques à développer, au départ des évaluations externes, une réflexion sur la qualité ou l'efficacité de leur action ou de leur fonctionnement. Mais il ne l'exclut pas non plus.

Conclusion générale

La Communauté française de Belgique, comme cela a été rappelé en introduction, possède un système éducatif fortement décentralisé, où les réseaux disposent d'une forte autonomie pédagogique, et où les parents ont le libre choix de l'établissement (système de quasi-marché). Ce système, avec ses particularités, n'échappe cependant pas à la tendance générale qui marque les politiques éducatives depuis le début des années 1990 à un renforcement du pilotage des systèmes éducatifs, en particulier, l'évolution vers une logique d'obligation de résultats. Mais, comme le souligne Dupriez (2004), « ces transformations adoptent des voies diverses en fonction des pays considérés et de l'histoire des systèmes scolaires ». Entre 1994 et 2006 s'est ainsi progressivement mis en place un dispositif original qui permet une forme « douce » de pilotage du système. Comme cet article l'a montré, cette mise en place a nécessité plusieurs étapes : d'abord la mise en place d'un référentiel commun (condition préalable à une évaluation externe), ensuite un premier dispositif centré sur l'enseignant, enfin un dispositif davantage tourné sur les établissements et le système éducatif dans son ensemble. Des débats intervenus autour des dispositifs d'évaluation se dégagent un souci de renforcement de l'équité – avec l'adoption d'une épreuve certificative commune en fin d'enseignement primaire – et un souci parallèle, très vif, de ne pas renforcer encore les disparités ou la concurrence entre établissements, souci lisible dans les textes de loi successifs qui insistent tous sur la non-publication des résultats par établissement.

Quelle que soit la volonté des décideurs, la Communauté française doit composer avec la donne que constitue l'autonomie des pouvoirs organisateurs, elle est donc condamnée à des compromis et à des avancées qui, en matière de pilotage, ne peuvent être que prudentes. Malgré le caractère progressif de la mise en place du pilotage, le côté plus incitatif que prescriptif des mesures[8], il n'empêche qu'en l'espace d'une dizaine d'années des avancées significatives ont été réalisées. Ceci peut paraître long, mais c'est sans doute un temps nécessaire « à la lueur de la complexité effective des formes de coordination dans ce système scolaire » (Dupriez, 2004, p. 11). Le Décret récemment adopté, comme on l'a montré, constitue un infléchissement vers une logique de pilotage par les résultats, sans toutefois s'inscrire résolument dans cette voie. Certains ne manqueront pas de s'inquiéter, voire de se désespérer, de cette lenteur alors que les indicateurs d'équité calculés au départ de l'enquête PISA pointent la Communauté française comme le champion toutes catégories de l'inéquité

[8] Cette tendance à laisser les acteurs/évaluateurs se saisir du dispositif pour lui donner son orientation.

sociale (Lafontaine *et al.*, 2003 ; Baye *et al.*, 2006 ; Gerese, 2003) ; à la lueur de l'âpreté des débats qui ont entouré le projet de Contrat pour l'école de la Ministre de l'Enseignement[9], on peut toutefois se demander si la Communauté française peut rêver d'un autre horizon que la politique des tout petits pas…

Références bibliographiques

Anderson, J.-C. & Wall, D. (1993). Does Washback exist ? *Applied linguistics*, vol. 14, *2*, 115-129.

Baye, A., Demonty, I., Fagnant, A., Matoul, A., Lafontaine, D., Monseur, C. (2006). *Les compétences des jeunes de 15 ans en Communauté française en mathématiques, en lecture et en sciences. Résultats de l'enquête PISA 2003*. Université de Liège : Cahiers du Service de Pédagogie expérimentale, *19-20*, 225 pages.

Dupriez, V. (2004). La place de l'évaluation comme ressource pour le pilotage des systèmes scolaires : état des lieux en Belgique francophone et en Angleterre. *Les Cahiers de Recherche en Education et Formation, 35*, 1-15.

Gerese (2003). *L'équité des systèmes éducatifs européens. Un ensemble d'indicateurs.* Liège : Service de pédagogie expérimentale et théorique. En ligne sur la page http://www.snesup.fr/navigation/general/page.php?ndoc=71

Grisay, A. (1984a). Les mirages de l'évaluation scolaire (I). Rendement en français, notes et échecs à l'école primaire. *Revue de la Direction générale de l'Organisation des Etudes*, *19*, *5*, mai, 29-42.

Grisay, A. (1984b). Les mirages de l'évaluation scolaire (II). Rendement en français, notes et échecs à l'école primaire. *Revue de la Direction générale de l'Organisation des Etudes*, *19*, *6*, mai, 9-23.

Lafontaine, D. (1999). Les évaluations externes en Communauté française de Belgique : quel sens donner à la mesure ? *Actes du XXIIIe colloque de l'Admée.* Dijon.

[9] En 2004, le projet de Contrat stratégique pour l'école, prenant la mesure de l'inefficacité et surtout de l'inéquité du système éducatif en Communauté française de Belgique mises en évidence par les enquêtes PISA 2000 et PISA 2003, a proposé une série de mesures ambitieuse visant à augmenter l'efficacité et à réduire l'inéquité de ce système. Ce projet a fait l'objet de très longs débats – plusieurs débats publics ont été organisés en présence de la Ministre de l'Enseignement Marie Arena, qui ont mis en évidence combien le monde enseignant était divisé sur la nature même du projet que la Communauté française peut ou doit se donner pour son école.

Lafontaine, D. (2001). Le dispositif d'évaluation externe en Communauté française de Belgique. *Cahiers du Service de pédagogie expérimentale*. Université de Liège. *7-8*, 195-221.

Lafontaine, D., Baye, A., Burton, R., Demonty, I., Matoul, A., Monseur, C. (2003). *Les compétences des jeunes de 15 ans en Communauté française en lecture, en mathématiques et en sciences*. Université de Liège : Cahiers du Service de Pédagogie expérimentale, *13-14*, 230 pages.

Lafontaine, D. & Blondin, C. (2004). *Regards sur les acquis des élèves en Communauté française Apports des enquêtes de l'I.E.A, de PISA et des évaluations externes*. Bruxelles : De Boeck.

Lafontaine, D. (2005). Relativiser les performances des établissements : quelle valeur ajoutée ? In M. Demeuse, A. Baye, M.-H. Straeten, J. Nicaise, A. Matoul (Eds), *Vers une école juste et efficace. 26 contributions sur les systèmes d'enseignement et de formation*. Bruxelles : De Boeck, Coll. Economie, Société, Région, 519-539.

Ocdé (1991). *Les systèmes éducatifs en Belgique : similitudes et divergences.* Bruxelles : Ministère de l'Education, de la Recherche et de la Formation.

SYMPOSIUM 6 : LES DISPOSITIFS INNOVANTS D'ÉVALUATION DES STAGES ET LEUR PRISE EN COMPTE PAR LES SUPERVISEURS ET LES FORMATEURS DE TERRAIN DANS UN CONTEXTE DE PROFESSIONALISATION : DIVERSES PERSPECTIVES

Louise M. Bélair, Université du Québec à Trois-Rivières

Résumé

Ce symposium regroupant des chercheures du Québec, de France, de Suisse et de Belgique se propose d'étudier et d'analyser divers dispositifs d'évaluation des stagiaires en enseignement. Il s'inscrit dans l'axe des **nouvelles pratiques d'évaluation ; développement d'outils**.

L'évaluation des compétences en enseignement pose certains défis aux institutions de formation. Au travers la variété de dispositifs mis en œuvre, on décèle certains invariants et interrogations communes, traduisant par là même des enjeux et des contraintes liées entre autres aux identités enseignantes, aux cultures des établissements appelés à se redéfinir, en fonction des implications soulevées par l'approche par compétence (Bru et Donnay, 2002 ; Gauthier et Mellouki, 2006 ; Lessard et Gervais, 2000 ; Rogiers, 1997).

À cet effet et pour lancer le débat, **Louise Bélair** (UQTR) pose un regard transversal sur plusieurs dispositifs d'évaluation des stages utilisés au Québec pour répondre aux exigences d'un même référentiel de compétence prescrit par le Ministère de l'éducation, des loisirs et des sports. De cet examen, **Christine Lebel** et **Louise Bélair** (UQTR), *L'implantation graduelle d'un dispositif d'évaluation des compétences en stage à L'UQTR : enjeux et limites*, font ressortir les choix théoriques sous-jacents à un dispositif particulier de stage pour les enseignants primaires à Trois-Rivières. Pour leur part, **Françoise Campanale** et **Hélène Gondrand** (IUFM de Grenoble), *Un dispositif d'évaluation formative pour les stages de formation professionnelle initiale de professeurs d'écoles : quelles régulations pour quels effets ?*, proposent l'analyse d'une innovation mise en place pour l'évaluation des stages en responsabilité des PE2 de l'IUFM de Grenoble, au regard des modèles théoriques et ce, afin de voir l'évolution des capacités réflexives des stagiaires. De leur coté, **Catherine Van Nieuvenhoven et Éric Bothy**, (ISPG, Belgique), *Un dispositif d'accompagnement formatif des étudiants futurs instituteurs dans leur pratique de stage : quels apports au niveau réflexif ?*, présentent un dispositif innovant d'accompagnement formatif des stagiaires de dernière année, afin d'examiner les apports au niveau réflexif.

Dans une autre perspective, **Colette Gervais** et **Michel Lepage** (U. de Montréal), *Recherche sur la construction des compétences : défis dans la révision des dispositifs*

d'évaluation des stages, discutent de l'argumentation pratique afin de rendre explicite la mobilisation souvent plus ou moins consciente des ressources par les stagiaires du secondaire. Enfin, **Lucie Mottier-Lopez** et **Walther Tessaro** (U. de Genève), *Entretiens de coévaluation instrumentés par un portfolio : pour quelle ré-interprétation de la progression des compétences professionnelles ?*, étudient comment les médiations instrumentales offertes par le portfolio dans le cadre d'entretiens de coévaluation, peuvent alimenter et offrir de nouvelles possibilités dans le cadre de l'évaluation des stages.

Ces six communications sont suivies d'une discussion soulevée par Anne Jorro, Université de Toulouse le Mirail 2, Département des Sciences de l'éducation.

Références

Bru, M. & Donnay, J. (2002). *Recherches, pratiques et savoirs en éducation*, Bruxelles : De Boeck

Gauthier, C. & Mellouki, M. (2006). *La formation des enseignants au Québec à la croisée des chemins*, Québec : PUL.

Lessard, C. & Gervais, C. (2000) *L'évaluation des nouveaux programmes de formation des maîtres: une compétence à développer.* Actes du sixième colloque de l'AQUFOM. Montréal: Faculté des sciences de l'éducation.

Rogiers, X. (1997). *Analyser une action d'éducation ou de formation.* Bruxelles : De Boeck.

ENTRETIENS DE COEVALUATION INSTRUMENTES PAR UN PORTFOLIO : POUR QUELLE RE-INTERPRETATION DE LA PROGRESSION DES COMPETENCES PROFESSIONNELLES ?

Lucie Mottier Lopez, Université de Genève
Walther Tessaro, Université de Genève

Mots-clés : portfolio, coévaluation, médiation

Résumé :
L'utilisation d'un portfolio est aujourd'hui fréquemment promue dans les programmes de formation professionnelle. Il représente une forme novatrice d'évaluation qui paraît pouvoir articuler les processus régulateurs de l'évaluation formative aux exigences de l'évaluation certificative. Dans une étude exploratoire, nous avons mis en évidence les médiations instrumentales que le portfolio peut offrir à des échanges de coévaluation (Mottier Lopez & Wegmuller, 2005). Notre hypothèse, fondée sur les travaux de la cognition située et distribuée, est que l'outil portfolio peut rendre possible l'établissement d'intercompréhensions, d'ajustements, de nouvelles interprétations et évaluations entre les acteurs qui se l'approprient. Le portfolio qui nous intéresse concerne un séminaire de formation à l'enseignement primaire et spécialisé. Il contient plusieurs types de traces : des démarches d'autoévaluation au sens strict, des bilans sommatifs de stages, des textes réflexifs liés à des problématiques définies dans le contexte de la formation universitaire. En fin de parcours de formation, des entretiens de coévaluation fondés sur le portfolio représentent des occasions propices à un (re)questionnement de la progression effectuée au fil des expériences dans le terrain. Nos résultats exposent les différentes médiations instrumentales, directes et indirectes, que le portfolio a pu offrir à sept entretiens de coévaluation entre un formateur et un étudiant. Deux portraits d'entretiens contrastés sont ensuite présentés, afin d'observer dans quelle mesure ils ont conduit à de nouvelles appréciations et interprétations de la construction (non terminée) des compétences professionnelles. Peu de recherches ont étudié, à notre connaissance, les dynamiques d'évaluation conjointe instrumentée par un portfolio. Il s'agit d'un des apports majeurs de notre recherche.

Introduction

L'utilisation d'un portfolio est fréquemment promue dans la formation professionnelle, vu comme une alternative à des évaluations traditionnelles. Il représente une forme novatrice d'évaluation qui articule les processus régulateurs de l'évaluation formative aux exigences de l'évaluation certificative (Vanhulle & Schillings, 2004). Dans une formation en alternance, il permet le recueil, l'appréciation et la mise en tension entre des traces issues des contextes de la formation théorique et pratique. Il offre, en tant que processus itératif, des modalités favorables à une évaluation de la progression des compétences professionnelles, à une

réflexion sur l'émergence de l'identité professionnelle, à un questionnement des stratégies de régulation mises en œuvre ou projetées. Le portfolio de développement professionnel, conçu dès la formation initiale, offre une instrumentation aux démarches d'autoévaluation et de réflexions métacognitives, propices à l'intégration des savoirs théoriques et des savoirs expérientiels. C'est sur la base de cette conception que nous allons examiner des interactions de coévaluation articulées à des portfolios réalisés dans le cadre d'une formation initiale à l'enseignement primaire et spécialisé. Plusieurs auteurs s'accordent à souligner le rôle majeur des processus sociaux dans l'élaboration d'un portfolio (e.g., Bélair, 1999 ; Vanhulle & Schillings, 2004). Il s'agit de promouvoir, entre autres, des pratiques d'*évaluation conjointe* à des fins de confrontation d'interprétations et d'appréciations entre pairs et/ou avec des formateurs. Pour notre part, nous centrerons notre étude sur des *entretiens de coévaluation* (un formateur universitaire – un étudiant) réalisés sur la base de portfolios présentés en fin de parcours de formation. Notre objectif est de questionner, dans une perspective d'*évaluation située*, les formes de médiation du portfolio ainsi que les apports de cette démarche à l'évaluation des stages en formation initiale. Mais commençons par circonscrire brièvement quelques éléments théoriques de l'évaluation située et de la médiation sociale. Nous présenterons ensuite le dispositif de recherche et les résultats.

Evaluation située : une relation indissociable entre activité et contexte

La « perspective située[1] » incite à reconceptualiser la relation entre l'individu-en-activité (membres qui évaluent/qui sont évalués) et le contexte de développement de cette activité, tant dans l'idée d'un contexte immédiat, contingent et local, que dans une relation à un contexte socioculturel, économique et historique plus large (Mottier Lopez, 2005). Dans cette conception, la cognition n'est plus vue comme une propriété individuelle, mais elle apparaît distribuée entre les personnes et les outils situés dans un monde socioculturel complexe (Salomon, 1993). Allal (à paraître) postule un rapport dialectique entre les facteurs contextuels de régulation (structure de la situation, interventions de l'enseignant, interactions entre pairs, outils) et les processus d'autorégulation propres aux apprenants. La recherche de Mottier Lopez a mis en évidence la *relation dialectique de régulation*[2] entre les plans individuel (apprentissages des élèves) et communautaire (pratiques et normes d'une microculture de classe) qui sont liées à des négociations collectives (plan interpersonnel) des significations considérées comme partagées entre un enseignant et ses élèves. Ces significations sont qualifiées de *taken-as-shared*, eu égard aux postulats des approches interprétatives des processus d'enseignement/apprentissage (Erickson, 1986). Autrement dit, on considère ces significations comme socialement reconnues et partagées entre les participants, bien qu'elles peuvent ne pas être effectivement communes à chaque individu. Il est néanmoins nécessaire de faire « comme si » elles l'étaient, afin de rendre possible les processus de communication et de compréhension partagée entre les participants (Cobb & Bauersfeld, 1995). D'une façon générale, la

[1] Terme générique emprunté à Greeno (1997) qui englobe les apports de la cognition située et distribuée et de l'apprentissage situé.

[2] Ou régulation « réflexive », dans la terminologie de Cobb et al. (1997) qui désigne la relation de co-constitution au sens fort entre plans individuel et collectif.

posture théorique de la perspective située incite à examiner attentivement les contextes sociaux, matériels, significationnels dans lesquels et par lesquels les activités évaluatives se constituent dans une relation de structuration réciproque – les activités évaluatives[3] contribuant elles-mêmes activement à la structuration des contextes et des communautés de pratique qui leur donnent sens. L'évaluation apparaît foncièrement liée aux processus de (re)négociation de valeurs et de significations partagées, dans une tension dynamique entre des relations de régulation entre les plans communautaires (e.g., pratiques et normes sociales), interpersonnels (e.g., confrontations de valeurs et de positionnements) et individuels (e.g., prises de conscience, auto-questionnements).

Le rôle prépondérant de la médiation sociale

Ailleurs, nous avons mis en évidence le rôle prépondérant des processus de médiation dans toute pratique d'évaluation située (Mottier Lopez, 2006). Rappelons seulement ici que, dans le cadre de travaux portant sur le portfolio en formation professionnelle, Vanhulle (2005) définit trois types de médiation : (1) la médiation sociale désignant les interactions structurées entre des acteurs en vue de construire des connaissances, (2) la médiation sémiotique notamment par la production de discours collectifs et individuels en lien avec des processus de formation, (3) la médiation instrumentale qui désigne, dans le cas auquel elle se réfère, l'outil portfolio qui vise à rendre compte d'une progression d'apprentissage. Ces médiations, distinguées à des fins de conceptualisation, sont intimement enchevêtrées en situation réelle ; elles sont également foncièrement liées aux pratiques sociales dans lesquelles et par lesquelles elles se constituent et prennent sens. En d'autres mots, elles sont socialement et culturellement situées (Lave & Wenger, 1991). Pour notre part, nous préférons parler de médiation *intersubjective* pour désigner les interactions structurées et intentionnelles des personnes dans des activités d'évaluation conjointes et réserver le terme de médiation sociale et culturelle (Bruner, 1983) pour englober l'ensemble des formes de médiations. Dans le cadre de la recherche présentée plus bas, nous nous intéresserons plus particulièrement aux *médiations instrumentales* rattachées à un portfolio. La notion d'*instrumentation* désigne à la fois l'outil/artefact et les schèmes d'usage développés par son utilisateur eu égard à son action finalisée par des buts (Rabardel, 1999). Relevons finalement que les éléments théoriques que nous convoquons dans la « perspective située » n'ont pas été développés pour étudier l'évaluation pédagogique en particulier. Comme d'autres auteurs avant nous, nous effectuons des emprunts aux théories psychologiques – et également anthropologiques ici – tout en les intégrant à des orientations conceptuelles existantes sur l'évaluation et la régulation des apprentissages. Il conviendrait évidemment de discuter les transformations que cette intégration fait subir aux notions empruntées, ainsi que les déplacements conceptuels que cette intégration produit dans le domaine de l'évaluation.

[3] Entre autres activités évidemment.

La recherche

Le contexte de la recherche

Notre recherche concerne un « séminaire d'intégration » qui fait partie d'un programme de formation à l'enseignement primaire et spécialisé. Un des buts principaux du séminaire est d'accompagner les étudiants dans leur travail d'appropriation et de mise en relation des différents apports de la formation théorique et pratique lors de leur dernière année de formation. L'évaluation certificative du séminaire porte sur l'élaboration d'un portfolio individuel, ainsi que sur un entretien mené en fin d'année académique avec le formateur universitaire. Les objectifs du portfolio sont de rendre visible la progression des compétences professionnelles construites au cours des stages, d'encourager une analyse conceptuelle des pratiques, d'envisager des moyens de régulation proactive en vue de nouvelles situations d'enseignement. La consigne écrite pour l'entretien est la suivante : « relevez ce que vous pensez avoir construit concernant vos compétences professionnelles au terme de votre formation initiale et ce qu'il est souhaitable d'anticiper dans votre future formation continue ». Parmi huit domaines d'objectifs définis et travaillés dans le séminaire d'intégration, l'étudiant doit sélectionner trois domaines à aborder dans l'entretien et cibler son questionnement sur un des objectifs de chaque domaine qu'il a choisi. Il lui est demandé de s'appuyer sur des exemples du terrain et de se référer à des concepts construits dans la formation. Il a encore été précisé que l'articulation entre le portfolio et l'entretien devrait permettre à l'étudiant de mettre en évidence sa progression dans la maîtrise des compétences professionnelles exercées en stage, et sur lesquelles il a mené une analyse réflexive. Sept portfolios individuels et sept entretiens menés par un même formateur sont concernés par notre étude.

Les questions de recherche

Deux questions de recherche sont investiguées ici :

1. Quelles formes de médiations instrumentales le portfolio offre-t-il aux échanges de coévaluation ?

Notre hypothèse est que l'outil portfolio peut rendre possible l'établissement d'intercompréhensions, d'ajustements, de nouvelles interprétations et évaluations entre les acteurs qui se l'approprient. En tant qu'artefact sociocognitif, le portfolio peut devenir un moyen de l'action pour les participants (Béguin & Rabardel, 2000). Notre étude identifiera, dans un premier temps, les différents usages du portfolio faits par le formateur et les étudiants au cours des échanges de coévaluation. Nous tenterons ensuite d'appréhender la façon dont les médiations instrumentales ont rendu possibles, ou non, la co-construction des discours évaluatifs.

2. Dans quelle mesure ces échanges ont-ils engagé de nouvelles appréciations des compétences professionnelles développées au cours des stages, déjà évaluées dans le cadre de bilans sommatifs ponctuels ?

Le portfolio qui nous intéresse contient différentes traces élaborées au cours de la dernière année de formation initiale à l'enseignement : des démarches d'autoévaluation, des rapports de stage sommatifs, des textes réflexifs liés à des problématiques spécifiques. Par le moyen du portfolio, les évaluations sommatives

devraient pouvoir être ré-intégrées dans un processus formatif plus large. En fin de parcours de formation, les échanges de coévaluation, fondés sur le portfolio, représentent des occasions propices à un (re)questionnement de la progression effectuée au fil des expériences dans le terrain. Une progression telle que perçue par chaque acteur et confrontée dans un processus de négociation de significations et de valeurs rattachées à des identités personnelles et professionnelles et tout à la fois à des normes, pratiques, discours valorisés dans des communautés de pratique données. La deuxième partie de nos résultats observe dans quelle mesure ces entretiens de coévaluation instrumentés par le portfolio ont conduit à de nouvelles appréciations et interprétations de la construction (non terminée) des compétences professionnelles. Deux portraits contrastés sont présentés.

Traitement du corpus

Notre questionnement de recherche nécessite un croisement entre des analyses de contenus des portfolios et des échanges de coévaluation. Trois axes principaux sont étudiés : (1) les contenus des portfolios confrontés aux objets discutés et évalués au cours des échanges de coévaluation, (2) les usages directs du portfolio dans les évaluations conjointes de la part du formateur et de l'étudiant, (3) la possible ré-interprétation de la progression des compétences professionnelles liées aux stages. Le principe méthodologique général est de travailler sur les transcriptions écrites des interactions verbales de coévaluation, puis par un croisement systématique, d'analyser les contenus des portfolios qui ont pu représenter une source de médiation instrumentale. Concrètement, une première étape a consisté en un découpage diachronique du verbatim de l'entretien de coévaluation, afin d'identifier les thèmes de discussion (*topic*) et les objets soumis à réflexion et à évaluation. Ensuite, différentes catégories d'analyse, délimitées dans une logique « inductive délibératoire » (Mottier Lopez, 2005), permettent d'observer les références et les usages du portfolio dans les échanges interpersonnels. A chaque fois qu'un des participants fait référence à un élément du portfolio, un code est attribué. Celui-ci se décide sur l'interprétation d'un ensemble de prises de parole qui permettent d'inférer l'intention – plus ou moins délibérée – du locuteur lorsqu'il se réfère au portfolio. A chaque changement d'usage finalisé du portfolio, un nouveau code est appliqué. Une confrontation d'interprétation des verbatim entre deux codeurs a été effectuée à des fins de validité scientifique. Enfin, nous avons systématiquement identifié les prises de parole relevant d'une évaluation des compétences professionnelles exercées pendant les stages. Les traits caractéristiques de chaque entretien ont été mis en évidence, servant de base à la rédaction d'un portrait pour chaque entretien. Notre but, à terme, est de dégager des *profils* de dynamique de coévaluation instrumentée par un portfolio.

Présentation de quelques résultats sur les médiations instrumentales du portfolio

Le formateur se réfère explicitement au portfolio (P) deux fois plus, en général, que les étudiantes (103 vs. 53). Des différences importantes apparaissent d'un

entretien à l'autre (tableau 1 : 12 occurrences vs. 34 occurrences, total formateur et étudiante par entretien).[4]

Tableau 1 – Nombre de références explicites au portfolio au cours des échanges de coévaluation

F	Doris	F	Vera	F	Diane	F	Noelle	F	Carine	F	Tam	F	Paule
20	14	11	8	23	11	15	9	11	3	14	5	9	3
Total													
34		19		34		24		14		19		12	

F : formateur

Parmi les différents éléments contenus dans le P, ce sont les *textes réflexifs* puis les *outils d'autoévaluation* qui sont le plus fréquemment mentionnés par les interactants. Le tableau 2 expose les différents usages du P qui ressortent de notre analyse de l'ensemble des entretiens.

Un constat tend à montrer que le contenu du P reste en *arrière-fond interprétatif* des échanges entre le formateur et l'étudiante (tableau 2, F-R : 9 occurrences ; E-R : 17 occurrences). Les participants ponctuent, en effet, leur discours de références brèves au P, établissant ainsi une relation entre ce qu'ils énoncent et ce qui a été écrit. Toutefois, on n'observe pas de développements particuliers dans l'échange interactif sur le contenu référé. Prenons pour exemple les propos suivants de Doris :

> J'ai encore eu de la peine à l'appliquer en stage (accueillir les élèves en classe avec des activités différenciées), parce que comme je l'ai relevé dans le portfolio, dans le cadre de la gestion de l'autorité, en stage on est quand même inscrits dans un contexte particulier (Ent-Doris).

Dans ce cas, Doris rappelle un contenu qu'elle a écrit dans son P concernant un autre thème (la gestion de l'autorité) que celui dont elle parle (accueil des jeunes élèves en classe). Elle sait que le formateur l'a lu. Ce contenu semble être considéré par l'étudiante comme suffisamment reconnu et partagé (*taken-as-shared*) pour qu'il reste dans un *non-dit* ou un *implicite* qui, par ailleurs, est accepté par le formateur. Ce dernier ne sollicite, en effet, pas de précisions particulières.

Tableau 2 – Usages du portfolio au cours des échanges de coévaluation

Usages par le formateur (F)			**Usages par l'étudiante (E)**		
Un indicateur du « taken-as-shared »					
F-R	pour faire brièvement référence à un contenu du P	9	E-R	pour faire brièvement référence à un contenu du P	17
Rattachés à une activité évaluative					
F-PF	pour inciter l'E à se situer « maintenant » et se projeter dans sa pratique prof.	23	E-Ev	pour évaluer sa progression sur des gestes professionnels identifiés, pour visibiliser cette progression	5
F-Ev	pour apprécier le P et la progression mise en mots dans le P	20			
F-	pour corroborer/valider les propos de	1			

[4] Par manque de place, nous ne les discuterons pas ici.

V F-CD	l'E pour contredire les propos de l'E	5 1			
Autres					
F-Q	pour inciter l'E à approfondir un objet	2 4	E-AO	pour approfondir une réflexion entamée sur un objet	1 7
F-AC	pour apporter des contenus par rapport à ce que dit l'E	1 1	E-J	pour justifier son point de vue	7
			E-Ex E-DQ	pour décrire situation expérimentée - sans problématisation - problématisée	3 4
	Total	1 03		Total	5 3

Ces contenus non-dits du P sont également perceptibles dans le discours du formateur (F-R : 9). Pour la suite de nos analyses interprétatives, nous proposons de retenir cette fonction de médiation indirecte du P qui offre une base – supposée – commune de connaissances et de compréhension entre le formateur et les étudiantes. Nos interprétations suivantes montreront que les contenus du P contraignent et rendent possible, en partie, le choix des thèmes discutés et des objets évalués ainsi que le développement des discours co-construits.

Médiation à des fins d'évaluation par le formateur

Parmi les différents usages du P exposés dans le tableau 2, nous choisissons de cibler ici nos résultats sur les médiations instrumentales liées à des *activités évaluatives*. Comparativement aux étudiantes, le formateur utilise clairement sa connaissance du P pour engager un discours évaluatif (59 occurrences vs. 5). La deuxième partie de nos résultats montrera que cela ne signifie pas que les étudiantes ne procèdent pas à des autoévaluations de la progression de leurs compétences professionnelles[5] ; mais pour ce faire, elles ne convoquent pas directement des écrits du P, s'appuyant plutôt sur l'évocation d'expériences vécues en stages notamment. Pour le formateur, le P représente, par contre, la source principale d'accès à cette progression. Nos analyses détaillées mettent en avant trois cas de figure concernant l'activité évaluative du formateur instrumentée par le P qui dénote d'une temporalité différente dans l' « ici et maintenant » de chaque entretien :

1. Le formateur communique à l'étudiante son interprétation et appréciation du P ou de la progression d'une compétence professionnelle sémiotisée par l'écrit (tableau 1, F-Ev : 20). L'évaluation porte ici sur l'objet P déjà constitué et sur ce qu'il témoigne en termes de développement professionnel dans une temporalité qui renvoie à des *expériences passées*.

> F cela (le P) vous permet de voir ce cheminement … cela vous permet de vous questionner ↑ / au début de la dernière année ↑ / et puis on retrouve un peu ces questionnements à la fin de votre P / donc il y a certaines compétences que vous avez reprises / ben voilà là où j'en suis aujourd'hui/ par rapport à une année … on voit votre progression par rapport à des compétences pour certaines qui semblaient déjà acquises / à cette époque ↑/ et puis pour d'autres il y avait

[5] Consigne de l'entretien, rappelons-le.

encore un cheminement à faire / donc ça ressort assez bien dans vos textes ↑ / j'ai relevé que // vous parlez en effet de la progression depuis le début de l'année et puis dans votre dernier texte / donc dans le texte 10 j'ai relevé à la page 53 ↑ // il y a deux aspects qui restent encore en questionnement // sur les différentes compétences ↑ / d'une part c'est la relation / aux partenaires à l'équipe aux enseignants ↑ et puis par rapport aux parents ↓ / alors bon / vous l'expliquez très bien ↑ / en disant que / il y a des choses que vous n'avez pas eu beaucoup l'occasion de tester (Ent-Noelle)[6]

2. Le formateur s'appuie sur ce qu'il connaît du P pour valider ou corroborer un propos de l'étudiante (tableau 1, F-V : 15). Dans ce cas de figure, la médiation du P est fonctionnelle en tant que support au feedback du formateur dans le *hic et nunc* interactif :

Tam au niveau du stage filé alors là / euh j'ai été presque surprise tellement cela s'est bien passé à ce niveau-là (…) donc je me suis sentie vraiment bien je n'ai eu aucun problème d'autorité et j'étais / vraiment contente de ça (…) c'est vrai que je sais que dès le départ je dois être ferme et me positionner comme telle / et puis ne pas essayer d'abord d'être sympa et tout ça ↑ mais que / une fois que j'ai fait attention à ça normalement / enfin je ne me fais plus autant de souci qu'avant en tout cas / donc euh / pour ça j'ai vraiment progressé à mon avis

F d'accord ↓ / donc il y a eu un cheminement / c'est vrai que dans les textes vous relevez / que le stage des quatre semaines a été déterminant par rapport à ça visiblement

Tam mm / [dans l'ensemble ouais

F [des paliers qui ont été franchis à ce moment-là ↑ / ce qui vous a mis sur une piste euh / sur le plan de l'estime de soi / aussi plus à l'aise plus confiante ↑

Tam ouais exactement ouais ouais et puis ça joue beaucoup en fait / le fait d'avoir plus d'assurance (Ent-Tam)

De tous les entretiens analysés, le formateur ne s'appuie qu'une seule fois sur le P pour contredire l'étudiante : un indice de la dynamique interactive des entretiens, avec un formateur qui s'aligne pour une grande part sur les propositions des étudiantes.

3. Le formateur prend appui sur ce qu'il a lu dans le P pour inciter l'étudiante à se « situer maintenant » (en fin de parcours de formation) et se projeter dans sa pratique professionnelle future (tableau 1, F-PF : 23). Notons que cette fonction « volitive » (Vanhulle, 2001), incitant les étudiantes à des projections, est particulièrement intéressante dans le cadre de ces entretiens qui représentent des lieux de transition entre la formation initiale et l'entrée dans le métier.

Véra la gestion de l'autorité ben justement j'en parle après // parce que / je trouve que c'est vraiment quelque chose qui est variable / en fonction de / de la classe ↓ / par exemple dans le premier stage justement j'ai eu un peu des difficultés par rapport / aux limites … (l'étudiante explique sa progression sur ce point)

F et comment vous vous situez par rapport à ça ↓ // parce que là vous dites dans le dernier document du portfolio qu'il est difficile de se positionner dans sa propre autorité dans les stages … / comment vous l'anticipez vous quand vous allez prendre votre première classe ↑

[6] ↑ : le ton monte ; ↓ : le ton baisse ; / : bref arrêt ; /// : arrêt de moins de 4 secondes. Nous avons choisi de supprimer ces conventions de transcription lorsque les citations sont insérées dans le texte.

> Véra explique qu'elle pense construire les règles de la classe avec les élèves dès le début de l'année scolaire, tout en restant ferme sur les règles convenues (Ent-Véra)

Ce faisant, le formateur incite l'étudiante à introduire de nouveaux éléments comparativement à ce qui a été énoncé et déjà évalué dans le P. Dans certaines situations, les échanges semblent avoir été propices à des régulations proactives potentielles.

Médiation à des fins d'autoévaluation par les étudiantes

Comme dit précédemment, peu d'étudiantes se sont explicitement appuyées sur des éléments formalisés dans leur P pour développer un discours autoévaluatif. Esquissons deux pistes interprétatives autres que la fonction de médiation indirecte du P citée plus haut : (a) les étudiantes préfèrent se référer directement à des expériences vécues, considérant peut-être que tout ne se dit pas dans la « mise en mots » effectuée dans le P ; (b) pour certaines étudiantes, il peut y avoir une prise de risque non négligeable de se référer directement à des éléments écrits, formalisés et connus par le formateur ; des stratégies d'évitement peuvent avoir lieu dans la situation foncièrement paradoxale (Paquay, Darras & Saussez, 2002) que représente l'entretien de coévaluation. Sans entrer dans les détails ici, on note que la référence explicite au P (ou à un de ses contenus) dans le discours des étudiantes vise essentiellement à approfondir une réflexion ou à puiser des exemples de situations vécues dans les stages.

Exemple de deux entretiens : Tam et Noëlle

Les deux étudiantes ont obtenu des évaluations satisfaisantes à tous leurs stages. Tam a postulé à une place de titulaire pour la prochaine rentrée solaire. Noëlle n'a pas demandé de classe bien qu'elle pense enseigner dans un avenir proche. La seconde étudiante possède de nombreuses expériences de remplacements en plus des stages liés à la formation initiale. Au début des deux entretiens, le formateur annonce que le P répond aux critères d'évaluation et que la certification du séminaire est obtenue. Il incite ensuite chaque étudiante à développer un discours à propos des objets préalablement choisis sur la base du référentiel de compétences professionnelles socialement instituées, reconnues et travaillées.

Entretien de Tam

L'étudiante a choisi d'aborder trois objets conformément aux consignes distribuées au séminaire : la gestion de l'autorité, la gestion de la classe, la conception de l'enseignement. Tam précise : *je vais faire une espèce d'historique de comment j'ai évolué à travers ces compétences*. Notre analyse met en évidence une dynamique similaire dans les échanges interactifs de coévaluation concernant ces trois objets. Dans un premier temps, l'étudiante s'exprime longuement et le formateur acquiesce régulièrement sans l'interrompre. Comme annoncé, Tam illustre chacune des compétences professionnelles qu'elle a choisi d'évaluer par des situations de stage. Sont ainsi systématiquement évoqués les trois stages de dernière année de formation

et le stage – formatif – de l'année précédente, parfois de façon chronologique. Concernant la gestion de l'autorité par exemple, la progression évoquée prend sa source lors du premier stage (*c'est là qu'on est pour la première fois confronté à ce genre de chose*), puis elle est relevée à deux reprises comme un objectif sensible (*j'ai noté ça comme point à améliorer à partir du stage de quatre semaines* ; *j'avais un peu des inquiétudes par rapport à l'autorité par rapport à toutes les idées que les classes spécialisées véhiculent*). Finalement, Tam pense que la compétence en question est acquise en fin de formation (*je me suis sentie vraiment bien je n'ai eu aucun problème d'autorité et j'étais vraiment contente de ça*). Ce faisant, l'étudiante cite volontiers des moments d'autoévaluation et de coévaluation vécus avec les formateurs de terrain qui l'ont amené à prendre conscience de son cheminement. D'une façon générale, les situations de stage évoquées par l'étudiante sont présentes dans son P sous forme de rapports d'évaluation sommative et de textes réflexifs qui s'y rapportent en partie. Mais Tam ne les convoque pas directement dans son discours : ils restent dans un implicite que le formateur accepte (médiation indirecte du P). Par contre, ce dernier utilise plus explicitement sa connaissance du P pour interagir avec Tam, par exemple pour valider (F-V) et compléter les constats de l'étudiante (F-AC). Cela lui permet de relever, entre autres, l'aspect non linéaire de la progression de Tam et de pointer un moment-clé de son parcours (F-Ev). Les situations spécifiques que Tam a mises en relation de façon détaillée et chronologique offrent au formateur l'opportunité d'introduire des éléments – qui nourrissent la réflexion métacognitive de l'étudiante – tout en se positionnant dans son rôle institutionnel d'évaluateur.

Notre comparaison entre les contenus du P et les échanges interactifs de l'entretien montre que les textes réflexifs de Tam sur la progression de ses compétences professionnelles relèvent d'une démarche auto-évaluative très générale. L'étudiante y précise pour le premier stage certificatif : *je manquais de confiance en moi … notamment par rapport à la gestion de l'autorité et à la conduite de l'enseignement.* Elle se limite ensuite à mentionner que les stages suivants se sont bien déroulés (et donc qu'elle a progressé). Lors de l'entretien, cette démarche d'autoévaluation est plus fréquente et plus précise. Elle porte sur des situations vécues ainsi que sur des documents du P qui demandaient une autoévaluation de sa progression :

> La conception de l'enseignement / dans la grille avec les domaines d'objectifs ↑ / j'ai chaque fois marqué que c'était très positif / dans le sens où / même si cela me prenait beaucoup de temps ↑ / enfin à la maison pour préparer quand j'arrivais dans la classe c'était toujours des activités / structurées et / j'ai pas eu de problèmes devant les élèves en fait

Après les trois thèmes introduits par l'étudiante, le formateur choisit de commenter un extrait d'un texte réflexif de Tam afin d'engager la discussion sur le rôle des stages pratiques dans la formation initiale. L'entretien se termine dans une toute autre dynamique interactive qui se caractérise, notamment, par l'expression par le formateur de valeurs personnelles et de valeurs institutionnelles[7] concernant la pratique du P et de son rôle attendu dans la circulation et l'intégration des savoirs théoriques et des savoirs d'expérience en formation initiale.

---

[7] Ou, en d'autres mots, de valeurs de la communauté de la formation institutionnelle qu'il représente.

Entretien de Noëlle

L'entretien de Noëlle se distingue de celui de Tam sur plusieurs points. Au début de l'interaction, l'étudiante ne répond pas aux consignes formulées par le formateur. Elle ne propose pas d'objets de réflexion mais elle demande au formateur d'initier lui-même la discussion en lui posant des questions sur le contenu de son P. Elle se justifie comme suit : *j'ai réfléchi un petit peu à ce que je voulais dire simplement je préfère peut-être que vous me posiez des questions parce qu'en fait c'est répéter déjà ce qui a été dit*.[8] En réponse à cette demande, le formateur commence par rappeler longuement le but de la pratique du P (F-AC) – témoignant de la construction nécessaire d'une conception minimalement partagée – puis il porte une appréciation globale sur les textes réflexifs de Noëlle (F-Ev) : *l'idée du P c'est bien … de voir un peu votre progression par rapport à des compétences pour certaines qui semblaient déjà acquises à cette époque et puis d'autres il y avait encore un cheminement à faire. Donc ça ressort assez bien dans vos textes*. Il rend compte de deux aspects qui, selon lui, restent en questionnement chez l'étudiante : la relation aux parents et la collaboration avec les collègues. Il demande à l'étudiante de se positionner « maintenant » (en fin de formation initiale) par rapport à ces deux objets (F-PF). La demande est immédiatement acceptée par Noëlle qui développe un discours fourni et critique sur ces différents points, en alternant des références à des expériences vécues dans les stages et un regard autoévaluatif sur son cursus en s'appuyant, notamment, sur des éléments formalisés dans son P (E-Ev). Ce faisant, un discours se co-construit, entre le formateur et l'étudiante, sur la qualité de la formation initiale telle que perçue par Noëlle.

La seconde partie de l'entretien est également initiée par le formateur. A partir d'un contenu relevé dans un texte réflexif (*vous disiez que le P vous a permis de faire une démarche intéressante, le regret que vous aviez, c'était que cela n'avait pas été débuté plus tôt*), le formateur demande à Noëlle comment la démarche mise en place dans le séminaire pourrait être proposée en amont de la dernière année d'études. On pourrait interpréter les remarques complaisantes que formule l'étudiante concernant une pratique valorisée par l'institution de formation comme un biais de la démarche de coévaluation. La réflexion qu'elle développe dans l'entretien à ce sujet et les exemples évoqués soulignent cependant l'importance accordée par l'étudiante à la démarche réflexive suscitée par le P. Noëlle considère néanmoins que la réflexion métacognitive sur le développement professionnel devrait être faite sur une période plus longue que la dernière année de formation (*je pense que j'ai énormément progressé jusqu'avant cette année et puis cette* (dernière) *année, ben j'ai aussi progressé un petit peu mais finalement peut-être pas, enfin c'est moins visible*), à l'image des pratiques observées dans les écoles, dit-elle. Depuis le début de sa formation, Noëlle tient d'ailleurs un journal dans lequel elle note des réflexions, des questions, des constats. C'est cet outil qui l'a aidée, de son point de vue, à envisager une progression à plus long terme de ses compétences. En évoquant ensuite un usage éventuel du P comme dossier de présentation à un futur employeur, Noëlle relève

[8] Demande révélatrice d'une conception de l'entretien comme un « examen » dans lequel il faut « répondre aux questions du formateur » ? Ou d'une stratégie d'évitement de la part de l'étudiante qui préfère, par exemple, s'ajuster à l'interprétation par le formateur de son P plutôt que d'avancer sa propre appréciation interprétative sans connaître d'abord l'avis du formateur ?

l'importance des rapports sommatifs des stages qui y figurent. Elle prend l'exemple d'un de ses rapports pour expliquer qu'il ne reflète pas, à son avis, l'état de ses compétences *à la fin* du stage et plus encore en fin de formation. En ce sens, il pourrait être injustement pénalisant.

D'une façon générale, la dynamique interactive entre le formateur et Noëlle prend ici une autre forme que dans l'interaction avec Tam. Si la cohérence de cet entretien est presque entièrement construite par le formateur, elle découle cependant de la posture initiale de l'étudiante exprimée dans le P. Une posture réflexive que le formateur dit être de qualité et perçue comme utile par l'étudiante. On note que le P joue un rôle important de médiation aux interventions du formateur, notamment pour initier les thèmes de discussion et pour exprimer des jugements évaluatifs étayés. Ceux-ci incitent l'étudiante à dépasser les constats effectués dans le P et à poursuivre la réflexion (F-Q/E-AO dans tableau 2). C'est l'occasion pour Noëlle, par exemple, d'évoquer les conséquences *négatives* d'une tendance qu'elle a à la perfection, évoquée pourtant à plusieurs reprises comme une qualité par ses formatrices de terrain. L'étudiante établit un parallèle avec l'élaboration d'un P qu'elle pourrait mettre en place avec ses futurs élèves :

F	vous vous fixez des objectifs élevés pour atteindre la perfection
Noëlle	je suis un peu trop perfectionniste donc même en stage … pour le portfolio c'est une bonne idée (d'en faire un avec les élèves) mais simplement / il faut pas que quand je commence + / que je veuille faire tout pour toutes les branches / et tout super bien parce que / je ne vais pas y arriver

Globalement, cet entretien se caractérise par le fait que les discours se sont co-construits *à partir* du portfolio et *sur* le portfolio, dans ses différentes fonctions, usages et contextes sociaux.

Eléments de conclusion

Nos résultats mettent en évidence que le portfolio, d'une façon générale, a été davantage utilisé par le formateur en tant qu'outil de médiation directe à ses interventions *évaluatives* dans les entretiens. Les étudiantes, quant à elles, ont préféré s'appuyer sur l'évocation d'expériences vécues en stage notamment, afin d'autoévaluer leur progression professionnelle. Certaines ont articulé judicieusement les différentes sources pour développer leur discours. Trois étudiantes sur sept, dont Tam, semblent avoir évité de faire référence explicitement aux contenus de leur portfolio. Nos analyses montrent que celui-ci reste cependant en arrière-fond interprétatif des échanges entre l'étudiante et le formateur, ce dernier l'utilisant aussi comme ressource pour initier les thématiques et inciter un approfondissement des réflexions initialement menées dans le portfolio. On relève un jeu subtil entre ce qui se dit explicitement du portfolio et les non-dits dont on fait l'hypothèse qu'ils sont vus comme reconnus et partagés entre le formateur et l'étudiante (une dimension qui est également fonctionnelle aux processus de communication, Cobb & Bauersfeld, 1995). D'une façon générale, les dynamiques interactives des entretiens ont été favorables à l'émergence de nouveaux questionnements et à l'expression de nouvelles évaluations et interprétations de la progression de l'étudiante dans les différents stages, avec l'émergence d'aspects non directement mis en évidence dans le portfolio.

Des différents entretiens, il ressort des représentations et valeurs collectives qui ont été construites en amont et redites pendant les échanges de coévaluation concernant le développement professionnel : celui-ci ne se termine pas en fin de formation initiale, mais l'obtention de la certification de la formation initiale signifie que l'on dispose d'un niveau de compétence jugé suffisant pour entrer dans le métier. Des projets de formation continue, ainsi que des stratégies de collaboration avec les collègues, sont non seulement à anticiper mais font partie intégrante de l'identité professionnelle. L'élaboration d'un portfolio, ici lié aux stages de formation pratique et associé aux entretiens de coévaluation, semble avoir joué un rôle important de médiation à la construction de ces valeurs rattachées non seulement à des positionnements individuels et interpersonnels mais également aux communautés de pratique professionnelle concernées.

Références

Allal, L. (à paraître). Régulations des apprentissages : orientations conceptuelles pour la recherche et la pratique en éducation. In L. Allal & L. Mottier Lopez (Ed.), *Régulation des apprentissages en situation scolaire et en formation.* Bruxelles : De Boeck.

Bélair, L. M. (1999). *L'évaluation dans l'école, nouvelles pratiques.* Paris : ESF Editeur.

Béguin, P. & Rabardel, P. (2000). Designing for instrument mediated activity. *Scandinavian Journal of Information Systems, 12,* 173-190.

Bruner, J. S. (1983). *Le développement de l'enfant, savoir dire, savoir faire*. Paris : PUF.

Cobb, P. & Bauersfeld, H. (1995). Introduction : The coordination of psychological and sociological perspectives in mathematics education. In P. Cobb & H. Bauersfeld (Ed.), *The emergence of mathematical meaning* : *Interaction in classroom cultures*, 1-16. Hillsdale, NJ : Lawrence Erlbaum Associates Publishers.

Cobb, P., Gravemeijer, K., Yackel, E., McClain, K. & Whitenack, J. (1997). Mathematizing and symbolizing : The emergence of chains of signification in one first-grade classroom. In D. Kirshner & J. A. Whitson (Ed.), *Situated cognition, social, semiotic, and psychological perspectives*, 151-233). Mahwah, NJ : Lawrence Erlbaum Associates Publishers.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Merlin (Ed.), *Handboock of research on teaching,* 119-161. New York : Macmillan Publishing Company.

Greeno, J. G. (1997). On claims that answer the wrong questions. *Educational Researcher, 26* (1), 5-17.

Lave, J. & Wenger, E. (1991). *Situated learning* : *Legitimate peripheral participation.* Cambridge : Cambridge University Press.

Mottier Lopez, L. (2005*). Co-constitution de la microculture de classe dans une perspective située : étude d'activités de résolution de problèmes mathématiques en troisième année primaire*. Thèse de doctorat en sciences de l'éducation, Université de Genève.

Mottier Lopez, L. (2006). Evaluation « située » des apprentissages : le rôle fondamental de la médiation sociale. In G. Figari & L. Mottier Lopez (Ed.), *Recherche sur l'évaluation en éducation. Problématiques, méthodologies et épistémologie*, 82-89. Paris : L'Harmattan.

Mottier Lopez, L. & Wegmuller, E. (2005). *Le portfolio en formation initiale d'enseignants : étude d'entretiens de coévaluation et des médiations instrumentales de l'outil*. Communication au 18ème colloque international de l'ADMEE-Europe, IUFM Champagne Ardenne, France

Paquay, L., Darras, E. & Saussez, F. (2001). Les représentations de l'autoévaluation. In G. Figari & M. Achouche (Ed.), *L'activité évaluative réinterrogée, regards scolaires et socioprofessionnels,* 119-133. Bruxelles : DeBoeck Université.

Rabardel, P. (1999). Le langage comme instrument ? Eléments pour une théorie instrumentale étendue. In Y. Clot (Ed.), *Avec Vygotski,* 241-265. Paris : La Dispute.

Vanhulle, S. (2001). *Des savoirs en jeu au savoir en « je ». Médiations sociales et processus de subjectivation en formation initiale d'enseignants.* Thèse de doctorat en Sciences de l'éducation, Université de Liège.

Vanhulle, S. (2005). L'écriture réflexive, une inlassable transformation sociale de soi. *Repères, revue de didactique du français, 30,* 13-31.

Vanhulle, S. & Schillings, A. (2004). *Former des enseignants compétents en lecture-écriture : interactions sociales, écriture réflexive et portfolio.* Bruxelles : Labor édition.

TRANSFORMATION D'UN DISPOSITIF D'EVALUATION DES STAGES PROFESSIONNELS : INCLURE LE CERTIFICATIF DANS UN PROCESSUS FORMATIF

Françoise Campanale, IUFM et LSE-UPMF – Grenoble
Hélène Gondrand, IUFM – Grenoble

Mots-clés : stages professionnels, évaluation par compétences, dispositif de régulations et de certification de la formation.

Résumé :
L'exigence de professionnalisation du métier d'enseignant amène à considérer les stages professionnels comme des points nodaux dans la formation des enseignants stagiaires. Leurs évaluations déterminent à la fois régulations de la formation et validation certificative. Dans les IUFM, leur organisation vient d'être modifiée pour les professeurs des écoles stagiaires, ce qui amène à repenser le dispositif de leur évaluation. La communication présente le contexte dans lequel intervient cette modification, ainsi que le dispositif en vigueur jusqu'à juin 2006 à l'IUFM de Grenoble pour l'évaluation des stages des professeurs d'écoles stagiaires. L'évolution des conceptions et des pratiques d'évaluation des acquis permet de situer ce dispositif dans une conception où évaluations formative et sommative sont bien distinctes et où la certification se fonde sur l'évaluation des performances. Les caractéristiques de l'évaluation par compétences sont à l'origine des principes qui régissent le nouveau dispositif, mis en chantier pour 2006-2007. Il vise à évaluer, au fil de leur développement, les compétences en construction et à articuler évaluation formative et évaluation certificative, dans un processus de professionnalisation qui ouvre sur la formation continue. Le basculement d'un dispositif à l'autre nécessite l'adhésion des formateurs à cette nouvelle approche, qui les éclaire par ailleurs sur de nouvelles formes d'évaluation dans les classes. Mais, la multiplication des postures qu'ils auront à adopter, les nouveaux outils à construire, l'accentuation de l'importance de l'attitude réflexive des enseignants stagiaires comme des formateurs ne va pas sans poser problème.

Introduction

En France, la formation professionnelle initiale des enseignants du premier degré fait alterner stages en responsabilité et modules de formation à l'IUFM. L'exigence de professionnalisation du métier d'enseignant (Paquay, Altet, Perrenoud, 1996), le modèle du *praticien réflexif* (Schön, 1994), qui orientent la formation des enseignants dans les IUFM, et le refus de modèles à appliquer que cela implique, font des stages le point nodal de la formation. En effet, les enseignements dispensés à l'IUFM n'ont

de sens que s'ils servent de *grilles de lecture de l'expérience* (Perrenoud, 1996)[1], s'ils trouvent écho dans la pratique (Belair, 1996). La conception d'une articulation pratique-théorie-pratique (Altet, 2000) dans laquelle les stages et leur évaluation jouent un rôle central vise à permettre aux stagiaires et aux formateurs de personnaliser la formation. La formation débouche aussi sur la validation et la certification. Quel dispositif d'évaluation des stages mettre en œuvre pour que les *actes de travail*, réalisés en stages, se fassent *actes de formation* (Barbier, 1996), c'est-à-dire pour cette évaluation interagisse avec la formation à l'IUFM et contribue à la construction des compétences professionnelles des enseignants stagiaires ? Comment un dispositif d'évaluation formative peut-il s'articuler avec l'évaluation certificative des compétences professionnelles, dans le cadre de la nouvelle organisation des stages imposée, pour la rentrée scolaire 2006, par le ministère de l'éducation nationale ?

Nous nous appuierons sur l'expérience de l'IUFM de Grenoble, en présentant le contexte des stages, en explicitant les références théoriques qui ont conduit à l'élaboration d'un nouveau dispositif d'évaluation des stages, que pour finir nous décrirons de façon succincte en terminant par les questions à travailler.

1. Un contexte des stages professionnels pour les professeurs des écoles en évolution

1.1 Les IUFM, seuls responsables de la formation professionnelle et de son évaluation.

En France, la formation initiale professionnelle des enseignants est entièrement confiée aux Instituts Universitaires de Formation des Maîtres (IUFM), par le ministère de l'éducation nationale qui définit à la fois le référentiel de compétences du PE et le cahier des charges de la formation. Il n'y a pas de répartition des rôles, comme dans d'autres branches professionnelle, entre un centre de formation chargé des enseignements théoriques et une entreprise, où se déroule le stage, chargée de l'acquisition de savoir faire et de savoir être (comportements et attitudes relationnelles en situation de travail).

L'évaluation est entièrement de la responsabilité de l'IUFM. Elle se fait en référence à un référentiel de compétences de l'enseignant débutant, joint au cahier des charges de la formation. C'est le directeur de l'IUFM qui prononce la validation de la formation, transmet au Recteur de l'Académie un avis pour déterminer la certification de l'acquisition des compétences nécessaires à l'exercice du métier, et ce dernier alors propose alors la titularisation du stagiaire au ministre. L'évaluation qui fonde validation et certification consiste en un bilan portant sur 3 pôles : le stage en responsabilité, le mémoire professionnel et les enseignements[2].

---

1 Perrenoud, P. (1996). Le travail sur l'habitus ans la formation des enseignants. In Paquay *et al., Former des enseignants professionnels* (pp. 181-207). Bruxelles : De Boeck.

2 Circulaire du 4 avril 2002 parue au B.O. n°15 du 11 avril 2002 (F – La validation de la formation) :
http://www.education.gouv.fr/bo/2002/15/default.htm

1.2 Un référentiel difficile à utiliser

Le « Référentiel des compétences du professeur des écoles stagiaire en fin de formation initiale » date de 1994[3]. Il énonce 4 catégories de compétences relatives 1) à la polyvalence et maîtrise des savoirs disciplinaires, 2) à la conception, mise en œuvre et analyse de situations d'apprentissage, 3) à la conduite de la classe et à la prise en compte de la diversité des élèves, 4) à la responsabilité éducative et à l'éthique professionnelle. Dans chaque catégorie, les compétences générales sont déclinées en multiples composantes : la compétence « *Concevoir une situation d'apprentissage* » est, par exemple, déclinée en : **«** *définir un objectif d'apprentissage…, inscrire la situation dans une progression, définir l'activité proposée à l'élève, identifier les obstacles qui peuvent exister, concevoir des évaluations aux différents moments de l'apprentissage* », etc.).
Ce référentiel de 4 pages est davantage un référentiel du professeur d'école confirmé que du débutant. Peut-il en être autrement d'ailleurs ? En effet, que le PE soit débutant ou confirmé, ces compétences sont nécessairement à mettre en œuvre dans l'exercice du métier en responsabilité. Elles ne peuvent d'ailleurs être considérées comme définitivement acquises dans toutes les situations. Elles sont à développer sans cesse au cours de la carrière.
La forme adoptée par ce référentiel, avec sa présentation en déclinaison de compétences, semble inviter à élaborer une grille avec les composantes transformées en critères à cocher, or bien évidemment les composantes et même les compétences générales se recoupent et sont à combiner dans l'exercice du métier. En outre, l'évaluation perd de son sens si l'on cherche à isoler des compétences ou des composantes de compétences (ex. Quel sens aurait de concevoir une situation d'apprentissage sans en penser en même temps les modalités d'évaluation ?). A fortiori, il serait artificiel (illusoire ?) de vouloir quantifier avec précision des degrés de maîtrise dans les différentes catégories de compétences, compte tenu de la variété des situations de stages dans lesquelles elles peuvent se manifester.

1.3 Un nouveau référentiel en gestation

De plus, un projet de nouveau cahier des charges pour les IUFM, assorti d'un nouveau référentiel, est bien avancé et devrait paraître courant 2007[4]. Il inscrit la formation professionnelle initiale dans la perspective de la formation tout au long de la vie, en l'articulant avec la formation continue à travers la mise en place d'un portefeuille de compétences de l'enseignant. Dans cette optique, l'IUFM devrait, en fin de formation initiale, informer chaque professeur des compétences qu'il lui reste à développer ou à confirmer, en vue de la formation continuée dans le cadre de l'Accompagnement à l'Entrée dans le Métier (AEM). D'ailleurs, le nouveau référentiel (commun aux enseignants du premier, du second degré et aux conseillers

[3] *Cf.* Annexe III, BO n° 43 du 24 novembre 1994.
[4] Ce projet de la commission Piétryck est consultable sur le site du snesup : http://www.snesup.fr/navigation/general/page.php?ndoc=2274

d'éducation) est annoncé comme, à la fois, base pour orienter la formation et référence pour l'évaluation des enseignants stagiaires en fin de formation et des néotitulaires en première et deuxième années d'exercice. On y retrouve les grandes lignes du référentiel de 1994, avec un accent toujours fort sur l'organisation des apprentissages et plus affirmé sur l'organisation de situations d'évaluation et la prise en compte de l'hétérogénéité des élèves, sur la dimension éthique et le travail en équipe.

1.4 De nouvelles modalités d'alternance

La formation professionnelle initiale (36 semaines) des professeurs d'écoles stagiaires (PE2) repose sur l'alternance de stages en responsabilité (et en tutelle) et de modules de formation à l'IUFM. Jusqu'en juin 2006, chaque PE2 effectuait 4 stages groupés de 2 ou 3 semaines répartis sur l'année. Depuis la rentrée scolaire 2006[5], les stages sont de deux types : 1 stage filé d'une journée par semaine dans une même classe d'une même école et 2 stages groupés de trois semaines, à des moments différents de l'année, chacun dans une classe d'un cycle différent de celui du stage filé.

1.5 Un dispositif d'évaluation des stages à modifier

Dans le dispositif d'évaluation des stages des PE2 de l'IUFM de Grenoble, en vigueur jusqu'à fin juin 2006 (*Cf.* annexe 1), la période formative est distincte du temps de l'évaluation de validation qui pèse lourd dans la décision de certification.

Dans la période formative (l'année moins le dernier mois), les PE2 effectuent 3 stages formatifs, chacun dans un cycle différent de l'école primaire. L'évaluateur, formateur de l'IUFM ou Professeur d'Ecole Maître Formateur (PEMF), est chaque fois différent, mais il fait partie de l'équipe qui intervient dans sa formation. Chaque retour de stage donne lieu à un conseil des formateurs, appelé conseil de formation, qui, au vu des bulletins de visite, propose éventuellement au stagiaire un complément de formation (court stage supplémentaire en pratique accompagnée, module facultatif corps et voix par exemple). La posture dominante des évaluateurs se veut être celle du conseiller (Jorro, 2000) qui aide le stagiaire à analyser les compétences qu'il maîtrise, ses difficultés et à chercher comment les surmonter.

Cette régulation externe rétroactive (Allal, à paraître) est accompagnée par un tuteur qui aide chaque stagiaire à déterminer des choix dans son parcours de formation (dossier de suivi de la formation). Le tuteur, dont la posture oscille du conseiller au consultant (id.), facilite l'autoévaluation, l'interrogation des conceptions implicites qui régissent la pratique. L'action conjuguée du tuteur et des séminaires d'analyse des pratiques contribue à développer une attitude réflexive chez le stagiaire.

Pour assurer une progressivité dans l'évaluation des stages, on constate des essais divers de la part des formateurs : définir en équipes de formateurs des axes d'observation qui évoluent au fil des stages ou faire déterminer par les stagiaires,

[5] *Cf.* Circulaire du 11 mai 2006.

avant le stage, 2 ou 3 points forts et 2 points faibles à travailler particulièrement, points auxquels s'attachera plus spécifiquement l'évaluation formative.

Enfin, le dernier stage de l'année donne lieu à une évaluation en vue de la validation, réalisée par un formateur de l'IUFM et un PEMF extérieurs à l'équipe qui forme le stagiaire et éventuellement par l'inspecteur de la circonscription où se déroule le stage. A l'issue de la visite, les évaluateurs rédigent un rapport qui précise le contexte d'exercice, se prononce sur le développement des grandes compétences et émet un avis global soit positif ou négatif, en déclinant l'avis positif sur une échelle à 3 degrés (assez favorable / favorable / très favorable).

Ce dispositif, compte tenu de la circulaire ministérielle modifiant l'organisation des stages, n'est plus possible. De plus, si le dispositif est central dans la régulation de la formation, la grande majorité des formateurs n'approuve plus cette rupture entre le formatif et le certificatif. Toutes les informations qu'ils ont pu recueillir au cours de l'année sur la progression du stagiaire n'interviennent pas, sauf en cas de problème, au moment de la validation déterminée par des évaluateurs qui ne connaissent pas le stagiaire. Le stage terminal fait figure d'épreuve à réussir. Deux performances, à l'occasion des deux visites pour la validation peuvent-elles être considérées comme une évaluation des compétences professionnelles ?

2. L'évolution des conceptions et des pratiques dans l'évaluation

Scallon (2004) met en relation l'évolution des pratiques pédagogiques et celle des pratiques d'évaluation. Nous référerons les pratiques d'évaluation aux modèles d'évaluation de Bonniol et Vial (1997). Cette évolution correspond aussi à une évolution dans la hiérarchie des qualités d'une évaluation (De Ketele & Gérard, 2005) ; elle peut nous éclairer sur l'évolution de l'évaluation des stages professionnels, à l'IUFM de Grenoble.

2.1 La logique du contrôle et de la mesure

A la pédagogie par les contenus, conçue dans une finalité de reproduction, correspond le contrôle des connaissances et de leur utilisation pertinente, la mesure des performances avec un seuil quantitatif, en vue de discriminer en fonction des différences individuelles, voire de classer. Tous les élèves/stagiaires suivent le même programme ou itinéraire et il s'agit, pour eux, d'obtenir une note égale ou supérieure à la note signifiant un niveau moyen de performance (10/20). L'idéal premier, que l'on sait vain, de cette pratique d'évaluation est l'objectivité de la mesure, sa fidélité quel que soit l'évaluateur.

Certains formateurs des IUFM sont encore très imprégnés de cet idéal premier, même si leurs pratiques de formation ont changé. Ils abandonnent difficilement la posture du contrôleur (Jorro, 2000).

2.2 La logique de la régulation externe et de l'interprétation critériée de la performance

L'émergence de l'évaluation formative, avec la mise en évidence de l'importance du feed back pour les apprenants et pour la régulation du dispositif d'enseignement, amène la pédagogie de la maîtrise ou la pédagogie par les objectifs avec son exigence

de rationalité et sa référence au behaviorisme. Les référentiels déclinent de façon détaillée les objectifs généraux des apprentissages ou des formations. Apparaissent des évaluations qualitatives pratiquant l'interprétation critériée de savoir faire. Les pratiques distinguent évaluation formative en cours d'apprentissage (ou de formation) et évaluation sommative, qui se réduit à un contrôle de performances, devenu analytique, appuyé sur des critères et indicateurs. L'élève s'exerce, puis prouve qu'il sait faire. La notation a tendance à disparaître à l'école primaire en France au profit d'une appréciation (Acquis / En cours d'acquisition / Non acquis) en regard d'items déclinant de façon détaillée des savoir faire plus que des compétences. Et cela ne va pas sans poser problème à beaucoup d'enseignants au point que souvent existe un double système : appréciations d'acquis et traditionnelle notation chiffrée. Toutefois, on peut dire que le qualitatif est entré dans l'évaluation.

Le souci premier est alors la validité de l'évaluation (La tâche d'évaluation évalue-t-elle bien ce qui est visé ?) et émerge l'importance de la pertinence du feed back et donc de la pertinence des modalités d'évaluation (Permettront-elles de bien identifier ce qui va / ce qui dysfonctionne ?).

Le dispositif d'évaluation des stages des PE2, que nous venons de présenter, avec sa distinction entre stages formatifs et stage de validation, tient de ce modèle de la « gestion des apprentissages » (Bonniol & Vial 1997) pour réguler les apprentissages. Mais, par certains aspects, déjà il le dépasse à travers les modalités de régulation qui mettent en œuvre un questionnement du sens des pratiques et de leur interaction avec les réactions observées chez les élèves, qui touchent le modèle de la « problématique du sens » (*id.*).

2.3. La logique de l'évaluation, au fil de leur développement, des compétences en construction ou le certificatif comme un moment dans le formatif

Des ouvrages récents[6] posent la question de l'évaluation des compétences des élèves et, ce faisant, éclairent ce terme énigmatique : la compétence (Dolz & Ollagnier, 2000). Rogiers (2004) tente, entre autres, de concilier évaluation des compétences et mesure en quantifiant des indicateurs, Scallon (2004) propose une estimation sur une échelle réduite, assortie d'appréciations. Crahay & Detheux (2005) se demandent si l'entreprise n'est pas impossible.

Ils s'accordent pour rattacher l'approche par les compétences à une pédagogie socioconstructiviste, pédagogie des situations-problèmes, pédagogie de projets. Des différentes définitions du terme « compétence », il ressort l'idée d'une combinaison de savoirs de différents types, d'un « savoir agir » responsable et validé face à des tâches scolaires/des situations professionnelles complexes et événementielles, qui nécessitent de savoir mobiliser des ressources internes et externes (Le Boterf, 2000). Cela se traduit pour des élèves par savoir mobiliser ces deux types de ressources, de façon autonome, savoir déterminer une stratégie afin de résoudre un problème/mener à bien un projet, planifier son action, réguler en cours d'action, et à la fin évaluer les effets de son action, la démarche conduite (Perrenoud, 2001 ; Scallon, 2004 ; Crahay & Detheux, 2005). La compétence permet de faire face, de façon efficace, à une famille de situations de même type, grâce à la construction de schèmes opératoires complexes (Gillet, 1991; Le Boterf, *id.*). Le contexte favorise plus ou moins l'expression de la compétence (Meignant, 2003).

[6] Nombre d'ouvrages ont traité précédemment de l'évaluation des compétences professionnelles mais surtout de celle des salariés en entreprises ou dans le cadre de la formation continue. Pour cette raison, ce sont plutôt les ouvrages traitant de la formation initiale qui nous ont été utiles.

Evaluer le développement de compétences chez les élèves suppose de les placer en situation de réaliser des tâches complexes, une situation « authentique » (Bélair, 1999) avec un enjeu social, de leur permettre de s'approprier les critères d'évaluation, d'évaluer autant le produit que la démarche. Cela ne dispense pas de l'apprentissage et de l'évaluation de connaissances et habiletés nécessaires aux compétences, en situations décrochées. Le doute concernant l'évaluation des compétences des élèves vient, d'une part, de la masse d'opérations mentales à effectuer en regard de la capacité limitée de la mémoire de travail et, d'autre part, de l'écart entre la place accordée le plus souvent dans l'enseignement à la métacognition, à l'apprentissage de stratégies et de régulations et l'importance de ces démarches en situation de tâche complexe (Crahay & Detheux, *id.*). Complexité pour l'élève et complexité pour l'enseignant !

L'évaluation des compétences est surtout qualitative. Elle nécessite une interactivité entre évalué et évaluateur, l'implication de l'apprenant dans l'évaluation.

Elle se traduit par un bilan, qui s'appuie sur les résultats des évaluations ponctuelles qui formulent des appréciations en regard de critères explicités, utilisent éventuellement des échelles descriptives d'appréciation (Bélair, *id.* ; Scallon, *id.*). L'observation des comportements qui se stabilisent au fil des réalisations prime sur une évaluation détaillée de l'activité et de son efficacité ponctuelle. Au-delà des évaluations ponctuelles, c'est un jugement global sur les compétences développées et sur l'histoire de leur développement qui est attendu (Scallon, *id.*). Il s'agit d'établir des profils en évolution. C'est alors la pertinence de l'évaluation qui prime sur sa validité et sa fiabilité (De Ketele & Gérard, 2005). L'évaluation des compétences réhabilite le jugement de l'expert, explicité en regard d'un référentiel. L'objectivité n'est plus dépendante de la précision des outils mais de la dimension plurielle de l'évaluation (deux experts valent mieux qu'un !). Toutefois, organiser les situations d'apprentissage, observer et choisir les informations nécessaires pour suivre l'évolution des compétences, mettre en place des régulations efficaces, tout cela à la fois risque de déborder aussi des capacités de travail de l'enseignant. Il peut préférer par nécessité d'économie la régulation *on line* des apprentissages à l'évaluation des compétences, d'ailleurs non dénuée du risque de dérive techniciste que provoque la plupart du temps une nouvelle approche. D'où la nécessité du travail en équipes des enseignants et de tâches assez longues à réaliser par les élèves (ce qui laisse du temps pour l'observation), du travail en groupes dans lesquels peuvent déjà se produire des régulations sans que nécessairement l'enseignant intervienne, de situations d'évaluation mutuelle (Allal, 1999 ; Campanale, 2003).

Au niveau des professeurs stagiaires, on retient de cette brève présentation de l'évaluation des compétences, outre que c'est dans l'action (en exercice du métier) que l'on peut inférer des compétences, 4 caractéristiques :

- C'est un processus continu dans lequel l'évaluation certificative n'est qu'un arrêt sur image, fixant très provisoirement un profil engageant, à partir de l'amont, un itinéraire aval de formation.
- Les évaluateurs les mieux à même d'évaluer le développement des compétences des enseignants stagiaires sont ceux qui les côtoient et dialoguent avec eux sur une longue période, qui peuvent observer leur évolution, autrement dit leurs formateurs et non des évaluateurs extérieurs. « Fondées sur les *mêmes données*, mais utilisées dans des postures et à des fins différentes » (Perrenoud, 2001, p.5), évaluations formative et certificative peuvent alors s'articuler.
- Compte tenu de la variété des situations de stages, l'évaluation en vue de la certification vise prioritairement à fonder une décision pertinente, en montrant que le stagiaire maîtrise suffisamment les compétences minimales requises (De Ketele & Gérard, 2005) pour

exercer le métier. Cela signifie identifier ces compétences minimales, sans renoncer pour autant à apprécier les autres qui se manifestent.

- Les capacités d'analyse de la situation et de ses stratégies sont aussi importantes à évaluer que les performances en stage, qui peuvent ponctuellement s'avérer décevantes, par défaut occasionnel d'efficacité.

3. Le nouveau dispositif de l'IUFM de Grenoble

Le schéma en annexe présente la nouvelle organisation des stages en responsabilité, qui renforce le temps de présence sur le terrain des stagiaires : un stage filé d'une journée par semaine sur quasiment toute la durée de l'année scolaire dans une classe – cycle - et deux stages en responsabilité de 3 semaines chacun dans les deux autres cycles de l'école primaire

Cette modification introduit une logique d'évaluation différente puisqu'elle remet en cause la distinction entre « stages formatifs » et « stage terminal support de validation », et par là même pose la question du positionnement des formateurs/évaluateurs. Elle oblige à basculer d'une évaluation sommative appréciant un « état final » à une évaluation au fil du développement des compétences en construction.

3.1 Ses principes

La formation professionnelle des professeurs des écoles stagiaires est organisée pour articuler étroitement les pratiques en stage (filé, groupés, en tutelle) et les apports des modules de formation. Elle ancre fortement la professionnalisation dans l'analyse des pratiques : pratiques observées lors des visites, analysées lors des entretiens ou des séminaires d'analyse des pratiques, décrites et théorisées en modules. Les stagiaires sont fortement impliqués dans les choix de leur itinéraire de formation et dans la construction de leurs compétences professionnelles à travers l'interaction mise en œuvre pour l'élaboration du dossier de suivi de formation.

Une équipe de formateurs fixe prend en charge la formation, le tutorat, les visites et l'évaluation pour une section de 25 à 30 PE2. Le tuteur ne fait pas de visite en stage à son tutoré.

L'évaluation du volet « stages » prend en compte les trois stages dans leur dynamique, ce qui suppose un repérage des besoins[7] et un suivi de l'évolution des compétences professionnelles du stagiaire au cours de l'année. Au cours des différents stages, l'évaluation est formative et cette dimension se traduit dans les entretiens et bulletins de visite. L'évaluateur cherche à repérer des éléments de compétences stables et ceux qui sont prioritairement à renforcer, à la fois à travers ce qu'il observe en classe et au cours de l'entretien.

L'équipe de formateurs, réunie en conseil de formation, fait régulièrement le point sur les besoins de formation des stagiaires, sur les remédiations ou les aides individuelles à proposer (stages en tutelles, modules spécifiques), sur les régulations de la formation à mettre en place au niveau des modules. Le PE2 a un feed back. En fin d'année, ces bilans ponctuels débouchent sur un bilan final des compétences

[7] Ce repérage des besoins est d'ailleurs inscrit dans le projet du nouveau cahier des charges des IUFM.

observées dans des contextes d'exercice variés (type d'écoles, cycles, …) et de celles à travailler ultérieurement. Le bilan est assorti d'un avis positif ou négatif.

L'évaluation des stages vise ainsi à la fois une régulation de la formation au niveau de la section de PE2, une régulation individuelle du parcours de chaque stagiaire, et à fonder l'avis final.

3.2 Ses modalités

Chaque PE reçoit 5 visites au cours de l'année. Selon les cas, le nombre de visite par stagiaire peut varier afin de prendre en compte les situations qui nécessitent un suivi particulier. Un travail sur le référentiel de compétences, mené conjointement par le stagiaire et les formateurs visiteurs, permet d'identifier, au fil des visites de stage, les compétences professionnelles maîtrisées par le stagiaire. Chaque visite donne lieu à la rédaction d'un bulletin s'appuyant sur le même outil d'évaluation (*Cf.* Référentiel minimal en annexe).

Le stagiaire a connaissance du bulletin de visite, du bilan ponctuel du conseil de formation et en discute avec son tuteur.

En fin d'année, le conseil de formation dresse le bilan des stages en responsabilité (filé+groupés) : au vu des différents bulletins de visite, il rédige un rapport précisant les compétences construites, décrivant l'évolution du stagiaire, et propose un avis : positif ou négatif. L'avis négatif est justifié par ce qui, d'un commun accord, est jugé non acceptable de la part d'un enseignant, fonctionnaire du service public d'éducation : ne pas assurer la sécurité physique et morale des élèves, déroger aux règles de la déontologie et du respect des personnes, ne pas préparer les situations d'apprentissage, rejeter systématiquement les conseils des formateurs, ne pas évaluer le travail des élèves, renoncer à différencier, marginaliser les élèves différents,…

3.3 Les questions à approfondir

- Quels usages du référentiel ?

Comment faire du référentiel un outil de formation lors des visites et à l'IUFM et non un outil de contrôle qui comptabilise au fil des visites les compétences construites et celles à renforcer pour établir l'évaluation sommative terminale ? Autrement dit, comment éviter de faire du référentiel minimal, complété d'indicateurs prédéterminés, un outil de contrôle pour conformer et non un outil d'évaluation au service de régulations externes et internes de la formation ?

Au début de l'année, les formateurs proposent un travail sur le référentiel ministériel et le référentiel minimal commun. Le but est de permettre aux stagiaires de prendre conscience de la complexité du métier de PE, de pointer ce qu'ils pensent savoir déjà faire, d'exprimer des besoins de formation, qui seront pris en compte dans la régulation des modules de formation et dans les conseils donnés sur les choix d'itinéraire.

Mais, quelle progression peut être envisagée ensuite au niveau du développement des compétences ? La marge de manœuvre est étroite dans la mesure où les PE en responsabilité dans les classes sont d'emblée confrontés à la complexité du métier et

doivent mettre en œuvre la plupart des compétences du référentiel. Aussi, c'est la façon dont le PE mettra en œuvre les compétences qui va varier selon des situations de stages rencontrées et s'enrichir au fil de la construction de sa pratique. Cela suppose que les formateurs, que ce soit au cours de la préparation des stages, de l'entretien formatif lors d'une visite, de l'analyse a posteriori de la pratique, suscitent la réflexion des stagiaires sur la façon dont ils traduisent les gestes du métier en gestes professionnels qui leur sont propres en fonction de la situation dans laquelle ils s'exercent (Jorro, 2002). Il leur faut mettre en œuvre une démarche de co-évaluation qui aide les stagiaires à identifier les indicateurs révélateurs de telle ou telle compétence en fonction de la situation dans laquelle elle a été requise. Toutefois, à partir du référentiel minimal, les formateurs s'entendent sur les compétences prioritaires à observer au fil de l'année en fonction du cycle de classes et sur des indicateurs sensibles relatifs à des critères de réussite des tâches-clés de l'enseignant.

En fin d'année, le bilan terminal des compétences observées en stages s'appuiera sur le référentiel minimal. A défaut de pouvoir les mesurer, il les mentionnera en les associant aux caractéristiques des situations dans lesquelles elles ont été observées.

- Comment engager le PE2 dans une *réelle* démarche d'autoévaluation ?

Au cours de l'année de PE2, les stagiaires vivent une transformation identitaire radicale en passant du statut d'étudiant à celui d'enseignant. Inciter le stagiaire à entrer dans un processus d'autoévaluation (Campanale, 1997), c'est l'aider à s'engager dans une dynamique, à mobiliser de multiples référents pour concevoir, mettre en œuvre et évaluer les situations de classe, à se construire un répertoire de pratiques structuré (Gondrand & Pierrard, 2001 ; Gondrand, 2004). Il s'agit tout autant d'éviter les effets *caméléon* (Paquay, 1994) (« je me conforme à ce qu'on attend de moi, et ensuite je ferai bien comme je voudrai »), que les effets de dépendance (« dites-moi si/que c'est bien ! ») ou de démotivation. Cette démarche d'autoévaluation est d'autant plus nécessaire que le stagiaire pourrait se sentir enfermé dans ce suivi « rapproché » et qu'il s'agit de lui préserver un espace de liberté, où son expression soit entendue. Elle se traduira dans son dossier de suivi de la formation qu'il élabore avec son tuteur et qui est communiqué au conseil de formation.

Conclusion : entre continuité et rupture, sens du basculement

Même s'il conserve les modalités de régulations de la formation du dispositif précédent, le nouveau dispositif marque une rupture dans l'évaluation des compétences professionnelles des PE. Le renoncement à une évaluation en termes de mesure d'écarts entre compétences observées et compétences figurant dans le référentiel complet, en fonction d'indicateurs préétablis et stables, le renoncement à l'établissement d'une estimation graduée de développement des compétences pour un jugement global débouchant sur un avis positif ou négatif, marquent le basculement d'une logique de contrôle, dont l'objectivité est censée être garantie par l'extériorité de l'évaluateur et d'une logique d'administration de la preuve, appuyée sur outils dont on reconnaissait le défaut de fidélité, à une évaluation qui réhabilite le jugement du

formateur, pose avec acuité la question de la nature de l'entretien, de la forme du bulletin de visite et de la forme du rapport final. La subjectivité inévitable des jugements est compensée par la pluralité des évaluateurs au sein de la même équipe. C'est la pertinence de l'évaluation, appuyée sur la connaissance du stagiaire et l'interactivité avec lui qui priment pour dessiner un profil de développement des compétences. Les explicitations de démarches, qui peuvent témoigner de schèmes opératoires efficaces, sont aussi importantes que les performances (un stagiaire peut rater une séance, mais pouvoir dire comment il aurait dû faire, ce qui signifie qu'il sera capable de réguler sa pratique et que la compétence est construite, même si elle ne s'est pas traduite sur le moment par une action efficace). L'avis prononcé en fin de formation est davantage de l'ordre du pronostic favorable que de la preuve incontestable (ce qui d'ailleurs était déjà le cas dans les dispositifs précédents). L'articulation de l'évaluation des stages avec les évaluations des deux autres volets (mémoire professionnel et enseignements) reste une question à travailler. Les équipes de formateurs ont une responsabilité et une charge de travail accrues : sauront-ils les assumer ? L'enjeu est la professionnalisation : préparer les enseignants stagiaires à assumer, en l'assumant soi-même en tant que formateur, l'incertitude inhérente à l'évaluation dans des situations complexes à forte charge d'imprévus. Restent à élaborer très rapidement les modalités d'évaluation de la mise en œuvre de ce nouveau dispositif.

Références

Allal, L. (à paraître). Régulations des apprentissages : orientations conceptuelles pour la recherche et la pratique en éducation. In L. Allal & L. Mottier Lopez (Ed.), *Régulation des apprentissages en situation scolaire et en formation.* Bruxelles : De Boeck.

Allal, L. (1999). Impliquer l'apprenant dans le processus d'évaluation : promesses et pièges de l'autoévaluation. In C. Depover & B. Noël, *L'évaluation des compétences et des processus cognitifs*, 35-56. Bruxelles : De Boeck.

Altet, M. (2000). L'analyse de pratiques : une démarche de formation professionnalisante ? *Recherche et Formation*, n° 30, 25-41.

Barbier, J.M. (1996). Introduction. In J.M. Barbier (sous la dir. de), *Savoirs théoriques et Savoirs d'action*.1-17. Paris : Presses Universitaires de France

Bélair, L. M. (1996). La formation à la complexité du métier de l'enseignant. In Paquay *et al., Former des enseignants professionnels*, 63-75. Bruxelles : De Boeck.

Bélair, L. M. (1999). *L'évaluation dans l'école, nouvelles pratiques.* Paris : ESF Editeur.

Bonniol, J.J. & Vial, M. (1997). *Les modèles de l'évaluation – Textes fondateurs avec commentaires*. Paris-Bruxelles : De Boeck.

Campanale, F. (1997). Autoévaluation et transformation de pratiques pédagogiques. *Mesure et évaluation en éducation*, Vol. 20 (n°1), 1-24.

Campanale, F. (2003). L'évaluation mutuelle interrogée par les principes d'efficacité et d'équité. *Actes du colloque international de l'Admée-Europe « Evaluation entre efficacité et équité »*, Université de Liège, 4-6 septembre, *Les Cahiers du Service de Pédagogie expérimentale*, n°s 15 et 16, cdrom, 251-262.

Crahay, M. & Detheux M. (2005). L'évaluation des compétences, une entreprise impossible ? *Mesure et Evaluation en Education*, 2005, Vol. 28 (n°1), 57-78.

De Ketele, J.M. & Gérard, F.M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences. *Mesure et Evaluation en Education*, 2005, Vol. 28 (n°3), 1-26.

Dolz, J. & Ollagnier, E. (Eds) (2000). *L'énigme de la compétence en éducation.* Paris-Bruxelles : De Boeck

Gillet, P. (1991). *Construire la formation : Outils pour les enseignants et les formateurs*. Paris : ESF.

Jorro, A. (2000). *L'enseignant et l'évaluation : des gestes évaluatifs en question*. Bruxelles : De Boeck.

Jorro, A. (2002). *Professionnaliser le métier d'enseignant.* Paris : ESF.

Gondrand, H. & Pierrard, A. (2001). *Contribution à la réflexion sur la professionnalisation des enseignants du premier degré en formation initiale. Thèse de doctorat en Sciences de l'éducation*. Université Lyon II - Louis Lumière.

Gondrand, H. (2004). La mémoire des pratiques : son rôle dans la construction de la professionnalité des professeurs d'école en formation initiale. *Recherche et Formation*, n°47, 127-140.

Le Boterf, G. (2000). *Construire les compétences individuelles et collectives.* Paris : Éditions d'organisation.

Meignant, A. (2003). *Manager la formation*. Rueil-Malmaison : Editions Liaisons.

Paquay, L. (1994). Prescrire l'auto-évaluation? Oui, mais... ! *Bulletin de l'Admée*, n˚94/2 – octobre.

Perrenoud, P. (1996). Le travail sur l'habitus ans la formation des enseignants. In L. Paquay *et al.*, *Former des enseignants professionnels*, 181-207. Bruxelles : De Boeck.

Perrenoud, P. (2001). Évaluation formative et évaluation certificative : postures contradictoires ou complémentaires ? *Formation professionnelle suiss*e, n° 4, 25-28. Article accessible à l'adresse suivante : [http://www.unige.ch/fapse/SSE/teachers/perrenoud/php_main/evaluation.html#A2001-2005]

Rogiers, X. (2004). *L'école et l'évaluation – Des situations pour évaluer les compétences des élèves*. Bruxelles : De Boeck.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences.* Bruxelles : De Boeck.

Schön, D. A. (1994). *Le praticien réflexif*. Montréal : Editions Logiques.

Référentiel minimal pour le repérage des compétences professionnelles en construction

(à partir du Référentiel du professeur des écoles stagiaire en fin de formation initiale).

Document de travail 18/01/2007

1. Concevoir une situation d'apprentissage en parvenant à :

a. définir un objectif d'apprentissage lié aux programmes et aux acquis des élèves ;

b. inscrire la situation dans une progression en anticipant les obstacles possibles ;

c. mobiliser les connaissances disciplinaires et didactiques adéquates ;

d. définir l'activité proposée à l'élève (durée, consignes, supports) ;

e. concevoir des évaluations critériées aux différents moments de l'apprentissage ;

f. à partir de l'analyse des résultats constatés, prévoir des activités de remédiation et d'approfondissement ;

g. concevoir des activités de consolidation (entraînement, mémorisation, réinvestissement).

2. Mettre en œuvre une situation d'apprentissage en parvenant à :

a. organiser l'espace scolaire en fonction des activités prévues ;

b. gérer le temps en fonction des rythmes d'apprentissage des élèves et des activités prévues (alternance entre les différentes phases de la séance) ;

c. gérer des modalités pédagogiques différentes en fonction de la diversité des élèves, des tâches et des formes de travail ;

d. utiliser de façon appropriée les supports, outils, aides diverses (tableau, documents écrits, audiovisuel, TICE) ;

e. choisir des formes de communication adaptées (posture, intervention orale, vérification des consignes) à la situation d'apprentissage et à la compréhension des élèves ;

f. permettre aux élèves de donner du sens aux apprentissages.

3. Maîtriser la relation pédagogique, et notamment l'exercice des responsabilités et de l'autorité, en parvenant à :

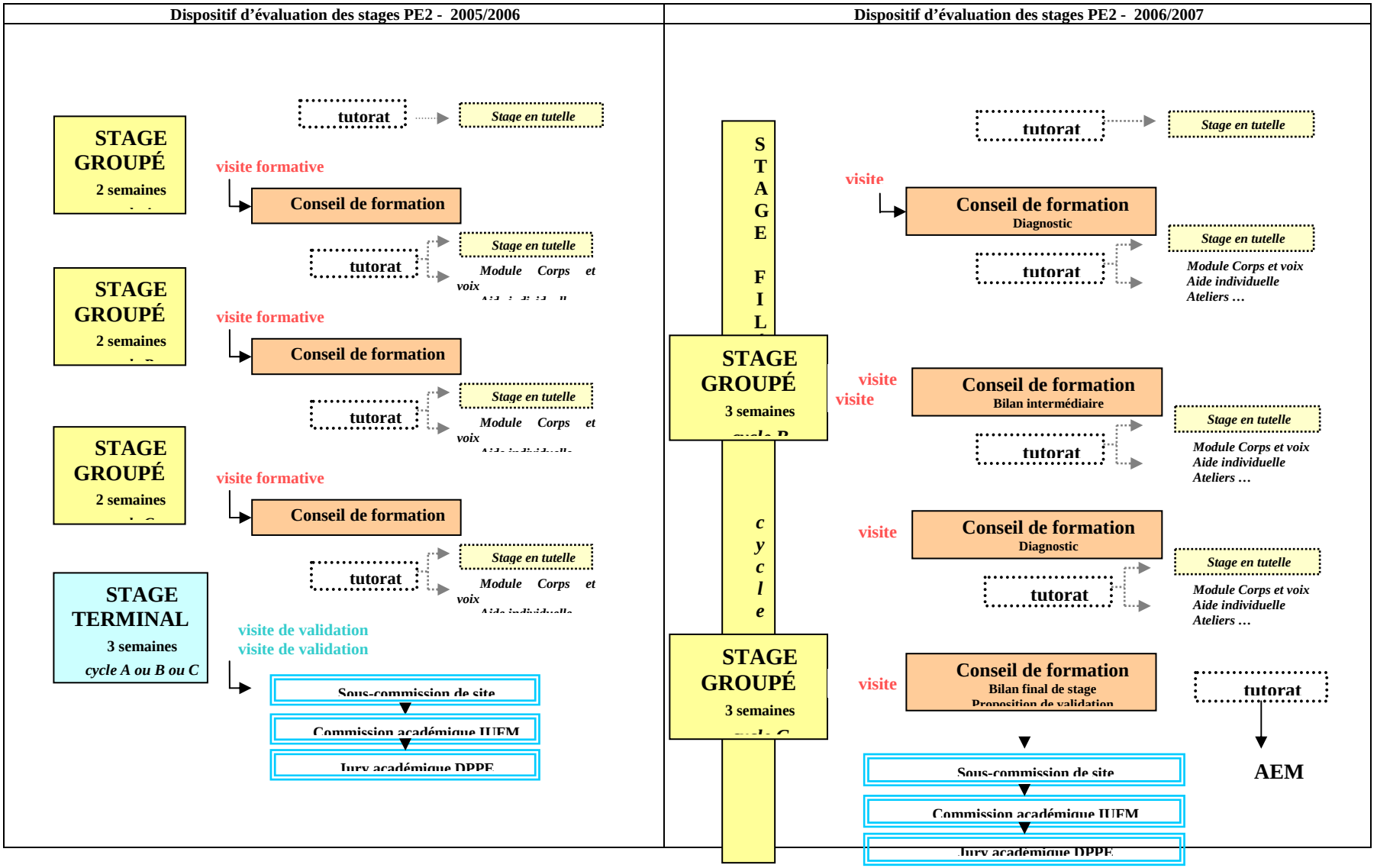
a. élaborer, faire comprendre et respecter les règles de vie collective (selon l'âge des enfants) ;
b. réguler la mise en œuvre et les réactions des élèves pour prévenir la survenue de situations mettant en jeu leur sécurité ;
c. faire comprendre et respecter les consignes de travail ;
d. savoir situer les élèves dans une dynamique de progrès, et les responsabiliser ;
e. se mettre à l'écoute des élèves et développer une écoute mutuelle dans la classe.

4. Evaluer et réguler une situation d'apprentissage en étant capable de :

a. être attentif aux réactions et aux réponses des élèves : modifier les activités prévues, relancer l'intérêt, exploiter une proposition intéressante, etc. ;
b. tirer parti collectivement de certaines erreurs favorables à l'apprentissage ;
c. analyser les erreurs observées en distinguant des causes possibles (obstacles épistémologiques, didactiques ou pédagogiques) ;
d. mettre en relation les différents aspects d'une situation d'apprentissage, notamment l'objectif, le déroulement, les résultats obtenus et le comportement des élèves ;
e. apprécier l'efficacité de son action, analyser sa pratique et en tenir compte pour les réguler *(formulation différente de celle du référentiel ministériel).*

Remarque

Ce référentiel minimal reprend la partie centrale du référentiel ministériel. Il ne reprend pas les compétences liées à la maîtrise des savoirs académiques et à l'éthique professionnelle, les considérant comme devant être d'emblée maîtrisées à la suite de la réussite a concours de recrutement.


Dispositif d'évaluation des stages PE2 - 2005/2006
STAGE GROUPÉ
2 semaines
tutorat
Stage en tutelle
visite formative
Conseil de formation
tutorat
Stage en tutelle
Module Corps et voix
STAGE GROUPÉ
2 semaines
visite formative
Conseil de formation
tutorat
Stage en tutelle
Module Corps et voix
STAGE GROUPÉ
2 semaines
visite formative
Conseil de formation
tutorat
Stage en tutelle
Module Corps et voix
STAGE TERMINAL
3 semaines
cycle A ou B ou C
visite de validation
visite de validation
Sous-commission de site
Commission académique IUFM
Jury académique DPPE
Dispositif d'évaluation des stages PE2 - 2006/2007
STAGE FILé cycle
STAGE GROUPÉ
3 semaines
tutorat
Stage en tutelle
visite
Conseil de formation
Diagnostic
tutorat
Stage en tutelle
Module Corps et voix
Aide individuelle
Ateliers …
visite
visite
Conseil de formation
Bilan intermédiaire
tutorat
Stage en tutelle
Module Corps et voix
Aide individuelle
Ateliers …
STAGE GROUPÉ
3 semaines
visite
Conseil de formation
Diagnostic
tutorat
Stage en tutelle
Module Corps et voix
Aide individuelle
Ateliers …
visite
Conseil de formation
Bilan final de stage
Proposition de validation
tutorat
AEM
Sous-commission de site
Commission académique IUFM
Jury académique DPPE

RECHERCHE SUR LA CONSTRUCTION DES COMPÉTENCES : DÉFIS DANS LA RÉVISION DES DISPOSITIFS D'ÉVALUATION DES STAGES

Colette Gervais, CRIFPE-Université de Montréal
Michel Lepage, CRIFPE-Université de Montréal

Mots-clés : construction de compétences; argumentation pratique; dispositif d'évaluation

Résumé :

Une compétence se définit comme la mobilisation de ressources en contexte, permettant de répondre adéquatement aux besoins de la situation (Le Boterf, 2001). Son évaluation ne peut donc être strictement basée sur un listage de comportements attendus ou d'indicateurs. Au Québec, la formation à l'enseignement est étalée sur plusieurs années, tout au long d'un programme comportant quatre stages. Que peut-on attendre comme manifestation des compétences pendant leur construction? Notre étude veut explorer comment se construisent les compétences pendant la formation. Quelles sont les ressources auxquelles les stagiaires disent faire appel? Comment ces stagiaires inexpérimentés les organisent-ils? La seule observation d'actions de stagiaires en classe ne peut suffire à rendre compte du développement des compétences : il faut avoir accès à leur raisonnement pédagogique sous-jacent. Nous avons utilisé l'approche de l'argumentation pratique (Fenstermacher, 1994) qui cherche à rendre explicites les prémisses d'action des praticiens et la mobilisation plus ou moins consciente des ressources dans l'action. Au total, 22 stagiaires, répartis de la 2e à la 4e année de formation à l'enseignement secondaire, ont participé à l'étude, de même que leurs enseignants associés et leurs superviseurs universitaires. Une approche vidéoscopique d'analyse de la pratique a été utilisée. Des rencontres collectives, par groupes d'acteurs d'un même stage (stagiaires, enseignants associés, superviseurs), ont contribué à faire un bilan de l'étude. L'analyse de l'ensemble des données recueillies permet de dégager certaines séquences dans la construction des compétences. La prise en compte de ces résultats dans l'évaluation des stages par les enseignants associés et les superviseurs suggère une révision des dispositifs afin de permettre l'explicitation des motifs de l'agir chez les enseignants en formation.

Problématique

Constituant un moment de formation charnière entre les milieux universitaire et professionnel, l'expérience de stage pose le problème de l'évaluation de manière particulière. Mentionnons quelques caractéristiques de ce contexte : encadrement par un formateur de chacun des deux milieux de formation; relation souvent étroite établie entre le stagiaire et le formateur « de terrain »; présence restreinte et ponctuelle du formateur universitaire, diminuant sa crédibilité dans le jugement final (Beauchesne, 2000) : diversité des contextes de stage, d'où la crainte d'une injustice à l'égard d'un stagiaire particulier (Toohey, Ryan et Hughes, 1996); difficultés liées

aux observations d'une pratique complexe en contexte. Outre la difficile conciliation d'un rôle de formateur et d'évaluateur bien documentée par la recherche (Ilott et Murphy, 1997 ; Pelpel, 1991), plusieurs défis sont posés aux évaluateurs : difficulté de faire des commentaires critiques au stagiaire (McIntyre, Byrd et Foxx, 1996; Martin, 2002), les praticiens ayant peu l'habitude de verbaliser leur pratique (Faingold, 1996); manque de précision à propos de l'objet à évaluer et difficulté à définir précisément le seuil attendu (Ilott et Murphy, 1997); rôle des valeurs dans la définition d'une bonne pratique (Myers, 1996) et subjectivité inhérente à l'évaluation de pratique, souvent tributaire d'impressions générales (Alexander, 1996). Ce dernier point paraît particulièrement sensible chez plusieurs formateurs, à la recherche d'instruments *objectifs*.

S'ajoute à cela la difficulté de l'évaluation de compétences professionnelles, la compétence étant définie comme la mobilisation en contexte de ressources permettant de répondre adéquatement aux besoins de la situation (Le Boterf, 2001). L'évaluation ne peut donc reposer strictement sur une liste de comportements attendus ou d'indicateurs. Les ressources mobilisées dans la pratique n'étant pas nécessairement observables, il faut prévoir un espace pour la verbalisation du stagiaire à ce sujet : cela pose la question, au moment de l'évaluation, de la part qui doit revenir au stagiaire qui se forme. Il semble que l'autoévaluation, souvent préconisée par le portfolio (Tellez, 1996), est de fait rarement utilisée dans la notation (Toohey *et al*, 1996).

Au Québec, les stages sont étalés sur quatre années, à raison d'un stage par année de formation. Une difficulté supplémentaire apparaît alors pour les évaluateurs : bien cerner et reconnaître ce qui peut être manifesté par un stagiaire à divers moments de la construction des compétences. Actuellement au Québec, les enseignants et les superviseurs universitaires accompagnent des stagiaires dans cette construction de compétences, doivent en attester la maîtrise avant l'entrée dans la profession, sans qu'on sache précisément quels indices relever. Quels repères peut-on fournir aux formateurs qui ont à porter un jugement? Il nous a semblé important de connaître à quelles ressources les stagiaires disent faire appel dans leur agir en classe, comment ils organisent ces ressources pour agir. La seule observation d'actions de stagiaires en classe ne peut suffire à rendre compte du développement des compétences. Il faut avoir accès à leur raisonnement pédagogique sous-jacent. Dans le cadre d'un programme de recherche, nous visons à comprendre comment se construisent les compétences pendant la formation. Dans cette communication, nous explorons certains défis soulevés par nos résultats quant aux dispositifs d'évaluation.

Cadre

Le Boterf (2001) fait remarquer que la recherche sur la construction de compétences porte sur des situations de travail ou de formation continue, non pas sur des situations de formation initiale. Un praticien organise les ressources dont il dispose à partir de schèmes d'analyse et d'action développés dans l'exercice du métier. La recherche a déjà montré que les novices n'ont qu'une vision partielle de la classe, qu'ils ont de la difficulté à distinguer l'essentiel de l'accessoire (Kagan, 1992).

Le Boterf (2002) fait l'hypothèse que leurs ressources, encore partielles, ne sont pas organisées en blocs mobilisables dans l'action.

Dans ce contexte, notre étude a voulu tenter d'identifier une séquence appréhendée de développement des compétences, à partir du bagage de ressources que transportent les stagiaires au cours de leurs différents stages. Le compte-rendu d'expériences pratiques vécues par les stagiaires eux-mêmes et leur explicitation nous sont apparus comme étant le moyen le plus légitime pour arriver à nos fins. Dans cette communication, nous faisons d'abord état des résultats obtenus dans cette étude avant de discuter leur implication dans la révision des dispositifs d'évaluation.

Méthodologie

Pour tenter d'avoir accès au raisonnement pédagogique des stagiaires, nous avons utilisé l'approche de l'argumentation pratique (Fenstermacher, 1987) qui cherche à rendre explicites les prémisses d'action des praticiens, ou les ressources plus ou moins conscientes qu'ils mobilisent dans l'action. Ces prémisses d'action peuvent être constituées des valeurs des praticiens, de savoirs théoriques qu'ils ont acquis, de savoirs tirés de l'expérience ou d'éléments du contexte dont ils tiennent compte pour intervenir.

Un total de 22 stagiaires, répartis de la 2e à la 4e année de formation à l'enseignement secondaire, ont participé à l'étude, de même que la plupart de leurs enseignants associés (15/19[1]) et leurs superviseurs universitaires (6). Les stagiaires ont enregistré leur enseignement sur vidéo et ont sélectionné des séquences qu'ils jugeaient signifiantes pour l'analyse en groupe. Ces extraits devaient être en lien avec certaines compétences regroupées sous l'appellation *acte d'enseigner* du référentiel de compétences (MEQ, 2001) à la base des programmes de formation à l'enseignement au Québec, plus particulièrement la *conception de situations d'enseignement-apprentissage*, le *pilotage* de ces situations et la *gestion du fonctionnement du groupe-classe*. Pour un stage donné, les rencontres d'analyse regroupaient tous les stagiaires d'une même école (entre 3 et 5 stagiaires), leurs enseignants associés, le superviseur universitaire et les chercheurs. Chaque stagiaire présentait ses séquences; les interventions du groupe l'amenaient ensuite à dégager les raisons pour lesquelles il avait agi de telle manière. Par la suite, des rencontres collectives, par groupes d'acteurs d'un même stage (stagiaires, enseignants associés, superviseurs), ont contribué à faire un bilan de l'étude, dans le cadre d'entrevues semi-structurées.

Résultats

L'analyse des données permet de dégager certaines séquences dans la construction des compétences, ainsi que d'identifier des *pivots* de leur développement. À titre d'exemple, la première expérience d'enseignement en contexte réel est l'occasion pour les stagiaires de progresser en prenant conscience des contingences de la réalité

[1] Les stages de 2e année se font parfois en dyades, deux stagiaires étant jumelés au même enseignant associé, ce qui était le cas de 6 stagiaires parmi les participants de cette étude.

du métier et surtout, de la nécessité de s'y adapter, l'*intelligence de la situation* (Le Boterf, 2001) qui se développe tout d'abord, qui se poursuit dans les stages ultérieurs par la prise en compte d'autres dimensions. Nous avons ainsi pu identifier une certaine séquence dans la construction des trois compétences étudiées.

Séquence dans la construction des compétences

Nos résultats nous mènent à poser l'hypothèse de phases, chez les participants à notre étude, dans la construction des trois compétences professionnelles retenues de l'*acte d'enseigner* : *concevoir* des situations d'enseignement-apprentissage, *piloter* ces situations, ainsi que *planifier, organiser et superviser* le fonctionnement de la classe. Ces phases seraient les suivantes :

1. Renouvellement de la compréhension de la compétence
2. Développement de schèmes d'action et de routines
3. Décentration graduelle du contenu à enseigner
4. Consolidation de la compétence
5. Compréhension raffinée de la compétence
6. Identité professionnelle assumée

Ainsi, l'expérience en contexte permet la confrontation à la réalité de la pratique de l'enseignement~~,~~ : les tentatives du stagiaire sont suivies de résultats chez les élèves, ce qui lui permet de développer une perception de plus en plus adéquate des conséquences de son action. Il apprend graduellement, dans un mouvement de décentration, à prélever des indices et à traiter cette information (Faingold, 1996). Par ailleurs, cette confrontation ne contribue pas seulement à des interventions de plus en plus adaptées au contexte mais également à une compréhension de plus en plus raffinée et pertinente de chacune des compétences. Se produirait alors, selon Tardif (2006), une réorganisation cognitive des connaissances qui conduit à des changements irréversibles soutenant la génération de nouvelles manifestations encore plus adaptées de la compétence.

Toutefois, l'expérience seule ne semble pas suffisante pour permettre la construction des compétences : elle doit s'accompagner d'une démarche réflexive. Une compétence pouvant se définir comme l'orchestration de ressources diverses, en situation (Perrenoud, Altet, Charlier et Paquay, 1996), la construction des compétences suppose une capacité de régulation à partir de l'expérience, à la fois par les rétroactions perçues de la situation elle-même, les commentaires d'observateurs ainsi que l'analyse que peut en faire le stagiaire. Ce passage d'un niveau de manifestation à un autre semble lié aux possibilités d'analyse et de recul, à travers diverses ressources auxquelles les stagiaires font appel.

Ces résultats nous suggèrent une re-conceptualisation de l'articulation entre divers dispositifs de manière à replacer au centre de l'évaluation des stages le discours de la personne qui se forme, sa sensibilité au contexte du stage et sa capacité à rendre explicites les motifs de son agir. À titre d'exemple, demander aux stagiaires comment

leur compréhension des compétences a évolué pourrait permettre de tenir compte non seulement d'habiletés qu'ils développent mais également des réorganisations cognitives qu'ils doivent faire, conduisant à concevoir des manières de faire encore plus adaptées (Tardif, 2006). Ainsi, les réflexions des stagiaires, sollicitées jusqu'ici à titre de manifestations de la compétence réflexive, pourraient être réorientées de manière à leur faire verbaliser explicitement les ressources qu'ils mobilisent dans leur pratique et à en décrire la mobilisation.

Parmi les pistes à explorer, nous privilégions, dans cette communication, l'axe des ressources mobilisées dans l'action. Les phases de la séquence proposée ci-dessus proviennent en grande partie des témoignages des stagiaires eux-mêmes sur les ressources qu'ils ont utilisées dans l'action : il semble que sur certains points, ils soient les seuls à pouvoir en témoigner. À l'instar de Tardif (2006), il nous semble que cette conscience des ressources mobilisées et éventuellement mobilisables soit essentielle dans le développement de la compétence, et donc de son évaluation. Nous présentons les ressources évoquées par les stagiaires à propos de la compétence *planifier, organiser et superviser* le fonctionnement de la classe, lors des stages de 3e et de 4e années.

Ressources évoquées

Un premier indice de l'importance de la verbalisation par les stagiaires eux-mêmes des ressources qu'ils disent mobiliser provient d'une comparaison des fréquences des énoncés portant sur les ressources par les participants de l'étude. Lors du stage de 3e et de 4e années, les stagiaires ont fait trois fois plus d'énoncés que leurs enseignants et leurs superviseurs, mentionnant également une plus grande variété de ressources. Il semble normal que les stagiaires soient les mieux placés pour parler des ressources qu'ils mobilisent, mais il est aussi plutôt rassurant de constater qu'ils sont conscients et capables de rendre compte des ressources mobilisées en faisant eux-mêmes référence à leur « trajectoire de développement » (Tardif, 2006).

Les stagiaires évoquent principalement l'importance du contact avec le milieu de pratique, « *le terrain…le meilleur endroit pour apprendre* », grâce notamment à la variété des situations qu'il présente et qu'il permet au stagiaire de rencontrer. Les stagiaires sont aussi amenés à vivre des expériences auprès d'une diversité d'élèves (d'écoles et de niveaux différents) dans des milieux aux cultures parfois même situées aux antipodes (écoles privée de banlieue et écoles urbaine multiethnique, par exemple), ce qui les oblige à adapter leur approche : « *J'ai appris une nouvelle sorte de gestion, parce que la façon que ça fonctionne ici, c'est vraiment différent de ce que j'ai vécu avant.* »

L'enseignant associé demeure, on ne s'en étonnera pas, une personne ressource privilégiée, autant dans le lien irremplaçable qui est créé avec lui que dans l'indispensable appui inscrit dans ses rôles même de guide et de modèle que rapporte aussi la littérature : supervision des planifications, suggestions, questions, rétroactions quotidiennes, encouragement, etc. Un stagiaire le souligne en ces termes: « *Une des contributions que je trouve primordiales, c'est que lui va te montrer les trucs, ce qui marche bien ou moins bien : regarde dans tel livre, tel site… Ça t'amène à devenir un*

meilleur enseignant plus vite… C'est sûr qu'après, c'est à toi à faire les démarches, à t'inspirer, mais au moins tu as une direction donnée par l'enseignant associé. » Celui-ci s'accommode, en fait, assez bien de ce rôle de première ressource, si on se réfère aux propos recueillis chez certains enseignants ayant participé à notre étude.

Même si les activités proposées par le superviseur universitaire (séminaires, visites en classe, rencontres de rétroaction) sont souvent vues comme une forme de contrainte rattachée au stage, il reste que les commentaires des stagiaires manifestent à leur égard une attitude plutôt positive, loin de la résignation obligée à laquelle certains auteurs font allusion (Borko, Mayfield, 1995). Les échanges sont vus comme des occasions de partage d'expériences, des explorations de nouvelles pistes complémentaires, ou encore comme des moments privilégiés pour constater sa propre évolution : « *C'était intéressant de parler ensemble, en séminaire, des problèmes de tous et chacun et des fois des bons coups aussi. Partager des expériences, des trucs, des activités, des stratégies.* » Les superviseurs interrogés ont, eux aussi, tendance à considérer les activités qu'ils organisent comme des moments importants de réflexion, entre pairs, « *l'occasion pour les stagiaires de profiter de l'expérience des autres stagiaires et de celle de 10 enseignants associés* », des moments encadrés par une personne ressource précieuse puisque bien au fait des contextes d'apprentissage de l'université comme du milieu scolaire.

Les cours théoriques ont aussi leur place dans la réflexion des stagiaires. Même s'ils jugent que les notions théoriques en gestion de classe font partie de leur bagage, les stagiaires affirment qu'en contexte, ils ne sont pas nécessairement capables de les utiliser, réagissant souvent par impulsion. Ils doivent y réfléchir et ce n'est qu'après coup que les liens apparaissent. Ces savoirs issus de la théorie sont surtout mentionnés lors des premiers stages, moins spontanément lors du dernier. Un stagiaire finissant affirme tout de même : « *C'est la formation théorique qui m'a amené à me poser ces questions. Mais, je ne me souviens pas exactement dans quel cours j'ai appris ça ou quelle est la théorie reliée à ça, mais ça soulève des questions.* »

Certaines ressources, enfin, appartiennent à ce qu'il serait convenu d'appeler le domaine personnel. Des stagiaires notent qu'une expérience préalable en animation de groupe, dans des colonies de vacances, par exemple, a été utile en certaines circonstances difficiles en facilitant une prise de décision rapide en classe.

Les enseignants semblent particulièrement sensibles aux ressources personnelles, mentionnant par exemple comme ressource importante, en gestion de classe, la confiance que manifestent certains stagiaires et à laquelle les élèves sont sensibles : « *Maturité, confiance en soi, solidité…ça transparaît* ».

Enfin, les stagiaires ont mentionné comme utiles au développement des compétences des conditions reliées aux stages lui-même et au projet de recherche. À titre d'exemple, un stage long comme celui de la 4e année (10 semaines, une étape scolaire complète) les oblige à régler les problèmes disciplinaires, donc à développer des stratégies efficaces. De même, l'occasion de se voir sur vidéo, d'observer des collègues stagiaires en action et d'échanger sur des situations concrètes a été jugée comme importante dans leur démarche réflexive.

Réflexion sur l'évaluation des compétences

Si les stagiaires sont les mieux placés pour parler des ressources qu'ils mobilisent en classe et pour en donner des détails, leurs formateurs peuvent aussi en identifier certaines. Les étudiants qui ont participé à l'étude sont très conscients, à la fin de leur formation, que leurs compétences ne sont pas entièrement développées. Ils se disent capables de concevoir et gérer des situations dans un certain contexte, celui qu'ils ont connu lors du stage : ils croient qu'un changement de milieu ou de types d'élèves, par exemple, pourrait les replacer dans la situation de devoir développer de nouvelles routines d'intervention avant d'atteindre le niveau de compétence manifesté à la fin de leur dernier stage. Ce qui remet un peu en question une affirmation du référentiel proposé par le MEQ (2001) pour la formation initiale des enseignants qui considère que la compétence devrait être manifestée de façon récurrente en fin de formation.

Conclusion

Les résultats de l'étude suggèrent qu'une réorganisation cognitive semble se produire chez les stagiaires durant leur cheminement vers la construction des compétences. Les réflexions sur la pratique, exigées des stagiaires dans l'évaluation de leur expérience de stage, devraient intégrer davantage la prise de conscience des ressources qu'ils mobilisent de manière à favoriser le renouvellement de la compréhension de la compétence. Comment les formateurs peuvent-ils évaluer si cette étape est franchie, à partir de quels indices?

L'exploration des impacts des résultats de notre étude sur les dispositifs d'évaluation n'est pas terminée, nous devons encore analyser certaines données et poursuivre la recherche, par exemple, pour déterminer les ressources auxquelles devraient normalement faire appel les stagiaires, dans l'action, ou encore pour mieux circonscrire les contextes et conditions de stage favorables à la construction de compétences. Par ailleurs, les étudiants qui ont participé à l'étude ont été particulièrement honnêtes et ouverts dans leur autoévaluation du niveau de compétence atteint. Le contexte de l'étude leur offrait des garanties quant à la confidentialité des propos et l'absence d'impact sur leur évaluation du stage en cours. Il faut se demander comment réagiront les étudiants si l'explicitation des ressources devient une dimension du portfolio à élaborer : feront-ils une réflexion réelle ou essaieront-ils de répondre aux attentes des formateurs, par exemple, en alignant des références théoriques ayant plus ou moins de liens avec leur pratique?

Références

Alexander, H.A. (1996). Physiotherapy student clinical education : the influence of subjective judgements on observational assessments. *Assessment and Evaluation in Higher Education, 21*(4), 357-366.

Beauchesne, A. (2000). Tirer parti de l'évaluation, mais mater le dispositif et le message : perspectives de stagiaires sur l'évaluation sommative des compétences. *In* C. Lessard et C. Gervais (dir.), *L'évaluation des nouveaux programmes de formation des maîtres*, 239-262. Publications de la Faculté des sciences de l'éducation de l'Université de Montréal.

Borko, H. et Mayfield, V. (1995). The roles of the cooperating teacher and university supervisor in tearning to teach. *Teacher and Teacher education, 11*(5), 501-518.

Clot, Y. (1999). *La fonction psychologique du travail.* Paris : Presses Unversitaires de France.

Faingold, N. (1996). Du stagiaire à l'expert : construire les compétences professionnelles. *In* L. Paquay, M. Altet, É. Charlier et P. Perrenoud (dir.) *Former des enseignants professionnels. Quelles stratégies? Quelles compétences?*, 137-153. Bruxelles : De Boeck.

Fenstermacher, G.D. (1987). Prologue to my critics, and a reply to my critics. *Educational Theory, 3*, 357-360, 413-422.

Fenstermacher, G.D. (1994). The knower and the known. The nature of knowledge in research on teaching. *In* L. Darling-Hammond (dir.), *Review of researche on education,* 1-54. Washington, DC : American Educational Research Association.

Ilott, I. et Murphy, R. (1997). Feelings and failing in professional training : The assessor's dilemma. *Assessment and Evaluation in Higher Education, 22*(3), 307-316.

Kagan, D,M. (1992). Professional growth among preservice and beginning teachers. *Review od Educational Research, 62*(2), 129-169.

Le Boterf, G. (2001). *Construire les compétences individuelles et collectives* (2e édition). Paris : Éditions d'Organisation.

Le Boterf, G. (2002). *Développer la compétence des professionnels* (4e éd.). Paris : Éditions d'Organisation.

Lepage, M. (2004). *Cadre de référence d'enseignants associés révélé lors de l'accompagnement et de l'évaluation de stagiaires en difficulté ou en échec.* Thèse de doctorat en psychopédagogie. Université de Montréal.

Martin, D. (2002). Terrain et théories dans la réforme de la formation des maîtres. *In* M. Carbonneau et M. Tardif (dir.), *Les réformes en éducation, leurs impacts sur l'école,* 95-109. Sherbrooke : Éditions du CRP.

McIntyre, D.J., Byrd, D.M. et Foxx, S.M. (1996). Field and laboratory experiences. *In* J. Sikula, T.J. Buttery et E. Guyton (dir..), *Handbook of Research on Teacher Education,* 171-193. New York : Macmillan.

Ministère de l'Éducation du Québec (2001). *La formation à l'enseignement. Les orientations. Les compétences.* Québec : Gouvernement du Québec.

Myers, C.B. (1997). *University-school collaboration : A need to reconceptualize schools as professional learning communities instead of partnerships.* Communication présentée à l'American Educational and Research Association, New York.

Pelpel, P. (1991). *Conseil et formation: guide pour les conseillers pédagogiques en IUFM.* Paris: CDRP.

Perrenoud, P., Altet, M., Charlier, É. et Paquay, L. (1996). Fécondes incerttudes. Ou comment former des enseignants avant d'avoir toutes les réponses. . *In* L. Paquay, M. Altet, É. Charlier et P. Perrenoud (dir.) *Former des enseignants professionnels. Quelles stratégies? Quelles compétences? ,* 240-253. Bruxelles : De Boeck.

Tardif J. (2006). *L'évaluation des compétences. Documenter le parcours de développement.* Montréal : Chenelière Éducation.

Tellez, K. (1996). Authentic assessment. *In* J. Sekula, T. J. Buttery et E. Guyton (dir.), *Handbook on Research on Teacher Education* , 704-721. New York : Simon and Shuster MacMillan.

Toohey, S., Ryan, G. et Hughes, C. (1996). Assessing the practicun. *Assessment and Evaluationin in Higher Education, 21*, 215-227.

UN DISPOSITIF FORMATIF D'ACCOMPAGNEMENT DES STAGES DES FUTURS INSTITUTEURS : QUELS APPORTS AU NIVEAU REFLEXIF ?

Catherine Van Nieuwenhoven, UCL-ISPG et Eric Bothy, ISPG

Mots clés : supervision, professionnalisation, analyse réflexive

Résumé :
Depuis la réforme de la formation initiale des enseignants en Belgique francophone, les pratiques d'enseignement dans le supérieur sont centrées sur le développement de compétences professionnelles autour d'un référentiel commun.

Un dispositif centré sur l'alternance n'étant formateur qu'à la condition d'une véritable articulation théorie-pratique, la remise en cause de la nature et du statut des stages est devenue incontournable. Ainsi, pour soutenir l'étudiant dans sa démarche active d'intégration, un dispositif d'accompagnement formatif et personnalisé de la pratique a été mis en place, axé sur le développement de la pratique réflexive. Ce dispositif s'inscrit globalement dans le courant des recherches innovantes sur le développement de dispositifs professionnalisants et autour du modèle du praticien réflexif (Dejemeppe et Dezutter, 2002; Paquay, Altet, Charlier et Perrenoud, 1998, Donnay & Charlier, 2006 ; ...). Ce dispositif a été mené pour la première année auprès de 70 étudiants et a fait l'objet d'une évaluation par questionnaire auprès des étudiants et des formateurs. Ces résultats seront discutés et le dispositif questionné. Plusieurs questions se profilent déjà: comment gérer l'ambiguïté des rôles des deux types de partenaires ? Le superviseur formatif n'est-il pas en définitive le mieux placé pour évaluer la maîtrise des compétences de l'étudiant ? Quelles expertises pour les formateurs et quelles modalités d'accompagnement prévoir ? Ce dispositif permet-il le développement d'un praticien-réflexif ? (Gervais, 2005).

Problématique

Un des défis majeurs de la formation initiale des enseignants est d'assurer une cohésion entre la formation théorique organisée par l'institution de formation et le champ diversifié de la pratique. Selon Charlier (2001), l'enseignant peut apprendre à partir de la pratique dans la mesure où celle-ci constitue le point de départ et le support de sa réflexion (réflexion sur l'action), que ce soit sa propre pratique ou celle de ses collègues. Le futur enseignant apprend par la pratique. Confronté à la réalité qui lui résiste, l'enseignant se positionne en acteur, c'est-à-dire comme quelqu'un qui peut agir sur les caractéristiques de la situation, expérimenter des conduites nouvelles et découvrir des solutions adaptées à la situation. Dans le même sens, Schön (1994) souligne que le professionnel développe essentiellement ses compétences dans la pratique et à partir de la pratique.

Selon Perrenoud (2001), la notion d'alternance n'interdit pas de penser et surtout de vivre les rapports entre les stages et les autres moments de la formation sur le mode d'une pure juxtaposition. Bikorindagara (2002) parle d'alternance de

type « intégration » quand la théorie et la pratique alternent dans un rapport interactif au sein d'une relation « circulaire ». L'étudiant est encadré par un praticien réflexif et des « garants » du cadre théorique. Nous pensons que c'est ce type d'alternance qu'il nous faut privilégier dans la formation initiale des instituteurs. En effet, si l'on veut des instituteurs capables d'adapter leur enseignement aux événements de la vie de la classe, d'avoir un regard critique sur leur pratique, d'expliciter leur pratique et de rechercher des solutions aux problèmes qu'ils rencontrent à l'aide de la théorie, seule une alternance de type intégration par ses allers-retours entre la pratique et la théorie nous semble pouvoir développer de tels gestes professionnels. Ce modèle d'alternance a donc servi de fil conducteur à notre réflexion et a stimulé la remise en cause de la nature et du statut des stages. La question des rôles des évaluateurs a directement été posée : est-il un accompagnateur, un évaluateur externe, un coach, un catalyseur d'émotions, un garant de l'évaluation certificative ? Peut-il raisonnablement se positionner dans une posture d'évaluateur formatif et prévoir des pistes d'intervention tout en délivrant une note chiffrée qui atteste d'une réussite ou d'un échec du même stage ?

L'accompagnement des stages

En Belgique, la formation des enseignants qui se destinent à l'enseignement fondamental se réalise en trois ans dans le cadre des Hautes écoles. Depuis six ans une réforme[1] importante de l'enseignement supérieur pédagogique a vu le jour. Cette réforme s'articule autour du développement de 13 compétences au coeur du métier d'enseignant. Depuis lors, des instituteurs de terrain sont venus rejoindre l'équipe des formateurs dans le cadre des ateliers de formation pratique (AFP) organisés pour favoriser les liens entre la théorie et la pratique.

Perrenoud (2001a) distingue quatre modèles de stages qui peuvent coexister durant le parcours de formation et même au sein d'un même stage. Pour lui, le stage peut être :

1. Un terrain d'application sur lequel les étudiants nantis d'une « doxa » et de théories vont appliquer les principes appris lors des cours théoriques,
2. Un moment de socialisation professionnelle où l'étudiant apprendrait sur le tas, par essais et erreurs.
3. Une épreuve initiatique dans laquelle l'étudiant vivrait une expérience formatrice.
4. Une composante d'une démarche clinique de formation où l'étudiant pourra, en adoptant une posture réflexive, construire en situation des solutions aux problèmes qu'il rencontrera en classe. L'étudiant développant ses compétences au fur et à mesure de l'avancement des stages, de la pratique réflexive qu'il met en place et de l'apport théorique que lui fournissent les formateurs sur les problèmes qu'il a rencontrés.

En accord avec Perrenoud (2001), seul le quatrième modèle est défendable dans l'optique de la professionnalisation et de la pratique réflexive. «Les situations

[1] Décret du 12 décembre 2000 de la communauté française de Belgique.

complexes, toujours singulières, appellent une démarche de résolution de problème plutôt que l'application d'un répertoire de recettes, ou de réponses préprogrammées, ou encore le recours à un algorithme fondé sur une connaissance théorique générale. Toute normalisation de la réponse, qu'elle s'ancre dans une doxa, une théorie ou une tradition, entraîne un affaiblissement de la capacité d'action et de réaction du praticien» (Perrenoud, 2001, p8).

Présentation du dispositif

L'ensemble de ces réflexions a poussé un nombre important de professeurs à réfléchir à la nature de l'évaluation des stages. Très rapidement est apparue la nécessité de séparer les rôles formatifs et certificatifs et de prévoir des interlocuteurs différents pour chaque étudiant. C'est ainsi que deux rôles différents ont été établis au sein des intervenants de l'institut de formation, l'un (le superviseur) centré sur l'accompagnement formatif des étudiants tout au long de l'année et l'autre (l'évaluateur) qui atteste la maîtrise des compétences visées par chaque stage.

Différentes étapes ont également été définies pour baliser le processus d'accompagnement de la pratique : la définition d'objectifs personnalisés avant le stage lors d'un premier entretien, le suivi sur le terrain pendant le stage y compris un entretien « à chaud » après la visite, un entretien réflexif et conseil d'évaluation après le stage, mettant les différents partenaires en présence.

Le premier entretien vise à faire connaissance et à créer un climat de confiance entre le superviseur et le supervisé. Cette rencontre doit permettre l'expression des appréhensions de départ, des attentes mutuelles, des préoccupations et des besoins en plus de l'échange d'informations (Tardif et Ziarko, 1997). Elle doit témoigner de l'engagement conscient à cheminer ensemble dans une relation ouverte, complexe et impliquante (Pelpel, 2002). La qualité de la relation (partage, confiance, collaboration, entraide) conditionne l'efficacité du processus de rétroaction; le superviseur doit convaincre de son désir de venir en aide avant même d'en prouver sa capacité. La communication personnalisée réelle et profonde s'établit grâce à une écoute active permettant à chacun de se sentir sécurisé et valorisé. La dimension synergique fait appel aux ressources réciproques : le superviseur se place lui-même en situation d'amélioration de sa supervision, donc en position d'être aidé par le supervisé dans une relation de rétroaction réciproque exempte de jugement.

Lors des visites, le superviseur donne un premier feed-back à l'étudiant. Il tente d'amener l'étudiant, par un questionnement orienté, à une prise de conscience de son fonctionnement et de son action pédagogique. L'entretien réflexif après le stage est une étape-clé du dispositif. Il s'agit d'un temps d'accompagnement du stagiaire qui l'aide à analyser une pratique professionnelle qu'il a vécue en situation dans le but de comprendre le déroulement de l'activité, de prendre conscience des effets de ses interventions et, in fine, de dégager les priorités pour ajuster ses pratiques futures (Wibaut, Dourte et Paquay, 2002).

Quatre périodes se dessinent lors de l'entretien avec l'étudiant:

- une description objective de la situation, des activités (questionnement ouvert sur les

faits, la démarche, les interactions, l' « observable »),
- une analyse de la situation (aider à mettre des mots sur ce qui s'est passé, aider l'étudiant à expliciter ce qu'il perçoit qui a bien ou moins bien fonctionné),
- une interprétation de la situation (explicitation des choix et comprendre les effets des actions menées),
- une étape créative de mise en projet (dégager les priorités pour l'activité d'enseignement future en précisant les critères d'observation-évaluation).

Le conseil d'évaluation est une modalité de supervision qui met en présence le stagiaire, le superviseur, le maître de stage et l'évaluateur dans le cadre du bilan de stage. Cela permet de diminuer la subjectivité et de permettre une concertation entre les acteurs. Pour l'ensemble de la démarche d'accompagnement, le superviseur assume auprès du stagiaire un rôle de médiateur inscrit dans la durée (1 an) pour lui permettre de construire son identité professionnelle. Il s'agit bien d'un processus d'accompagnement d'une personne en développement professionnel. Il importe donc qu'une interaction positive s'établisse rapidement.

Une attention particulière est portée au concept de supervision. Selon Legendre (1993), il s'agit d'un processus systématique d'amélioration de la qualité de l'enseignement qui se centre sur des opérations critiques d'observation, d'analyse et d'interprétation des pratiques. Pour Gervais et Laporte (1992), c'est un processus d'accompagnement et d'aide à une personne en développement professionnel, prenant en considération des dimensions cognitives de l'ordre du savoir-faire et des dimensions affectives. Enfin, pour Rousseau et Saint-Pierre (2002), il s'agit d'une activité de coopération complexe d'observation critique, d'analyse et d'interprétation de la pratique de l'enseignement des stagiaires dans le respect des politiques institutionnelles et des finalités de l'éducation et d'un acte de construction et de co-construction de la pratique de l'enseignement qui repose sur l'encadrement, le mentoring et la rétroaction à saveur pratique et théorique.

Dans la littérature, les rôles d'un superviseur se définissent comme suit : favoriser la démarche réflexive et la résolution de problème (Gervais, 1997), favoriser la réflexivité sur la dynamique d'un parcours (Hensler, Garant et Dumoulin, 2001), prendre un rôle de médiateur entre l'école primaire et l'institution de formation (Gervais,1997), être un support à la pratique réflexive (Donnay et Charlier, 2000), se positionner comme un « ami critique », sans jugement, dans un objectif de régulation et d'amélioration (De Ketele, 1999 ; Costa & Kallick, 1993), prévoir un accompagnement personnalisé des étudiants (Behrens,2002) et être le garant de l'articulation théorie-pratique (Maes, 2006). Les résultats recueillis nous permettront de vérifier les rôles perçus par les différents acteurs.

Recherche empirique

Cadre méthodologique

Ce dispositif a été testé pour la première fois auprès d'étudiants de troisième année. Il a fait l'objet d'une évaluation par questionnaire auprès des étudiants et des formateurs en fin d'année. L'échantillon est constitué de 48 étudiants sur 68 et de tous les superviseurs (N=8). Deux questions de recherche ont guidé nos analyses dans le cadre de cette communication :

- analyser le système d'encadrement des stages des futurs instituteurs primaires et l'influence de celui-ci sur le profil de formation mis en place,
- évaluer les changements perçus par les acteurs (étudiants et superviseurs) quant au nouveau dispositif d'accompagnement des stages.

Présentation des résultats

Hypothèse 1 : Les étudiants et les superviseurs perçoivent le rôle formatif du superviseur comme condition nécessaire au développement d'un praticien réflexif

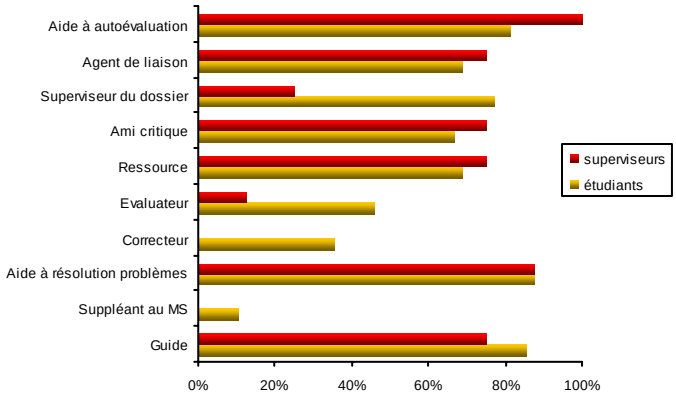


Figure 1 : Perception du rôle formatif du superviseur par les acteurs

Pour les étudiants, les principaux rôles du superviseur sont l'aide à la résolution des problèmes rencontrés, le rôle de guide, d'aide à l'autoévaluation de l'étudiant et de superviseur du dossier. Pour les superviseurs, leurs fonctions principales sont celles d'aide à l'autoévaluation de l'étudiant, d'aide à la résolution des problèmes rencontrés, d'ami critique, de ressource pour l'étudiant, d'agent de liaison entre la haute école et l'école primaire et de guide pour l'étudiant.

Ces résultats sont en accord avec le quatrième rôle attribué au stage par Perrenoud à savoir une démarche clinique de formation où l'étudiant peut, en adoptant une posture réflexive (auto-évaluation), construire en situation des solutions aux problèmes qu'il rencontre avec l'aide du superviseur (aide à la résolution de problème). Le rôle formatif du superviseur (guide) est clairement identifié par les deux acteurs.

Les différents rôles dévolus aux superviseurs dans la littérature nous conduisent à isoler trois dimensions :

- La première sera la dimension relationnelle qui unit le superviseur et le stagiaire dans le dispositif d'accompagnement des stages,
- La seconde sera la dimension méthodologique qui correspond au travail d'aide pédagogique que fournit l'accompagnateur de stage,
- La dernière sera la dimension institutionnelle, qui regroupe les tâches inhérentes à l'organisation des stages comme agent de liaison sur les rôles d'un superviseur

Lorsque l'on demande aux étudiants de définir par un verbe d'action le rôle principal des superviseurs, ils nous répondent : motiver (25%) et écouter (37%) dans la dimension relationnelle de l'accompagnement et guider (50%), faire réfléchir (37%), encadrer (25%), observer (25%), analyser (12%) et critiquer (12%) dans la dimension méthodologique, aucun verbe de la dimension institutionnelle n'étant cité. Quant à eux, les superviseurs pensent soutenir (23%) et écouter (19%) les étudiants dans la dimension relationnelle et conseiller (52%), guider (40%), aider (42%), critiquer (17%), accompagner (12%) et critiquer (12%) au niveau méthodologique. Dans la dimension institutionnelle intervient avec un pourcentage plus faible un rôle de contrôle (12%) et d'information (12%).

Ces résultats sont en accord avec les qualités d'un bon superviseur identifiées par les acteurs : à l'écoute (46%), disponible (40%), ouvert (33%), juste (29%) et motivé (10%) pour les étudiants dans la dimension relationnelle et empathique (50%), rigoureux (37%), disponible (25%), critique (25%) et confiant (12,5%) pour les superviseurs

Globalement, les trois dimensions relevées dans la littérature caractérisant les rôles d'un superviseur qui s'inscrit dans un projet de développement du praticien réflexif chez les étudiants se retrouvent bien mais à géométrie variable. Les acteurs s'accordent sur la nécessaire qualité de la relation établie entre les deux partenaires (Gervais, 2005; Pelpel, 2002) pour garantir l'efficacité de l'accompagnement réflexif. Quant à la dimension méthodologique, elle est relevée massivement par les étudiants et les superviseurs au niveau des verbes d'action mais est seulement mise en évidence par les superviseurs au niveau des qualificatifs. L'aide à la résolution de problème et à la démarche réflexive (Gervais, 1997, 2005), le support à la pratique réflexive (Donnay & Charlier, 2000, 2006), le rôle « critique » proposé par De Ketele (1999) caractérisant les rôles méthodologiques du superviseur par les auteurs sont perçus par les acteurs du dispositif. La dimension institutionnelle est, quant à elle, peu présente du côté du superviseur. Dans notre dispositif, ce rôle est plutôt dévolu à l'évaluateur ou à l'administration.

Les étudiants reconnaissent que le superviseur leur est indispensable pour progresser dans leur pratique. Certains étudiants ont même confié que la présence du superviseur leur manquera l'année prochaine et qu'ils aimeraient pouvoir retourner de temps en temps dans l'institution pour y être conseillé. Serait-ce les prémisses d'un lien possible entre la formation initiale et continue ? N'est-ce pas un indice clair de reconnaissance du rôle formatif du superviseur ?

Place de l'évaluation certificative

La dimension formative du rôle des superviseurs étant confirmée, qu'en est-il de leur éventuelle implication au niveau du volet certificatif ? Rappelons qu'à l'origine du projet, la responsabilité de la validation des compétences est du ressort unique de l'évaluateur.

Selon les superviseurs et les étudiants, les indicateurs servant de base à l'évaluation finale du stage d'un étudiant sont principalement le rapport du maître de

stage, l'avis du superviseur, l'autoévaluation de l'étudiant, la farde de préparation de l'étudiant, le rapport de visite de l'évaluateur. Ne semblent pas pris en compte pour fixer la cote du stage : le carnet de bord de l'étudiant, considéré comme un des outils d'évaluation formative et les entretiens de l'évaluateur avec l'étudiant. Les étudiants et les superviseurs semblent donc être en accord sur la plupart des indicateurs d'évaluation du stage. Le rôle certificatif de l'évaluateur est clairement identifié et les supports également (rapport de visite, rapport du maître de stage et farde de stage). Néanmoins, les avis divergent sur l'importance de l'autoévaluation soulignée essentiellement par les superviseurs (87,5%) et peu par les étudiants (60%) et le conseil d'évaluation identifié comme critère certificatif pour 62,5% des superviseurs et seulement 31 % des étudiants.

Cette cohérence globale observée entre les étudiants et les superviseurs concernant les critères d'évaluation certificative est encourageante et témoigne de la précision de ceux-ci. La divergence observée au niveau de l'auto-évaluation et le rôle du conseil d'évaluation est révélatrice d'un manque de clarté au niveau de l'implication et la responsabilité des superviseurs et questionne le dispositif sur l'ambiguïté de leur statut.

Dans la mesure où l'évaluation formative a pour principal objectif la régulation des apprentissages de l'étudiant, on comprend aisément qu'elle soit le moteur du dispositif d'accompagnement par les superviseurs mais pourquoi ne pourrions-nous pas en fin de parcours mettre au service de l'évaluation certificative la qualité des observations menées tant par l'étudiant, que par les pairs et les superviseurs ? Le superviseur, par son contact étroit avec l'étudiant, est véritablement le mieux placé pour certifier les compétences finales de l'étudiant. Il a effectivement, selon Perrenoud, eu tout loisir de construire une représentation précise et analytique des ressources qu'il a accumulées aussi bien que de sa capacité à les mobiliser en situation, puisqu'il l'a vu maintes fois en train de réfléchir et d'agir, aux prises avec des tâches diversifiées, dans les conditions les plus habituelles du travail manuel ou intellectuel (2001, p.27). Pour garantir l'efficacité et l'équité de l'évaluation certificative, il est essentiel par contre de définir un contrat clair avec les différents intervenants dès le démarrage du dispositif, comme premier ciment de la collaboration.

L'auto-évaluation est un des ingrédients réflexifs clés du dispositif dans une perspective formative, peut-elle dès lors constituer un support d'évaluation certificative ? N'est-ce pas un excellent révélateur de la progression de l'étudiant ? Une des compétences du décret étant centrée sur le développement de l'analyse réflexive, n'est-ce pas l'outil idéal ? Qui plus que le superviseur est le mieux placé pour valider la maîtrise de cette compétence tout au long de l'année ? Quel est le lieu d'expression du superviseur pour communiquer cette information sur les compétences réflexives de l'étudiant, si ce n'est le conseil d'évaluation ? Est-ce incompatible de porter les deux casquettes ? La question reste ouverte.

En regard de l'hypothèse 2, nous allons maintenant vérifier que les ingrédients

réflexifs du dispositif d'encadrement des stages sont perçus comme tels par les étudiants et les superviseurs.

Entretiens préalables au stage

Avant le stage et suite aux observations menées par l'étudiant, celui-ci, le maître de stage et le superviseur se rencontrent à l'école de stage pour finaliser le contrat de stage. Celui-ci comporte deux volets : l'un non négociable et l'autre personnalisable et permet à l'étudiant d'adapter son stage en fonction de ses besoins de formation, de son projet personnel, ... Ce volet n'a pas fait l'objet d'évaluation dans l'enquête mais est reconnu par les deux acteurs comme essentiel pour démarrer la collaboration sur de bonnes bases, renforcer le maître de stage dans son rôle et établir un contrat clair de collaboration.

Visite et entretien à chaud

23% des étudiants estiment recevoir un feed-back écrit de leur superviseur après la visite alors que la moitié des superviseurs prétend le faire à chaque visite de stage. Par contre, la majorité des étudiants estime toujours recevoir un feed-back oral après la supervision d'une de leur leçon et la totalité des superviseurs dit toujours en donner un. Nous pensons pourtant que les deux ont leur place dans un dispositif d'accompagnement des stagiaires. L'entretien oral est interactif, centré sur la dimension relationnelle et méthodologique tandis que l'écrit permet une réflexion plus approfondie et surtout laisse une trace aux partenaires.

Entretien réflexif

Le tableau ci-dessous met en parallèle les réponses des étudiants et des superviseurs concernant les objectifs poursuivis par l'entretien réflexif.

Tableau 1 : Objectifs poursuivis par l'entretien réflexif

Propositions	**Avis des étudiants**	**Avis des superviseurs**
Reformuler les propos des étudiants	56 %	62%
Valoriser les tentatives d'innovation	75%	100%
Indiquer d'autres pistes possibles	73%	75%
Soutenir moralement les étudiants	79%	87%
Mettre à l'aise	92%	100%
Nommer les préoccupations des étudiants	69%	100%
Ne pas émettre de jugement	71%	87%
Faire le point	87%	90%

Les dimensions relationnelles et méthodologiques de l'accompagnement formatif sont à nouveau relevées par les étudiants et les superviseurs avec toutefois de légères différences d'appréciation. Les interventions du superviseurs soulignées sont en phase avec les quatre étapes visées par l'entretien réflexif: une description objective de la situation, une analyse de la situation, une interprétation de la situation et une étape créative de mise en projet pour identifier les perspectives

d'amélioration. A la question de l'utilité de l'entretien réflexif, les étudiants répondent: voir ce qu'il faut améliorer (31%), avoir un retour réflexif (21%) et faire le point (21 %).

Le conseil d'évaluation

85% des étudiants pensent que le conseil d'évaluation joue son rôle réflexif. 80% des étudiants pensent que le conseil d'évaluation les aide à définir les objectifs d'amélioration de leurs pratiques professionnelles (33% en sont convaincus, 47 le pensent). La majorité des superviseurs relèvent à nouveau l'ambiguïté de leur rôle lors de ce moment à la fois formatif (pistes d'amélioration) et certificatif (points attribués).

Globalement, les étudiants disent bien percevoir l'importance de la supervision dans leur formation (92%). 19% disent néanmoins qu'ils n'ont pu profiter pleinement de la supervision formative par peur de la certification. L'ambiguïté des rôles du superviseur se profile à nouveau. 52% des étudiants sont tout à fait convaincus que le dispositif d'encadrement favorise une approche réflexive et critique de leur pratique. 43% pense que le dispositif de stage y contribue. Pour les étudiants, le dispositif vise essentiellement à les aider à identifier les points à améliorer. L'accent sur la régulation comme élément-clé de l'évaluation formative est donc bien perçu !

Hypothèse 3 : Les étudiants perçoivent le changement dans le dispositif d'accompagnement des stages à la Haute Ecole Galilée en faveur du praticien réflexif.

Questionnés sur les améliorations du dispositif d'encadrement des stages, 46% des étudiants trouvent que le nouveau système développe plus le paradigme du praticien réflexif, 27% d'entre eux pensent le contraire et 27% restent indécis. Malgré ce résultat mitigé, à 98% des étudiants continueraient le nouveau projet d'encadrement des stages. Quelles interprétations donner à ce résultat ? 92% d'entre eux disent bien percevoir l'importance de la supervision dans les stages pour favoriser une approche réflexive et critique de leur pratique et par ailleurs, seul la moitié d'entre eux identifie positivement le changement de dispositif en faveur du praticien réflexif. Or, les ingrédients réflexifs du dispositif sont largement soulignés par les étudiants un accompagnateur formatif jouant clairement la carte de la réflexivité, des visites de stage formatives, une autoévaluation de l'étudiant, un carnet du maître de stage jouant le rôle de journal de bord, la présence d'entretien réflexif, un conseil d'évaluation pendant lequel l'étudiant est de nouveau amené à adopter une posture réflexive…

Deux hypothèses se dégagent directement : les étudiants ne sont pas assez familiers avec le modèle de professionnalité proposé par Paquay (1994) et n'identifie pas les caractéristiques du praticien réflexif.ou le dispositif a été vécu pour la première année et les acteurs ayant vécu l'ancien dispositif ont besoin d'un temps d'adaptation au changement.

Le temps du bilan

Identification des faiblesses du dispositif

Des raisons ont été évoquées par les acteurs pour justifier les faiblesses relevées dans le dispositif :

- *Dispositif en rodage* : le dispositif est jeune et à sa première année d'utilisation.
- *Dispositif « pensé » et « agit » en même temps* : le constat d'inadéquation du dispositif précédent a donné lieu directement à un changement important de perspectives à la veille de la rentrée académique sous l'impulsion des 8 superviseurs volontaires.
- *Résistances des étudiants au changement* : les étudiants ayant connu un autre dispositif d'accompagnement de leur stage, certains ont eu des difficultés à prendre la posture du praticien réflexif qui implique un investissement important de leur part. Les étudiants identifiés comme bons techniciens ou maîtres artisan (Paquay, 1994), par exemple avaient toujours eu d'excellents résultats en stage et se trouvent par contre en difficulté pour verbaliser leurs pratiques et prendre du recul. Certains d'entre eux avouent avoir eu des difficultés à identifier le projet.
- *Manque d'outils pour soutenir l'analyse réflexive* : les superviseurs reconnaissent ne pas toujours disposer des outils nécessaires pour accompagner les étudiants dans la réflexion critique et le passage à l'écrit. En effet, la posture de l'étudiant réflexif ne s'improvise pas et l'écriture ne constitue pas en soi un outil de réflexivité si elle n'est pas travaillée dans cette perspective.
- *Accompagnement des superviseurs* : les superviseurs ont identifié le changement de rôle en tant qu'accompagnateur formatif d'un processus réflexif et ne se sentent pas tous compétents et à l'aise dans leur nouveau costume. La formation de départ des superviseurs est variée (psychopédagogue, didacticiens et disciplinaires) et ne prépare pas nécessairement aux dimensions tant relationnelles (gestion de la relation, de la communication, …) que méthodologiques (aide à la résolution de problème, ami-critique, …).
- *Ambiguïté du rôle du superviseur* : une clarification des rôles du superviseur s'impose. En tant qu'accompagnateur formatif, son attention est portée sur les progrès de l'étudiant dans une perspective de régulation mais tant pour les étudiants que pour eux-mêmes, ils estiment être bien placés pour se positionner sur la maîtrise finale des compétences de l'étudiant en lien avec le référentiel. Comment gérer alors au quotidien la double casquette du superviseur ?

Propositions d'amélioration du dispositif

Les étudiants ont proposé plusieurs pistes pour augmenter l'efficacité du dispositif : augmenter la présence des superviseurs sur le terrain, prévoir plus de temps pour le feed-back sur place, clarifier le rôle du superviseur, impliquer le superviseur dans l'évaluation certificative et avoir le choix de son superviseur. Les superviseurs, quant à eux, ont dégagé plusieurs axes de travail pour amender le dispositif avant de le généraliser aux autres sections : augmenter leur temps de visites,

clarifier le rôle du superviseur, établir des liens entre les superviseurs de 2ème et 3ème année, former les superviseurs à leur nouveau rôle, favoriser les entretiens réflexifs en petits groupes (Wolf et Kietz, 1998 ; Behrens, 2002) et construire ou adapter des outils d'accompagnement réflexif.

Pistes de discussion

Les résultats recueillis ont apporté certaines réponses à nos questions de départ mais certaines d'entre elles restent au cœur du débat : Si l'on considère que le superviseur reste le mieux placé pour évaluer la maîtrise des compétences professionnelles de l'étudiant et sa progression, comment éviter que les enjeux certificatifs ne pervertissent le caractère formatif de ce dispositif ? Comment encourager les étudiants à l'auto-évaluation, ingrédient essentiel du praticien réflexif, sans la rendre obligatoire ? (Paquay et Saussez, 1994). Faut-il viser l'uniformisation des compétences (profil type) des superviseurs ou au contraire mettre en valeur des spécificités de chacun dans l'accompagnement relationnel et méthodologique de l'étudiant ? Les superviseurs sont-ils tous prêts à jouer le rôle de « coach » (Gervais, 2005), d'« ami critique » (De Ketele, 1999) ? Sont-ils bien formés ? Ne devraient-ils pas aussi être engagés dans un processus d'analyse de leurs pratiques ? Toutes ces questions restent ouvertes et nous motivent à aller de l'avant.

Références bibliographiques

Behrens, M. (2002). Questions vives : évaluation et formation, Les Sciences de l'éducation en Question, vol 1, 1.

Bélair, L. (2001). La formation à la complexité du métier d'enseignant. In L. Paquay, M. Altet, E. Charlier & P. Perrenoud (éds). *Former des enseignants professionnels. Quelles stratégies ? Quelles compétences ?* Bruxelles : De Boeck Université (63-75).

Bikorindagara, R. (2002). *Etude de quelques conditions d'organisation d'une alternance dans la formation des enseignants du Burundi.* Louvain-la-Neuve : UCL, Faculté de Psychologie et des Sciences de l'éducation, thèse de doctorat (non publiée).

Bujold, N. (2002). La supervision pédagogique. Vue d'ensemble. In M. Boutet & N. Rousseau (dir.). *Les enjeux de la supervision pédagogique des stages.* Sainte-Foy : Presses de l'Université du Québec (9-22).

Charlier, E. (2001). Former des enseignants-professionnels pour une formation continuée articulée à la pratique. In L. Paquay, M. Altet, E. Charlier & P. Perrenoud (éds). *Former des enseignants professionnels. Quelles stratégies ? Quelles compétences ?* Bruxelles : De Boeck Université (97-117).

Collectif de recherche sur l'initiation à la pratique de l'enseignement de l'Université de Sherbrooke (2002). *Accompagner les stagiaires dans l'apprentissage de l'enseignement. Guide général.* Sherbrooke : Université de Sherbrooke, Faculté d'éducation, CRIPE, version préliminaire de guide d'accompagnement des stagiaires (non publié).

Costa, A. & Kallick, B. (1993). Through the lens of critical friend, *Educational Leadership*, 51, 49-51.

Dejemeppe, X. & Dezutter, O. (2001). Quels écrits réflexifs autour des stages professionnels ? , *Recherche et formation*, 36, 89-111.

De Ketele, J.M. (1999). Les concepts d'ami critique et de maillage au cœur de l'accompagnement, du suivi et de l'évaluation. In *L'approche par les compétences au service d'une éducation de qualité pour tous : communication donnée au 5e colloque francophone des Villes-Santé et des Villes et Villages en santé de l'OMS, Sousse, 4-7 décembre 1999.* Tunisie : UNICEF.

Donnay, J. & Charlier, E. (2006). *Apprendre par l'analyse des pratiques. Initiation au compagnognage réflexif*. Namur : Editions du CRP.

Dupuis, F. (2001). *Circulaire du 7 juin 2001 concernant la nouvelle formation des instituteurs et des régents.* In Moniteur. *Site du Ministère de la Justice* (en ligne).

Gervais, C. (1997). Spécificité du rôle du superviseur universitaire en stage. In M. Tardif & H. Ziarko (dirs). *Continuités et ruptures dans la formation des maîtres au Québec.* Québec : Presses de l'Université Laval (69-89).

Gervais, C. & Laporte, A. (1992). Etude exploratoire de la représentation du rôle du superviseur de stage. In *Colloque sur les domaines d'application de la statistique* : *communication présentée au 60e congrès de l'ACFAS, Montréal, 11-14 mai 1992.*

Hensler, H., Garant, C. & Dumoulin, M-J. (2001). La pratique réflexive, pour un cadre de référence partagé par les acteurs de la formation, , *Recherche et formation*, 36, 29-42.

Legendre, R. (1993). *Dictionnaire actuel de l'éducation.* Montréal : Guérin.

Maes, O. (2006). L'influence de l'accompagnement des TFE en formation initiale des enseignants du primaire sur le modèle de professionnalité développé. Mémoire inédit, Université catholique de Louvain : Louvain-la-Neuve.

Paquay, L. (1994). Vers un référentiel des compétences professionnelles de l'enseignant. *Recherche et formation*, 16, 7-33.

Paquay, L., Altet, M., Charlier, E. & Perrenoud, P. (éds) (2001). *Former des enseignants professionnels. Quelles stratégies ? Quelles compétences ?* Bruxelles : De Boeck Université.

Paquay, L., Parmentier, P., Wagner, M.C., De Cock, G., Brogniet, J.M., Dejemeppe, X., Rahier, N., Closset, D., Bourtembourg, M.P., Weustein, C., Coupremanne, M., Dourte, F., Gadisseux, C., Jamotte, P. & Warnier, A. (2001). *Accompagner les stagiaires. Un fichier d'autoanalyse à l'usage de maîtres de stage et de formateurs.* Louvain-la-Neuve : UCL (non publié).

Paquay, L. & Wagner, M.C. (2001). Compétences professionnelles privilégiées dans les stages et en vidéo-formation. In L. Paquay, M. Altet, E. Charlier & P. Perrenoud (éds). *Former des enseignants professionnels. Quelles stratégies ? Quelles compétences ?* Bruxelles : De Boeck Université (151-179).

Perrenoud, P. (2001a). *Articulation théorie-pratique et formation de praticiens réflexifs en alternance.* In Université de Genève. Faculté de psychologie et des sciences de l'éducation. Philippe Perrenoud. *Site de la Faculté de psychologie et des sciences de l'éducation de l'Université de Genève* (en ligne).

Perrenoud, P. (2001b). Le travail sur l'habitus dans la formation des enseignants. Analyse des pratiques et prise de conscience. In L. Paquay, M. Altet, E. Charlier & P. Perrenoud (éds). *Former des enseignants professionnels. Quelles stratégies ? Quelles compétences ?* Bruxelles : De Boeck Université (181-207).

Perrenoud, Ph. (2002). *Développer la pratique réflexive dans la métier d'enseignant.* Paris : ESF.

Rousseau, N. & Saint-Pierre, L. (2002). Redéfinition des rôles du superviseur de stage. Collaboration université et milieu scolaire. In M. Boutet & N. Rousseau (dir.). *Les enjeux de la supervision pédagogique des stages.* Sainte-Foy : Presses de l'Université du Québec (37-51).

Saussez, F. & Allal, L. (1998). Réfléchir sur sa pratique : le rôle de l'autoévaluation. In L. Paquay (1998). L'autoévaluation au cœur de la formation initiale d'enseignants professionnels ? Espoirs, écueils, perspectives. In *Recherche et formation des enseignants : actes du colloque organisé par l'IUFM de Grenoble, IUFM, Grenoble, 5-7 février 1998.* Grenoble : IUFM., 4.

Saussez, F. & Paquay, L. (1994). *La coévaluation en question(s). Le point de vue de l'étudiant écartelé entre « apprendre et réussir ».* Nivelles : ICADOP, Doc. DPF-94.02.

Schön, D.A. (1994). *Le praticien réflexif : à la recherche du savoir caché dans l'agir professionnel*. Montréal : Logiques.

Tardif et Ziarko (1997). *Continuités et ruptures dans la formation des maîtres au Québec.* Québec : Presses de l'Université Laval

L'IMPLANTATION GRADUELLE D'UN DISPOSITIF D'ÉVALUATION DES COMPÉTENCES EN STAGE À L'UQTR : ENJEUX ET LIMITES

Christine Lebel, Louise M. Bélair, UQTR

Mots-clés : évaluation des compétences professionnelles, formation pratique, outils

Résumé :

Les orientations de la formation initiale à l'enseignement au Québec formulées dans le sillage des transformations en cours, aux fins d'une professionnalisation accrue du métier, prévoient une formation initiale axée sur une solide articulation théorie – pratique au sein de laquelle douze compétences professionnelles sont à développer par les étudiants (Gouvernement du Québec, 2001). Il revient ainsi aux universités de traduire ce référentiel en des dispositifs et démarches propres à faciliter la construction des compétences et l'évaluation de celles-ci au long de la formation. Ces transformations appellent indéniablement des modalités d'évaluation susceptibles de rendre compte du développement effectif des compétences, ce qui nécessite un travail de régulation interactive et de co-analyse de pratiques entre le stagiaire et le formateur qui l'accompagne (Perrenoud, 2004).

Dans le cadre de la refonte du programme du baccalauréat en enseignement préscolaire et primaire (Bepep) de l'université du Québec à Trois-Rivières, le dispositif d'évaluation du volet pratique de la formation a été réaménagé en fonction des nouvelles orientations et du référentiel de compétences. Les guides de stage ont donc été analysés, révisés et expérimentés afin de répondre aux deux objectifs suivants : Déterminer l'ordre logique et chronologique des compétences en y associant les compétences obligatoires, optionnelles et non-exigées au travers les 4 stages de la formation et allouer un espace facilitateur à l'autorégulation et aux régulations proactive, interactive et rétroactive des stagiaires. La communication présentera le bilan des diverses collectes de données qui mettent en relief les choix inhérents aux compétences, la grille descriptive et la grille d'accompagnement réflexif qui seront examinés de nouveau en automne 2006 lors du dernier stage de la formation.

Contexte de la formation initiale des enseignants

Les orientations de la formation initiale à l'enseignement au Québec formulées dans le sillage des transformations en cours, aux fins d'une professionnalisation accrue du métier, prévoient une formation initiale axée sur une solide articulation théorie – pratique au sein de laquelle douze compétences professionnelles sont à développer par les étudiants (Gouvernement du Québec, 2001). Il revient ainsi aux universités de traduire ce référentiel en des dispositifs et démarches propres à faciliter la construction des compétences et l'évaluation de celles-ci au long de la formation. Les stages sont de plus en plus envisagés de manière à prolonger les activités de

formation mises en œuvre dans l'institution de formation. À l'intérieur de ce dispositif, on tente de réduire la dichotomie entre les savoir-faire pratiques et les savoirs théoriques en articulant et en mettant en tension les différents moments de la formation (Geay et Sallaberry, 1999). Cette forme de structure permet des interactions entre savoirs théoriques et savoirs pratiques que le stagiaire s'approprie, construit et transforme en compétences professionnelles. Ce type d'alternance intégrative (Malglaive 1993) s'incarne dans un dispositif de formation qui cherche à conjuguer l'action pratique et sa compréhension conceptuelle, ce qui induit la notion de guidance qui fait nécessairement appel à la mise en place d'un suivi pédagogique. Ces transformations appellent indéniablement des modalités d'évaluation susceptibles de rendre compte du développement effectif des compétences, ce qui nécessite un travail de régulation interactive et de co-analyse de pratiques entre le stagiaire et le formateur qui l'accompagne (Perrenoud, 2004). Le développement des compétences professionnelles en enseignement exige conséquemment une mobilisation de différents types de ressources en situation de même que des regards pluriels, lucides et concertés de la part de ceux qui sont appelés à évaluer afin d'une part de mieux s'approprier ce développement et d'autre part mieux le comprendre (Prud'homme et Lebel, 2006).

Le concept de compétence professionnelle en formation des enseignants

Le concept de compétence professionnelle a fait couler beaucoup d'encre dans les trois dernières décennies. Présenté comme une réponse adaptée à la conjoncture complexe face à laquelle les enseignants sont confrontés dans l'exercice de leurs fonctions (Perrenoud, 1996) une compétence recouvre des savoirs professionnels et des schèmes d'action et d'attitudes, qui sont mobilisés dans l'exercice du métier. Les compétences professionnelles sont donc d'ordre cognitif, affectif et pratique (Altet, 1998) et renvoient à des « savoir-faire de haut niveau, qui exigent l'intégration de multiples ressources cognitives dans le traitement de situations complexes » (Perrenoud, 1995a, p. 3). Elles ne se présentent donc pas comme un répertoire de savoir-faire spécialisés permettant de fournir des réponses à l'ensemble des situations possibles. Cependant, elles donnent les moyens de mettre en œuvre des réponses particulières à des situations inédites, par le biais d'une culture scientifique de base, de savoirs professionnels, d'une capacité d'analyse définie par Altet (1998) comme le « savoir-analyser » et d'une expérience construite au fil de la pratique (Bosman et al. 2000 ; Le Boterf, 1999). Une compétence étant à la fois complexe, évolutive et indissociable du contexte dans lequel elle s'exerce (Legendre, à paraître), son évaluation requiert des implications multiples dont l'implantation est encore à l'état embryonnaire tant les éclairages issus de la littérature scientifique à cet égard sont encore trop peu nombreux.

Le concept d'évaluation dans une approche par compétences en formation pratique

Au sein d'une approche par compétence, les étapes de la démarche évaluative soit l'intention, la mesure, le jugement et la décision (Cardinet, 1986) s'illustrent quelque peu différemment. Une compétence étant évolutive et complexe, la décision du niveau de compétences ne peut être uniquement basée sur des collectes ponctuelles de données. Dans cette perspective, Tardif (2006) insiste sur l'importance de mettre

l'accent sur le caractère authentique de l'évaluation au sens où ce type d'évaluation alternative permet l'analyse des savoirs-agir et autorise l'explicitation des «savoirs en actes» (p. 103). Il propose ainsi, dans la lignée de plusieurs auteurs, de rendre compte d'une progression dans une trajectoire de développement, procédure pouvant s'incarner dans l'élaboration de grilles sous forme de rubriques (Bélair, 1999 ; Wiggins, 1998) ou encore selon des indicateurs et des profils de développement de compétences (Bednarski, 2003). Cette alternative permet effectivement de cibler des catégories d'ordre procédural (Popham, 2002) et facilite par conséquent la prise en compte de la progression. L'évaluation des stages doit être envisagée de telle sorte qu'elle soit centrée sur l'évolution des compétences et de leurs composantes au travers le cursus de formation, tout en conservant le caractère scientifique de la démarche évaluative.

Le programme de l'UQTR

Dans le cadre de la refonte du programme du baccalauréat en enseignement préscolaire et primaire (BEPEP) de l'Université du Québec à Trois-Rivières (UQTR), le dispositif d'évaluation du volet pratique de la formation a été réaménagé en fonction des nouvelles orientations et du référentiel de compétences. Supportée par plusieurs écrits en évaluation (Bélair, 2005a; Rogiers, 2004 ; Scallon, 2004; Tardif, 2006), cette réorganisation du dispositif, entamée depuis l'automne 2004 s'effectue de manière graduelle tant au niveau de la chronologie des compétences à développer qu'au plan des espaces formels pour l'autorégulation par le stagiaire et les régulations proactive, interactive et rétroactive (Allal et Wegmuller, 2004) par le superviseur et l'enseignant-associé.

Le dispositif

À L'UQTR, le programme BEPEP d'une durée de 4 années comporte 4 stages allant de l'immersion d'une semaine à l'internat de 3 mois, entrecoupés de différents cours en milieu universitaire. Ces stages sont effectués dans des classes et des écoles diverses sous la supervision de 4 superviseurs différents (un par stage). Ces superviseurs animent les séminaires de préparation et de rétroaction des stages, observent les stagiaires dans les classes, mènent des entretiens avec le stagiaire et l'enseignant associé (enseignant hôte) et enfin, attribuent la note finale suite aux visites, aux commentaires de l'enseignant associé et à la lecture du dossier progressif du stagiaire. Ces superviseurs, à l'interface de l'université et de la salle de classe, sont par conséquent au premier plan de la formation et de l'évaluation des compétences professionnelles. Pour chacun des stages, l'outil de référence par excellence étant les guides de stage, et ce, autant pour les stagiaires que pour les superviseurs et les enseignants associés, ces outils ont donc été analysés, révisés et expérimentés afin de répondre aux deux objectifs suivants :

1. Déterminer l'ordre logique et chronologique des compétences en y associant les compétences obligatoires, optionnelles et non exigées au travers les 4 stages de la formation
2. Allouer un espace facilitateur à l'autorégulation et aux régulations proactive, interactive et rétroactive des stagiaires.

Pour ce faire, une hiérarchie des compétences du référentiel et leurs composantes a été élaborée au travers les 4 stages. Également, des grilles descriptives indiquant le niveau de compétence développé au travers ces stages ont été bâties. Finalement, un dispositif d'autorégulation des compétences a été expérimenté à l'hiver 2006 dans l'optique de permettre au stagiaire de faire des liens plus explicites entre sa formation à l'université, ses séminaires de stage et la pratique en tant que telle. Ce dispositif comporte deux volets. La grille d'accompagnement réflexif et le dossier progressif du stagiaire qui par ailleurs, ne sera pas abordé dans le cadre de cette présentation.

Méthodologie

La révision des guides de stage et les différents outils d'évaluation formative et sommative des compétences s'est amorcée à automne 2005. Un questionnaire suivi de divers entretiens avec des superviseurs, des enseignants-associés et des stagiaires ont permis d'identifier les faiblesses et les forces des outils élaborés. En parallèle à ceci, un groupe de travail a établi la logique du développement des compétences et de leurs composantes au travers les 4 stages, en fonction des moments des stages dans la formation et ce, afin de resserrer le plus possible les liens entre les moments de formation à l'université et ceux qui ont cours en milieu de pratique. De même, une grille descriptive d'évaluation des niveaux de compétences (4 niveaux) a été développée et expérimentée en hiver 2006 (Bélair, 2005b, 2000 ; Lebel et Bélair, 2004). Enfin, une grille d'accompagnement réflexif a été conçue pour favoriser l'autorégulation par le stagiaire et a été mise à l'essai au printemps 2006. Actuellement, la collecte de données des entrevues et des questionnaires vise à analyser dans quelle mesure l'ordonnancement des compétences est pertinent et adapté aux divers stages. Elle vise également à analyser par le biais de groupes de discussions avec des superviseurs et d'entretiens auprès d'enseignants associés, en quoi cette grille permet de bien cerner le profil du stagiaire.

Résultats

Cette partie présente le déroulement entourant la mise en œuvre du dispositif d'évaluation des stages ainsi qu'un exemple de l'évolution des choix des compétences du référentiel à faire développer dans chaque stage. Finalement, l'utilisation progressive des rubriques et des niveaux de compétence est présentée et discutée.

Déroulement de l'expérimentation

Phase 1. 2004-2005

Cette phase a consisté en l'adaptation des guides en fonction du référentiel, suivie du choix des compétences à développer dans les stages au regard des cours qui les précèdent et de la durée des 4 stages. Ces phases ont été suivies de réflexions entre diverses personnes impliquées dans ce processus face à un mode d'évaluation respectant les principes d'évaluation formative et sommative ainsi que face à l'élaboration de grilles avec 4 niveaux de développement. Cette première année s'est terminée par l'expérimentation de ce nouvel outil.

Phase 2. 2005-2006

Cette seconde phase a permis l'articulation des modes d'évaluation formative et sommative des compétences à la suite des échanges et des premières expérimentations. Elle a également servi d'une part au réajustement des grilles d'évaluation afin que le bilan du 4e stage regroupe l'ensemble des compétences, et d'autre part au développement d'une grille d'accompagnement réflexif pour les stagiaires (autorégulation) et à l'expérimentation de cette seconde version.

Phase 3. 2006-2007

Cette phase consiste actuellement à intensifier la collaboration au sein de l'équipe des superviseurs et à définir l'ensemble des composantes liées aux compétences dans chacun des stages.

À propos de l'intention de l'évaluation : les enjeux des diverses interprétations du référentiel de compétences.

Dans la première phase, il a semblé nécessaire d'échelonner les 12 compétences du référentiel au travers les 4 stages de la formation. Les évaluations sommatives (certificatives) portaient alors sur les compétences tandis que les évaluations formatives, vues comme des étapes dans le développement de la compétence, ciblaient certaines composantes du référentiel québécois. De même, les compétences relatives *aux fondements (F)* et à *l'identité professionnelle (IP)* se retrouvaient surtout dans les 2 premiers stages, celles relatives à *l'acte d'enseigner (AE)* débutaient dès le 2e stage et enfin celles reliées au *contexte social et scolaire (CS)* se situaient surtout dans les 3e et 4e stage.

Tableau 1- Exemples de compétence en 2004

Compétences	Stage 1	Stage 2	Stage3	Stage

				4
AE5. Évaluer la progression des apprentissages et le degré d'acquisition des compétences des élèves pour les contenus à faire apprendre.		Porter un jugement sur les acquis des élèves en repérant les forces et les faiblesses.	**Évaluer la progression des apprentissages et le degré d'acquisition des compétences des élèves pour les contenus à faire apprendre**	
IP11. S'engager dans une démarche individuelle et collective de développement professionnel.	**S'engager dans une démarche individuelle et collective de développement professionnel**.			

Ces deux exemples montrent bien qu'une compétence évaluée de manière certificative à un stage n'était pas réévaluée dans les stages subséquents. Or, même si ce choix s'insère dans les principes méthodologiques relatifs à l'articulation entre l'évaluation formative et certificative, lesquels évoquent le fait que l'évaluation certificative se réalise à la fin d'un apprentissage (Scallon, 2004), il demeure que cette première option a crée certains problèmes. De fait, des compétences, telle la # 11 (*Engagement professionnel*), n'étaient plus évaluées par la suite, ce qui ouvrait la porte à maintes situations délicates. Ces constats nous ont donc amené à revoir le dispositif d'ensemble.

Également, le 4e stage étant la synthèse de l'ensemble de la formation, il fallait impérativement évaluer de manière certificative les douze compétences du référentiel. Or, de la façon dont le dispositif était organisé, les compétences ayant déjà été évaluées de manière certificative dans les stages précédents n'étaient pas réévaluées en fin de parcours. Ainsi un étudiant plutôt faible à une compétence donnée au 3e stage, ne pouvait pas démontrer explicitement qu'il s'était amélioré dans le 4e stage.

Il fallait donc remanier le dispositif afin de ternir compte de la complexité inhérente aux compétences, mais surtout de leur caractère évolutif, non seulement dans un seul stage, mais au travers les 4 années de formation. En deuxième phase, il a

donc été décidé de remettre toutes les compétences au 4e stage, ce qui a débouché sur une révision complète des compétences ciblées dans les trois autres stages afin d'y inclure la gradation des composantes.

Actuellement, l'expérimentation porte sur une 3e articulation du référentiel qui tient compte à la fois des composantes et des compétences, tout en respectant la complémentarité avec les cours offerts à l'université. Dans cet esprit et puisqu'il faut évaluer chaque stage de manière certificative, il a semblé logique de cibler certaines composantes des compétences dans les 4 stages et d'insérer les douze compétences au 4e stage. Voici un exemple pour les compétences # 5 et # 11.

Tableau 2- Exemples de compétences en 2006.

Compétences	Stage 1	Stage 2	Stage3	Stage 4
AE5. Évaluer la progression des apprentissages et le degré d'acquisition des compétences des élèves pour les contenus à faire apprendre.		5.1 En situation d'apprentissage, prendre des informations afin de repérer les forces et les difficultés des élèves ainsi que de revoir et d'adapter l'enseignement en vue de favoriser la progression des apprentissages	5.2 Établir un bilan des acquis afin de porter un jugement sur le degré d'acquisition des compétences. 5.3 Construire ou employer des outils permettant d'évaluer la progression et l'acquisition des compétences.	***5. Évaluer la progression des apprentissages et le degré d'acquisition des compétences des élèves pour les contenus à faire apprendre.*** 5.4 Communiquer aux élèves et aux parents, de façon claire et explicite, les résultats attendus ainsi que les rétroactions au regard de la progression des apprentissages et de l'acquisition des compétences. 5.5 Collaborer avec l'équipe pédagogique à la détermination du rythme et des étapes de progression souhaitées à l'intérieur du cycle de formation.
IP11. S'engager dans une démarche individuelle et collective de développement professionnel.	11.1 Établir un bilan de ses compétences et mettre en œuvre les moyens pour les développer en utilisant les ressources disponibles	11.3 Réfléchir sur sa pratique (analyse réflexive) et réinvestir les résultats de sa réflexion dans l'action	11.2 échanger des idées avec ses collèges quant à al pertinence de ses choix pédagogiques et didactiques 11.4 Mener des projets pédagogiques pour résoudre des problèmes d'enseignement	***11. S'engager dans une démarche individuelle et collective de développement professionnel.*** 11.5 Faire participer ses pairs à des démarches de recherche liées à la maîtrise des compétences visées dans le programme de formation et aux objectifs éducatifs de l'école.

Ce nouveau format permet effectivement de mieux répondre au caractère évolutif et progressif de la compétence tout en garantissant que toutes les composantes de ces 12 compétences sont couvertes au travers les 4 stages et que la certification globale du niveau de développement des compétences se fasse au dernier stage. Ceci permet aux étudiants de progresser dans le développement de chacune d'elles. Cette façon de faire s'inscrit également dans les nouvelles orientations du MELS dans la mesure où le développement effectif des compétences est facilité à la fois par les grilles descriptives et les indicateurs et à la fois par un travail de régulation interactive et de co-analyse de pratiques entre le stagiaire et le formateur qui l'accompagne.

À propos de la mesure, du jugement et de la décision : les enjeux des divers outils de collecte de données et de jugement.

Selon les principes de Bélair (1999) une grille descriptive de niveaux de compétence a été développée en 2005. Elle mettait en parallèle diverses composantes d'une compétence et un profil de développement de la compétence en 4 temps. Voici un exemple pour la compétence # 5 (*évaluation des apprentissages*) au 3e stage.

Tableau 3- Grille d'évaluation des compétences en 2005

	Utilise des outils d'évaluation formative ou sommative 1	**Planifie des activités variées d'évaluation en fonction des compétences 2**	**Intègre l'évaluation des compétences à l'apprentissage et à l'enseignement 3**	**S'approprie tout le processus d'évaluation des compétences 4**
• Construire et employer des outils permettant d'évaluer la progression et l'acquisition des compétences.				
• Contribuer avec l'enseignante à la préparation du matériel d'évaluation des compétences.				
• Communiquer à l'élève les résultats d'un processus d'évaluation diagnostique et les remédiations envisagées.				

Lors des expérimentations, nous avons été rapidement appelées à réaliser qu'il y avait confusion entre les descriptions des niveaux et les composantes, de même que certaines incompatibilités entre les colonnes et les lignes. De fait, à l'examen de cet exemple, on remarque la difficulté d'arrimer *communiquer à l'élève*… avec le niveau 3 *intègre l'évaluation.*… Quoique élaborés au départ dans l'optique d'éclaircir les différences entre les profils des 4 niveaux de développement, ces indicateurs étaient de fait, des composantes de compétences et non des descripteurs. C'est donc dans la lignée des travaux de Bélair (2005), Lebel et Bélair (2004) et de Tardif (2006) et suite

aux résultats de l'expérimentation de la première mouture, qu'un ajustement de ces grilles a été apporté.

Pour 2006-07, ces grilles ont été remaniées afin de respecter l'idée de profil de niveau de développement de compétences. Dans l'exemple de la compétences 5, on remarque que les éléments en caractère gras indiquent la progression et que l'enseignant associé, comme le superviseur et le stagiaire peuvent en quelque sorte statuer du niveau de compétence développé en fonction de chacune des composantes.

Tableau 4- Grille d'évaluation de compétence en 2006

<u>Compétence 5</u> : **Évaluer la progression des apprentissages et le degré d'acquisition des compétences des élèves pour les contenus à faire apprendre.**

		Niveau 1	**Niveau 2**	**Niveau 3**	**Niveau 4**
COMPOSANTE 5.4 **Communiquer aux élèves et aux parents, de façon claire et explicite, les résultats attendus ainsi que les rétroactions au regard de la progression des apprentissages et de l'acquisition des compétences.**	Niveau minimal non atteint ❑	Procure aux élèves des occasions et des moyens de s'autoévaluer ; fournit aux parents une appréciation de l'évolution de leur enfant. ❑	Procure aux élèves des occasions et des moyens de s'autoévaluer ; fournit aux parents une appréciation **fondée** de l'évolution de leur enfant. ❑	Procure aux élèves des occasions et des moyens de s'autoévaluer **et de participer à l'interprétation du résultat**; fournit aux parents une appréciation fondée de l'évolution de leur enfant. ❑	Procure des occasions et des moyens de s'autoévaluer et de participer à l'interprétation du résultat; fournit aux parents une appréciation fondée de l'évolution de leur enfant **et leur suggère des moyens pour le soutenir.** ❑

		Niveau 1	Niveau 2	Niveau 3	Niveau 4
COMPOSANTE 5.5 **Collaborer avec l'équipe pédagogique à la détermination du rythme et des étapes de progression souhaitées à l'intérieur du cycle de formation.**	Niveau minimal non atteint ❑	En collaboration avec l'équipe pédagogique, détermine le rythme et les étapes de la progression des apprentissages. ❑	En collaboration avec l'équipe pédagogique, détermine le rythme et les étapes de la progression des apprentissages **et en assure quelques aspects**. ❑	En collaboration avec l'équipe pédagogique, détermine le rythme et les étapes de la progression des apprentissages **et en assure le respect.** ❑	En collaboration avec l'équipe pédagogique, détermine **et propose** le rythme et les étapes de la progression des apprentissages et en assure le respect. ❑

Les éléments de cette grille veulent répondre à ce Tardif (2006) nomme les indicateurs de développement de compétence (p.155), selon l'idée que le développement se construit dans le temps, que des étapes doivent être aménagées pour réaliser l'apprentissage et que l'évolution au cours des 4 années de formation doit être prise en compte. L'expérimentation actuelle vise ainsi à valider ces grilles, non seulement au regard de leur faisabilité, mais surtout dans l'articulation des diverses composantes au fil des 4 stages.

À propos de l'implication du stagiaire

Afin de favoriser l'autorégulation par le stagiaire, une grille d'accompagnement a été conçue et ce, afin que le stagiaire puisse s'y référer dès son premier séminaire avant chaque stage. Outre les composantes obligatoires, le stagiaire doit indiquer dans chaque case, les éléments qu'il juge nécessaires de développer dans son stage et doit faire approuver ce choix par son superviseur. Ainsi, dans l'idée de parcours de formation individualisé, un stagiaire peut se reconnaître des besoins particuliers de formation dans l'une ou l'autre des compétences, même si celles-ci n'ont pas été ciblées comme tel dans les grilles d'évaluation. Voici un exemple pour la compétence # 6 au 4e stage.

Au terme de la formation initiale, l'étudiante ou l'étudiant doit être en mesure :
✓ de mettre en place et de maintenir des routines permettant un déroulement efficace des activités usuelles de la classe; ✓ de repérer et de corriger des problèmes de déroulement qui nuisent au fonctionnement du groupe-classe; ✓ d'anticiper des problèmes de déroulement des activités de la classe et de planifier des mesures en vue de les prévenir;

✓ de déterminer et d'appliquer des moyens permettant de régler des problèmes avec les élèves présentant des comportements inappropriés.	
La ou le stagiaire :	**Stage 4**
➢ fait preuve de respect envers tous les élèves et développe une relation positive avec eux;	❑
➢ met en place et maintient un fonctionnement collectif efficace et harmonieux dans la classe en fonction des apprentissages à effectuer et des activités usuelles de la classe, en impliquant les élèves; responsabilise les élèves à la gestion de divers aspects de la vie de classe, de l'école;	❑
➢ anticipe, met en place et gère les modes d'interactions;	❑
➢ favorise l'autorégulation entre pairs lors d'écarts de conduite ou de conflits interpersonnels;	❑
➢ adapte sa planification selon les caractéristiques du groupe et des besoins;	❑
➢ reconnaît les comportements inappropriés et intervient efficacement;	❑
➢ fait preuve de constance, de cohérence et de fermeté dans l'application de ses exigences;	❑
➢ encourage les manifestations appropriées, les initiatives;	❑
➢ tient compte des besoins affectifs;	❑
➢ etc.	❑

La recherche est en cours. Un bilan des diverses collectes de données qui mettent en relief les choix inhérents aux diverses composantes et compétences ciblées sera mené. De même, la grille descriptive et la grille d'accompagnement réflexif seront examinées de nouveau en automne 2006 lors du dernier stage de la formation.

Conclusion

Le processus d'évaluation des compétences en milieu de pratique mis en œuvre au BEPEP dans le souci de mieux répondre aux nouvelles orientations ministérielles (Gouvernement du Québec, 2001) en matière d'approche par compétences et d'évaluation de celles-ci, a permis au fil des expérimentations, de mieux cibler les difficultés des stagiaires et de les aider à se réguler pour progresser dans les stages subséquents. Cela a également permis d'ouvrir le dialogue entre les partenaires de ces

stages, particulièrement entre superviseurs et enseignants associés. Le consensus reste à établir, mais des pas sont faits et le travail en ce sens se poursuit notamment par une interrogation sur de nouvelles avenues au plan de l'organisation des stages susceptibles d'accentuer l'articulation théorie-pratique. Pour la suite, il nous restera, à la lumière des propositions de Legendre (2006, à paraître) à demeurer particulièrement attentifs face à toute forme de rationalisation excessive de la démarche d'évaluation des compétences, laquelle pourrait tôt ou tard conduire à occulter l'une ou l'autre des dimensions qu'elle renferme et qui ne peuvent être appréhendées de manière individuelle puisque les compétences sont par définition complexes.

Références

Allal, L. & Wegmuller, E. (2004). Finalités et fonctions de l'évaluation. *Educateur* (numéro spécial 04), 4-7.

Altet, M. (1998). Les compétences de l'enseignant-professionnel : entre savoirs, schèmes d'action et adaptation, le savoir analyser. Dans Paquay, L., Altet, M., Charlier, E. & Perrenoud, Ph. (Eds.). *Former des enseignants professionnels*, 27-40. Bruxelles : De Boeck.

Bednarski, M. (2003). *Assessing Performance Tasks*. The Science Teacher, vol 70 (4), 34-38.

Bélair, L.M. (1999). *L'évaluation dans l'école*. Paris : ESF.

Bélair, L.M. (2000). Savoir construire la relation éducative. In M. Théberge (dir.). *Former à la profession enseignante*, 23-52. Montréal : Éditions Logiques.

Bélair, L.M. (2005a). Quels sont les obstacles épistémologiques à l'évaluation des compétences professionnelles ? un état des lieux. *Questions de pédagogie dans l'enseignement supérieur*. Actes du 3e colloque de l'école centrale de Lille, 176-181. Lille : École centrale de Lille.

Bélair, L.M. (2005b). Les outils d'observation. In N. Rousseau (dir.). *Se former pour mieux superviser*, 125-144. Montréal : Guérin.

Bosman, C., Gérard, F.-M. & Roegiers, X. (Éds) (2000). *Quel avenir pour les compétences ?* Bruxelles : De Boeck.

Cardinet, J. (1986) *Évaluation scolaire et mesure*. Bruxelles : De Boeck.

Geay et Sallaberry, (1999). La didactique en alternance ou comment enseigner dans l'alternance ? *Revue Française de Pédagogie*, no. 128, 7-15

Gouvernement du Québec, (2001), *Orientations des programmes de formation à l'enseignement*. Québec : MELS.

Le Boterf, (1999). *Compétence et navigation professionnelle*. Paris : Les éditions d'Organisation.

Lebel, C. et Bélair, L.M. (2004). Évaluer pour trans-former, évaluer pour qualifier. *Évaluer pour mieux rendre compte*, actes du colloque AQPC (107-116). Montréal : AQPC.

Legendre, M.-F. (à paraître). Quelques caractéristiques des compétences et leurs incidences sur l'évaluation. In, L.M. Bélair, D. Laveault et C. Lebel (eds). *Valider une professionalité : quel(s) rôle(s) pour l'évaluation ? le cas de la profession en éducation*. Ottawa : PUO.

Malglaive, G. (1993). Alternance et compétences, *Les cahiers pédagogiques*, no. 320. P. 28

Perrenoud, Ph. (2004). Évaluer des compétences. *L'éducateur*, 8-11.

Perrenoud, Ph. (1996). Formation continue et développement de compétences professionnelles, In, Éducateur, no. 9, 28-33.

Perrenoud, Ph. (1995). *La pédagogie à l'école des différences.* Paris : ESF.

Popham, W. J. (2002). *Classroom Assessment. What Teachers Need to Know.*, Boston : Allyn & Bacon.

Prud'homme, L. et Lebel, C. (2006). *Rompre avec le modèle universitaire prédominant.* Texte d'une communication présentée dans le cadre du symposium Évaluation des compétences professionnelles, Lausanne, mai 2006

Rogiers, X (2004). *L'école et l'évaluation*. Bruxelles : De Boeck.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Montréal : ERPI.

Tardif, J. (2006). *L'évaluation des compétences ; documenter le parcours de développement*. Montréal : De la Chenelière-éducation.

Wiggins, G. (1998). *Educative Assessment. Designing Assessments to Inform and Improve Student Performance*. San Francisco : Jossey-Bass.

SYMPOSIUM 7 : EVALUATION DES ACQUIS EN LANGUES ET DANS D'AUTRES DOMAINES : QUELQUES PROBLEMES POSES PAR LEURS COMPOSANTES LINGUISTIQUES ET QUELQUES SOLUTIONS POUR LES RESOUDRE

Monique Reichert, Université du Luxembourg
Christian Monseur, Université de Liège
Ariane Baye, Université de Liège
Claude Houssemand, Université du Luxembourg
Franck Desroches, CCIP-DRI/E
Marc Demeuse, Université de Mons-Hainaut

Résumé

Qu'il s'agisse d'évaluer la maîtrise d'une langue particulière ou d'évaluer une compétence non linguistique, les tests comportent indéniablement une composante linguistique ou, plus exactement, des composantes linguistiques multiples.

Une première manière d'approcher le problème est de considérer les évaluations internationales, quel qu'en soit le sujet principal. Dès que des études de ce type, comme PISA (Programme international pour le suivi des acquis des élèves, OCDE) ou les enquêtes menés par l'IEA (Association internationale pour l'évaluation du rendement scolaire) sont planifiées, se posent des questions relatives à la traduction et au maintien de l'équivalence des différentes versions linguistiques, indépendamment de la langue utilisée pour interroger les individus.

Une première communication, intitulée « impact de tests traduits sur les performances des élèves » situe le problème et présente les difficultés et les moyens d'y remédier (Baye et Lafontaine).

La seconde communication « procédures destinées à minimiser, a priori, les biais susceptibles d'altérer l'unidimensionnalité internationale des échelles de rendement » envisage, plus en détail, les procédures de vérification qui sont appliquées lors de la construction des épreuves, y compris les analyses statistiques de détection de biais (en particulier l'examen du fonctionnement différentiel des items selon la langue de test) (Monseur).

La troisième communication, contrairement à la précédente, examine a posteriori les biais, notamment linguistiques, qui semblent subsister après la publication des résultats sur la base d'analyses secondaires et qualitatives des résultats du PISA 2000 et 2003 (Houssemand).

Une seconde approche, développée dans ce symposium, consiste à examiner des tests à visée proprement linguistique et à les interroger quand à leurs qualités et à l'absence de biais.

La quatrième communication aborde la mise au point d'évaluations linguistiques à la fois valides, fidèles et économiques à travers le C-Test. Après avoir discuté de la qualité de cet instrument dans la littérature, la contribution aborde la comparaison de cette forme d'évaluation avec un test de forme plus classique, administré à des bacheliers luxembourgeois (Reichert, Keller et Martin).

La cinquième communication envisage, pour sa part, la construction d'un test de français langue étrangère (le TEF) et la manière dont, a priori, les biais linguistiques et culturels sont traqués lors de cette étape (Crendal et al.)

La dernière communication, en quelque sorte à l'autre bout de la chaîne, pose le problème de la compréhension des résultats à des tests linguistiques à travers la comparaison des données obtenues à travers une échelle d'autoévaluation (DIALANG) et un test critérié à visée certificative (le TEF) construits sur le même référentiel (Demeuse et al.).

LE C-TEST : UN NOUVEAU CHEMIN IÉAL POUR UNE ÉVALUATION VALIDE DE COMPÉTENCES LINGUISTIQUES ?

Monique Reichert, Ullrich Keller, Romain Martin, Université du Luxembourg

Résumé :

Actuellement un grand nombre d'institutions (p.ex. les systèmes scolaires, les universités, les entreprises

commerciales) se voient confrontées aux défis d'une hétérogénéité croissante dans le domaine des arrière-fonds et de la maîtrise des langues auprès de leur(s) population(s) cible(s). En même temps, il y existe souvent un manque en informations notamment sur les compétences linguistiques des individus visés, et ce entre autres à cause d'un manque de comparabilité entre les différents outils d'évaluation existants, d'une validité des outils souvent insuffisante, ainsi qu'à cause des coûts élevés qu'une évaluation des compétences linguistiques engendre.

Le but de la présente communication est de savoir à quel point le *C-Test*, un outil d'évaluation dont on affirme qu'il mesure une compétence linguistique générale, et connu pour son économie en termes de construction et de passation, offre des avantages importants en ce qui concerne les points de la comparabilité, de la validité, ainsi que des coûts mentionnés.

Afin de donner un premier aperçu sur les forces et faiblesses de cet outil, une synthèse de résultats d'un nombre d'études scientifiques récentes ayant utilisé le C-Test sera présentée dans un premier lieu.

Dans un deuxième lieu, les résultats de notre propre étude sur la validité du C-Test seront présentés.

Ceux-ci comprennent les résultats d'un C-Test appliqué auprès de bacheliers du Luxembourg, qui sont mis en relation avec les résultats des mêmes élèves dans le *Test de Connaissance du Français* (TCF).

Etant donné que le TCF est lié à l'échelle de compétence du Conseil de l'Europe (CECR), les résultats du C-Test peuvent indirectement être positionnés sur cette échelle. Une discussion sur les possibilités de relier le C-Test avec cette échelle sera par la suite entamée en misant particulièrement sur des caractéristiques psychométriques déterminantes (p.ex. niveau de difficulté, force discriminatoire des items), ainsi que sur les différences qui pourraient résulter à cet égard pour des C-Test de différentes langues.

Le C-Test : caractéristiques générales

Le C-Test est un test écrit dont un nombre d'études a pu montrer qu'il mesure une compétence linguistique

générale (Grotjahn, 2002). La caractéristique essentielle définissant ce type de test est la réduction de la

redondance d'un texte donné. Un C-Test typique comprend quatre à six textes ; chacun de ces textes commence et termine avec une phrase intacte, présentant un minimum en information concernant le contenu du texte respectif. A partir de la seconde phrase, la seconde moitié de chaque second mot est éliminée. Ce sont ces mots mutilés qui sont à compléter par les sujets lors d'une passation.

Au début d'une passation, il suffit en général d'indiquer en quelques mots que les expressions mutilées sont à compléter. Pour chaque texte comprenant entre 20 et 25 items, quatre à cinq minutes sont nécessaires. Une passation complète a donc une durée de 20 à 30 minutes.

Les caractéristiques psychométriques du C-Test

L'attractivité du C-Test est, outre l'économie dans sa passation, en grande partie due à des niveaux de fidélité et de validité considérables. Ainsi, les alphas de Cronbach se situent en général entre .80 et .90. La fidélité testretest varie entre .70 et .85 (Grotjahn, 2002).

En ce qui concerne la validité, des corrélations significatives ont été obtenues avec la compréhension et

l'expression écrite et orale (voir, p.ex., Arras, Eckes & Grotjahn, 2002). De même, le C-Test a déjà pu être relié avec des tests plus élaborés, comme notamment le TOEFL (Hastings, 2002), le TestDaF, ou le « English Placement Test » (EPT) (Arras, Eckes & Grotjahn, 2002 ; Jafarpur, 2002). Par exemple, les corrélations avec l'EPT, qui mesure la compréhension orale, la grammaire, le vocabulaire et la lecture, se situent entre .84 (grammaire) et .87 (compréhension orale) (Jafarpur, 2002). Le fait que le C-Test a pu être relié à cette hétérogénéité de dimensions a renforcé la supposition que cet outil mesure une compétence linguistique globale.

Une autre caractéristique psychométrique qui vaut d'être mentionnée est celle que les données de C-Tests

peuvent également être traitées en recourant à un échelonnage de Rasch (Arras, Eckes & Grotjahn, 2002).

Le C-Test et les cinq dimensions du TCF

Un des objectifs de notre propre étude fut de vérifier si des coefficients équivalents à ceux reportés dans la

littérature peuvent être présentés pour les corrélations d'un C-Test donné avec les dimensions du TCF. Celles-ci comprennent l'expression et la compréhension écrite, la maîtrise des structures de la langue (mesurant surtout des capacités et connaissances grammaticales et lexicales), ainsi que l'expression et la compréhension orale.

Un C-Test de langue française fut créé, regroupant neuf textes de différents niveaux de difficulté. Ceux-ci furent présentés à un total de 237 bacheliers de différentes écoles luxembourgeoises.

Une analyse préliminaire des données nous fournit les informations suivantes:

Tout d'abord, une analyse factorielle « maximum-likelihood » indique qu'un seul facteur suffit pour reproduire la structure de covariance de l'échelle C-Test (X2[27,N = 236] = 39.08, p = 0.062).

Le calcul de l'alpha de Cronbach, calculé sur les neuf items, donne 0.92.

Les résultats des études mentionnées plus haut sont confirmés par nos propres données, avec des coefficients un peu moins élevés: des corrélations significatives sont obtenues avec les cinq dimensions du TCF, se situant entre .53 pour l'expression orale, et .70 pour les structures de la langue. Enfin, la corrélation avec un score global du TCF, regroupant toutes les cinq dimensions, se situe au-delà de ces coefficients mentionnés, avec r=.8 (p<.001).

Le TCF et le CECR

Une première analyse discriminatoire des données indique que plus de 65% des sujets peuvent être affectés

correctement à un des niveaux du CECR, alors que tous les autres – à l'exception de trois élèves – sont classés au niveau voisin. Des analyses supplémentaires à ce sujet vont être présentées en misant particulièrement sur les différentes caractéristiques psychométriques du C-Test, ainsi que sur les différences qui pourraient y résulter pour des C-Test de différentes langues.

Conclusion

Les données présentées supportent les affirmations généralement énoncées par rapport au C-Test, soulignant que les avantages de cet outil se situent clairement dans son économie en termes de construction et de passation, tout en profitant du fait qu'il mesure une compétence linguistique globale de manière valide et fidèle. En outre, les données peuvent être affectées sur l'échelle du CECR, néanmoins, la question si cela s'applique aux C-Tests en général nécessite de la transparence concernant les caractéristiques psychométriques des tests ainsi que les différences qui pourraient y résulter pour de langues différentes.

Bibliographie

Arras, U., Eckes, T. & Grotjahn, R. (2002). C-tests im Rahmen des "Test Deutsch als Fremdsprache" (TestDaF):

Erste Forschungsergebnisse. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische*

Anwendungen (Bd. 4, 175-209). Bochum: AKS-Verlag.

Grotjahn, R. (2002). Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis. In R. Grotjahn (Ed.),

Der C-Test. Theoretische Grundlagen und praktische Anwendungen (Bd. 4, 211-225). Bochum: AKSVerlag.

Hastings, A.J. (2002). In defense of C-testing. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und*

praktische Anwendungen (Bd. 4, 11-25). Bochum: AKS-Verlag.

Jafarpur, A. (2002). A comparative study of a C-Test and a cloze test. In R. Grotjahn (Ed.), *Der C-Test.*

Theoretische Grundlagen und praktische Anwendungen 4, 31-51). Bochum: AKS-Verlag.

Journal of **E**ducational **M**easurement and **A**pplied **C**ognitive **S**ciences 1
http://www.jemacs.uni.lu/ 2007 volume 1, No 1

PROCÉDURES DESTINÉES À MINIMISER, À PRIORI, LES BIAS SUSCEPTIBLES D'ALTÉRER L'UNIDIMENSIONALITÉ INTERNATIONALE DES ÉCHELLES DE RENDEMENT.

Christian Monseur, Université de Liège
Béatrice Halleux, HallStat SPRL/ ACER

Résumé :

Depuis le début des années 90, les études internationales en sciences de l'éducation utilisent des modèles de réponse à l'item pour la construction des échelles de rendement. De nos jours, deux modèles sont principalement employés : le modèle de Rasch ou modèle à un paramètre dans le cadre des études PISA et le modèle à trois paramètres pour les études TIMSS et PIRLS. Depuis PISA2000, ces modèles de réponse à l'item sont également employés pour la construction d'échelles contextuelles au départ de données recueillies par l'intermédiaire des questionnaires élèves et directeurs.

Différentes procédures sont mises en œuvre par les responsables internationaux de ces études pour assurer la comparabilité ou l'unidimensionalité internationales de la mesure. Différents éléments peuvent altérer la difficulté relative des items :

(i) Les tests de rendement sont transmis au pays, soit en anglais, soit en anglais et en français et les centres nationaux doivent traduire ces tests dans les différentes langues nationales. Des biais de traductions peuvent dès lors être commis, rendant les questions soit plus faciles soit plus difficiles.

(ii) De plus, des différences de curriculum ou des spécificités culturelles peuvent aussi modifier la difficulté relative des questions.

(iii) Actuellement, ces épreuves de rendement comportent plus pou moins 50% de questions ouvertes, dont un nombre non négligeable nécessite une évaluation. Des manuels de codage très détaillés sont transmis aux centres nationaux pour uniformiser autant que faire se peut cette étape mais nul ne peut exclure l'introduction d'un biais lors de l'évaluation des questions ouvertes.

La présentation se focalisera sur les procédures mises en œuvre par les centres de coordination de ces études pour minimiser les biais susceptibles d'altérer l'unidimensionalité internationale des échelles de rendement.

Tout d'abord, nous décrirons les étapes de construction de ces épreuves, les vérifications mises en œuvre pour minimiser les biais de traduction et de codage. Ensuite, la présentation détaillera les analyses statistiques de détection de biais, et en particulier les *Differential Item Functioning* lors de la construction de ces échelles internationales.

L'IMPACT DE TESTS TRADUITS SUR LES PERFORMANCES DES ÉLÈVES

Ariane Baye, Dominique Lafontaine
Service de Pédagogie expérimentale, Université de Liège, Belgique
ariane.baye@ulg.ac.be, dlafontaine@ulg.ac.be

Mots clés : traduction, difficulté items, évaluations internationales

Résumé :
Depuis les premières évaluations internationales organisées par l'Association internationale pour l'évaluation du rendement scolaire (IEA), de nombreux progrès ont été réalisés pour assurer la qualité des traductions du matériel de test et pour obtenir un relatif équilibre linguistique des supports d'évaluation. L'équivalence des tests est en effet une condition nécessaire à l'égalité des pays participant à l'évaluation. Entre le début des années 70 les évaluations les plus récentes, telles que le Programme International pour le Suivi des Acquis des élèves (PISA), des dispositifs de plus en plus complexes et contraignants ont été mis en place en amont, au niveau de la conception des instruments, pour harmoniser et améliorer la qualité des traductions. Malgré ces progrès, cet aspect des évaluations internationales reste encore un point sensible, car aucune analyse statistique ne peut prouver l'indépendance des difficultés linguistique et cognitive dans ce type de test. L'objectif de la contribution est, d'une part, de décrire les évolutions en termes de procédures de traduction et d'équilibre linguistique des supports d'évaluation lors de la préparation des instruments, et, d'autre part, de montrer comment certaines techniques statistiques disponibles peuvent aider à déterminer *a posteriori* l'influence d'un matériel traduit sur la difficulté d'un test. La première partie de l'exposé s'appuiera sur les évaluations de la lecture organisées par l'IEA et l'OCDE des années 70 à nos jours. Pour la seconde partie, le corpus est constitué des bases de données cognitives de PISA 2000 et 2003, en lecture, mathématiques et sciences, afin de confronter les résultats obtenus à différents moments et dans différents domaines cognitifs.

ANALYSE SECONDAIRE ET QUALITATIVE DE L'ÉTUDE INTERNATIONNALE PISA

Claude Houssemand, Université du Luxembourg

Résumé :

La présente communication se réfère à une étude réalisée par l'Université du Luxembourg et l'Université de Moncton (Canada) et relative à une analyse secondaire et qualitative de l'étude internationale PISA 2000 (certains éléments de PISA 2003 ont également été utilisés). Ce travail a permis de mettre en évidence que, malgré les volontés affichées par l'OCDE, les scores obtenus aux items de ce type d'enquête ne sont pas indépendants de caractéristiques culturelles pouvant être définies comme des biais potentiels

des measures

.

Un premier travail a permis de montrer que la familiarité avec des modes de questionnements différents peut, en partie, être explicative de la performance à des items de nature différente. Une comparaison entre deux pays proches selon des critères linguistiques et socio-économiques mais éloignés suivant la réussite moyenne à l'enquête PISA 2000 a permis de montrer que ceux ayant le mieux réussi sont également les plus habitués à ce mode de questions et de réflexions sur un support écrit. Ce point a été confirmé en comparant les sentiments d'habitudes d'élèves provenant de régime d'enseignement différents et utilisant des modes pédagogiques différents. Un second travail a mis en évidence une divergence entre la structure

théorique cognitive définie a priori pour la constitution des items de l'enquête PISA 2000 et celle obtenue selon les performances empiriques des élèves interrogés puisque les différents scores obtenus se référeraient tous à un facteur général cognitif de traitement de l'information pouvant être rapproché du facteur d'intelligence fluide Gf de la littérature psychologique. Ce point n'est en aucun cas en contradiction avec les résultats précédents puisqu'il est maintenant acquis que cette dimension de la compétence cognitive n'est nullement indépendante de facteurs environnementaux. Un troisième et dernier travail a montré que des facteurs relatifs à des caractéristiques sociales, familiales, économiques et personnelles peuvent être rattachés aux performances cognitives des élèves. Ainsi, certaines particularités peuvent être retenues pour le Luxembourg. En effet, les facteurs les plus discriminatifs cognitivement sont la langue utilisée dans le foyer avec une différence positive en faveur du luxembourgeois et de l'allemand et les résultats des élèves lors des enseignements de l'allemand. Les autres facteurs explicatifs sont d'une importance légèrement inférieure.

LES BIAIS CULTURELS DANS LES TESTS INTERNATIONAUX DE FRANÇAIS LANGUE ÉTRANGÈRE : DIAGNOSTIC ET ANALYSE

DESROCHES Franck, Responsable du TEF, Chambre de commerce et d'industrie de Paris, Paris, France, fdesroches@ccip.fr

Chambre de commerce et d'industrie de Paris, Direction des relations internationales de l'enseignement, 28, rue de l'Abbé Grégoire, 75279 PARIS Cedex 06, France.

CRENDAL Alexandra, Responsable pédagogique, Chambre de commerce et d'industrie de Paris, Paris, France, acrendal@ccip.fr

RENAUD François, Responsable pédagogique, Chambre de commerce et d'industrie de Paris, Paris, France, frenaud@ccip.fr

CASANOVA Dominique, Responsable pédagogique, Chambre de commerce et d'industrie de Paris, Paris, France, dcasanova@ccip.fr

DEMEUSE Marc, Professeur, Institut d'Administration scolaire, Université de Mons-Hainaut, Mons, Belgique, marc.demeuse@umh.ac.be

ARTUS Frédérique, Chercheuse, Institut d'Administration scolaire, Université de Mons-Hainaut, Mons, Belgique, frederique.artus@umh.ac.be

Mots clés : évaluation, français langue étrangère, biais culturels

Introduction

Les tests internationaux de français langue étrangère (FLE) sont récents : les premiers (CGM 62, SOL test) sont apparus dans les années 70. Ils se caractérisent par une standardisation des conditions et des consignes de passation, première exigence pour assurer la qualité d'un test, et par un renvoi au référentiel de l'époque, le Français Fondamental (Gougenheim et al., 1956). Leur diffusion à l'étranger aura finalement peu de portée. Les années 80 voient un tournant avec l'avènement de nouvelles tendances en didactique des langues : les approches communicatives ; les tests de langues sont alors relégués au second plan. A cette époque, peu d'études sont menées sur les biais culturels.

Ce n'est qu'à la fin des années 90, en 1998, que le premier Test d'Évaluation de Français (TEF) pour non francophones est mis au point par une équipe de linguistes et

didacticiens de la Chambre de commerce et d'industrie de Paris (CCIP)[1], rejointe par un Comité Scientifique pluridisciplinaire (CCIP, 2006). Sa conception repose sur une multiréférentialisation (Demeuse, 2004), les ouvrages de référence étant le Cadre européen commun de référence pour les langues (CECR) du Conseil de l'Europe (2000), échelle de niveaux et de compétences s'appuyant sur la perspective actionnelle, et les Niveaux de compétence linguistique canadiens (CIC, 2002)[2]. Le TEF se base sur les approches communicatives. Deux visions, *a priori* opposées, semblent coexister : d'une part une langue normée pour les besoins de standardisation du test et, d'autre part, une langue vivante qui évolue au fil des usages que les natifs en font et qui s'inscrit dans les approches communicatives.

Face à ce paradoxe, quel choix doit préexister à la création d'un test de FLE destiné à des candidats de langues maternelles différentes dont, selon G. Mounin (1975), la pratique et la maîtrise « déterminent, prédéterminent même, notre vision du monde » ? C'est précisément la question que les auteurs du TEF se sont posée lors du processus d'élaboration du test. D'autres interrogations, secondaires, ont également fait l'objet de discussions : un accent neutre peut-il être retenu pour l'épreuve de compréhension orale ? Quels outils développer pour éviter le "franco-français" et les préjugés culturels ? Comment faire en sorte que les questions liées à la partialité soient rejetées ? Comment rester objectif en expression écrite et orale ? Autant de questions auxquelles les concepteurs du test répondent à travers les spécifications du test. Ces interrogations ne sont pas sans lien avec la notion de biais culturels (CCIP-DRI/E, 2003) qui a fait l'objet d'un contrôle particulier de la qualité des items.

L'étude sur les biais culturels présentée ci-après se proposera de présenter une typologisation des biais culturels rencontrés aux différentes étapes de conception, validations pédagogique et psychométrique, leur fréquence et les outils créés pour prévenir leur apparition dans les épreuves du TEF. Pour la démonstration, les auteurs s'appuieront sur un corpus de données empiriques recueillies en amont et en aval des pré-tests et s'étalant sur la période de janvier 2004 à juin 2006. L'exposé proposera deux approches complémentaires : une approche psychométrique complétée par une lecture pédagogique critériée.

[1] Consultable sur http://www.fda.ccip.fr

[2] Cette dénomination est récente et remplace désormais celle utilisée de 2002 à 2006 : les Standards linguistiques canadiens. Ces niveaux sont consultables sur http://www.language.ca/display_page.asp?page_id=533

1. Vers une sociolinguistique du français. Etude de cas : le Test d'Evaluation de Français (TEF)

La question du choix d'une langue pour la conception d'un test de français langue étrangère oblige les auteurs du test à faire une analyse de besoins, étude qui ne peut être réalisée qu'après avoir déterminé précisément les objectifs et le public ciblés par le test.

1.1. De la commande institutionnelle à l'analyse des besoins du public

Le Test d'Évaluation de Français a été initialement créé pour répondre à deux commandes : l'une des écoles de la Chambre de commerce et d'industrie de Paris, l'autre d'établissements d'enseignement à l'étranger dispensant des cours de français ou en français à un public non francophone. Les besoins langagiers de ces publics ont donc été analysés et ont révélé des situations de communication authentiques variées ; celles communes à l'ensemble de ces publics ont permis à la DRI/E-CCIP d'inscrire les domaines (personnel, éducatif, etc.) et thèmes de communication (l'environnement, les achats, etc.) dans les spécifications du test en précisant leur niveau en regard du CECR.

Outre ces premiers choix, la DRI/E-CCIP a dû s'orienter vers une variété du français, celle la plus adaptée au public et objectifs du test. Parlée dans plusieurs pays, la langue française possède différents statuts selon les régions ou les pays francophones. En outre, elle subit constamment des actualisations rendues nécessaires par l'évolution socio-économique des groupes sociaux dont elle reflète l'identité. Le statut de la langue et le passé historique du pays expliquent aussi les évolutions de cette langue que les candidats au test sont amenés à rencontrer aussi bien en contexte formel, en classe de langue, qu'informel, dans la rue. Le test s'adressant à un public large amené à rencontrer toutes les variétés du français, aussi bien en contexte endolingue qu'exolingue, le choix définitif s'est porté sur le français standard de France, des médias.

Ce choix nécessaire à la standardisation n'est pas sans influencer la passation du TEF dans des pays francophones. Par exemple, en compréhension orale, il arrive que des étudiants soient surpris, en Belgique, d'entendre « quatre-vingt-dix » et non « nonante » ; en lexique / structure, des assertions comportant le verbe « dîner » pourraient induire en erreur des candidats passant le test au Québec et associant la notion de dîner à déjeuner. À cette variété écrite s'ajoute celle de l'oralité rencontrée dans les épreuves de compréhension et production orales : au-delà du texte support, les éléments segmentaux et suprasegmentaux viennent complexifier le niveau d'évaluation. Afin d'atténuer cette variante, la DRI/E-CCIP porte un intérêt particulier aux transcriptions transmises par les concepteurs et aux normes d'enregistrement de ces supports.

Au-delà du choix des spécifications du test qui affirment la recherche d'une standardisation de la langue, la DRI/E-CCIP reste vigilante quant à la production de

batteries de tests qui doivent rester équivalentes au fil du temps pour assurer la fiabilité du test (Demeuse et al., 2005) et ne pas favoriser un groupe particulier. Là intervient la notion de culture, « un ensemble de caractéristiques propres à une société donnée », selon la définition donnée par Robert Galisson. Le TEF cherche constamment à neutraliser les aspects singuliers d'une culture donnée afin d'éviter de biaiser le résultat des candidats au test. Ces spécificités culturelles sont à la source de ce que nous nommerons dans cette étude « biais culturels ».

1.2. Présentation des spécifications du test

Le Test d'Évaluation de Français (TEF) est un test de français langue étrangère qui s'adresse à des non francophones. Il se compose de plusieurs épreuves : la compréhension écrite, la compréhension orale, le lexique / structure, l'expression écrite et l'expression orale, dont les spécifications générales sont présentées dans le tableau 1.

Tableau 1 – Les épreuves du TEF

Indexation sur le CECR	Epreuves obligatoires	Durée	Structure	Score sur
Réception	**Compréhension écrite**	60 min	50 questions	300 points
Réception	**Compréhension orale**	40 min	60 questions	360 points
Linguistique	**Lexique / structure**	30 min	40 questions	240 points
	TOTAL	2H10	150 questions	900 points
Indexation sur le CECR	**Epreuves facultatives**	**Durée**	**Structure**	**Score sur**
Production	**Expression écrite**	60 min	2 sujets à traiter	450 points
Interaction	**Expression orale**	35 min	2 sujets à traiter	450 points

Source : D'après la brochure de présentation du Test d'Evaluation de Français, 2006.

Pour les deux dernières épreuves, ouvertes et plus exposées à la subjectivité, des grilles d'évaluation ont été élaborées afin de réduire les écarts entre examinateurs et de rendre objective l'évaluation de ce type d'épreuve. Conçues à partir du CECR, ces deux grilles comptent chacune douze critères d'évaluation et 7 niveaux TEF indexés sur les six niveaux du CECR.

Basé sur les approches communicatives, le TEF s'appuie sur des documents authentiques et présente, pour ses épreuves de production écrite et orale, des sujets, des situations que les candidats au test peuvent rencontrer dans la vie quotidienne. Attrayantes et motivantes pour le candidat, ces situations de communication authentiques ne sont pas sans complexifier le travail des concepteurs. Elles requièrent toute l'attention du Comité d'Évaluation et de Validation, chargé de « traquer » toute imperfection au niveau de l'item ou du support, et notamment les biais culturels.

Les spécifications du TEF et leurs mises en œuvre s'appuient sur des critères de scientificité légués par la recherche dans les domaines de la linguistique et de la psychométrie. Ainsi, parmi les réflexions et recherches sur lesquelles reposent la création du test, président les travaux en sociolinguistique de Louis-Jean Calvet (1987) et les références données par Gilbert De Landsheere sur la définition scientifique d'un test (De Landsheere, 1992).

2. Définition et essai de typologisation des biais

Précisons la définition de biais culturel. De Landsheere (1992) indique qu'« une mesure est *biaisée* si elle conduit systématiquement à des résultats inexacts dus à des défectuosités inhérentes à l'instrument utilisé, à la façon d'utiliser un instrument ou à la façon d'interpréter les résultats ». Pour ce qui concerne le Test d'Évaluation de Français (TEF), la question du biais culturel se pose de multiples façons, à toutes les phases critiques de l'acte d'évaluer un candidat avec le TEF : depuis la conception des items de test et des sujets d'épreuves de production, jusqu'à l'évaluation proprement dite de la performance du candidat par l'examinateur d'expression orale.

Un biais culturel surgissant lors de la phase de conception du test renvoie, dans la définition de De Landsheere, à *« des défectuosités inhérentes à l'instrument utilisé »*. En regard des activités du test, on pourrait reformuler la définition comme suit : **erreur de conception touchant** un ou plusieurs **éléments constitutifs d'une unité** d'évaluation (consigne, document support, amorce, clé, distracteur) **susceptible d'influencer la réponse** du candidat (vers une réponse fausse, vers une réponse vraie, vers l'abstention) **pour des raisons d'appartenance** à un « groupe » culturel (genre, orientation politique, religieuse, groupe social, groupe national, âge, etc.) **ou de connaissances** culturelles (culture générale non universellement partagée sur des faits culturels ou des caractéristiques historiques, géographiques, sociales, économiques, religieuses, etc. auxquels peuvent renvoyer un ou plusieurs éléments constitutifs d'une unité d'évaluation) **indépendamment de sa compétence réelle à communiquer** (et qui est, normalement, la seule compétence évaluée).

La typologie d'erreurs de conception TEF, dont un extrait est présenté dans le tableau 2, développée à l'intention du Comité d'Évaluation et de Validation (chargé de valider pédagogiquement les items proposés par les concepteurs) et des concepteurs eux-mêmes, prévoit deux catégories d'erreurs sous la rubrique « Culture ».

Tableau 2 – Extrait de la typologie d'erreurs de conception

Typologie	Définition	Remédiation
Culture	Thématique culturellement / politiquement sensible pour certains candidats.	*Neutraliser*
	Référence à un particularisme culturel / politique / social / économique non explicité.	*Rendre universel*

Source : Document extrait de la typologie d'erreur (CTEF-LIS-15)

Ces deux catégories renvoient, d'une part, à l'appartenance du candidat à un groupe culturel donné et, d'autre part, aux connaissances culturelles que le candidat possède. Naturellement, appartenance culturelle et connaissances culturelles varient considérablement d'un candidat à l'autre. Cette catégorisation binaire des biais

culturels a un caractère général. Chaque catégorie renvoie en fait à des réalités multiples et complexes. Ainsi, lorsque l'on parle de groupe culturel, on peut définir ce groupe de différentes façons. Mais ce sont les mêmes mécanismes qui sont à l'œuvre pour constituer le biais culturel dans l'outil d'évaluation conçu. Lorsqu'une unité de test peut donner lieu, d'un candidat à l'autre, à une différence de traitement (traitement de la tâche par le candidat ou traitement/évaluation du candidat par le correcteur/examinateur) en raison de leur appartenance à tel ou tel groupe culturel, le biais culturel est établi.

Dans un premier temps, on peut définir le « groupe culturel » comme un ensemble d'individus liés par leur appartenance nationale (ou régionale), par leurs convictions idéologiques (politiques ou religieuses), par leur statut social ou professionnel, par leur culture éducative, etc. Ainsi, une unité de test peut être porteuse d'un biais culturel lié à l'appartenance du candidat à un groupe de conviction politique ou idéologique car elle impose un point de vue distinct ou opposé à celui du groupe. Par exemple, la figure 1 présente un exemple d'unité qui a été rejetée par le Comité d'Évaluation et de Validation. La situation proposée dans cette unité est très marquée idéologiquement : elle impose un point de vue qui n'est pas forcément celui du candidat. Il éprouvera plus de difficultés à défendre un point de vue contraire à ses convictions qu'un candidat partageant la prise de position imposée par la consigne.

ES511005 - CEB « Rébellion-Energétique. L'EPR contre-attaque »

Consigne : « Vous êtes sûr(e) que **l'énergie éolienne est une bonne énergie** et que le nucléaire n'est pas une bonne énergie »

Figure 1 – Item présentant un risque de biais culturel lié aux valeurs

De la même manière, une unité de test contenant des assertions négatives sur une culture nationale ou régionale, peut entraîner une gêne déstabilisante chez un candidat ressortissant de cette culture. Par exemple, l'unité présentée en figure 2 a été corrigée par le Comité d'Évaluation et de Validation pour cette raison. En effet, le document fait état d'un jugement de valeur sur un fait culturel (ici, américain) en l'opposant à une autre culture (ici, française). Ce jugement de valeur négatif est susceptible de gêner un candidat américain et de perturber sa concentration pour la suite du test.

SP0205-04 CEB « Histoire d'une rencontre »

Document : « J'adore le pain français. Je souligne français, parce que je suis américain, et que le pain dans mon pays est généralement **immonde**. »

Figure 2 – Item présentant un risque de biais culturel lié à l'image

Si l'on intègre à la définition de « groupe culturel » des caractéristiques comme le sexe ou l'âge (dans leur dimension culturelle), le même type de biais peut survenir. Ces caractéristiques peuvent représenter une source de biais en elles-mêmes (lorsque

la situation d'évaluation exige, par exemple, qu'un candidat masculin joue le rôle d'une femme), ou bien agir comme un renforcement de caractéristiques autres (sociales, religieuses, etc.) et donc renforcer le biais induit par l'unité porteuse de l'erreur de conception.

De la même manière, si l'on intègre à la notion de « groupe culturel » des caractéristiques plus individuelles, relevant peut-être davantage de la psychologie individuelle ou de l'histoire personnelle que de l'imprégnation culturelle, de nouvelles sources de biais apparaissent, qui établiraient une inégalité face à la tâche en jouant sur les différences de sensibilité des candidats face à une situation morbide ou macabre. Tel est le cas de l'unité présentée dans la figure 3. Cette unité a été rejetée par le Comité d'Évaluation et de Validation pour cette raison. Pour peu que le sujet développé dans le document fasse écho chez le candidat à une angoisse ou à un traumatisme personnel, sa capacité à répondre risque d'être perturbée.

2TF0104-12 CEC « Disparitions »

Document : «°Il y a des histoires qui fascinent parce qu'elles naviguent dans ces **territoires troubles** où l'espérance et la **peur** n'en finissent pas de se confondre : ce sont les affaires de **disparitions**. »

Figure 3 – Item présentant un risque de biais culturel lié à la sensibilité personnelle

Lorsque l'on parle de connaissances culturelles, il peut s'agir d'une compétence culturelle générale (culture générale), forcément variable d'un candidat à l'autre, et qui peut servir de pierre d'achoppement à un biais culturel lorsque l'unité en cause propose un item auquel il est plus facile de répondre si l'on possède tel ou tel élément de culture générale. La figure 4 en est une bonne illustration. Les options A et B renvoient à la distinction entre un nom et un prénom japonais. Cette distinction n'est pas évidente, voire impossible pour un candidat qui n'a aucune connaissance particulière de la culture japonaise. En l'occurrence, il ne s'agit pas seulement de savoir à quoi renvoient les termes « nom » et « prénom » (compétence linguistique), mais également si Akiko est un nom ou un prénom (connaissance culturelle).

ES230504-LSA « Mon *** est Akiko »

Item : 1) Mon…… est Akiko.

A) **prénom**

B) **nom**

Figure 4 – Item présentant un risque de biais culturel lié à des connaissances culturelles

Cela vaut quelles que soient les langues ou les cultures concernées. A fortiori lorsqu'il s'agit d'une culture française (de France). Dans ce cas, les candidats appartenant à des cultures « proches » (occidentales, européennes) sont avantagés par rapport à des univers culturels plus éloignés. Un exemple symptomatique de biais

culturel de ce type est présenté dans la figure 5. Par exemple, l'enseigne d'une pharmacie n'est pas symbolisée dans tous les pays par une croix. Un dessin particulièrement signifiant pour les uns ne le sera donc pas forcément pour les autres.

SP0106-10 COA « Itinéraire en ville »

Dessin A : Une pharmacie qu'on peut identifier d'après son enseigne : une **croix**

Figure 5 – Item présentant un risque de biais culturel lié à des connaissances culturelles

De la même manière, l'erreur de conception dans l'unité peut renvoyer à une réalité quotidienne (probablement celle du concepteur) mais très localisée. Il est très possible qu'un candidat n'ait aucune expérience de ce dont on parle. Si, comme c'est le cas dans l'unité présentée dans la figure 6, le candidat est censé identifier le lieu où ce type de message est diffusé, le fait qu'un système de métro existe dans sa ville ou dans son pays représente un aspect facilitateur de la compréhension du message par rapport à un candidat qui n'a jamais eu l'expérience de ce type de transport.

ES480905-COB « Nous vous informons qu'en raison d'un incident technique »

Document : « Nous vous informons qu'en raison d'un incident technique, la circulation des **rames** sur les **lignes** 2 et 9 est fortement perturbée ».

Figure 6 – Item présentant un risque de biais culturel lié à des connaissances culturelles

La *Typologie d'erreurs* évoquée précédemment (cf. tableau 2) s'est appuyée sur le recensement des erreurs de conception les plus fréquemment relevées par le Comité d'Évaluation et de Validation. Pour les besoins de l'étude présentée ici, un nouveau recensement des erreurs de conception relevant du type d'erreur « Culture » a été effectué. Les deux catégories prévues dans la typologie ont été développées en sous catégories permettant d'identifier plus précisément les erreurs de conception relatives aux biais culturels. Elles sont présentées dans le tableau 3.

Tableau 3 – Catégories d'erreurs de conception risquant d'entraîner un biais culturel

Appartenance à un groupe culturel		**Connaissances culturelles**	
A1	Atteinte aux valeurs du groupe	0C1	Renvoi à des faits culturels localisés (non universels)
A 2	Atteinte à l'image du groupe	**C2**	Appel à des savoirs ou savoir-faire spécialisés
A 3	Atteinte à la sensibilité personnelle		

Les catégories « Appartenance… » et « Connaissances… » renvoient au profil culturel du candidat. Les sous-catégories d'erreurs numérotées A1, A2, A3 ou C1, C2, C3 sont la résultante de l'interaction « biaisante » entre le profil culturel du candidat et le profil culturel attendu par l'unité de test erronée.

3. Les outils de prévention pour éviter les biais culturels

Pour prévenir les risques d'apparition de biais culturels dans la production des concepteurs, plusieurs outils ont été élaborés : ils sont utilisés aux différents moments de la phase de conception des unités de test.

En amont, dès la phase de recrutement des concepteurs d'items de test et des correcteurs d'épreuves d'expression écrite, la procédure prévoit de déterminer si le candidat est sensibilisé à la question des biais culturels en conception et en correction. Ainsi, dans les tests de recrutement pour la conception, un exercice consiste à identifier les éléments erronés dans une unité de test. Les biais culturels font partie des erreurs de conception à identifier (type C2). Et lors de la commande d'essai (2e phase du recrutement), les annotations portées sur la commande à corriger après la 1re validation pédagogique, pointent, le cas échéant, les erreurs de conception liées à un biais culturel et proposent des pistes de correction. Dans la procédure de recrutement pour la correction des épreuves écrites du TEF, le candidat est interrogé sur sa connaissance des critères parasites de l'évaluation dont les biais culturels font partie. Lorsqu'un candidat à la conception a subi avec succès les tests de recrutement et qu'il intègre le vivier des concepteurs d'épreuves, le Guide du concepteur (cf. tableau 4) qui lui est remis identifie les différents risques de biais auxquels il doit être attentif et indique, pour chaque phase de son travail de conception (choix des documents, rédaction des consignes, etc.) comment les réduire.

En aval, lors des phases de validation pédagogique des commandes, les membres du Comité d'Évaluation et de Validation peuvent s'appuyer, pour mémoire, sur la typologie d'erreurs (cf. tableau 2) qui contient la rubrique « Culture » déjà citée. L'expérience interculturelle des membres du Comité (plusieurs années d'expérience de vie et d'enseignement à l'étranger, dans des contextes culturels variés) vient renforcer leur capacité à identifier des biais culturels dans la production des concepteurs. Lorsque le biais culturel est établi, le Comité d'Évaluation et de Validation propose soit la correction de l'erreur lorsque cela est possible, soit le rejet de l'unité en cause.

Enfin, suite à l'analyse psychométrique des items pré-testés, la DRI/E-CCIP identifie et, lorsque cela est possible, neutralise les items comportant une erreur de conception entraînant un biais culturel.

Tableau 4 – Extrait du Guide du concepteur sur les biais culturels donnés (CTEF-GUID-01)

p.	Chapitre	Citation	Type	Commentaires
26	Types de textes en CE	• documents assez courts tirés de la presse francophone généraliste traitant de sujets variés	C2 C3	• francophone : pour éviter le franco-français • généraliste : pour éviter la langue de spécialité
72	LS Section C	• Éviter les tournures familières : préférez un registre standard ou soutenu (pour les niveaux avancés).	C2	• tournures familières : souvent l'expression de particularismes linguistiques ou culturels régionaux • tournures familières : langue non stabilisée
76	EE Section A	• Éviter cependant les thèmes macabres ou trop sensationnels • Éviter les thématiques qui pourraient gêner des candidats pour des raisons culturelles, politiques ou religieuses	A3 A1/A2	• macabres : risque d'atteinte à la sensibilité individuelle • gêner : risque de préjugés, d'atteinte à l'image, aux valeurs
78	EE Section B	• (Permettre) le débat sans choquer ni exalter les candidats. • L'opinion du candidat n'a aucune importance : seul compte la façon dont il l'exprime et l'articule • Préférer un sujet peu original mais fécond en arguments à un sujet si inattendu qu'il laissera le candidat coi !	A1 C1	• risque d'atteinte à l'image, aux valeurs • risque de disparités liées à la culture générale
86	Sélection des documents (écrits et oraux)	Variété des sources • non seulement français, mais aussi plus largement francophones Authenticité des documents • masquer les marques commerciales et les noms de « célébrités » cités dans les documents originels par des pseudonymes plausibles Thématique • renvoyant à des réalités francophones pouvant intéresser tous les candidats, pas ou peu spécialisés • rejeter des documents faisant référence, de manière explicite ou implicite, à des réalités culturelles, sociales ou politiques nationales, françaises ou parisiennes • proposer des documents « politiquement corrects », c'est-à-dire ne portant pas atteinte à la sensibilité nationale, politique ou religieuse des candidats • veiller à la neutralité ou l'universalité des thèmes exploités Pérennité • thème choisi ne renvoie pas à un événement trop ponctuel et […] reste d'actualité à un horizon de trois ans	 C2 C2 C3 C2 A1/A2 A/C C2	 • francophone : pour éviter le franco-français • une marque, une célébrité peut n'être connue que localement • thématique spécialisée : savoirs et savoir-faire spécialisés • éviter des références franco-françaises probablement inconnues des candidats • ne risquant pas, pour une raison quelconque, de les perturber ou de les blesser • L'actualité « brûlante » est souvent liée au contexte culturel

88	L'exploitation d'un document long écrit ou oral (Compréhension)	• [Les questions] ne doivent pas non plus porter sur des implicites culturels ni même des connaissances culturelles si celles-ci ne sont pas explicitées dans le document • Souvenez-vous que la règle d'or est que le candidat ne doit se déterminer qu'en fonction de la seule compréhension qu'il a du document et de l'item	C1/C2 C	• L'explicitation est une remédiation à l'implicite culturel • La tâche ne doit pas être réalisable en faisant appel à d'autres compétences que celle visée par l'objectif
90	Principes de rédaction d'items	• Lorsque les items s'appuient sur un texte, il ne doit pas être possible d'y apporter une réponse juste seulement fondée sur la culture générale sans avoir lu et compris le texte	C1	• risque de disparités liées à la culture générale
94	Critères de validation	Thématique • Le document proposé développe une thématique acceptable	A	• risque de préjugés, d'atteinte à l'image, aux valeurs
95	Typologie d'erreurs	Culture • Thématique culturellement / politiquement sensible pour certains candidats. • Référence à un particularisme culturel / politique / social / économique non explicité.	A C	

4. De l'évaluation quantitative des biais à leur correction

4.1. Au niveau de la conception et de la validation pédagogiques

Les données statistiques recueillies pour cette étude résultent du relevé des erreurs de conception de la catégorie « Culture » commentées par le Comité d'Évaluation et de Validation lors des phases de validation pédagogique des commandes. Elles concernent, d'une part, les « commandes d'essai » proposées aux candidats à la conception d'items et, d'autre part, les commandes TEF « normales » passées auprès des concepteurs du TEF. Il a semblé instructif de relever, au delà du nombre d'unités et d'items contenant des erreurs, la localisation de ces erreurs, soit sur le document support (lorsque l'unité en comporte un), soit sur l'item (une unité peut en compter plusieurs). Ces données apportent un éclairage sur l'efficacité des outils de prévention à disposition.

On constate, par exemple, que le taux d'erreurs de conception va en s'amenuisant, au fur et à mesure que les concepteurs acquièrent de l'expérience. Ainsi pour les commandes d'essai réalisées par des concepteurs « novices », les taux d'erreurs sont deux à trois fois plus élevés que ceux des commandes TEF, réalisées par des concepteurs réguliers. Cette réduction du taux d'erreurs entre commandes d'essai et commandes TEF s'explique également par le fait que les outils de formation tels que le Guide du concepteur n'est mis à la disposition des concepteurs que lorsque ces derniers sont effectivement engagés et donc tenus par leur engagement de confidentialité. On fait également le constat que le taux d'erreurs diminue entre la première et la seconde validation pédagogique des commandes (commandes d'essai et commandes TEF), ce qui s'explique par le fait que les erreurs relevées par le Comité d'Évaluation et de Validation en première lecture sont généralement corrigées par le concepteur avant la seconde lecture.

Les données collectées permettent également d'établir des taux d'occurrence d'erreurs de conception unité-type (toute unité-type renvoie à un niveau, un objectif, une tâche) par unité-type et concepteur par concepteur. Il est ainsi possible de déterminer précisément quelles unités présentent le plus de risques d'erreurs et pour quel concepteur. Ces informations viennent donc réalimenter les outils de suivi et de formation mis en place pour le contrôle de la qualité. A titre d'exemple, on peut se pencher sur les épreuves d'expression écrite et orale. Pour l'épreuve d'expression écrite, l'unité-type de la section A, qui correspond à l'objectif « rédiger un récit » présente un risque d'erreur sensiblement plus élevé que celle de la section B (argumenter un point de vue). Pour l'épreuve d'expression orale, c'est l'inverse : les unités les plus « risquées » sont celles de la section B dont l'objectif est de présenter un document et de soutenir son point de vue.

D'une manière générale, on peut émettre l'hypothèse que les unités faisant appel à un document-support à caractère authentique sont plus porteuses de risques que les unités dépourvues de document-support ou s'appuyant sur des documents-supports didactisés.

4.2. Au niveau des traitements psychométriques

Lors de l'analyse des traitements psychométriques des résultats aux pré-tests, on constate que le manque de discrimination ou le pourcentage de réussite de certains items peuvent s'expliquer par la présence de biais culturels. Le tableau 5 propose un recensement des biais culturels décelés à l'issue des pré-tests pour les 12 derniers jeux.

Tableau 5 – Relevé des erreurs de conception de type culturel
Pré-tests - Période de janvier 2004 à juin 2006

Nombre de pré-tests	**12**
Biais culturels en compréhension écrite	**4**
Biais culturels en compréhension orale	**4**
Biais culturels en lexique et structure	**3**
Nombre total de biais culturels	**11**
Nombre total d'items	**1386**
Fréquence des biais	**0,8%**

Il permet d'estimer à 1% (soit en moyenne un item par jeu de pré-test) la part des items pré-testés présentant des biais culturels. Contrairement à la phase de conception, ces items se répartissent de façon générale entre les trois épreuves obligatoires. Sur ces onze items, huit avaient un indice de discrimination (R.bis) compris entre –0,22 et 0,19. Les trois autres items avaient un indice de discrimination compris entre 0,31 et 0,37, mais le taux de bonnes réponses était nettement supérieur ou nettement inférieur au taux escompté compte tenu du niveau présumé de l'item. Voici un exemple d'item de compréhension orale (cf. figure 7) qui présente un biais culturel décelé lors de l'analyse psychométrique : dans ce cas précis, il semble qu'acheter un billet d'avion dans un grand magasin ne soit pas une valeur partagée par tous.

2TF0405 – COB
Script :
Partez en vacances avec l'agence de voyages d'Uniprix ! Préparez vos vacances d'été ! Vous désirez partir au Sahara ? En Espagne ? Pas de problème ! Cette semaine, nous offrons jusqu'à 75 euros de réduction sur les billets Air France. Au premier étage, près de l'espace librairie, nos hôtesses vous accueillent pour répondre à vos questions ! Alors, ne dépensez plus des fortunes pour partir !

Ce message est diffusé
A. dans un aéroport
B. dans un grand magasin
C. au guichet d'Air France
D. dans une agence de voyage

Figure 7 – Item de compréhension orale présentant un biais culturel décelé à l'analyse psychométrique

4.3. Au niveau du retour d'information des centres agréés

Un troisième contrôle de la qualité est également opéré sur les unités du TEF. Ce contrôle est réalisé grâce à des analyses a posteriori sur les épreuves obligatoires et facultatives (calcul de la corrélation inter-juges pour les correcteurs et les examinateurs). Le retour d'informations des centres agréés faisant passer le TEF est également un moyen efficace de remédier à ces biais culturels et de rétroagir. Ce retour d'informations se fait sur des feuilles d'anomalies que les centres agréés TEF retournent à l'issue de la session à la DRI/E-CCIP.

Les retours d'information les plus fréquents concernent l'épreuve d'expression orale du TEF. L'épreuve étant passée dans le centre agréé TEF et animée par des examinateurs locaux, les retours d'informations sur cette épreuve sont donc plus nombreux. Par exemple, l'unité que constitue la section B de l'épreuve (2e sujet de l'épreuve) peut poser quelques difficultés d'ordre culturel pour des candidats peu habitués à défendre leur point de vue. Dans ce cas précis, les habitudes éducatives peuvent influencer la capacité du candidat à s'exprimer. Les biais culturels sont également décelables dans l'échange oral guidé par l'examinateur et dans son interprétation des critères inscrits dans la grille d'évaluation. Par exemple, on constate que les écarts entre deux examinateurs sont plus importants sur les critères communicatifs que sur les critères linguistiques.

Conclusion

Les résultats de cette étude sur les biais culturels révèlent que les biais peuvent être de plusieurs natures : ils peuvent être liés aux valeurs, à l'image, à la sensibilité personnelle et aux connaissances culturelles du candidat, du correcteur et de l'examinateur. Les conclusions établies par les auteurs quant à la catégorisation des biais culturels à partir d'une base de données d'items rejoignent celles proposées par Bachman (1995).

Ce premier regard posé sur les biais culturels dans les évaluations internationales de français langue étrangère a permis de mettre en évidence les nombreux paramètres qui pourraient influencer l'évaluation du candidat, du correcteur et de l'examinateur. Des études complémentaires sont actuellement en cours : il s'agit à l'aide de la méthode DIF (differential item functioning) de poursuivre ces travaux sur les biais culturels en croisant le regard du linguiste avec celui du psychométricien.

L'enjeu de la démonstration est de taille : au delà de la nécessaire standardisation, il s'agit bien de proposer un test équitable pour tous les candidats. Cette notion d'équité est essentielle pour les politiques d'accueil et de suivi des flux de populations allophones dans un contexte de plus en plus marqué par l'internationalisation des cursus d'études supérieures et des parcours professionnels.

Bibliographie

BACHMAN, L. (1995). *Fundamental Considerations in Language Testing* (3e édition). Oxford : Oxford University Press.

CALVET, L.-J. (1987). *La guerre des langues et les politiques linguistiques*. Collection Pluriel. Paris : Hachette Littératures.

CENTRE DES NIVEAUX DE COMPÉTENCE LINGUISTIQUE CANADIENS (2002). *Standards linguistiques canadiens 2002*. Ottawa : Citoyenneté et Immigration Canada.

CENTRE DES NIVEAUX DE COMPÉTENCE LINGUISTIQUE CANADIENS (2006). *Niveaux de compétence linguistique canadiens 2006 : Français langue seconde pour adultes*. Ottawa : Citoyenneté et Immigration Canada.

CHAMBRE DE COMMERCE ET D'INDUSTRIE DE PARIS (2006) : *Le Test d'Évaluation de Français de la Chambre de commerce et d'industrie de Paris.* Paris : Chambre de commerce et d'industrie de Paris, Centre de Langue de la Direction des relations internationales de l'Enseignement.

CHAMBRE DE COMMERCE ET D'INDUSTRIE DE PARIS – DIRECTION DES RELATIONS INTERNATIONALES DE L'ENSEIGNEMENT (2003) : *Adéquation du Test d'Évaluation de Français aux critères d'évaluation de Citoyenneté et Immigration Canada.* Paris : Chambre de commerce et d'industrie de Paris, Centre de Langue de la Direction des relations internationales de l'Enseignement.

CONSEIL DE L'EUROPE (2000). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer.* Paris : Didier.

DE LANDSHEERE, G. (1992) : *Dictionnaire de l'évaluation et de la recherche en éducation* (2e édition). Paris : Presses Universitaires de France.

DEMEUSE, M., CRENDAL, A., DESROCHES, F., OSTER, P., RENAUD, F., LEROUX, X. (2004). *L'évaluation des compétences linguistiques des adultes en français langue étrangère dans une perspective de multi-référentialisation. L'exemple du Test d'Évaluation de Français (TEF) de la Chambre de commerce et d'industrie de Paris (CCIP).* In Actes du 17e colloque international de l'Association pour le Développement des Méthodologies d'*É*valuation en Education (ADMEE-Europe). Lisbonne, 18-20 novembre 2004

DEMEUSE, M., DESROCHES, F., CRENDAL, A., RENAUD, F., CASANOVA, D., ARTUS, F. (2005). *La fiabilité de l'évaluation des compétences linguistiques pour des adultes non francophones : présentation d'un protocole d'évaluation*. In Actes du 18e colloque international de l'Association pour le Développement des Méthodologies d'*É*valuation en Education (ADMEE-Europe), Reims, 24-26 octobre 2005.

DESROCHES, F., CASANOVA, D., CRENDAL, A., RENAUD, F., MAHY, C., DEMEUSE, M. (2005). *Mise en équivalence des versions papier-crayon et électronique d'un test de français langue étrangère et mise en perspective des résultats par une procédure*

d'autoévaluation, In Actes du 18e colloque international de l'Association pour le Développement des Méthodologies d'*É*valuation en Education (ADMEE-Europe), Reims, 24-26 octobre 2005

GOUGENHEIM, G., MICHEA, R., RIVENC, P., SAUVAGEOT, A. (1956). *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base.* Paris : Didier.

MOUNIN, G. (1975). *Clefs pour la linguistique*, Editions Seghers.

Evaluation et autoévaluation en français langue étrangère : approche empirique du *Cadre européen commun de référence pour les langues* (CECR) du Conseil de l'Europe à travers le TEF et l'échelle d'autoévaluation DIALANG

DEMEUSE Marc, Professeur, Institut d'Administration scolaire, Université de Mons-Hainaut, Mons, Belgique, marc.demeuse@umh.ac.be

Institut d'Administration scolaire, Faculté de Psychologie et des Sciences de l'Education, Université de Mons-Hainaut, Académie universitaire Wallonie-Bruxelles, Place du Parc, 18, B-7000 Mons, Belgium.

ARTUS Frédérique, Chercheuse, Institut d'Administration scolaire, Université de Mons-Hainaut, Mons, Belgique, frederique.artus@umh.ac.be

CASANOVA Dominique, Responsable pédagogique, Chambre de Commerce et d'Industrie de Paris, Paris, France, dcasanova@ccip.fr

CRENDAL Alexandra, Responsable pédagogique, Chambre de Commerce et d'Industrie de Paris, Paris, France, acrendal@ccip.fr

DESROCHES Franck, Responsable du TEF, Chambre de Commerce et d'Industrie de Paris, Paris, France, fdesroches@ccip.fr

RENAUD François, Responsable pédagogique, Chambre de Commerce et d'Industrie de Paris, Paris, France, frenaud@ccip.fr

---

Introduction

La mise en équivalence des niveaux linguistiques en langue seconde dans différentes langues est aujourd'hui une question politique et économique importante qui touche aussi bien le problème du recrutement de fonctionnaires fédéraux bilingues en Belgique que l'accès à l'immigration économique dans un pays bilingue comme le Canada (Citoyenneté et Immigration Canada, 2002). A côté du développement d'outils d'évaluation critériés à finalité certificative, il importe, principalement lorsque les résultats comportent de forts enjeux pour les candidats (Engle & Engle, 2004), de leur permettre d'avoir à la fois une bonne compréhension de la manière dont leurs compétences sont appréciées et une bonne capacité à s'autoévaluer en regard du référentiel, en particulier le *Cadre européen commun de référence pour les langues* (CECR) du *Conseil de l'Europe* (2000).

Le CECR constitue en effet un outil de référence de plus en plus important et qui devient même incontournable dans le secteur des langues étrangères, du moins en Europe. Il constitue la référence aussi bien en matière d'élaboration de tests de langue critériés (Conseil de l'Europe, Division des politiques linguistiques, 2003) que dans le

domaine de la production de manuels d'apprentissage de langues étrangères. Les tests de français langue étrangère s'y réfèrent et le *Test d'Evaluation de Français* (TEF)[1] y est donc lié. Une grille de correspondance permet de faire concorder les scores du TEF répartis en sept niveaux de compétences avec les six niveaux de l'échelle du CECR (Noël-Jothy & Sampsonis, 2006).

Le CECR sert également de référence aux différents modèles nationaux de portfolios européens des langues. Ces derniers présentent aux apprenants d'âges et de milieux variés, des propositions leur permettant de faire une autoévaluation de leurs connaissances et savoir-faire dans quatre compétences communicatives (compréhension orale, compréhension écrite, expression orale, expression écrite)[2].

DIALANG[3], premier système majeur d'évaluation mis gratuitement à disposition sur internet depuis mars 2004, est par ailleurs un outil entièrement conçu à partir du CECR et développé par plus de 20 universités européennes avec le soutien financier de la Commission européenne, Direction-Générale de l'Education et de la Culture (Programme SOCRATES, LINGUA Action D). Cet outil permet d'établir un diagnostic équivalent dans 14 langues de l'UE : l'allemand, l'anglais, le danois, l'espagnol, le finnois, le français, le grec, l'irlandais, l'islandais, l'italien, le néerlandais, le norvégien, le portugais, le suédois. DIALANG couvre cinq compétences : l'écoute, la lecture, la rédaction, la grammaire et le vocabulaire.

Ce test n'a pas de vocation certificative (Danish Technological Institute, Nov. 2004) mais il fournit des résultats et commentaires dès la fin du passage du test. Ce test est précédé par deux étapes (pré-tests) qui sont proposées aux candidats afin de procéder à une première estimation de leurs compétences et de choisir le niveau de difficulté adéquat pour le test qui en présente trois. Ce pré-positionnement des candidats comporte deux parties distinctes :

1. Un test de reconnaissance des mots (Test de Niveau sur l'étendue du Vocabulaire) : une liste de mots (verbes) qui doit être validée dans la langue cible (existe ou n'existe pas). Ce pré-test permet de poser un premier diagnostic, néanmoins assez sommaire, sur le niveau.
2. Une deuxième étape est un questionnaire d'autoévaluation qui permet d'apprendre à évaluer soi-même ses compétences acquises au cours de l'apprentissage d'une langue étrangère. Cette activité est disponible pour les compétences de lecture, d'écoute, ou de rédaction.

La partie autoévaluation de DIALANG est une composante centrale de l'outil, non seulement dans la phase de pré-positionnement mais également dans le système de rapport des résultats et du test en lui-même (Alderson, 2005a). Les descripteurs proviennent en grande majorité du CECR, certains ayant été adaptés, d'autres

[1] Le *Test d'Evaluation de Français* (TEF) a été mis au point par la *Chambre de Commerce et d'Industrie de Paris* dès 1998 - http://www.fda.ccip.fr.

[2] Ces modèles de portfolio, tout en étant calibrés sur le CECR, présentent des descripteurs ciblés en fonction du public auquel ils sont adressés et doivent être validés par le Conseil de l'Europe pour être homologués.

[3] http://www.dialang.org/project/francais/index.html

rajoutés, selon les besoins spécifiques de l'épreuve. Tous ont été calibrés par le premier Centre de Coordination du projet DIALANG (96-99) en Finlande dont les résultats ont démontré des indices de fidélités très élevés (North, 2000 et Kaftandjieva, 2002).

La philosophie qui sous-tend l'intérêt de développer chez l'apprenant une auto-évaluation de ses capacités linguistiques et communicatives est que l'autoévaluation non seulement permet une meilleure compréhension de son niveau de compétences, mais contribue également à évaluer ses propres besoins de formation ultérieurs et donc au développement de son autonomie. Dans la démarche du test diagnostique de DIALANG qui propose de faire le lien entre les résultats de l'autoévaluation et ceux du test proprement dit, le but poursuivi n'est pas de démontrer les possibles disparités, voire de suggérer l'imprécision de l'autoévaluation, mais bien d'aider l'apprenant à prendre conscience des éventuelles divergences, d'en comprendre les raisons et de l'aider à poursuivre sa réflexion sur les implications de celles-ci. (Alderson, 2005a).

La recherche en cours

Dans le domaine du français langue étrangère, les auteurs qui contribuent à l'élaboration du *Test d'Evaluation de Français* (TEF) ont souhaité s'assurer de la bonne compréhension des informations fournies aux candidats à l'issue d'un test à forts enjeux[4] (Demeuse *et al*, 2004 et 2005). Pour ce faire, l'idée qui a été poursuivie est de mettre en relation les résultats obtenus au test critérié (TEF) et la manière dont s'apprécient les candidats, partant du principe que les résultats au test seront d'autant mieux compris que la manière dont les candidats s'autoévaluent correspond à ces résultats. L'échelle d'autoévaluation mise au point dans le cadre du projet DIALANG a été adoptée dans cette expérimentation. Cette échelle est présentée dans la langue maternelle des candidats de manière à éviter les problèmes de compréhension et de vocabulaire auprès des candidats les plus faibles.

Les résultats

Une première expérimentation a été menée auprès de 36 candidats hispanophones, inscrits dans un même centre de langue agréé en France, avant une passation de la version électronique du TEF, baptisée e-TEF (Desroches *et al*, 2005). L'autoévaluation a été proposée dans sa version courte (utilisée dans les tests pilotes en compréhensions écrite (CE) et orale (CO)) de DIALANG (Alderson, 2005b). Le niveau de chaque candidat, sur l'échelle en six niveaux du CECR, a alors été calculé à partir de l'échelle d'autoévaluation d'une part et du TEF d'autre part, puis comparé. Cette approche n'a pas conduit à des résultats satisfaisants, au sens où la concordance entre évaluation et autoévaluation est très limitée (Desroches *et al*, 2005). Le manque de correspondance constaté entre ces deux instruments, dans une perspective d'autodiagnostic, met en évidence les difficultés qui peuvent subsister face à un référentiel lorsqu'il doit être maîtrisé par des usagers qui ne sont pas des experts du domaine linguistique, mais qui possèdent néanmoins une certaine compétence dans au moins une autre langue que leur langue maternelle.

De manière à mieux comprendre les problèmes posés par la mise en correspondance des compétences « objectivées » par un test et des compétences telles que les sujets les évaluent eux-mêmes, l'expérience est poursuivie. Celle-ci porte sur une extension de l'échantillon à d'autres locuteurs qu'hispanophones et sur la prise en compte de l'échelle DIALANG dans sa version longue et non plus dans sa version réduite, soit 36 items en compréhension écrite (CE) et 44 en compréhension orale (CO). L'échantillon hétérogène se compose de 6 autoévaluations en anglais dont 4 non-

[4] Le TEF est en effet, notamment, reconnu par Citoyenneté et Immigration Canada et est utilisé pour l'admission officielle des étudiants non francophones dans les premiers cycles des universités françaises.

anglophones, 5 allemands, 6 espagnols et 1 portugais. Les candidats ont passé pour la plupart l'autoévaluation dans leur propre langue, excepté lorsque la langue maternelle n'était pas disponible parmi les 14 langues proposées par DIALANG, en l'occurrence pour quatre candidats. Le protocole est le même que pour la première expérimentation (Desroches *et al*, 2005).

Pour rappel, la consistance interne de chaque outil est assez élevée, avec un α de Cronbach de 0,80 (CE) et de 0,82 (CO) pour DIALANG, et de 0,80 (CE) et 0,67 (CO) pour le TEF. Par contre, les corrélations entre les résultats obtenus aux épreuves du TEF et le nombre de réponses positives à l'autoévaluation sont toujours positives, mais faibles (r_{36}=0,14 en CE et r_{44}=0,30 en CO) et donc non significatives ($p<0,51$ en CE et $p<0,35$ en CO).

Si l'on compare nos résultats à ceux recueillis par Alderson (2005b) dans une étude confrontant l'autoévaluation aux épreuves du test DIALANG en langue anglaise sur un échantillon plus grand (n=718 en CE et n=606 en CO), les corrélations ne sont pas plus satisfaisantes. Elle s'élève à 0,54 en CE entre l'autoévaluation globale[5] et l'épreuve elle-même et est encore plus modeste entre l'autoévaluation par habilité et l'épreuve équivalente en CE (r=0,49). Concernant la CO, la corrélation est de 0,47 entre l'autoévaluation globale et l'épreuve correspondante, elle est du même ordre de grandeur (r=0,49) entre l'autoévaluation par habilité en CO et l'épreuve (Alderson, 2005a).

La deuxième expérience ayant été basée sur la version longue d'autoévaluation proposée par DIALANG alors que la première expérience reprenait la version courte, nous avons repris les items d'autoévaluation communs aux deux versions afin de les comparer sur un ensemble identique d'items d'autoévaluation. Dans la version longue de DIALANG, les corrélations r_{36} et r_{44} correspondent aux corrélations entre les 36 items pour la CE et 44 items pour la CO de la version longue et les épreuves du TEF correspondantes. Ces corrélations sont comparées aux corrélations respectives (r_{18}) des 18 items de la version courte de DIALANG et les résultats aux épreuves correspondantes du TEF. Les résultats restent globalement stables :

- pour la compréhension écrite (CE) : $r_{36} = 0,14 \rightarrow r_{18} = 0,16$
- pour la compréhension orale (CO) : $r_{44} = 0,30 \rightarrow r_{18} = 0,24$.

L'échantillon sélectionné pour la deuxième expérience présente des candidats d'origines plus variées que lors de la première, réalisée sur un échantillon homogène (composé exclusivement de candidats hispanophones). Malgré cette plus grande variété de langues, si on prend en compte l'accroissement du nombre d'items dans la seconde version des autoévaluations, la dispersion des résultats obtenus n'est pas beaucoup plus large que dans la première expérimentation en CE (écart type de 1,73

[5] L'étude pilote de DIALANG inclut une autoévaluation globale et une autoévaluation par habilité ; seule la seconde sera retenue dans la version mise en ligne (Alderson, 2005a).

pour la première contre 4,11 pour la deuxième). La différence se creuse en CO (écart type de 2,38 pour la première contre 7,24 pour la deuxième).

Ces résultats ne sont donc pas plus satisfaisants et les conclusions se maintiennent à propos de la faible liaison statistique observée entre les deux outils (TEF et DIALANG), pourtant liés au même référentiel. Les candidats apparaissent, comme relevé par Alderson, ne pas avoir une bonne perception de leurs propres compétences, en regard des évaluations « objectives » qui leur sont administrées. Cette conclusion, qui se maintient lorsque l'échantillon est plus large, pose bien les limites du recours à des descripteurs tels que ceux du CECR lorsque les usagers ne maîtrisent pas bien le référentiel dans son ensemble.

Les perspectives

Les résultats des deux expérimentations nous conduisent à réfléchir à la manière de mieux communiquer les résultats aux candidats et à les informer, lors de la formation linguistique, du référentiel qui est utilisé pour les positionner sur l'échelle de niveaux pour chacune des compétences linguistiques et communicatives. Si dans le cadre de DIALANG, test sans enjeu et à vocation purement diagnostique, l'autoévaluation est utilisée pour alimenter les commentaires et les conseils dans une perspective de formation, les divergences de jugement et de niveau font partie intégrante de la démarche diagnostique et se justifient dans le choix d'une orientation formative. Par contre, le recours à l'autoévaluation dans le cadre d'un test à visée certificative tel que le TEF dont la fonction principale n'est pas, *a priori*, formative, doit être interrogé : l'outil d'autoévaluation est-il le plus adéquat dans une démarche de compréhension et d'interprétation des résultats du TEF ? Ne faudrait-il pas plutôt étayer les commentaires fournis sur l'attestation afin de préciser les liens entre le score obtenu, le niveau sur l'échelle du CECR auquel il correspond et ce que cela traduit en termes de compétences et d'habilités ? Le recours à l'autoévaluation ne se justifierait alors sans doute que lorsque le TEF est utilisé comme test de positionnement dans le cadre d'un plan de formation.

Par ailleurs, une réflexion approfondie à propos des items d'autoévaluation, de l'utilisation de la langue maternelle ou non, ou encore de la réceptivité et compréhension de l'outil par ses utilisateurs, mérite d'être poursuivie. Les travaux rapportés par Alderson (2005a) appellent au recueil de données empiriques supplémentaires d'une part, et d'autre part à une réflexion sur l'amélioration de l'outil d'autoévaluation de DIALANG sans en altérer le lien au référentiel européen. Nos conclusions vont dans la même direction.

BIBLIOGRAPHIE

ALDERSON, J.C. (2005a) : *Diagnosing Foreign Language Proficiency – The Interface between Learning and Assessment*, London & New York, Continuum.

ALDERSON J.C., HUHTA A. (2005b) : *The Development of a suite of computer-based diagnostic tests based on the Common European Framework*, Language Testing 2005-22 (3)

CHAMBRE DE COMMERCE ET D'INDUSTRIE DE PARIS (2005) : *Le Test d'Evaluation de Français de la Chambre de Commerce et d'Industrie de Paris.* Paris : Chambre de Commerce et d'Industrie de Paris, Centre de Langue de la Direction des relations internationales de l'Enseignement.

CITOYENNETE ET IMMIGRATION CANADA (2002) : *Standards Linguistiques Canadiens 2002*.

CONSEIL DE L'EUROPE (2000) : *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*, Paris, Didier.

CONSEIL DE L'EUROPE, DIVISION DES POLITIQUES LINGUISTIQUES (2003) : *Relier les examens de langue au Cadre européen de référence pour les langues : apprendre, enseigner, évaluer (CECR) – Manuel avant-projet*, Strasbourg : DGIV/EDU/LANG.

DEMEUSE, M., CRENDAL, A., DESROSCHES, F., OSTER, P., RENAUD, F., LEROUX, X. (2004) : *L'évaluation des compétences linguistiques des adultes en français langue étrangère dans une perspective de multi-référentialisation. L'exemple du Test d'Evaluation de Français (TEF) de la Chambre de Commerce et d'Industrie de Paris (CCIP)*, In Actes du 17e colloque international de l'Association pour le Développement des Méthodologies d'Evaluation en Education (ADMEE-Europe), Lisbonne, 18-20 novembre 2004

DEMEUSE, M., DESROSCHES, F., CRENDAL, A., RENAUD, F., CASANOVA, D., ARTUS, F. (2005) : *La fiabilité de l'évaluation des compétences linguistiques pour des adultes non francophones : présentation d'un protocole d'évaluation*, In Actes du 18e colloque international de l'Association pour le Développement des Méthodologies d'Evaluation en Education (ADMEE-Europe), Reims, 24-26 octobre 2005.

DESROSCHES, F., CASANOVA, D., CRENDAL, A., RENAUD, F., MAHY, C., DEMEUSE, M. (2005) : *Mise en équivalence des versions papier-crayon et électronique d'un test de français langue étrangère et mise en perspective des résultats par une procédure d'autoévaluation*, In Actes du 18e colloque international de l'Association pour le

Développement des Méthodologies d'Evaluation en Education (ADMEE-Europe), Reims, 24-26 octobre 2005.

ENGLE, L., ENGLE, J. (2004) : *Assessing Language Acquisition and Intercultural Sensitivity Development in Relation to Study Abroad Program Design, Frontiers*, in The
Interdisciplinary Journal of Study Abroad, Volume X, Carlile, PA.

KAFTANDJIEVA F., TAKALA S. (2002) : *Council of Europe scales of language proficiency: a validation study*. In Alderson, 2002, Common European Framework of References for languages (2002): Learning, Teaching, Assessment - Case studies : pp.106-29, Strasbourg, Council of Europe.

NOËL-JOTHY F., SAMPSONIS B. (2006) : *Certifications et Outils d'évaluation en FLE*, Paris, Hachette.

NORTH B. (2000) : *The development of a common framework scale of laguage proficiency*, New York, Peter Lang.

SITES INTERNET

Citoyenneté et Immigration Canada: http://www.cic.gc.ca/french/skilled/qual-3-1.html

Chambre de Commerce et d'industrie de Paris : http://www.fda.ccip.fr

Conseil d'Europe: http://www.coe.int

"Defining a Strategy for the Direct Assessment of Skills", Danish Technological Institute (Nov.2004): http://europa.eu.int/comm/education/doc/reports/doc/skills.pdf

DIALANG: http://www.dialang.org/project/francais/index.html

SYMPOSIUM 8 : L'EVALUATION DE PROGRAMME : NOTIONS, MODELES ET METHODOLOGIES

Gérard Figari, Université de Grenoble II
Marthe Hurteau, Université du Québec à Montréal
Nathalie Younes, Université Pierre Mendes France
Maria Palmira Alves, Université du Minho
Marie-Françoise Fave-Bonnet, Université Paris X Nanterre

Résumé

Evaluation de programme, évaluation de politiques éducatives, évaluation de projets, évaluation de systèmes, évaluation de curricula, évaluation institutionnelle, assurance-qualité, autant de formulations de formes, de théorisations, de méthodologies et d'objets d'évaluation que l'on trouve en Amérique du Nord et en Europe et dont le moins que l'on puisse dire est qu'elles comportent des éléments communs. Les travaux afférents à ces thématiques sont également très éclatés, par disciplines (sciences politiques et de gestion, sciences de l'éducation, psychologies et didactiques) et par objets (cf. supra), distinguant macro et micro, quantitatif et qualitatif, théorique et opérationnel.

Ce thème d'investigation est particulièrement prégnant au moment où l'ADMEE internationale réunissant les deux ADMEE existantes est en gestation et gagnerait à initier dès à présent un réseau de travail commun sur une question qui les intéresse également et au moment où les activités européennes consécutives au développement du processus de Bologne sont confrontées aux questions de "programmes" identifiables, comparables, échangeables, etc.

Equipes concernées

Quatre équipes provenant de trois pays différents sont intéressées (sur la base de travaux terminés ou en cours) :

- Département Education et pédagogie, UQAM, Montréal (Hurteau, M. ; Houle, S.)
- Laboratoire des Sciences de l'Education, Université P. Mendès France, Grenoble et ENESA Dijon (Figari, G. ; Tourmen, C. ;Younès, N.)
- Institut de recherche en éducation, Université du Minho, Braga (Alves, P. ; Machado, A.)
- Centre de Recherche Education et Formation (CREF), Université de Paris X Nanterre (Fave-Bonnet, M-F., Macarie-Florea, M.)

Objectifs

Déjà sous-jacents dans l'exposé des motifs, ils consistent à confronter, dans les travaux qui vont être présentés par les 4 équipes, les "notions, modèles et méthodes" de la recherche sur l'évaluation de programme au sens large englobant les autres formulations présentées supra.

Cette confrontation devrait tout d'abord pouvoir déboucher sur des conceptualisations sinon consensuelles du moins clairement comparables et échangeables.

Elle devrait permettre ensuite de formuler des problématiques de recherche nouvelles et donner lieu à un programme de recherche à plus long terme de l'ADMEE internationale.

MODÉLISER L'ÉVALUATION DE PROGRAMME, ENTRE THÉORIE ET PRATIQUE
- VERS UNE COMPARAISON DES APPROCHES AU QUÉBEC ET EN FRANCE–

Gerard Figari, Université Pierre Mendès France
Claire Tourmen, Université Grenoble 2

Mots clés : Evaluation de programme – dispositif - référentialisation

Résumé :

Nous proposons d'identifier, en nous inspirant des théorisations de l'évaluation, des éléments invariants qui seraient constitutifs de l'activité évaluative. Nous tentons ensuite de comparer la façon dont ils sont utilisés et nommés dans des pratiques d'évaluation de "programme/dispositif" en France et au Québec. Nous proposons enfin une modélisation de l'évaluation de « programme/dispositif » qui modélise à la fois l'objet et l'activité d'évaluation sur la base de ses invariants (démarche théorique) et peut se décliner, dans la pratique, sous forme de protocole méthodologique : la « référentialisation ».

Présentation

Evaluer des « programmes », des « dispositifs », des « politiques publiques », des « projets »… Si nous postulons que ces termes différents recouvrent des « objets » très proches que l'on désignerait, selon les cultures, avec la notion de « programme » en Amérique du Nord et avec la notion de « dispositif », notamment en France, nous disposons alors d'une remarquable opportunité de comparaison. C'est sur ce postulat d'existence d'un objet, qu'on appellera provisoirement le « programme-dispositif », que nous allons conduire notre réflexion. Dans ce cadre, la diversité des approches et des concepts utilisés, au lieu de constituer le signe d'une confusion sémantique, peut devenir une source de confrontation théorique, méthodologique et pratique : tel est l'enjeu de cette contribution.

La comparaison de pratiques d'évaluation de « programmes-dispositifs » demande, tout d'abord, une clarification des concepts sous-jacents. Qu'est-ce qu'évaluer un « programme-dispositif » ? Nous ferons l'hypothèse que, comme toute forme d'évaluation, celle-ci se déroule selon des invariants qui se retrouvent quel que soit le type d'évaluation menée et quel que soit l'objet : il s'agit donc de commencer par un repérage de ces invariants, qui seraient des dimensions toujours présentes dans les activités évaluatives.

En second lieu, nous allons esquisser une réflexion comparative entre des approches québécoises et françaises. Il s'agit de tenter une comparaison entre les conceptions du rapport « théorie-pratique » en matière d'évaluation de « programme-dispositif » dans les deux contextes, avec l'hypothèse d'un possible enrichissement mutuel. Nous allons notamment comparer la façon dont les invariants de l'activité évaluative sont nommés

et pratiqués dans ces deux contextes, sur base de lectures comparatives et d'entretiens avec des évaluateurs.

Enfin, nous essayerons d'appliquer cette explicitation des éléments communs repérés à la présentation d'une modélisation de l'évaluation de « programme-dispositif », la « référentialisation » (Figari, 1994), qui revêt un double statut : elle modélise à la fois l'objet et l'activité d'évaluation sur la base de ses invariants (démarche théorique) et peut se décliner, dans la pratique, sous forme de protocole méthodologique. Cette réflexion pourrait constituer une contribution à des projets de recherche communs.

L'EVALUATION DE PROGRAMME : NOTIONS, MODELES ET METHODOLOGIES

Marthe Hurteau, Université du Québec à Montréal

Sylvain Houle, Université du Québec à Montréal

Mots-clés : pratique, logique de l'évaluation

Résumé:

L'évaluation de programme est régulièrement remise en question au niveau de sa fonction première, soit sa capacité à poser un jugement éclairé sur les programmes à l'étude et à dégager des recommandations pertinentes. À ce jour, certains auteurs ont émis des hypothèses d'explication du phénomène sans plus autant les appuyer par une démarche scientifique. L'équipe de recherche a décidé de se pencher sur la pratique en évaluation de programme afin de la définir et d'établir sa contribution au niveau des activités professionnelles. Scriven (1980) a été le premier à s'intéresser à la pratique de l'évaluation de l'évaluation qu'il explique au moyen de la logique de l'évaluation qui se résume au moyen de quatre opérations qui établissent en quelque sorte la spécificité de toute démarche évaluative. Par la suite, Hurteau (1991), Fournier (1995) et Stake (2004) ont poursuivi la réflexion amorcée par Scriven (1980) et cela a permis d'identifier les paramètres et le déroulement spécifique à toute démarche évaluative (Hurteau et Houle, 2005; Hurteau, Lachapelle et Houle, 2006). Ceux-ci constituent les fondements de la pratique en évaluation puisqu'ils en établissent la spécificité et ils doivent être distingués des contributions à cette démarche, pour ne citer que les modèles en évaluation de programme ainsi que les standards du Joint Committee. Ces considérations ont alimenté les travaux de Hurteau et Houle (2005) qui s'inscrivent dans le cadre d'un programme de recherche dont ils constituent la première phase. Plus précisément, ils portent sur la pratique de l'évaluation de programme et portent sur l'analyse de 69 articles tirés de revues scientifiques spécialisées en évaluation de programme en fonction des paramètres établis ainsi que du déroulement. Le présent exposé présente les résultats obtenus ainsi que les perspectives pour la phase deux du programme de recherche.

APPLICATION DU CADRE THEORIQUE DE L'EVALUATION DE PROGRAMME SELON SHADISH, COOK ET LEVITON AU CAS DE L'ENSEIGNEMENT UNIVERSITAIRE

Nathalie Younes, UPMF et Université d'Auvergne, France.

Résumé :

Dans cette communication nous explorons une application à l'enseignement universitaire du cadre théorique élaboré par Shadish, Cook et Leviton (1991) pour l'évaluation des programmes sociaux, en établissant des parallèles en termes de passages ou de limites entre ces deux champs d'évaluation. Les résultats d'une recherche doctorale sur les effets de l'implantation d'un dispositif d'évaluation de l'enseignement par les étudiants dans une université française sont notamment réinterrogés dans ce cadre d'après les cinq paramètres identifiés par ces auteurs : la programmation sociale, la construction de connaissance, l'attribution de valeur, l'utilisation de la connaissance et la pratique de l'évaluation.

Introduction

Les formations universitaires, en tant qu'ensembles d'activités sociales sous-tendues par des politiques et ayant pour but d'assurer une formation supérieure à une partie de la population, peuvent être considérées comme des programmes. Dans la littérature anglophone, la notion de programme est un cadre conceptuel d'étude global qui inclut des pratiques. Ainsi, elle dépasse largement ce qu'on appelle le « programme de formation » c'est à dire la liste des cours, contenus et activités prévus, mais inclut aussi ce qui est réellement fait, la question du public, des ressources ou autres ainsi que les effets de la formation en question.

Dans cette communication nous explorons une application à l'enseignement universitaire du cadre théorique élaboré par Shadish, Cook et Leviton (1991) pour l'évaluation des programmes sociaux, en établissant des parallèles en termes de passages ou de limites entre ces deux champs d'évaluation. Nous reprenons le terme de « théorie » au sens employé par les anglo-saxons à la fois comme cadre général théorique et comme cadre de modélisation d'une pratique. Les résultats d'une recherche doctorale sur les effets de l'implantation d'un dispositif d'évaluation de l'enseignement par les étudiants dans une université française sont notamment réinterrogés dans ce cadre d'après les cinq paramètres identifiés par ces auteurs.

I - Cadre théorique de l'évaluation de programme selon Shadish, Cook et Leviton

Shadish, Cook et Leviton sont des théoriciens de l'évaluation de programme américains qui, dans leur ouvrage *Foundations of Program Evaluation* paru chez Sage en 1991, ont distingué trois générations de « théories » de l'évaluation produites depuis les années 60 jusqu'aux années 90[1] (principalement aux Etats-Unis). Ce que nous reprenons ici plus particulièrement, c'est le cadre théorique qu'ils ont élaboré pour l'évaluation de programme en insistant sur le rôle déterminant d'une théorie de l'évaluation. Car si l'évaluation des programmes sociaux est un champ conduit par la

[1] La première génération de théories dans les années 60 est représentée par Scriven et Campbell qui misent avant tout sur la rigueur scientifique en cherchant une connaissance causale valide des effets des programmes. Dans les années 70, la deuxième génération, (autour de Weiss, Wholey et Stake), devant l'échec de la première génération à produire de la connaissance utile pour améliorer les programmes, propose de nouveaux modèles plus pragmatiques visant l'utilisation de la connaissance produite par l'évaluation pour améliorer les programmes. Une troisième génération, à partir des années 80, représentée par Cronbach et Rossi, cherche à concilier la recherche d'une connaissance descriptive concernant l'usage et la programmation sociale avec les exigences d'une connaissance valide.

pratique, la théorie est fondamentale en ce qu'elle développe les raisons théoriques des prescriptions pratiques. Ils déplorent que de nombreux ouvrages sur l'évaluation soient des listes a-théoriques de méthodes. Selon eux, le but fondamental d'une théorie de l'évaluation est de spécifier des pratiques faisables par les évaluateurs pour construire une connaissance de la valeur des programmes qui puisse être utilisée afin d'améliorer la résolution des problèmes sociaux pour lesquels les programmes sont pertinents. Ainsi, suivant cette définition, une théorie de l'évaluation doit explicitement décrire et justifier les 5 composantes suivantes : la programmation sociale, la construction de connaissance, l'attribution de valeur, l'utilisation de la connaissance et la pratique d'évaluation (1991, p.36).

1. **La programmation sociale** rend compte des façons dont les politiques et programmes sociaux se développent, s'améliorent et changent, particulièrement par rapport aux problèmes sociaux. Cette composante est divisée en trois :

- . la structure interne du programme et son fonctionnement : équipe, destinataires, ressources, buts, normes, organisation,
- . les contraintes externes qui forment et contraignent le programme : pressions des politiques et des différents porteurs d'enjeux, valeurs sociales…,
- la façon dont le changement social a lieu, la façon dont les programmes varient et contribuent au changement social.

Au terme des trois générations de théories analysées, Shadish, Cook et Leviton considèrent que les problèmes sociaux sont mal définis et qu'il y a des désaccords sur les priorités. D'une part, différents porteurs d'enjeux aux intérêts contradictoires sont

impliqués dans ces programmes. D'autre part, les buts des programmes sont vagues et contradictoires et le monde social n'est pas rationnel.

Par ailleurs, ils mettent en jeu de façon plus ou moins explicite des modèles sociaux, des systèmes d'organisation institutionnelle et des positionnements politiques. Ils soulignent que les programmes sont tellement imbriqués aux politiques que les résultats d'évaluation contribuent peu à les initier ou les arrêter. Ils identifient néanmoins trois types de stratégies de changement. Des petits changements sont censés améliorer les choses mais on en critique le caractère modeste voire conservateur. On peut aussi utiliser des projets de démonstration localisés pour tester des innovations qui pourront être intégrées plus tard à une nouvelle politique. Mais il est difficile de les estimer correctement en tant que tels (la motivation de l'équipe de l'innovation étant notamment en général supérieure à celle des autres). Une autre stratégie est de questionner les hypothèses fondamentales des politiques et programmes existants et de chercher des alternatives radicales.

2. La **construction de la connaissance** traduit les façons dont les chercheurs apprennent sur les activités sociales. Une théorie de la connaissance se pose la question du type de connaissance élaborée par les évaluateurs et comment ils construisent cette connaissance. Cette composante décrit : une ontologie, ou étude de la nature ultime de la réalité, une épistémologie, ou étude de la nature, des origines et des limites de la connaissance et une méthodologie, ou étude des techniques pour la construction de la connaissance.

Les théoriciens des trois générations postulent l'existence d'un monde réel et inscrivent l'évaluation de programme comme une pratique dans la tradition des sciences sociales en considérant que différents types de connaissance doivent être

construits (causalité, généralisation, description de l'implantation…) mais qu'ils n'auront pas tous la même priorité suivant les études. Selon eux, aucune méthode de science sociale ne doit être exclue du répertoire de l'évaluateur mais chacune est faillible. Pour construire une connaissance fiable et valide, il vaut mieux des résumés de nombreuses études car la qualité de la connaissance augmente avec son examen minutieux et critique. Mais à chaque niveau, ontologique, épistémologique et méthodologique, des divergences existent, que ce soit à propos de la complexité du monde, la mesure dans laquelle il peut être connu et les conséquences de sa simplification ; ou à propos du type de connaissance prioritaire ou bien encore à propos des méthodes à utiliser.

Mais la distinction entre une position interprétative et critique et une position objectiviste n'est dans la réalité pas si claire. Les recherches scientifiques reposent aussi sur la subjectivité des chercheurs, elle même liée à celle d'une époque. En outre, l'apparente simplicité et objectivité des chiffres et indicateurs peut masquer les contradictions de la réalité. Ainsi, Campbell est d'après Shadish et al. (1991, p.143), le seul théoricien de l'évaluation qui bien qu'ayant largement favorisé les méthodes expérimentales est resté néanmoins critique à leur endroit. L'histoire de la science l'a convaincu que les chercheurs comme les politiques partagent des « biais », de tels biais étant en général tellement incorporés à la recherche que, la plupart du temps, l'on ne peut pas les détecter (Feyerabend 1979, Kuhn, 1970). Il a développé une méthodologie d'évaluation reposant sur 3 méthodes principales liant une position objectiviste et une position interprétative : expérimentation, indicateurs sociaux et rapports d'activité, de satisfaction (1988, p.298, dans Shadish, Cook et Leviton, 1991). Bien qu'il soit connu comme l'avocat des méthodes expérimentales et quasi-

expérimentales, ses travaux ont aussi été utilisés pour justifier une pratique d'évaluation centrée sur les études de cas qualitatives (Yin, 1984, cité par Shadish, Cook et Leviton, 1991). Il renie toute fondation sûre de la connaissance mais il croit que l'organisation de la science en communautés de controverses s'approchera plus de la vérité, les évaluations pluralistes, critiques, conduisant à des inférences intersubjectivement vérifiées. Globalement, l'évaluation doit favoriser la discussion ouverte et la critique. Il préconise de faire plusieurs petites études d'un programme, de réanalyser les évaluations complètes de manière critique et d'encourager la critique interne d'un programme ou d'une évaluation.

3. **L'attribution de valeur** décrit comment la valeur peut être attachée aux descriptions du programme. Les valeurs sont omniprésentes dans la programmation sociale. Ce qui est en question, c'est comment les évaluateurs peuvent expliciter les problèmes de valeur, traiter ouvertement avec ceux-ci et produire une analyse sensible des implications des programmes de ce point de vue. Trois types de conceptions de la valeur sont repérés :

- une méta-théorie décrivant la nature et la justification des jugements de valeur. Scriven (1980, 1983) a particulièrement développé cet aspect,
- une théorie prescriptive défendant la primauté de certaines valeurs (ex : Rawls, 1971),
- une théorie descriptive utilisée par la plus grande partie des évaluateurs décrivant les valeurs sans en défendre une en particulier (ex : Rossi & Freeman, 1985).

Les théoriciens des trois générations se sont tous positionnés sur l'importance de la valeur. Ils s'accordent sur le fait qu'elle est consubstantielle à l'évaluation, à savoir

que l'évaluation réalise nécessairement les trois premières étapes de la logique de l'évaluation de Scriven mais pas nécessairement la dernière (synthétiser les résultats en un simple jugement de valeur). Enfin pour tous, considérer les intérêts des différents porteurs d'enjeux augmente les chances d'inclure toutes les perspectives de valeur pertinentes. En revanche, il n'y a pas d'unanimité sur les critères de mérite par lesquels peuvent être jugés les programmes sociaux. Les positions diffèrent quant au fait que ces critères doivent être comparatifs ou absolus, ainsi que sur le fait de savoir si les théories prescriptives éthiques doivent jouer un rôle significatif dans la sélection des critères de mérite.

Un mouvement en faveur de la prise en compte du point de vue des bénéficiaires du programme s'est dessiné dans les années 70-80 aux Etats-Unis. En 1971, Campbell souligne l'importance de traiter les clients des programmes non pas comme des sujets mais comme des experts indispensables à propos de leurs besoins et de leurs expériences du programme. De plus, suivant Stake, ce sont aux porteurs d'enjeux d'émettre les jugements de valeur et il faut accorder une attention particulière aux positions minoritaires. Dans l'enseignement, les étudiants, les enseignants et d'autres observateurs exerceraient les jugements critiques les plus pertinents, que leurs critères soient ou non explicites (Stake, 1980, p. 86).

4. **L'utilisation de la connaissance** rend compte des façons dont l'information produite est utilisée pour modifier les programmes et politiques. L'utilisation des résultats de l'évaluation est un problème capital et récurrent tant est patent que bien souvent ils restent lettre morte, sont biaisés ou mal utilisés. La question est de savoir comment fournir des résultats utiles pour la résolution des problèmes sociaux. Les théoriciens des trois générations reconnaissent deux types possibles d'utilisation de la

connaissance produite par l'évaluation, une utilisation instrumentale ou conceptuelle. Ils considèrent en général que c'est par l'utilisation à court terme que l'évaluation est la plus justifiée bien qu'ils reconnaissent que les évaluations déterminent rarement les décisions prises au sujet des programmes. En revanche, ils ne privilégient pas tous le même usage instrumental ou conceptuel de l'évaluation. Ils ne sont pas tous attentifs à l'identification des utilisateurs potentiels et ils ne s'accordent pas non plus sur les utilisateurs à privilégier pour l'évaluation, ni sur ce qui favorise l'utilisation instrumentale de l'évaluation : modalités de relation avec les utilisateurs potentiels et de communication des résultats.

En distinguant l'évaluation sommative et formative, Scriven a été le premier théoricien majeur à décrire différents types d'usage. Mais, c'est plus une théorie de comment les résultats peuvent être utilisés qu'une théorie de comment ils le sont dans le monde réel. Campbell quant à luiredoute principalement deux mauvaises utilisations des résultats de l'évaluation. Premièrement, l'utilisation de résultats faux sur la base d'évaluations biaisées. Deuxièmement, le contrôle des résultats ou de leur diffusion par des parties intéressées. Pour contrer ces risques, il recommande plus d'étude publique des résultats, plus de support public et légal à ceux qui tirent la sonnette d'alarme et plus de débat public sur les résultats. Pour lui, l'utilisation n'est pas du ressort des évaluateurs mais des politiques. Mais sa position, suivant laquelle on ne doit utiliser que les résultats ayant résisté aux tests les plus rigoureux, risque de produire un élitisme scientifique bien loin des réalités du terrain où le critère n'est pas tant la vérité que l'utilité pour l'action. C'est ce qui a conduit les théoriciens comme Wholey, Stake ou Weiss, à préconiser d'identifier les usagers spécifiques pour chaque évaluation et de travailler de manière très proche avec eux pour créer de l'information

utile, et ce au niveau local que ce soit à l'intention des responsables (Wholey) ou des porteurs d'enjeux locaux (Stake, 1975, 1980).

5. la **pratique de l'évaluation** traduit les tactiques et stratégies professionnelles des évaluateurs, étant données les contraintes auxquelles ils sont confrontés. Les théoriciens des trois générations considèrent que la pratique de l'évaluation est difficile. Elle a lieu dans des conditions de temps et de ressources qui requièrent de délicats échanges. Les évaluateurs ne sont en général pas les bienvenus, une simple étude est insuffisante et les évaluateurs doivent être actifs pour faciliter l'utilisation de l'évaluation. En revanche, il existe des désaccords entre les différents théoriciens quant au rôle de l'évaluateur, quant aux valeurs qui doivent être représentées dans l'évaluation, quant aux questions à se poser, et quant aux méthodes à employer.

II – Parallèle avec l'enseignement universitaire : mise en perspective des effets de l'introduction d'un dispositif d'EEE

Le parallèle avec l'enseignement universitaire envisagé est mené suivant les cinq composantes d'une théorie de l'évaluation des programmes.

1. Programmation

La programmation sociale rend compte des façons dont les programmes se développent, s'améliorent et changent : de leur structure interne, des contraintes externes et du changement social. En ce qui concerne l'enseignement universitaire, la structure traditionnellement définie par les exigences des disciplines et la liberté académique reste une affaire plutôt individuelle dont les modèles pédagogiques sont largement implicites. Les récentes évolutions de l'université (massification, professionnalisation, européanisation) avec notamment les procédures d'assurance

qualité remettent en question ce modèle puisqu'elles exigent de rendre compte de ce qui se passe dans l'enseignement et de s'appuyer sur des démarches collectives. Elles constituent de nouvelles contraintes de changement de l'enseignement supérieur.

La recherche menée à l'IUT de l'Université d'Auvergne met en évidence les paradoxes de l'institution universitaire. Alors qu'il y a une injonction institutionnelle d'évaluation de l'enseignement, ce dernier reste cependant peu valorisé par l'institution elle-même puisqu'il a peu d'impact sur la carrière, qu'il n'y a pas de formation à la pédagogie non plus qu'un système d'accompagnement pédagogique consécutif aux résultats de l'évaluation. De plus, alors que la promotion de la culture générale des étudiants versus celle de la spécialisation, l'élitisme versus la démocratie et l'autonomie de l'université versus son instrumentalisation sont toujours en question, nous avons pu constater à quel point les cadres de référence pédagogiques ne sont ni identifiés ni débattus par les enseignants, voire considérés de manière contradictoire ce qui met en évidence les contradictions entre un système d'enseignement dont les modèles et pratiques pédagogiques sont largement implicites opaques et individuelles et l'EEE qui exige un système d'enseignement explicite, visible et collectif.

2. Nature de la connaissance construite par l'évaluation de l'enseignement par les étudiants

Une *« théorie » de la connaissance* pose la question du type de connaissance élaborée par les évaluateurs et comment ils construisent cette connaissance. En ce qui concerne l'évaluation de l'enseignement universitaire deux positions épistémologiques et méthodologiques existent, une position interprétative et critique et une position

objectiviste. Etant donné le peu de connaissance objective de l'enseignement supérieur, Cranton argumente pour l'inscription de l'évaluation de l'enseignement dans un paradigme interprétatif et critique. A l'opposé de cette approche interprétative et critique de l'évaluation de l'enseignement, une autre conception vise à objectiver l'évaluation. Selon certains chercheurs, les approches quantitatives sont considérées comme les seules qui soient fiables parce que plus objectives et permettant des comparaisons (Rossi and Freeman, 1985, p. 127). La tendance générale dans tous les pays européens montre que le recours aux indicateurs va augmentant, à la fois comme mesure directe et comme partie des évaluations qualitatives d'experts (West et Trinczek, 1998, p. 185). Mais trois difficultés importantes dans l'utilisation des statistiques et des indicateurs dans le processus d'évaluation sont liées à leur fiabilité, à leur validité et à leur interprétation (West et Trinczek, 1998, p. 192-194).

Si nous questionnons le statut donné aux résultats de l'évaluation de l'enseignement par les étudiants, dans la recherche menée à l'UA, il apparaît que l'information produite par un dispositif d'évaluation de l'enseignement par les étudiants n'atteint pas réellement un statut de connaissance au sens de savoir sur l'enseignement puisqu'elle n'est pas reformulée, relativisée et débattue dans le cadre d'un processus de référentialisation (Figari, 1994).

3. Attribution de valeur et place donnée aux étudiants dans un dispositif d'EEE

L'attribution de valeur s'attache aux critères de valeur considérés pertinents d'un programme. Elle reste très relative et débattue. Il est apparu notamment que le mouvement en faveur de la prise en compte des bénéficiaires des programmes dans leur évaluation s'est dessiné aux Etats-Unis dans les années 70-80. La prise en compte

du point de vue des étudiants dans l'évaluation de l'enseignement exprime de même une évolution de valeur en ce sens que leur avis sur l'enseignement est sollicité.

A l'IUT de Clermont-Ferrand, une contradiction est évidente puisque même si l'EEE est mise en place, il apparaît que ce point de vue pose problème. Soit il devient une sorte de référent de la valeur de l'enseignement soit la capacité des étudiants à juger l'enseignement est contestée soit le dispositif d'évaluation lui-même est contesté.

4. Utilisation des résultats de l'évaluation

L'utilisation de la connaissance rend compte des façons dont l'information produite est utilisée pour modifier les programmes et politiques (Shadish, Cook et leviton, 1991). L'utilisation des résultats de l'évaluation est un problème capital et récurrent tant est patent que bien souvent ils restent lettre morte, sont biaisés ou mal utilisés. Dans les évaluations de l'enseignement universitaire on distingue deux types principaux d'utilisation, une utilisation pour contrôler et sanctionner à savoir plutôt ou démarche de contrôle (dite aussi sommative), ou une utilisation pour améliorer l'enseignement à savoir une démarche formative. Dans la recherche menée à l'Université d'Auvergne, nous avons étudié comment l'institution universitaire se positionne par rapport à l'utilisation des résultats de l'évaluation. Alors que l'objectif institutionnel affiché était de mettre en place une ressource formative, un glissement vers le contrôle de l'évaluation par les responsables de l'université a été manifeste. Les résultats n'ont pas été communiqués à l'ensemble du collectif pédagogique mais seulement aux responsables et aux enseignants concernés. De plus, l'EEE n'a été suivie d'aucun accompagnement pédagogique ce qui en limite la portée formative. Les différents types d'effets paradoxaux étudiés reflètent l'ambiguïté de l'institution comme des acteurs responsables et enseignants vis à vis de l'EEE qui a oscillé entre

trois positions antagonistes : la considérer comme une formalité administrative, comme une ressource formative pour améliorer l'enseignement ou comme un outil de contrôle. Ce qui pourrait expliquer qu'un dispositif d'EEE même à visée formative ait suscité plus d'inertie et d'opposition que d'adhésion chez les enseignants qui redoutaient son utilisation à des fins de contrôle. Ainsi il est clairement apparu à quel point l'utilisation des résultats de l'évaluation constitue un problème central particulièrement délicat.

5. Pratique d'évaluation et impact d'un dispositif informatisé d'évaluation de l'enseignement par les étudiants

La pratique d'évaluation traduit les tactiques et stratégies professionnelles des évaluateurs face aux contraintes externes.

En ce qui concerne la pratique de l'évaluation à l'Université d'Auvergne, il apparaît que si le dispositif informatisé d'évaluation de l'enseignement par les étudiants a facilité l'effectuation de l'évaluation dans ses aspects de recueil, de traitement et de mise à disposition de l'information, il a aussi été ce qui a cristallisé les résistances étant donné qu'il a rendu visibles les points de vue des étudiants sur les enseignements et sur les enseignants. Nous avons pu constater que l'EEE même inscrite dans une visée formative n'a pas suffi pour être un vecteur de valorisation de l'enseignement même si on peut considérer qu'elle y a contribué et qu'elle a signifié l'intérêt naissant pour la pédagogie de la part de l'institution universitaire ainsi que l'a pointé notamment Bru (2004, p.24). Mais elle a suscité une telle méfiance, inertie ou résistance qu'elle aurait exigé encore plus d'échanges entre les acteurs et d'autres ressources pédagogiques pour aider les enseignants à améliorer leur enseignement

ainsi que d'autres modalités d'évaluation de l'enseignement. L'évaluation de l'enseignement suppose qu'elle soit inscrite dans une politique globale universitaire pour que cette pratique soit effective et qu'elle ne masque pas les profonds désaccords quant aux questions, valeurs et méthodes qui y sont liées.

En conclusion : limites et passages

Les cinq composantes d'une théorie de l'évaluation de programme élaborées par Shadish, Cook et Leviton forment un cadre qui représente sans aucun doute une source de questionnement épistémologique important pour l'évaluation de l'enseignement universitaire que ce soit du point de vue de son organisation et de son évolution, des valeurs de l'enseignement privilégiées, de la connaissance produite par l'évaluation de l'enseignement par les étudiants, ainsi que de l'utilisation des résultats et de la pratique de l'évaluation. Un tel cadre liant théorie et pratique, heuristique pour les chercheurs et l'évaluation de l'évaluation reste cependant difficile à utiliser d'autant que sur le terrain pratique de l'évaluation les critères de l'utilité pour l'action l'emportent sur ceux de la vérité et que les résistances aux changements sont mues par des logiques contradictoires parfois opposées à l'objectif institutionnel d'évaluation de l'enseignement. Un point important à souligner dans le parrallèle que nous tentons. Dans les travaux sur l'évaluation de programme analysés par Shadish, Cook et Leviton, l'évaluation est positionnée comme externe au programme. Ils soulignent également l'importance de mobiliser activement tous les acteurs du programme évalué. Dans notre étude, l'évaluateur est interne puisqu'il s'agit des étudiants. Il manque peut-être un intermédiaire externe entre les étudiants et les équipes pédagogiques ainsi que d'autres formes d'évaluation de l'enseignement étant donné le

changement impliqué par l'EEE dans le système d'enseignement universitaire et le changement requis dont l'EEE n'est qu'un élément.

Références

Bru, M. (2004). Pratiques enseignantes à l'Université : opportunité et intérêt des recherches. Dans E. Anoot & M.F. Fave-Bonnet (Eds.), *Pratiques pédagogiques dans l'enseignement supérieur : enseigner, apprendre, évaluer* , 17-36. Paris : L'Harmattan.

Campbell, D.T. (1988). *Methodology and epistemology for social science: Selected Papers.* In E.S. Overman (Ed.), Chicago : University of Chicago Press.

Cranton, P. (2001). Interpretative and Critical Evaluation, *New Directions for Teaching and Learning*, 88, 87- 97, San Francisco : Jossey-Bass.

Feyerabend, P. (1979). *Contre la méthode*. Paris : Seuil.

Figari, G. (1994). *Évaluer, Quel référentiel ?* Bruxelles : De Boeck.

Kuhn, T.S. (1972). *La structure des révolutions scientifiques*, Paris : Flammarion.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA : Harvard University Press.

Rossi, P. H. & Freeman, H.E. (1985). *Evaluation : A Systematic approach* (3ème ed.). Beverly-Hills CA : Sage.

Scriven, M. (1980). *The logic of evaluation. Inverness*, CA : Edgepress.

Scriven, M. (1983). Evaluation Ideologies. In G.F. Madaus, M. Scriven, D.L. Stufflebeam (Eds.), *Evaluation Models : Viewpoints on educational and human services evaluation* , 229-260. Boston : Kluwer-Nijhoff.

Shadish, W.R., Cook, T.D. & Leviton, L.C. (1991). *Foundations of Program Evaluation*, Newbury Park, CA : Sage.

Stake, R. E. (1975). To evaluate an art program. In R.E. Stake (Ed.), *Evaluating the arts in education : A responsive approach* , 13-31. Columbus, OH : Merrill.

Stake, R. E. (1980). Program evaluation, particularly responsive evaluation. In W. Dockrell
& D. Hamilton (Eds.), *Rethinking educational research* , 72-87. London : Hodder and Stoughton.

Wholey, J.S. (1983). *Evaluation and public effective management.* Boston : Little : Brown.

Yin, R. K. (1984). *Case study research : Design and methods*. Beverly-Hills, CA : Sage.

Younes, N. (2006). *L'effet évaluation de l'enseignement supérieur par les étudiants*.

Thèse de Doctorat. Grenoble : Université Pierre Mendès France.

VERS UN REFERENTIEL D'EVALUATION D'UN CURRICULUM / PROGRAMME DE FORMATION DANS L'ENSEIGNEMENT SUPERIEUR

Maria Palmira Alves, Université du Minho

Eusébio André Machado, Université du Minho

Mots-clés : curriculum ; enseignement supérieur ; évaluation

Résumé :

Des nombreuses définitions conçoivent le curriculum comme un plan d'études, ce qui signifie l'établissement d'un inventaire et d'une séquence de contenus ou de disciplines (Tyler, 1947). D'autres définitions le conçoivent non seulement comme un programme d'enseignement avec une liste de thèmes d'un cours et d'orientations didactiques, mais aussi la façon dont les enseignants et les étudiants mettent en place les prescriptions programmatiques (Kelly, 1986). D'une façon général, notamment au niveau de l'enseignement supérieur, le curriculum d'un cours est vraiment un programme de formation qui poursuit un profil professionnel, et donc un "dispositif particulier" par lequel se définie « un ensemble de constructions et de mécanismes » qui se mettent en place pour programmer un enseignement, soit qu'il s'agit du curriculum construit par les décideurs, soit lequel qui est vécu par les étudiants (Figari, 1994). Dans ce contexte, nous présenterons une étude de cas développée à l'Université du Minho, qui a permis, à partir d'une analyse de contenu des entretiens, de dégager les logiques sous-jacentes à l'évaluation du curriculum et d'y dégager la façon dont l'évaluation du curriculum se développe, surtout en ce qui concerne les différentes variables du dispositif, les critères d'évaluation, les acteurs concernés, le traitement des donnés, etc.

Les catégories ont été axées dans les paradigmes d'évaluation du curriculum proposés par Alkin et Elliot (1985), aussi que dans les métanarratives de l'évaluation identifiées par Alves et Machado (2003): la performance et l'émancipation. En effet, les différents apports théoriques ont permis soit une analyse critique des processus de design du curriculum, sa mise en œuvre et sa gestion curriculaire dès le niveau micro – la classe – jusqu'au macro – la totalité du système éducatif, c'est-à-dire, le programme de formation, soit la proposition d'un référentiel d'évaluation.

ASSURANCE QUALITÉ ET ÉVALUATION DE PROGRAMME

Marie-Françoise Fave-Bonnet, Université Paris X Nanterre, France

Mots clés : Processus de Bologne, évaluation des programmes, assurance qualité

Résumé :

Dans la déclaration de Bologne de 1999, les ministres européens se sont engagés, pour 2010, à une structuration des niveaux de sortie dans l'enseignement supérieur à + 3 (licence ou *bachelor)*, + 5 (master) et + 8 (doctorat), à la mise en place du système ECTS, mais aussi à promouvoir une « assurance-qualité » des formations. La déclinaison française du processus de Bologne, le LMD, « traduit » donc cet objectif de *quality assurance.*

Il sera donc intéressant d'analyser les mots utilisés et leurs traductions (assurance de la qualité, garantie de la qualité, standards, référentiels, évaluation des formations, auto-évaluation, etc.) afin de comprendre les enjeux, les influences et les objectifs implicites et explicites à l'œuvre.

Cette analyse se base sur les documents recueillis dans le réseau TUNING (réseau européen qui, depuis 2000, rassemble une centaine d'universités pour mettre en œuvre, par disciplines, les axes du processus de Bologne, dont l'évaluation) et ceux recueillis en tant que *Promoteur de Bologne.*

Introduction : La notion d'assurance qualité dans le Processus de Bologne

Dans la déclaration de Bologne de 1999, les ministres européens se sont engagés, pour 2010, à une structuration des niveaux de sortie dans l'enseignement supérieur à + 3 (licence ou *bachelor)*, + 5 (master) et + 8 (doctorat), à la mise en place du système ECTS, mais aussi à promouvoir une « assurance de la qualité » des programmes.

La déclinaison française du processus de Bologne, le LMD, « traduit » donc cet objectif de *quality assurance*.

Les textes du Processus de Bologne, les conférences, les rapports, etc. lors de la période (Berlin, Bergen, etc.) sont en anglais, et s'inscrivent dans une culture anglo-saxonne (accountability, quality assurance, benchmarking, etc.) qui n'a pas d'équivalent en France.

Cette communication propose d'analyser les mots utilisés et leurs traductions (assurance de la qualité, garantie de la qualité, référentiels, évaluation des formations, auto-évaluation, etc.) afin de tenter de comprendre les enjeux, les influences et les objectifs implicites et explicites à l'œuvre. Plus qu'un problème de traduction linguistique, il s'agit de comprendre dans la déclinaison française du Processus de Bologne, le « LMD », comment l'évaluation des programmes[1] est **traduite en mots et en actes**.[2]

[1] Généralement sous le terme « d'évaluation des formations ».

[2] Le corpus comprend :

Sans entrer dans un long développement sémantique, nous postulons que chaque mot a un sens différent pour chacun d'entre nous : c'est une convention. A l'intérieur d'un même pays, c'est l'usage qui stabilise un signifiant. Lors d'une traduction, le phénomène est amplifié. *Evaluation* en anglais n'a pas le même sens qu'*évaluation* en français[3]…

1. Éléments de problématique

1.1. Détermination du champ

Dans les universités, l'évaluation des programmes, des formations, s'inscrit dans un processus plus large d'évaluation des établissements (EVALUE, 1998). Il s'agit, en effet, d'évaluer toutes les activités de l'établissement : enseignement, recherche, administration, etc. Deux instances principalement évaluent, en France, les universités : le Comité National d'Evaluation (depuis plus de 20 ans), et le Ministère (à l'occasion des contrats quadriennaux). Nous ne nous intéresserons ici qu'à ce qui concerne l'évaluation des enseignements et des formations.

Lors de ces évaluations, une évaluation interne (auto-évaluation) précède toujours l'évaluation externe. Nous nous centrerons sur l'évaluation interne des formations.

Enfin, afin d'améliorer la qualité des formations, les différents ministères qui se sont succédé incitent fortement les universités à procéder à une évaluation de leurs formations. Ces évaluations font partie du champ étudié.

1.2. Clarification sur l'usage, en France, de quelques mots :

- Le terme de « **programmes** » n'est pas utilisé en France lorsqu'il s'agit de formations dans l'enseignement supérieur (niveau tertiaire). On parlait de « programmes » au niveau primaire et

---

- les documents recueillis dans le réseau TUNING auquel je participe. « Tuning Educational Structures in Europe » depuis 2001, est un projet financé par la Commission Européenne pour étudier la mise en place de l'harmonisation européenne des diplômes. Projet regroupant plus d'une centaine d'universités en Europe (15 pays dans TUNING I, 40 dans TUNING III). Huit rencontres de deux jours dans différents pays. Enquêtes sur les compétences attendues à l'issue des licences et des masters, contribution sur les ECTS, les contenus de formation, l'assurance qualité, etc. www.relint.deusto.es/TuningProject/index.htm

- les documents de travail en tant que *Promoteur de Bologne.* Les promoteurs de Bologne, nommés par la Commission Européenne (15 pour la France) ont pour mission d'aider les établissements d'enseignement supérieur à la mise en place des recommandations de Bologne.

- les articles, rapports, textes ministériels, glossaires, etc. qui peuvent permettre de comprendre comment le ministère français s'est servi de cette réforme pour développer l'évaluation des formations (voir bibliographie).

[3] C'est une des difficultés que nous avons rencontrée au-début de la recherche EVALUE. Le projet EVALUE est recherche comparative sur l'évaluation et l'auto-évaluation des universités en Europe, qui a associé onze équipes de recherche dans huit pays (Allemagne, Espagne, Finlande, France, Italie, Norvège, Portugal, Royaume-Uni) et une quarantaine de chercheurs de différentes disciplines. Elle a été subventionnée par la Commission Européenne (DG XII, 1995-1998) et coordonnée par Pierre DUBOIS (Laboratoire Travail et Mobilités, Université Paris X Nanterre)

secondaire pour désigner les contenus, les connaissances que les élèves étaient censés avoir appris en fin d'année ou en fin de cycles. Ceux-ci étaient fixés par le ministère de l'éducation.

On parle de moins en moins de programmes en primaire, et l'on parle d' « objectifs », de « compétences attendues ». Seules les disciplines du secondaire continuent à fixer des programmes. Il s'agit là du vieux conflit français entre les partisans de l'éducation et ceux de l'instruction. Et je n'aborderai pas ici les polémiques actuelles à l'occasion de la mise en place du « *socle commun de connaissances* » à l'occasion de cette rentrée…

Dans le supérieur, on retrouve le clivage (enseigner/apprendre ?) avec le terme de **formation** opposé à enseignement. Dans la plupart des textes officiels depuis une quinzaine d'années, c'est le terme de « formations » qui est utilisé pour désigner un ensemble d'activités pédagogiques organisées autour d'un objectif, d'un diplôme, d'un programme (*la licence de…*, etc.) Il peut donc désigner des activités, mais aussi des structures pédagogiques ou administratives.

- On distingue, dans les textes officiels, depuis les arrêtés Bayou de 1997, **l'évaluation des formations et l'évaluation des enseignements.**

Notons d'abord que tous les textes lient évaluation des enseignements et évaluation des formations, et que, faute d'information sur les textes, et de formation à l'évaluation, règne une grande confusion sémantique sur le sujet. Dans la pratique, comme le montre un rapport récent (DEJEAN, 2002), il est difficile de séparer les deux : comment faire, en effet, une évaluation des formations (pertinence de l'organisation pédagogique, des contenus, des méthodes, des moyens, etc. et résultats obtenus en termes de réussite, voire d'insertion professionnelle) sans faire une évaluation des enseignements, c'est à dire sur ce qui se passe en cours ? Or c'est bien l'évaluation des enseignements, entre autres parce qu'elle doit faire appel à l'appréciation des étudiants, qui est redoutée, et parfois même rejetée par les universitaires.

- Autre constat : alors qu'en anglais on distingue clairement « teaching » (ce qu'on enseigne) et « learning » (ce qui est appris), en France, on sépare **l'évaluation des étudiants** (notes, examens, etc.) des procédures d'évaluation des enseignements et des formations. Les connaissances et compétences acquises par l'étudiant sont pourtant des indicateurs intéressants pour évaluer l'efficacité des formations (NAVARRO, 2006).

2. Historique de l'évaluation des formations en France

2.1. Le Comité National d'Evaluation

L'évaluation des formations apparaît institutionnellement avec le Comité National d'Evaluation, le **CNE** (créé en 1985). En 1995, par exemple, un rapport titre : *Évolution des universités, dynamique de l'évaluation.* Mais il s'agit, à cette époque, d'évaluations externes, qui portent sur tous les domaines d'activités de l'établissement. L'objectif est de développer l'évaluation comme levier de changement. Ces évaluations ne sont accompagnées d'aucune sanction.

En 20 ans, le CNE a affiné ses méthodes et ses outils et a abouti, en 2003, à la publication du *« Livre des références : les références de l'assurance de la qualité dans les établissements d'enseignement supérieur* » afin d'aider les établissements dans leur auto-évaluation. Nous y reviendrons.

2.2. Les injonctions ministérielles

C'est à l'occasion de la publication des arrêtés par Jack Lang, en mai-juin 1992, après des mois de discussion, qu'apparaissent les unités capitalisables, la semestrialisation **et l'évaluation des enseignements**. L'arrêté du 26 mai 1992 dans l'article 24 relatif au DEUG, à la licence et à la maîtrise préconise : *« une procédure d'évaluation des enseignements, faisant notamment appel à l'appréciation des étudiants, peut* (souligné par nous) *être établie par le conseil d'administration… »*

À son arrivée en avril 1993, le nouveau ministre (de droite), François Fillon déclare: *« il appartient aux universités de décider quand elles mettront en place cette rénovation* ». En 1994, la date limite de mise en application des dispositions des arrêtés de 1992 et de 1993, initialement prévue pour la rentrée universitaire 1993-1994, est repoussée à la rentrée universitaire 1996-1997.

Notons que le Conseil d'Etat du 13 mars 1996 a rejeté la demande d'annulation (concernant l'évaluation des enseignements) posée au titre de l'atteinte au principe d'indépendance des professeurs. C'est bien l'évaluation par les étudiants qui pose problème…

Par ailleurs, en janvier 1997, la Direction de l'Évaluation et de la Prospective (DEP) du Ministère publie pour la première fois un indicateur de «performance», université par université et formation par formation. Il concerne le taux d'accès en second cycle, taux réel et taux simulé. La publication rencontre un large écho dans la presse. Le Monde titre en première page: «*L'éducation nationale évalue les inégalités entre universités et la réussite des étudiants. Le taux de succès dans le premier cycle progresse depuis 1990*». Un autre article titre : *«les réserves des présidents d'établissement face à l'évaluation*».

Les « arrêtés BAYROU », en septembre 1997, entre autres dispositions, précisent la place respective de l'évaluation des enseignements et des formations :

- dans l'arrêté général relatif aux premier et second cycles universitaires, au titre II : « *un rapport annuel sur le fonctionnement de chaque formation est établi par le responsable de la formation et remis aux instances compétentes de l'établissement* (article 8).

- au titre IV : « habilitation et évaluation », pas de changement pour ce qui concerne l'habilitation des diplômes nationaux (article 22 et 24). L'article 23 est, par contre, totalement nouveau. *«Pour chaque cursus, est organisée une procédure d'évaluation des enseignements et de la formation. Cette évaluation, qui prend en compte l'appréciation des étudiants, se réfère aux objectifs de la formation et des enseignements »*. Cette procédure, garantie par une instruction ministérielle, a deux objectifs. Elle permet, d'une part, à chaque enseignant de prendre connaissance de l'appréciation des étudiants sur les éléments pédagogiques de son enseignement. Cette partie de l'évaluation est destinée à l'intéressé. La procédure permet, d'autre part, une évaluation de l'organisation des études dans la formation concernée, suivie pour chaque formation par une commission, nommée par le Président de l'Université et composée d'un nombre égal de représentants élus des étudiants et d'enseignants-chercheurs.

Le texte distingue donc bien « évaluation des enseignements », destinée à l'enseignant, et « évaluation des formations », destinée à l'établissement.

Près de 10 ans plus tard, il faut constater que ces évaluations sont loin d'être mises en œuvre dans les universités…

Mais les instances ministérielles vont revenir « à la charge » à l'occasion d'autres réformes…

2.3. L'évaluation des formations professionnelles

Dès 1999, est mise en place une commission d'évaluation des formations de gestion, qu'elles soient universitaires publiques ou privées, sous l'égide du ministère de l'Education nationale. Les formations de gestion sont sans doute celles où s'exerce le plus la compétition internationale et où la lisibilité de l'offre de formation en France et l'évaluation de la qualité étaient impératives.

C'est la création de la licence professionnelle qui réitère l'obligation à l'évaluation. Dans l'arrêté du 17 novembre 1999 relatif à la création de la licence professionnelle, est précisé, à l'article 13 : « *le dossier de demande d'habilitation décrit le projet de licence professionnelle sous ses différents aspects. Il doit préciser […] les modalités prévues, d'une part, pour le contrôle des connaissances des étudiants et, d'autre part, pour la mise en oeuvre de la procédure d'évaluation des enseignements et de la formation dans les conditions prévues par l'article 23 de l'arrêté du 9 avril 1997 susvisé* ». En bref, si l'évaluation n'est pas mise en place, la formation ne sera par habilitée par le ministère…

De plus, l'arrêté institue une commission nationale de suivi de la licence professionnelle et ses modalités de bilan (d'évaluation ?). Ce dispositif de commission nationale sera reprise lors de la mise en place du « LMD ».

2.4. La Déclaration de Bologne (19 juin 1999)

Un point de cette déclaration concerne l'évaluation : les ministres s'engagent à la « *promotion de la coopération européenne en matière d'évaluation de la qualité, dans la perspective de l'élaboration de critères et de méthodologies comparables »*[4] *(Promotion of European co-operation in quality assurance with a view to developing comparable criteria and methodologies).* Le terme est ***quality assurance*** est traduit par « *évaluation de la qualité* ». Il s'agit, en effet, de promouvoir une coopération des agences et organismes d'évaluation, ce qui sera pris en charge par l'ENQA[5].

A Prague (19 mai 2001) est réaffirmée l'importance donnée aux procédures d'évaluation de la qualité

2.5. La mise en œuvre en France du processus de Bologne : le LMD

De 1999 à 2002, le ministère français ne prend aucune disposition de mise en œuvre du processus de Bologne. Rien n'est dit, par exemple, sur la qualité et l'évaluation, dans le décret du 8 avril 2002 portant « *application au système français d'enseignement supérieur de la construction de l'espace européen de l'enseignement supérieur* ».

Mais la circulaire de mise en œuvre du 14 novembre 2002, adressée aux chefs d'établissements, n'a qu'un objectif : préciser «*les critères d'évaluation que le ministère entend mettre en oeuvre pour apprécier vos propositions* ».

Comme le souligne le texte, «*expliciter les critères d'évaluation constitue- en quelque sorte - une «première»*. Le texte explique les objectifs : « *Il convient de le faire pour deux raisons principales :*

En premier lieu, il faut fonder sur le dispositif d'évaluation la valeur du diplôme national,[6] *c'est-à-dire d'un diplôme dont la qualité est garantie par l'Etat et qui donne à l'étudiant l'assurance*

[4] Traduction de la DRIC du ministère de l'éducation nationale
[5] European Association for Quality Assurance in Higher Education
[6] Souligné dans le texte…

d'une qualification reconnue. La qualité d'un diplôme national ne saurait être reconnue sur la base de critères purement formels ; c'est la consistance réelle qui doit être appréciée, consistance des projets de formation comme des résultats déjà obtenus. Seulement ainsi de véritables garanties peuvent être données aux étudiants.

En second lieu, la construction de I'espace européen de I'enseignement supérieur repose sur la confiance mutuelle entre les divers systèmes nationaux. Cette confiance ne peut être durable que fondée sur des méthodes d'évaluation de la qualité des formations et des diplômes, solides et reconnues par l'ensemble des pays.

Le texte poursuit : *« II convient donc de vous dire comment vos propositions seront évaluées ce qui, tout en répondant à vos attentes naturelles et légitimes, vous permettra aussi de les finaliser avec la meilleure efficacité.* » Suivent, pendant 9 pages, les critères liés à l'espace européen, les critères de cohérence globale, les critères spécifiques pour la licence (amélioration pédagogique et professionnalisation) et, pour les masters, l'ancrage avec la recherche pour les masters recherche, et avec les milieux professionnels pour les masters professionnels.

Le ministère insiste sur l'autonomie laissée aux établissements pour construire leur offre de formation, tout en se posant en « arbitre » (avec des phrases telles que « *les jeux de concurrence inter-établissements auxquels on assiste parfois et que la « pause démographique » pourrait exacerber ne sont plus de mise et doivent donc cesser* »), ou en « gendarme » (*« L'attractivité de l'offre de la France est à ce prix. Le ministère y veillera.* »).

2.6. Les développements de la place de l'évaluation dans le Processus de Bologne

La conférence de Berlin va définitivement ancrer l'évaluation de la qualité dans le processus. Dans le communiqué de la conférence des ministres chargés de l'enseignement supérieur à Berlin le 19 septembre 2003, intitulé *Réaliser l'espace européen de l'enseignement supérieur*, c'est en premier point qu'est développée l'évaluation de la qualité : « *C'est un fait établi que la qualité de l'enseignement supérieur se trouve au cœur de la construction d'un Espace européen de l'enseignement supérieur. Les Ministres s'engagent à soutenir le développement de l'évaluation de la qualité au niveau institutionnel, national et européen. Ils soulignent la nécessité de développer des critères et des méthodologies reconnus mutuellement en matière d'évaluation de la qualité.*

Ils soulignent également que, conformément au principe de l'autonomie institutionnelle, la responsabilité première en matière d'évaluation de la qualité dans l'enseignement supérieur incombe ***à chaque institution elle-même***[7]*, ce qui fonde les bases d'une réelle responsabilité du système académique au sein du cadre national de la qualité.*

Ils conviennent donc que d'ici 2005 les systèmes nationaux d'évaluation de la qualité devraient inclure :

- *Une définition des responsabilités des organismes et institutions concernés*
- *Une évaluation des programmes ou des institutions, qui inclue l'évaluation interne, l'examen externe, la participation des étudiants et la publication des résultats*
- *Un système d'accréditation, de certification ou de procédures comparables*

[7] Souligné par nous

- *La participation, la coopération et la mise en réseau internationales.*

Au niveau européen, les Ministres demandent à l'ENQA, en coopération avec l'EUA, l'EURASHE et l'ESIB, de mettre au point une série de standards, de procédures et de lignes directrices pour l'évaluation de la qualité qui fassent l'objet d'un consensus, d'explorer les moyens d'assurer un système adéquat d'examen par les pairs pour l'évaluation de la qualité et/ou pour les agences et organismes chargés de l'accréditation, et d'en faire rapport aux Ministres en 2005 par l'intermédiaire du groupe de suivi. Les compétences des autres groupes et réseaux travaillant sur l'évaluation de la qualité devront être dûment prises en compte. »

Le rapport en question parait en mars 2005. Le titre original du rapport de l'ENQA (European Association for Quality Assurance in Higher Education) est *Standards and Guidelines for Quality Assurance in the European Higher Education Area*, et la traduction proposée (par le CNE, membre de l'ENQA) est « *Références et lignes directrices pour le management de la qualité dans l'espace européen de l'enseignement supérieur ».* Le CNE, dans son introduction à la traduction française (janvier 2006) précise : *« sans vouloir éviter un débat trop technique et probablement déjà dépassé sur la traduction la plus juste de l'expression quality assurance dans le domaine de la formation et plus précisément de l'enseignement supérieur (s'agit-il d'assurance de la qualité*[8]*, d'assurance qualité ou bien de garantie de la qualité ?), le CNÉ a choisi ici de traduire ladite expression par la notion plus englobante de "management de la qualité"*[9] *qui paraît plus conforme à l'esprit et à la portée des démarches proposées dans le rapport.*

De même, le terme anglais process a été traduit par le mot "démarche" qui paraît mieux exprimer la nature des exercices d'appréciation et d'amélioration de la qualité que celui de processus. »

Ces « précautions » linguistiques expriment (de la part du CNE) une « certaine » idée de l'évaluation qui se démarque de la vision anglo-saxonne de la *quality assurance.*

Il semble que les problèmes de traductions ne se posent pas seulement pour la France. Un des derniers colloques de l'ENQA, *The Language of European Quality Assurance* ((29-30 Juin 2006, à Warwick) avait pour objectif de travailler sur un certain nombre de termes :

Quality, standards, quality assurance/quality control, accountability, enhancement/improvements, principles/values, independence/autonomy/academic freedom, compliance/conformity/adherence, convergence/harmonisation/standardisation, assessment, evaluation, inspection, accreditation, audit, review, experts/peers, self-assessments/self evaluation, visits, reports, judgements, follow-up, recommendation.

---

[8] ***Assurance de la qualité :*** *Ensemble des activités préétablies et systématiques mises en œuvre dans le cadre du système qualité et démontrées en tant que de besoin, pour donner la confiance appropriée en ce qu'une entité satisfera aux exigences pour la qualité. (Source : Office québécois de la langue française, 2003)*

[9] **Management de la qualité :** Ensemble des activités de la fonction générale de management qui déterminent la politique qualité, les objectifs et les responsabilités et les mettent en œuvre par des moyens tels que la planification de la qualité, la maîtrise de la qualité, l'assurance de la qualité et l'amélioration de la qualité dans le cadre du système qualité. (Source : Office québécois de la langue française, 2003)

3. Les modèles en jeu

Plusieurs conclusions provisoires peuvent être avancées à l'issue de cette analyse de textes :

- La traduction « littérale » d'un certain nombre de termes trouve difficilement un équivalent en français courant. Ce problème n'est pas spécifique à la France ou au français : les termes utilisés pour parler d'évaluation restent encore dans une nébuleuse sémantique faute de conceptualisation et de définitions de ces termes. On peut supposer que ce flou est accentué, en France, par l'absence de « culture » de l'évaluation dans l'enseignement supérieur en France. Celle-ci est due à un manque de recherches sur la question, à une absence de formation à l'évaluation (tant dans le secondaire que dans le supérieur) et à une résistance à toute forme d'inspection (FAVE-BONNET, 2005).

- Dans la déclinaison française du processus de Bologne, le ministère s'est servi de cette réforme pour tenter de développer l'évaluation des formations et des enseignements. L'insistance sur la mise en place de procédures d'auto-évaluation des formations (arrêtés BAYROU, CNE, circulaire LMD) traduit le choix politique d'une évaluation « formative » pour l'établissement. Ce passage d'une évaluation externe sommative à une évaluation interne formative s'inscrit dans le mouvement d'autonomisation des établissements universitaires : il s'agit de passer d'une politique centralisée à une véritable autonomie des universités. L'évaluation est ici le levier de l'amélioration de la qualité des formations. On peut remarquer que d'autres politiques seraient possibles : augmentation des crédits pour un meilleur suivi des étudiants, formation pédagogique, etc. L'assurance qualité « à la française » reste cependant marquée par le contrôle de la qualité des formations par l'accréditation des diplômes par l'Etat.

Ces évolutions ne se font pas sans tension. Ces derniers mois, c'est le terme de « *performance* » qui est mis en avant par le ministère :

- la LOLF :

Initiée en 2001, effective depuis janvier 2006, levier de la réforme de l'État, la loi organique relative aux lois de finances (LOLF) enclenche un processus de transformation radicale des règles budgétaires et comptables de l'État. L'ambition de cette loi est de réformer la façon dont l'argent public est utilisé par chaque ministère. « *Cette réforme vise à instaurer une gestion plus démocratique et plus efficace des dépenses publiques* » selon la présentation du ministère.

Le budget général de l'État est désormais découpé en 34 missions, 132 programmes et près de 580 actions qui détaillent les finalités des politiques publiques. Auparavant, chaque ministère recevait une certaine somme. Désormais, des moyens sont alloués en vue d'une action précise.

Désormais, pour une politique donnée, l'État se fixe des objectifs précis à atteindre, avec des moyens alloués pour atteindre ces objectifs. *« La « performance » des services sera ainsi mesurée de façon plus concrète : une politique publique sera d'autant plus performante que les objectifs auront été atteints ou approchés grâce aux moyens alloués. Une vraie « culture de la performance » devrait ainsi s'instaurer au sein de l'administration.* »

« Les responsables de programme devront définir, chacun à leur niveau, les objectifs, les stratégies et les indicateurs de performances des politiques publiques dont ils sont chargés. Ils disposeront d'une plus grande liberté dans la gestion des moyens alloués, en contrepartie d'un engagement sur des objectifs de performance. Les gestionnaires auront l'obligation de rendre compte de leur gestion pour expliquer les éventuels écarts entre prévision et réalisation. »

La LOLF instaurant une culture du résultat (avec insistance sur les termes *d'efficacité et d'efficience*) , « *l'évaluation devient primordiale* ». Les indicateurs viendront aider à l'appréciation de l'utilisation des moyens par rapport aux finalités recherchées.

On voit bien ici que l'évaluation a un objectif sommatif et non formatif, même si, de façon indirecte, les *« responsables de programmes »* vont devoir mettre en œuvre des indicateurs afin d'évaluer leurs activités, ce qui ne manquera pas d'introduire une préoccupation d'efficacité et d'efficience dans la fonction publique, et donc dans les universités…

- depuis mai 2006, la DEP est devenue Direction de l'Evaluation, de la Prospective **et de la Performance**[10], dans laquelle on trouve une sous-direction de la performance de l'enseignement supérieur, de la recherche et de l'innovation.

Sommes-nous condamnés à toujours choisir entre Charybde et Scylla, entre évaluation formative et sommative ?

Les différentes enquêtes menées dans la recherche EVALUE (1998) nous ont permis de dégager les conditions d'un autre modèle d'évaluation. Comme nous avons tenté de le synthétiser (FAVE-BONNET, 2003), les modèles actuels d'évaluation ont montré leurs limites. L'évaluation de contrôle par l'Etat est légitime, car les universités sont un service public. Mais elle ne parvient pas à engager une dynamique de transformations dans les universités, même quand elle est associée à une politique de contractualisation, car elle est souvent contradictoire avec le développement de l'autonomie universitaire.

L'auto-évaluation s'essouffle assez vite parce qu'elle ne parvient pas à interagir avec les décisions et les financements externes.

L'évaluation managériale se heurte aux cultures universitaires : les universités ne sauraient fonctionner selon la seule logique du marché. Elles doivent être efficaces et efficientes, mais en diffusant des valeurs autres que celles du marché, en appliquant des réglementations, en respectant et en impliquant les enseignants, les personnels et les étudiants.

Une **évaluation *"pluraliste", "dynamique" et "contextuelle"*** semble être le modèle d'évaluation le plus pertinent pour la transformation des universités. Évaluation pluraliste parce qu'elle doit associer et prendre en compte les analyses et les points de vue de tous les acteurs et de tous les partenaires de l'Université, fussent-ils contradictoires. Évaluation dynamique en ce sens qu'elle doit comparer l'Université à elle-même, dans son évolution. Évaluation contextualisée, enfin, car elle doit être sensible aux différentes dimensions du contexte de chaque université, en particulier lors de comparaisons.

Ce modèle d'évaluation nous semble le plus pertinent pour l'évolution des universités contemporaines.

[10] Souligné par nous

Bibliographique

ANNOOT E., FAVE-BONNET M.-F (coord.) (2004) : *Pratiques pédagogiques dans l'enseignement supérieur : enseigner, apprendre, évaluer.* Paris, L'Harmattan, 328 pages.

BONNAFOUS S., DIZAMBOURG B., MENDEL G. & MOREAU J.-F. (1997) : *Changement et participation à l'Université, modernisation administrative: l'exemple de Paris XII,* Grenoble, Presses Universitaires de Grenoble.

Comité National d'Évaluation (1995) : *Évolution des universités, dynamique de l'évaluation : 1985-1995*, rapport au Président de la République, Paris, La Documentation Française.

Comité National d'Évaluation (2005) : *Livre des références : les références de l'assurance de la qualité dans les établissements d'enseignement supérieur*, Paris, 28 pages.

COMMISSION EUROPEENNE (1994) : *European pilot projects for evaluating quality in Higher Education. Guidelines for participating institutions*. Bruxelles, Bureau Erasmus

COMMISSION EUROPEENNE (1996) : *Note d'information sur les résultats des projets pilotes européens pour l'évaluation de la qualité dans l'enseignement supérieur,* Luxembourg, 12 p.

CROZIER M. (1990), président du groupe de travail : *L'évaluation des performances pédagogiques des établissements universitaires*, rapport au ministre, La Documentation Française.

DEJEAN J. (2002) : *L'évaluation de l'enseignement dans les universités françaises*, rapport établi pour le Haut Conseil de l'Evaluation de l'Ecole, Paris, février 2002, 108 pages

European Journal of Education (1997) : *Evaluation and Quality in Higher Education.* Vol. 32, n° 2, Paris: European Institute of Education and Social Policy, 221 p.

European Journal of Education (1999) : *Evaluation: the Institutions' Response,* Vol. 34, n° 3, Paris: European Institute of Education and Social Policy. 384 p.

EVALUE (Évaluation et auto-évaluation des Universités en Europe), Collectif sous la direction de Pierre Dubois (1998) *: Rapport final*. Commission Européenne (DG XII, TSER, Axe II, Contrat SOE2-CT 95-2004) juillet 1998, 233 pages (version en français).

FAVE-BONNET (2003) : L'évaluation dans les universités en Europe : une décennie de changements, *Revue des Sciences de l'Education* (Montréal), *Les transformations de l'Université : regards pluriels*, Vol XXIX, n°2, pages 319 à 336.

FAVE-BONNET M.-F.(2005) : L'évaluation des enseignements à l'Université en France : enjeux, pratiques et déontologie, *18ème Colloque international de l'Association pour le Développement des Méthodologies d'Evaluation en Education (ADMEE-Europe) :* Comment évaluer ? Outils, dispositifs et acteurs, Reims, 24-26 octobre 2005, CD Rom, 7 pages.

FAVE-BONNET M.-F. (2005) : La difficile mise en œuvre de l'évaluation des formations dans les universités en France : culture française, culture universitaire ou cultures disciplinaires ? , *Les Dossiers des Sciences de l'Education n° 13* : « Evaluation et cultures », Presses Universitaires du Mirail, pages 97 à 104

FAVE-BONNET M.-F. coord. (2006) : *Évaluer formations et enseignements dans les études supérieures*, Actes du Colloque inter- universitaire organisé par l'ADMES Grand-Ouest, Angers,

27 et 28 mars 2003, en collaboration avec Georges BERTIN, Richard LESCURE, Gabriela Ibeth NAVARRO, Presses Universitaires d'Angers, 250 pages.

FREVILLE Y. (2001) : *« Des universitaires mieux évalués, des universités plus responsables »*, Rapport d'information n° 54 sur la politique de recrutement et la gestion des universitaires et des chercheurs, Commission des Finances du Sénat, 522 pages.

MONTEIL J.-M. (2002) : Mise en oeuvre du schéma licence - master - doctorat (LMD), circulaire du 14 novembre 2002, Ministère de l'éducation nationale, Paris, 9 pages.

MUSSELIN C. (2001) : *La longue marche des universités françaises*, Paris, PUF, 218 pages.

NAVARRO G. (2006) : *L'évaluation des étudiants dans les universités françaises*, thèse de doctorat de Sciences de l'Education, sous la direction de Marie-Françoise FAVE-BONNET, Université Paris X Nanterre, mars 2006, 363 pages

THELOT C. (1993) : *L'évaluation du système éducatif : coûts, fonctionnement, résultats*, Paris, Nathan, 160 pages.