Mesure et Évaluation

Sous la direction de
GILLES RAÎCHE, KARINE PAQUETTE-CÔTÉ et DAVID MAGIS
Avec la collaboration de Diane Leduc et d'Hélène Meunier

DES MÉCANISMES POUR ASSURER LA VALIDITÉ DE L'INTERPRÉTATION DE LA MESURE EN ÉDUCATION

VOLUME 2 – L'évaluation

Presses de l'Université du Québec

COLLECTION
MESURE ET EVALUATION

Sous la direction de Gilles Raîche,
Université du Québec à Montréal

La collection **Mesure et évaluation** soutient la diffusion de recherches et de travaux fondamentaux, ainsi que de matériel didactique pour les niveaux collégial et universitaire, dans le domaine de la mesure et de l'évaluation en éducation et, plus largement, en sciences humaines.

Les nouveaux enjeux sociétaux et les besoins émergents des milieux de pratique demandent aux intervenants d'être informés des avancées récentes afin de les soutenir dans leur travail. Aussi, Mesure et évaluation offre aux chercheurs un moyen de partager les résultats de leurs travaux avec ces intervenants tout en faisant progresser la recherche, que ce soit en matière de mesure et d'évaluation des apprentissages, de programmes ou encore de méthodologie de recherche.

Les textes publiés sont soumis à un processus d'arbitrage avec le soutien d'évaluateurs externes. La collection **Mesure et évaluation** souscrit à l'adaptation canadienne-française, par la *Revue des sciences de l'éducation*, des règles de publication de l'American Psychological Association.

DES MÉCANISMES POUR ASSURER LA VALIDITÉ DE L'INTERPRÉTATION DE LA MESURE EN ÉDUCATION

VOLUME 2 – L'évaluation

DANS LA MÊME COLLECTION

DES MÉCANISMES POUR ASSURER LA VALIDITÉ DE L'INTERPRÉTATION DE LA MESURE EN ÉDUCATION, Volume 1 – La mesure
Sous la direction de Gilles Raîche, Karine Paquette-Côté et David Magis.
Avec la collaboration de Diane Leduc et d'Hélène Meunier
ISBN-978-2-7605-2685-3, 148 pages

Membre de L'ASSOCIATION NATIONALE DES ÉDITEURS DE LIVRES

Presses de l'Université du Québec
Le Delta I, 2875, boulevard Laurier, bureau 450, Québec (Québec) G1V 2M2
Téléphone : 418 657-4399 – Télécopieur : 418 657-2096
Courriel : puq@puq.ca – Internet : www.puq.ca

Diffusion/Distribution :

Canada et autres pays : Prologue inc., 1650, boulevard Lionel-Bertrand, Boisbriand (Québec) J7H 1N7 – Tél. : 450 434-0306/1 800 363-2864

France : Sodis, 128, av. du Maréchal de Lattre de Tassigny, 77403 Lagny, France – Tél. : 01 60 07 82 99

Afrique : Action pédagogique pour l'éducation et la formation, Angle des rues Jilali Taj Eddine et El Ghadfa, Maârif 20100, Casablanca, Maroc – Tél. : 212 (0) 22-23-12-22

Belgique : Patrimoine SPRL, 168, rue du Noyer, 1030 Bruxelles, Belgique – Tél. : 02 7366847

Suisse : Servidis SA, Chemin des Chalets, 1279 Chavannes-de-Bogis, Suisse – Tél. : 022 960.95.32

Sous la direction de
GILLES RAÎCHE, KARINE PAQUETTE-CÔTÉ et DAVID MAGIS
Avec la collaboration de Diane Leduc et d'Hélène Meunier

DES MÉCANISMES POUR ASSURER LA VALIDITÉ DE L'INTERPRÉTATION DE LA MESURE EN ÉDUCATION

VOLUME 2 – L'évaluation

Presses de l'Université du Québec

Catalogage avant publication de Bibliothèque et Archives nationales du Québec et Bibliothèque et Archives Canada

Vedette principale au titre :

Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation
Textes présentés lors d'un colloque tenu en mai 2009 à l'Université d'Ottawa,
dans le cadre du 77e Congrès de l'ACFAS.

Comprend des réf. bibliogr.

Sommaire: t. 1. La mesure – t. 2. L'évaluation.

ISBN 978-2-7605-2685-3 (v. 1)
ISBN 978-2-7605-2687-7 (v. 2)

1. Tests et mesures en éducation – Évaluation – Congrès. 2. Tests et mesures en éducation – Validité – Congrès.
3. Tests et mesures en éducation – Interprétation des résultats – Congrès. I. Raîche, Gilles, 1956- .
II. Paquette-Côté, Karine, 1983- . III. Magis, David. IV. Congrès de l'ACFAS (77e : 2009 : Université d'Ottawa).

LB3050.5.D47 2011 371.2601'3 C2011-940772-8

Les Presses de l'Université du Québec reconnaissent l'aide financière du gouvernement du Canada
par l'entremise du Fonds du livre du Canada et du Conseil des Arts du Canada pour leurs activités d'édition.

Elles remercient également la Société de développement des entreprises culturelles (SODEC)
pour son soutien financier.

Mise en pages : Info 1000 mots

Conception de la couverture : Richard Hodgson

Dépôt légal – 3e trimestre 2011 – Bibliothèque et Archives nationales du Québec / Bibliothèque et Archives Canada
Imprimé au Canada

TABLE DES MATIÈRES

Pour assurer la validité de l'interprétation de la mesure en éducation

Aspects pratiques

David Magis, Gilles Raîche et Karine Paquette-Côté

1. INTRODUCTION

Depuis ses premières définitions, vers 1950, la notion de validité a évolué, de sorte qu'on ne la considère plus aujourd'hui comme étant uniquement une caractéristique intrinsèque de la mesure, mais plutôt en relation avec l'utilisation et l'interprétation du score associé à la mesure. Au début, la validité était plutôt associée aux tests ; c'est pourquoi on a encore tendance à parler, à tort, de validité d'un test. Cette conception de la validité en situation de mesure en éducation était probablement tributaire de la prépondérance d'une fonction de prédiction et d'une forme de validité importante à ce moment-là, soit celle de validité de critère (*criterion-related validity*). La première édition d'un des classiques en mesure et évaluation en éducation, *Educational measurement*, par Linquist (1951), reflète bien cette situation, car la majeure partie du chapitre sur la validité (Cureton, 1951) traite de mesures critériées et de puissance de prédiction. Dans la deuxième édition de ce volume, Cronbach (1971) publie un chapitre intitulé *Test validation* dans lequel il met encore davantage l'accent sur le fait que la notion de validité est associée au test. Cependant, il met fortement en évidence la nécessité de tenir compte de la validité de contenu (*content validity*) et de la validité de concept (*construct validity*). Ces deux aspects de la validité, par la suite, vont être mis de l'avant pendant plusieurs années. Par exemple, la notion de validité de contenu sera très importante dans les opérations de définition du domaine en éducation pour planifier les tâches

d'évaluation des apprentissages. Un peu négligée avec l'introduction des approches par compétences en éducation, la validité de contenu pourrait regagner son importance pour soutenir la mise en œuvre des programmes élaborés selon ces approches. La notion de validité de concept, pour sa part, est primordiale lorsque vient le temps de vérifier ce que mesure véritablement un instrument de mesure. C'est encore Cronbach (1988) qui semble donner une nouvelle vie à la notion de validité en montrant la nécessité d'appuyer le processus de validation par des arguments. Par la suite, dans les troisième et quatrième éditions d'*Educational measurement*, Messick (1989) et Kane (2006) améliorent encore plus ce processus de validation conçu en tant que jugement évaluatif intégré du degré avec lequel les évidences empiriques et les justifications théoriques confirment la justesse et la pertinence des inférences et des actions basées sur les résultats obtenus à partir des instruments d'évaluation (Messick, 1989, p. 13). C'est Kane, toutefois, qui semble avoir vraiment proposé des approches plus structurées, fondées sur la logique de l'argumentation, pour défendre ce jugement évaluatif. Ces avancées autour de la notion de validité permettent de constater que les résultats des évaluations en éducation ont rarement une interprétation signifiante en eux-mêmes. Le score obtenu à un test n'a de sens que dans le cadre de référence utilisé pour l'interprétation et dans les inférences d'évaluation.

Cet ouvrage constitue, en quelque sorte, les actes d'un colloque tenu en mai 2009 à Ottawa à l'occasion du 77e congrès annuel de l'Association francophone pour le savoir (Acfas), qui invitait les chercheurs et les intervenants en éducation à porter un regard critique sur les mécanismes permettant d'assurer la validité de l'interprétation de la mesure en éducation afin d'en dégager des tendances, de trouver des solutions et d'alimenter les pratiques pédagogiques et le développement de la recherche dans le domaine de la mesure et de l'évaluation en éducation. Ce colloque comportait trois volets : 1) théorique : le développement des concepts, les avancées théoriques, les modèles de mesure, etc. ; 2) technique et technologique : instrumentation, méthodes de sélection des critères d'évaluation, méthodes de sélection des items d'un test, méthodes d'évaluation de la représentativité des items, modèles de réponse à l'item, utilisation des technologies dans la production du jugement, etc. ; 3) pratique : domaines et contextes d'application, exemples d'application dans le système d'éducation, formation des professionnels de l'évaluation, etc.

Le présent ouvrage, soit le volume 2, est consacré au troisième volet, celui de la pratique de l'évaluation. Les deux premiers volets, soit le volet théorique et le volet technique et technologique, ont été traités dans le premier volume.

2. CONTENU DE L'OUVRAGE

Cet ouvrage est divisé en sept chapitres regroupés en deux parties. Dans la première partie, l'accent est mis sur l'évaluation des apprentissages et les pratiques pédagogiques, tandis que la seconde partie aborde le jugement et l'argumentation de la validité en éducation.

Le premier chapitre est consacré au diagnostic des processus cognitifs utilisés pour répondre aux items d'un test. Dans ce contexte, le modèle DINA (Tatsuoka, 1990) permet d'intégrer à un modèle psychométrique de réponse à l'item une matrice Q établissant les liens entre les items du test et les processus cognitifs sous-jacents. L'utilisation d'une telle matrice dans les modèles psychométriques permet d'évaluer les processus cognitifs utilisés par les élèves lorsqu'ils répondent aux items du test. Cependant, il a été montré récemment (Rupp et Templin, 2008) qu'une formulation incorrecte de la matrice Q pouvait avoir un impact important sur l'estimation des paramètres du modèle. Dans ce premier chapitre, Loye, Caron, Pineault, Tessier-Baillargeon, Burney-Vincent et Gagnon proposent une méthode de validation des paramètres d'un modèle diagnostique cognitif et étudient sa capacité à porter un jugement qualitatif sur la matrice Q. Pour cela, les auteurs s'appuient sur des données empiriques recueillies lors du passage d'une épreuve de mathématiques à l'École Polytechnique de Montréal.

L'impact de l'utilisation du degré de certitude dans le cadre de l'évaluation diagnostique constitue le sujet du deuxième chapitre. Le degré de certitude peut être vu comme une caractéristique supplémentaire des items d'un test (Leclerq et Poumay, 2005), qui consiste à demander aux élèves d'évaluer leur degré de certitude de la justesse de leurs réponses aux items. Cependant, une mauvaise utilisation du degré de certitude pourrait affecter grandement l'analyse des réponses aux items des élèves, notamment lorsqu'ils surestiment la qualité de leurs réponses. De plus, l'interprétation des résultats pourrait varier en fonction des exigences cognitives, et donc être différente selon la propension à la surestimation de la qualité des réponses. Dans ce cadre, Boulé et Laveault étudient les liens existant entre le degré de certitude, et, par conséquent, le degré de réalisme des élèves, et le sexe et le niveau de performance de ces élèves. S'il appert que le degré de certitude ne varie pas en fonction du sexe, les auteurs montrent toutefois que plus l'élève a un niveau de performance élevé, plus il se fait une idée réaliste de la qualité de ses réponses.

Dans le troisième chapitre, l'accent est mis sur l'intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques. Longtemps, les pratiques d'évaluation des apprentissages ont été dissociées des pratiques pédagogiques. Cependant, les nouvelles

approches par compétences supposent une interaction étroite entre pratiques pédagogiques et pratiques d'évaluation des apprentissages (Biggs, 1995 ; Wiggins, 1998). Dans ce chapitre, Raîche, Leduc, Riopel et Isabelle s'appuient sur les données d'enquêtes nationales et internationales, et notamment sur les Tendances de l'enquête internationale sur les mathématiques et les sciences (TEIMS). Au moyen d'une analyse transversale des items, les auteurs mettent en relation les pratiques pédagogiques et les pratiques d'évaluation des apprentissages, dans un contexte d'approches par compétences.

Le quatrième chapitre, qui clôt la première partie, porte sur l'approche par compétences et la validité d'une situation d'apprentissage. Il est généralement admis (Brousseau, 2003 ; Scallon, 2004) que les compétences se manifestent lorsque le niveau de complexité est élevé et que l'on exige le recours à un bon nombre de ressources. Dans ce contexte, le concept de situation est un élément important du processus d'inférence en évaluation des apprentissages. En suivant le processus d'élaboration d'un modèle suggéré par Silvern (1972), Durand et Trépanier ont développé une grille d'analyse comprenant une échelle nominale. Cette grille a été établie pour l'évaluation de situations complexes en mathématiques, révisée suite aux résultats de cette évaluation, pour enfin être utilisée à nouveau dans d'autres situations en mathématiques.

Pour commencer la seconde partie, le cinquième chapitre propose une analyse empirique de la validité de l'interprétation des résultats de l'évaluation des apprentissages en éducation. Les modèles portant sur la validité de l'interprétation des résultats, dont le modèle ECD (*evidence-centred design* ; Mislevy, Almond et Lukas, 2004) et le modèle de structure d'argumentation interprétative de Kane (2006), ont été étudiés de façon théorique du point de vue de leur validité en mesure et évaluation (Lissitz, 2009). Toutefois, leur application pratique n'a pas fait l'objet d'études importantes. Dans ce chapitre, Paquette-Côté et Raîche appliquent la structure d'argumentation interprétative de Kane à l'analyse du contenu des politiques institutionnelles d'évaluation des apprentissages (PIEA) dans l'enseignement collégial. Des hypothèses quant à l'exhaustivité, l'exclusivité et la pertinence des catégories du modèle de Kane ont ainsi été envisagées.

Le sixième chapitre s'intéresse à la validité du jugement professionnel des enseignants du primaire dans un contexte d'approche par compétences. Dans cette étude, Ndinga propose un modèle permettant de se prononcer sur la validité du jugement professionnel des enseignants du primaire. Ce modèle fait suite à une analyse des principales références dans les encadrements du ministère de l'Éducation,

du Loisir et du Sport du Québec (MELS, 2006). Le modèle de Kane (2006), examiné au chapitre précédent, sert également de référence dans le modèle proposé par Ndinga. De plus, ce dernier modèle repose largement sur le travail en équipe de chaque cycle, où les membres de l'équipe agiraient comme des juges donnant un avis motivé, ce qui aide à assurer à la fois la validité du jugement et l'équité envers les élèves.

Finalement, le septième et dernier chapitre aborde la question de l'accord entre experts et de son impact sur la validité de la mesure. Il arrive en effet souvent que l'évaluation en éducation repose sur l'appréciation de juges ou d'experts. Il est dès lors nécessaire de pouvoir fournir une estimation du degré d'accord entre ces personnes et de quantifier l'impact d'un éventuel désaccord sur l'évaluation proposée. On doit donc disposer d'un outil permettant de repérer les experts proposant des classements nettement différents de ceux de la majorité, ainsi que de déterminer l'impact du retrait des données correspondantes. Pour ce faire, Blais, Charlin, Grondin, Lambert, Loye et Gagnon proposent une procédure reposant sur le modèle à facettes de Rasch (Linacre, 1989, 1994) pour déterminer le degré d'accord des classements établis par les experts, avant et après le retrait des données considérées comme étant problématiques pour la majorité des experts. On donne un exemple de la procédure à suivre à l'aide d'un test de concordance de script avec des cotes attribuées par des radio-oncologues de l'Université de Montréal.

3. CONCLUSION

Lorsque l'on porte un regard critique sur les mécanismes permettant d'assurer la validité de l'interprétation de la mesure en éducation, on se rend compte de la complexité des sujets entourant le concept de validité de la mesure en évaluation. Ainsi, si l'on s'en tient à la mesure, on doit tenir compte du diagnostic des processus cognitifs utilisés pour répondre aux items d'un test, de l'impact de l'utilisation du degré de certitude dans le contexte de l'évaluation à fonction diagnostique, ainsi que de l'impact sur la qualité de la mesure de l'accord entre les experts. En revanche, si l'on considère la dimension évaluative, on doit tenir compte des aspects suivants : les principes de l'intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques, le développement d'une grille pour l'évaluation de situations complexes en mathématiques, l'analyse empirique de la validité de l'interprétation des résultats de l'évaluation des apprentissages ainsi que la validité du jugement professionnel des enseignants du primaire dans un contexte d'approche par compétences. On présente donc dans

cet ouvrage un état des récents travaux de recherche dans ces différents domaines, et on ouvre la porte à de futurs développements et à des applications concrètes pour l'évaluation des apprentissages.

4. REMERCIEMENTS

La réalisation de cet ouvrage n'aurait pas été possible sans le soutien des chercheurs du Collectif pour le développement et les applications en mesure et évaluation (C*dame*). Aussi, toute notre gratitude va à Diane Leduc, postdoctorante en mesure et évaluation en éducation à l'Université de Montréal, et à Hélène Meunier, doctorante en mesure et évaluation à la Faculté des sciences de l'éducation de l'Université du Québec à Montréal, pour leur aide considérable à la rédaction de cet ouvrage, surtout en ce qui concerne le suivi avec les auteurs et avec la maison d'édition. Nous remercions aussi Gaëlle Joris pour le travail de révision linguistique des textes et de leur conformité avec les normes de l'APA adoptées par la *Revue des sciences de l'éducation* (Raîche et Noël-Gaudreault, 2009).

RÉFÉRENCES

Biggs, J. (1995). Assessing for learning : Some dimensions underlying new approaches to educational assessment. *Alberta journal of educational research, 41*(1), 1-17.

Brousseau, G. (2003). *Glossaire de quelques concepts de la théorie des situations didactiques en mathématiques*. Tiré du site : <http://math.unipa.it/~grim/Gloss_fr_Brousseau.pdf>. Mise à jour : 25/02/2003. Page consultée le 14 octobre 2007.

Kane, M. T. (2006). Validation. Dans R. L. Brennan (dir.), *Educational measurement* (4e éd.). Westport, Connecticut : Praeger Publishers.

Leclercq, D. et Poumay, M. (2005). *Degrés de certitude : épistémologie, méthodes et conséquences*. 18e Colloque international de l'ADMÉÉ-Europe, Reims, France.

Linacre, J. M. (1994). Constructing measurement with a many-facet Rasch model. Dans M. Wilson (dir.), *Objective measurement, theory into practice, volume 2*. Norwood, New Jersey : Ablex publishing.

Linacre, J. M. (1989). *Many-facet, Rasch measurement*. Chicago, Illinois : Mesa-Press.

Lindquist, E. F. (1951). *Educational measurement*. Washington, District of Columbia : American council on education.

Lissitz, R. W. (dir.) (2009). *The concept of validity : revisions, new directions, and applications*. Charlotte, Caroline du Nord : Information Age Publishing.

Messick, S. (1989). Validity. Dans R. L. Linn (dir.), *Educational measurement* (3e éd.). New York, New York : American council on éducation et Macmillan.

Ministère de l'Éducation, du Loisir et du Sport (MELS, 2006). *La valeur accordée au jugement professionnel des enseignants. Question et éléments de réponse – Principales références dans les encadrements ministériels*. Québec, Québec : Gouvernement du Québec.

Mislevy, R. J., Almond, R. G. et Lukas, J. F. (2004). *A brief introduction to evidence-centered design*. CSE Report 632. Los Angeles, Californie : The national center for research on evaluation, standards and student testing (CRESST), Center for the study of evaluation (CSE), University of California.

Raîche, G. et Noël-Gaudreault, M. (2009). Une adaptation, pour le Canada francophone, des règles de publication de l'APA : typographie et présentation des références. *Revue des sciences de l'éducation, 35*(1), 227-234.

Rupp, A. A., et Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and psychological measurement, 68*(1), 78-96.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Saint-Laurent, Québec : Éditions du Renouveau pédagogique.

Silvern, L. C. (1972). *System engineering applied to training*. Houston, Texas : Gulf publishing company.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. Dans N. Frederiksen, R. Glaser, A. Lesgold et M. G. Shafto (dir.), *Diagnostic monitoring of skill and knowledge acquisition*. Hillsdale, New Jersey : Erlbaum.

Wiggins, G. (1998). *Educative assessment : designing assessments to inform and improve student performance*. San Francisco, Californie : Jossey-Bass Publishers.

Partie 1

L'ÉVALUATION DES APPRENTISSAGES ET LES PRATIQUES PÉDAGOGIQUES

Chapitre 1

Validité du diagnostic issu d'un mariage entre didactique et mesure sur un test existant

Nathalie Loye, France Caron, Jenny Pineault,
Michèle Tessier-Baillargeon, Carole Burney-Vincent
et Michel Gagnon

Le diagnostic des processus cognitifs utilisés pour répondre aux items d'un test repose sur la structure cognitive sous-jacente aux items, identifiée par des experts et formalisée par une matrice Q. Le modèle DINA inclut deux indicateurs de la qualité de cette structure et, par conséquent, de la validité du diagnostic. Cette étude présente les résultats liés à la validité du diagnostic réalisé par des didacticiens lorsque la matrice Q est formée à partir du classement des items selon une ontologie des mathématiques et selon les compétences visées. Les données proviennent des résultats du test de mathématiques proposé en juin 2008 aux nouveaux étudiants de l'École Polytechnique de Montréal.

1. INTRODUCTION

En éducation, l'intérêt porté par les chercheurs à la fonction diagnostique de l'évaluation est relativement récent. Ce n'est que depuis les années 1980-1990 que la perspective de tirer parti de l'évaluation pour identifier les forces, mais surtout les faiblesses des sujets qui se soumettent à un test a pris une certaine importance et abouti en 1995 à un ouvrage collectif de Nichols, Chipman et Brennan faisant le point sur la situation. Plus récemment, Leighton et Gierl (2007) ont publié un livre présentant l'état de

développement de la théorie portant sur le diagnostic des processus cognitifs ainsi que diverses applications. C'est également en 2007 qu'est né le *Special interest group* (SIG) intitulé *Cognition and assessment* au sein de l'American educational research association (AERA). Ce groupe vise notamment à promouvoir les recherches qui associent cognition et psychométrie.

Pendant ce temps, les réformes implantées dans de nombreux pays entraînent des changements importants. Par exemple, le curriculum au Québec, développé dans une approche par compétences, vise le développement chez l'élève de la capacité à mobiliser et à utiliser efficacement un ensemble de ressources dans des situations d'une certaine complexité, et suppose une évolution des pratiques évaluatives des enseignants en classe, de même qu'une refonte des tests ministériels. En particulier, l'évaluation du degré de compétence doit porter *autant sur le processus ou les démarches adoptées par l'élève que sur les résultats auxquels il parvient* (MELS, 2006).

Il semble donc pertinent que les modèles de mesure en éducation évoluent afin de permettre l'évaluation des processus cognitifs utilisés par les sujets lorsqu'ils répondent aux items d'un test. C'est exactement l'objectif des chercheurs tels que Tatsuoka (1983, 2005), DiBello, Stout et Roussos (1995) ou Roussos, DiBello, Stout, Hartz, Henson et Templin (2007), qui ont mis au point des modèles statistiques afin d'estimer, pour chaque sujet, son degré de maîtrise des processus cognitifs requis par les items du test, à partir de son schéma de réponses. L'introduction dans le modèle mathématique d'une matrice Q assure l'interface entre les items du test et les processus visés par ces items. Cette matrice Q contenant des lignes représentant les items et des colonnes représentant les processus cognitifs a initialement été formalisée par Tatsuoka (1990). Lorsqu'un lien existe entre le processus et l'item, la valeur correspondante dans Q est 1 ; dans le cas contraire, cette valeur est 0.

L'état des connaissances actuelles sur les modèles cognitifs diagnostiques (MDC) laisse supposer qu'il peut s'avérer difficile de réaliser un diagnostic cognitif à partir d'un test qui aurait été conçu pour un autre usage et dans un cadre de référence non cognitif. Gorin (2006) ou Leighton et Gierl (2007) considèrent d'ailleurs qu'il est essentiel de définir un cadre de référence adéquat avant de réaliser un diagnostic portant sur les processus de pensée utilisés par les sujets lorsqu'ils répondent aux items d'un test. Les tests eux-mêmes vont donc devoir évoluer.

Quoi qu'il en soit, l'utilisation d'un modèle statistique pour inférer quels sont les processus que les sujets maîtrisent et ceux qui posent problème, à partir des schémas de réponses aux items d'un test,

exige de valider la structure cognitive de la matrice Q élaborée pour cet usage. Si la construction et la validation de la matrice Q peuvent être basées sur des processus externes au modèle statistique, comme le recours à des experts *a priori* ou *a posteriori*, ou encore à l'analyse du langage des sujets qui passent le test ou des traces qu'ils laissent, les paramètres d'items estimés par certains MDC fournissent une autre source de validation de la matrice Q. Rupp et Templin (2008) ont d'ailleurs étudié l'effet de certaines erreurs de spécifications dans la matrice Q sur les paramètres d'items et de sujets estimés à l'aide du modèle DINA (Junker et Sijtsma, 2001). Ils ont observé les variations des paramètres du modèle en fonction des erreurs qu'ils ont volontairement introduites dans la matrice Q en utilisant des données simulées. De la Torre (2008) fournit une procédure pour valider les liens de la matrice Q à partir des paramètres estimés par le modèle DINA et d'un paramètre composite qui en découle. Il a appliqué sa méthode à des données simulées, mais aussi à des données réelles (ne provenant pas d'un cadre de référence cognitif), et montré ainsi sa pertinence.

La qualité de la matrice Q est la condition *sine qua non* de l'utilisation d'un modèle cognitif ; pourtant, à part les recherches citées ci-dessus, peu d'études ont été réalisées sur ce sujet. En outre, la plupart des articles disponibles reposent sur des données simulées, et d'autres utilisent et réutilisent les mêmes données réelles (ex. : les données sur les soustractions de fractions utilisées par Tatsuoka en 1990 l'ont ensuite été par de la Torre et Douglas en 2004, puis encore par de la Torre en 2008). Enfin, et c'est le cas pour l'exemple donné, les processus cognitifs retenus sont souvent très spécifiques et en lien étroit avec des contenus disciplinaires élémentaires (ex. : simplifier ou réduire une fraction) et ne sont pas sans rappeler les objectifs de comportement associés à des connaissances procédurales, d'où l'inconvénient de ne rejoindre dans un même test qu'un nombre limité de contenus et d'évacuer à nouveau les tâches plus complexes, faisant appel à plusieurs étapes, concepts et processus, qui sont pourtant nécessaires à l'évaluation des compétences.

Il est donc utile de se demander si une procédure de validation basée sur les paramètres d'items d'un modèle cognitif diagnostique permet véritablement de porter un jugement sur la qualité de la matrice Q que l'on a développée, lorsqu'on utilise des données réelles de tests élaborés par des enseignants en vue d'un diagnostic et touchant un certain nombre de contenus disciplinaires plus ou moins diversifiés, et que les tâches associées aux items peuvent nécessiter plusieurs étapes de raisonnement et une combinaison inédite de concepts et de processus.

2. CONTEXTE

2.1. L'épreuve diagnostique de l'École Polytechnique de Montréal

Le test de mathématiques de l'École Polytechnique de Montréal fournit justement de telles données. Cette épreuve, qui date de 1989, fait suite à une refonte des cours de mathématiques de première année implantée à l'automne 1988. Conçue par deux professeurs de l'École mandatés par la Direction des études, elle vise à vérifier l'état des connaissances des étudiants sur des notions ou des techniques jugées nécessaires pour réussir les cours de mathématiques de première année. La taille des cohortes, alliée à la nécessité d'une correction rapide, a abouti à un test à choix multiple comprenant 60 items dont la durée de passation est de deux heures, sans calculatrice ni documentation. Pour chaque item, quatre réponses sont proposées, dont la bonne réponse. Une cinquième option *je ne sais pas* a été introduite afin de distinguer un étudiant qui ne connaît pas la réponse d'un étudiant qui n'a pas eu le temps de répondre à la question.

Depuis 1989, tout étudiant admis à l'École est invité à passer l'épreuve de mathématiques (son score lui donne d'ailleurs droit à une légère bonification de la note obtenue dans son premier cours de mathématiques obligatoire). Avec l'invitation, l'étudiant reçoit un fascicule qui décrit les notions et techniques mathématiques visées par l'épreuve ainsi qu'une liste de références afin de l'aider à se préparer. La moitié des 60 items du test porte sur des notions du niveau secondaire et l'autre moitié, sur des notions du niveau collégial. Les items sont rattachés à six domaines : fonctions élémentaires (13 items), trigonométrie (7 items), géométries plane et analytique (10 items), matrices et vecteurs (12 items), calcul différentiel (12 items) et calcul intégral (6 items). Les titres donnés aux domaines permettent de rattacher les notions et techniques à des cours de mathématiques du cursus scolaire québécois.

Dans les années 1990, une étude statistique liant les résultats de l'épreuve à la réussite des étudiants à l'École Polytechnique (les moyennes générales et les crédits obtenus, trimestre après trimestre) a conduit le Conseil académique à reconnaître une valeur diagnostique à cette épreuve de mathématiques et à y voir un moyen d'offrir à tout étudiant qui entre à l'École une évaluation de sa connaissance des différents domaines ciblés et de lui proposer des moyens d'accroître ses possibilités de réussite à l'École.

Les résultats transmis aux étudiants correspondent à un score établi de la façon suivante : 1 point pour une bonne réponse, une pénalité de –0,25 point pour une mauvaise réponse et 0 point pour l'option *je ne sais pas* ou en cas d'absence de réponse. L'étudiant reçoit un score pour chaque domaine ainsi qu'une note globale. Étant donné que les domaines correspondent à des cours bien précis, l'étudiant qui veut réviser certaines notions ou techniques peut le faire plus facilement.

Au fil des années, une banque d'items a été constituée. Elle contenait 60 items en mars 1989, puis 250 items en 1992 et enfin 985 items en 1998. Au total, on compte 343 items en fonctions élémentaires, 149 en trigonométrie, 199 en géométries plane et analytique, 76 en matrices et vecteurs, 138 en calcul différentiel et 80 en calcul intégral. Chaque item est classé selon 1) le domaine, 2) des indications pour identifier soit des sous-domaines, soit d'autres caractéristiques, 3) un numéro et 4) parfois, une lettre pour indiquer des variantes.

2.2. Classification didactique des items de l'épreuve

En dépit du fait qu'elles ne font appel qu'à un choix de réponse, les questions à choix multiple de l'épreuve diagnostique de mathématiques de l'École Polytechnique de Montréal sont équivalentes à des problèmes à résoudre, car les connaissances requises et la stratégie à adopter ne vont pas de soi. La résolution des items peut nécessiter plusieurs étapes (ex. : interprétation d'un graphique, utilisation des propriétés d'un concept sous-jacent, mise en équations, calculs algébriques) ; il ne s'agit donc pas de simples exercices d'application de procédures connues. Cette épreuve diagnostique demande de faire appel et d'utiliser, de façon efficace et souvent inédite pour les étudiants, un ensemble de ressources développées en mathématiques, en liant concepts et méthodes, dans la recherche d'une solution à un problème. On peut donc envisager de l'utiliser pour évaluer l'état des compétences en mathématiques des étudiants, compétences qu'ils devront continuer à développer et à utiliser dans l'apprentissage et l'application des nouveaux contenus de leur formation.

Il convient de préciser que si l'évaluation par questions à choix multiple a souvent été associée à une approche procédurale de l'enseignement, approche jugée incompatible avec le développement de compétences globales, cela n'invalide pas d'emblée l'utilisation des questions à choix multiple pour évaluer l'état des compétences d'étudiants d'un point de vue diagnostique. En effet, par son caractère externe, lié à une institution particulière (l'École Polytechnique), où le résultat ne compte pas pour l'admission, cette évaluation n'est pas

prise en compte par le niveau d'enseignement qui précède (au secondaire ou au collégial) et ne vient donc pas fausser l'apprentissage en le réduisant à d'éventuels schémas d'association entre les questions typiques et les procédures adéquates. Les problèmes gardent donc toute la fraîcheur et la complexité nécessaires à une véritable utilisation des compétences. Et lorsqu'ils sont bien conçus pour représenter les erreurs typiques qui peuvent se produire dans la modélisation, l'interprétation ou le raisonnement, les leurres peuvent être de puissants catalyseurs des compétences qui font défaut aux étudiants. L'épreuve paraît donc *a priori* constituer un contexte objectif d'observation des compétences mathématiques des étudiants admis en génie, si l'on prend la peine d'inférer de ce choix de réponses les processus utilisés dans la résolution des items.

Afin d'affiner le diagnostic (par exemple en appliquant un modèle cognitif diagnostique aux données de ce test), nous avons établi une caractérisation didactique des items. Pour ce faire, nous nous sommes servis d'une grille élaborée et validée pour analyser des textes d'étudiants universitaires en situation de résolution de problèmes (Caron, 2001). Utilisée pour caractériser les erreurs commises et inférer de ces erreurs les compétences utilisées, cette grille (voir l'annexe 1) est subdivisée selon les quatre phases du processus de résolution de problèmes de Polya (1945) (analyse, planification, exécution, retour), auxquelles on a ajouté une phase générale de contrôle (Schoenfeld, 1985) pour rendre compte des processus menant au passage d'une phase à l'autre, incluant les arrêts et les reprises. Pour chacune de ces phases, on a établi une liste d'éléments de compétence susceptibles d'être utilisés dans la résolution d'un problème de nature mathématique. Ces différents éléments de compétence ont ensuite été associés à trois types de compétences (De Terssac, 1996) que nous avons projetés sur les différents modes d'utilisation du savoir mathématique :

- les *compétences d'explicitation* (les *savoir-dire*) pour traduire ce qui est, ce qu'il y a à faire et ce qui a été fait. En mathématiques, cela suppose la maîtrise d'au moins trois langages différents (naturel, symbolique et graphique) (De Serres et Groleau, 1997) et du passage de l'un à l'autre, par la modélisation et l'interprétation ;
- les *compétences d'évaluation* (les *savoir-se-situer*) pour identifier, légitimer et valider tout ce qu'on engage dans l'action. En mathématiques, c'est, par exemple, la décomposition en sous-problèmes, l'identification des cas possibles, la reconnaissance des champs théoriques appropriés, la navigation dans le réseau des concepts, l'utilisation du raisonnement ;

– les *compétences d'intervention* (les *savoir-intervenir*) pour agir en mettant en situation les connaissances disponibles et en transformant les situations rencontrées. Dans le contexte mathématique, cela correspond surtout à l'utilisation des différentes méthodes (analytiques et algorithmiques) permettant de calculer, d'appliquer une transformation sur un objet, de résoudre un système d'équations, d'optimiser une fonction.

Dans la grille, chacun des différents éléments de compétence est suivi du symbole EX, EV et IN selon qu'il peut être associé, respectivement, aux compétences d'explicitation, d'évaluation ou d'intervention. Certains éléments ont été associés à plus d'un type ; dans ce cas, le premier symbole de la liste indique le type dominant.

Puisque avec l'épreuve diagnostique, nous ne disposons pas de données explicites sur la démarche utilisée par les étudiants, quelques lacunes concernant certains éléments de compétence de la grille peuvent être difficiles à repérer. En revanche, nous avons pris soin de cibler la forme de langage (naturel, graphique ou symbolique) utilisée dans l'interprétation correcte de l'énoncé, et nous avons aussi croisé cette analyse de compétences avec une analyse du contenu mathématique visé par les items, autant dans l'obtention de la bonne réponse que dans les erreurs pouvant expliquer le choix des leurres.

Cette analyse du contenu s'est d'abord largement appuyée sur l'étiquetage original des items. Les 105 types de contenus ainsi identifiés ont été associés à l'un ou l'autre des champs suivants : algèbre et fonctions, trigonométrie, géométrie, vecteurs, matrices, calcul différentiel, calcul intégral. Puisqu'il s'agit de problèmes à résoudre relativement complexes, il convient de signaler que la plupart des questions font appel à une combinaison de concepts et de processus qui ne relèvent pas tous du même champ. Les leurres peuvent aussi ramener à des types de contenu plus élémentaires (ex. : les calculs algébriques) que ceux qui sont visés par la question de départ. À partir des types de contenu identifiés, nous avons construit une ontologie du savoir mathématique visé par cette épreuve, de façon à prendre éventuellement en compte dans le diagnostic les relations de filiation entre ces types de contenu.

Chacune des réponses à chacun des 985 items a ensuite été associée à un maximum de cinq types de compétence et de cinq types de contenu. Le choix de la bonne réponse témoigne ainsi de la maîtrise des types de compétence et de contenu qui lui sont associés. Le choix d'un leurre ou de la réponse *je ne sais pas* révèle des types de compétence ou de contenu qui ne semblent pas maîtrisés.

2.3. Objectif de l'étude

Cette étude exploite les données 2008 de l'épreuve diagnostique de l'École Polytechnique de Montréal. Son objectif consiste à valider la classification didactique des bonnes réponses des items relativement aux compétences à l'aide d'un modèle cognitif diagnostique. Les compétences, traitées comme étant des processus visés par les items, sont utilisées pour élaborer *a posteriori* une matrice Q à partir du test. Le modèle cognitif diagnostique retenu est le modèle DINA (de la Torre et Douglas, 2004 ; Macready et Dayton, 1977), parce qu'il propose deux paramètres caractérisant les items tout en étant un modèle relativement simple et pour lequel les modélisations peuvent être réalisées à l'aide d'un algorithme programmé dans Ox (Doornik, 2002).

2.3.1. *Le modèle DINA*

Le modèle DINA est un modèle cognitif non compensatoire (une habileté liée à un processus donné ne peut pas compenser une difficulté liée à un autre processus) qui permet d'estimer sous forme dichotomique la maîtrise ou la non-maîtrise des processus cognitifs inclus dans la matrice Q, à partir des scores dichotomiques (réussite ou échec) des sujets aux items du test. En outre, ce modèle fournit l'estimation de deux paramètres d'items qui donnent une information sur la qualité de la description de chaque item par les processus qui lui sont reliés dans la ligne correspondante de la matrice Q. Le premier paramètre, g_j, se définit par la probabilité que les sujets devinent la réponse (*guessing*) de l'item *j* plutôt que de la trouver grâce aux processus ; le deuxième paramètre, s_j, se définit par la probabilité de donner une mauvaise réponse à l'item *j* alors que les processus nécessaires sont maîtrisés (*slipping*). Ces deux paramètres ont une valeur attendue d'autant plus petite que la matrice Q représente de manière valide et fidèle les processus utilisés par les sujets pour répondre à l'item (Rupp et Templin, 2008). En 2008, de la Torre a également montré à l'aide de données simulées qu'omettre un attribut requis par l'item *j* aboutit à une augmentation importante du paramètre s_j, mais également qu'inclure des attributs inutiles fait augmenter le paramètre g_j. Il semble donc logique d'utiliser ces deux paramètres comme des indicateurs de la qualité de la matrice Q dans une démarche de validation de la matrice. Notons toutefois que des petites valeurs pour ces paramètres forment une condition suffisante, mais non nécessaire à un bon ajustement du modèle (de la Torre, 2008).

Pour le modèle DINA, la ligne de la matrice Q correspondant à un item est correctement spécifiée si elle maximise la différence entre la probabilité de bien répondre à l'item selon que l'on possède $(1 - s_j)$ ou

pas (g_j) les attributs spécifiés. De la Torre (2008) définit ainsi un nouvel indicateur δ_j [$\delta_j = (1 - s_j) - g_j$] qui varie selon les attributs qui sont reliés à l'item *j* et peut être considéré comme un indice de discrimination. Plus la discrimination est grande, plus la valeur de δ_j est proche de 1. Le modèle DINA est donc un modèle relativement simple qui fournit l'estimation de deux paramètres et d'un indicateur qui en découle, permettant de vérifier la validité de la matrice Q. De la Torre (2008) propose également un indicateur d'ajustement global $\bar{\hat{g}} + \bar{\hat{s}}$, pour lequel une valeur d'environ 25 % correspond à un ajustement acceptable alors qu'une valeur de 60 % ou 70 % dénote un problème d'ajustement.

3. MÉTHODE

Les données sont constituées par les réponses dichotomisées (réussite ou échec) des 451 étudiants qui ont passé l'épreuve diagnostique en juin 2008. Les 60 items de ce test sont classés selon des éléments de compétences tels que définis et dégagés par les didacticiens de l'équipe. Cette classification offre la possibilité de déterminer plusieurs matrices Q dans lesquelles les éléments de compétences tiennent lieu de processus cognitifs. Les choix décrits donnent six matrices Q qui sont utilisées pour modéliser les données du test de 2008 à l'aide du modèle DINA. Le tableau 1.1 présente un récapitulatif des six matrices élaborées comme suit :

- Dans un premier temps, l'ensemble de toutes les compétences reliées aux items sont prises en considération afin de créer une première version de la matrice Q. De cette première matrice sont exclues les compétences qui sont reliées à moins de 4 items. Cette matrice appelée Q_0 contient 15 compétences ; les 60 items sont conservés.
- À partir de Q_0, la suppression des compétences reliées à moins de 10 items permet de limiter la taille de la matrice et donne la matrice Q_1, qui comporte 11 compétences. Suite à cette simplification, deux items ne sont plus reliés à aucune compétence et sont donc supprimés (N = 58).
- Dans le même souci de diminution de la taille de la matrice, la suppression des compétences reliées à moins de 18 items dans Q_0 donne la matrice Q_2, qui comporte 6 compétences (N = 58).
- La matrice Q_3 est obtenue à partir de Q_0 en veillant à toujours garder les compétences classées comme premier choix pour chaque item. Q_3 comporte alors 9 compétences et 58 items (N = 58).

– La matrice Q_4 est créée de la manière suivante : de la liste des compétences présentes dans la matrice Q_0, on retient les 6 compétences qui semblent présenter un intérêt diagnostique particulier aux didacticiens de notre équipe. Deux items sont alors éliminés (N = 58).
– Enfin la matrice notée Q_5 ne contient que les trois compétences (1) d'explicitation, (2) d'évaluation et (3) d'intervention. Elle est obtenue à partir des 15 compétences incluses dans Q_0 en établissant un lien dès qu'au moins une de ces compétences est classée EX, EV ou IN. Les 60 items sont conservés (N = 60).

Tableau 1.1
Récapitulatif des matrices Q utilisées

Matrice	Nombre d'attributs	Nombre d'items
Q_0	15	60
Q_1	11	60
Q_2	6	58
Q_3	9 (1^{er} choix)	58
Q_4	6 (choisis)	58
Q_5	3 (EX, EV, IN)	60

4. RÉSULTATS

Les données sont modélisées pour chacune des matrices Q du tableau 1.1 avec le modèle DINA en utilisant un algorithme programmé dans Ox (Doornik, 2002). Les résultats relatifs aux paramètres d'items sont présentés aux tableaux 1.2 et 1.3 relativement aux matrices Q_1, Q_2, Q_3, Q_4 et Q_5, aucune convergence n'ayant pu être obtenue pour la matrice Q_0. Le tableau 1.2 fournit les valeurs minimale, maximale et moyenne des paramètres g_j et s_j ainsi que de l'indicateur d_j. Le tableau 1.3 contient les valeurs de l'indicateur d'ajustement global $\bar{\hat{g}} + \bar{\hat{s}}$.

Tableau 1.2
Estimation des valeurs moyennes des paramètres s_j et g_j de l'indicateur δ_j

	Guessing g_j	***Slipping*** s_j	δ_j
Q1			
Minimum	0,0000	0,0000	0,0422
Maximum	0,9246	0,6404	0,9221
Moyenne	0,4762	0,2019	0,3219
Q_2			
Minimum	0,1381	0,0003	0,0443
Maximum	0,9161	0,6234	0,5728
Moyenne	0,5003	0,2113	0,2884
Q_3			
Minimum	0,1438	0,0042	0,0455
Maximum	0,9256	0,6180	0,5955
Moyenne	0,4990	0,1973	0,3037
Q_4			
Minimum	0,1825	0,0081	0,0543
Maximum	0,9209	0,6496	0,5269
Moyenne	0,5015	0,2130	0,2855
Q_5			
Minimum	0,0453	0,0074	0,0622
Maximum	0,9054	0,6512	0,6453
Moyenne	0,5081	0,2318	0,2600

Tableau 1.3
Estimation des valeurs $\bar{g} + \bar{s}$ pour chaque matrice Q

	Q_1	Q_2	Q_3	Q_4	Q_5
$\bar{g} + \bar{s}$	0,6781	0,7176	0,6963	0,7214	0,7400

5. INTERPRÉTATION

L'objectif de cette étude est de procéder à la validation du classement des items de l'épreuve diagnostique de 2008 selon les compétences. Les matrices Q utilisées pour les modélisations sont basées sur les compétences en lien avec la bonne réponse de chaque item et contiennent de 3 à 11 compétences. Le tableau 1.2 illustre le fait que peu importe la matrice choisie, les paramètres g_j du modèle DINA ont une moyenne d'environ 50 %. Cela signifie que les sujets qui ne possèdent pas les compétences visées par les items selon la matrice Q ont en moyenne une chance sur deux de répondre correctement aux items. Il existe

même un item pour lequel le paramètre g_j a une valeur supérieure à 90 % pour chacune des matrices Q. Pour cet item, il y a donc plus de 90 % de chances de répondre correctement même si on ne possède pas les compétences requises. En outre, le paramètre g_j de cinq items est supérieur à 80 %, indépendamment de la matrice utilisée. Ainsi, les paramètres g_j indiquent que les compétences reliées aux items du test 2008 ne permettent pas de les caractériser suffisamment, puisque la probabilité de répondre correctement alors qu'on ne possède pas les compétences spécifiques est grande.

Les paramètres liés au fait de ne pas répondre correctement alors que l'on possède les compétences (*slipping*) ont une valeur moyenne d'environ 20 % pour chacune des cinq matrices (voir le tableau 1.2). Cela signifie qu'en moyenne, environ 80 % des sujets qui possèdent les compétences requises répondent correctement. D'après ce paramètre, la classification semble donc acceptable.

L'indicateur de discrimination d_j a une moyenne d'environ 30 %. Étant donné que cet indicateur correspond à la différence entre la probabilité de répondre correctement à l'item selon que l'on possède $(1 - s_j)$ ou pas (g_j) les compétences spécifiques, une valeur de 30 % indique que le fait de maîtriser les compétences ne donne guère de chances supplémentaires de mieux répondre. Ainsi, en moyenne, les compétences spécifiées dans les cinq matrices Q ne sont pas vraiment discriminantes pour les items. Enfin, l'indicateur général d'ajustement $\bar{\hat{g}} + \bar{\hat{s}}$ a une valeur d'environ 70 % pour chacune des cinq matrices utilisées, ce qui dénote un ajustement médiocre.

6. DISCUSSION, LIMITES ET CONCLUSION

Les résultats obtenus pour chacun des paramètres sont tout à fait semblables pour l'ensemble des matrices, ce qui montre que quelle que soit la manière de spécifier les liens entre les items et les compétences, le diagnostic obtenu à partir du classement des items relativement aux compétences est peu fiable. Ces résultats peuvent être comparés à ceux d'une étude antérieure (Loye, 2008), dans laquelle les données de plusieurs épreuves de mathématiques de l'École Polytechnique de Montréal ont été modélisées à l'aide d'un modèle cognitif diagnostique. L'objectif consistait à élaborer plusieurs matrices Q pour chaque test en recourant à des experts en mathématiques auxquels on fournissait différentes informations. En utilisant celles-ci (par exemple le degré de difficulté de chaque item), les experts devaient identifier les processus cognitifs liés à chaque item, pour chacun de ces tests. L'application d'un modèle cognitif diagnostique permettait ensuite de

juger de la qualité des différentes matrices Q obtenues et d'en arriver à des recommandations sur la construction de matrices Q offrant un diagnostic juste et fiable des sujets. Les résultats ont montré que l'ajout d'informations au libellé des items n'aidait pas les experts à construire de meilleures matrices Q, voire que ces informations étaient parfois nuisibles. Ils ont également montré que, pour l'ensemble des matrices, le diagnostic posé était d'assez piètre qualité.

Les recherches récentes sur les modèles cognitifs diagnostiques indiquent que les tests élaborés dans une approche classique (souvent basée sur la théorie des réponses aux items) ne permettent pas d'obtenir un diagnostic valide des sujets relativement aux processus cognitifs. Ainsi, les résultats des évaluations à grande échelle (par exemple le Programme international pour le suivi des acquis des élèves [PISA] de l'OCDE) ne devraient pas être utilisés dans une visée diagnostique (voir, par exemple Gierl, Alves et Roberts, 2009 ; Willse, 2009). Notre étude montre que des problèmes semblables se posent avec des tests élaborés par des enseignants, même si l'intention est d'établir un diagnostic comme c'est le cas avec l'épreuve de l'École Polytechnique. Ainsi, nos résultats vont dans le même sens que les suggestions de Gorin (2006) ou de Leighton et Gierl (2007) quant à la nécessité de définir un nouveau cadre de référence pour élaborer des tests avant de pouvoir prétendre à la réalisation d'un diagnostic cognitif.

Toutefois, on peut se demander pourquoi ce que les Américains appellent le *retrofitting* (autrement dit, l'utilisation de données *a posteriori*) ne fonctionne pas. Plusieurs hypothèses peuvent être proposées dans le cadre de notre étude :

1. Le fait que les compétences utilisées sont probablement fortement corrélées peut affecter l'estimation des paramètres du modèle statistique.
2. L'existence potentielle de compétences non ciblées dans la matrice Q et qui jouent un rôle dans le processus de réponse peut également être une source de difficultés. Ainsi, chaque item nécessite des compétences très variées et trop peu d'entre elles sont captées par la matrice Q.
3. L'analyse didactique a également mis en évidence le fait que de nombreux items de la banque de données de l'épreuve de mathématiques de l'École Polytechnique peuvent être résolus de plusieurs manières. Or la matrice Q suppose une seule et même combinaison de compétences pour chaque item.
4. Il est possible que l'existence d'une caractéristique commune et inhérente à tous les items dilue le pouvoir diagnostique des processus (ou compétences).

5. Le format à choix multiple des items peut également influencer la manière de répondre des sujets, en permettant par exemple de procéder par élimination.
6. Les enjeux associés à la passation de l'épreuve de mathématiques pour les étudiants de l'École Polytechnique sont peu importants. Étant donné que ce test ne compte pas pour l'admission (un bonus dans leur premier cours de mathématique ne suffisant probablement pas à les motiver), les étudiants peuvent ne pas prendre le test au sérieux. Dans ce cas, les processus qui sont pris en considération dans la matrice Q ne sont peut-être pas appropriés parce que le hasard joue un rôle important dans les choix de réponses des étudiants.
7. Enfin, de façon plus fondamentale, il convient de rappeler que les compétences ne s'exercent pas à vide, mais reposent notamment sur des contenus. Ainsi le caractère générique des éléments de compétence retenus, s'il permet de rendre compte dans une certaine mesure de compétences mathématiques qui traversent les domaines de contenu (par exemple la maîtrise du langage mathématique), fait néanmoins en sorte que la compétence peut paraître maîtrisée pour un certain contenu et non maîtrisée pour un autre.

Nos résultats laissent clairement entrevoir qu'une épreuve qui combine librement, dans des tâches relativement complexes, des éléments variés de contenus et de compétences peut difficilement être associée *a posteriori* à une matrice Q en vue de fournir des indications sur les processus cognitifs maîtrisés ou non par les sujets. Deux pistes s'ouvrent maintenant à nous.

1. On peut chercher à développer de nouvelles méthodologies pour traiter conjointement compétences et contenus, en tirant notamment parti de l'information contenue dans le choix des leurres, potentiellement riche d'un point de vue diagnostique.
2. On peut également s'orienter vers la création de tests d'une nouvelle génération dont le but ultime serait de réaliser un diagnostic cognitif. Cela pourrait se faire en partant de tâches relativement simples et portant sur un ensemble limité de contenus, qui pourraient néanmoins recouvrir quelques domaines mathématiques. On pourrait ensuite combiner ces tâches simples afin de développer des tâches de plus en plus complexes. L'expansion du champ mathématique couvert par l'épreuve pourrait éventuellement se faire en ajoutant une dimension adaptative.

Cette deuxième voie paraît plus prometteuse à long terme.

RÉFÉRENCES

Caron, F. (2001). *Effets de la formation fondamentale sur les compétences d'étudiants universitaires dans la résolution de problèmes de mathématiques appliquées*. Montréal, Québec : Université de Montréal, Faculté des sciences de l'éducation.

De la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model : development and applications. *Journal of educational measurement, 45*(4), 343-362.

De la Torre, J. et Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353.

De Serres, M. et Groleau, J.-D. (1997). *Mathématiques et langages*. Montréal, Québec : Collège Jean-de-Brébeuf.

De Terssac, G. (1996). Savoirs, compétences et travail. Dans J.-M. Barbier (dir.), *Savoirs théoriques et savoirs d'action*. Paris, France : Presses universitaires de France.

DiBello, L. V., Stout, W. F. et Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. Dans P. D. Nichols, S. F. Chipman et R. L. Brennan (dir.), *Cognitively diagnostic assessment*. Hillsdale, New Jersey : Erlbaum.

Doornik, J. A. (2002). Object-oriented matrix programming using Ox (version 3.1). [Logiciel informatique]. Londres, Royaume-Uni : Timberlake Consultants Press.

Gierl, M. J., Alves, C., Roberts, M. et Gotzmann, A. (2009). *Using judgments from content specialists to develop cognitive models for diagnostic assessments*. Communication présentée au colloque durant l'assemblée annuelle (2009) du National Council on Measurement in Education, San Diego, Californie.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational measurement, issues and practices, 25*(4), 21-35.

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

Leighton, J. P. et Gierl, M. J. (dir.) (2007). *Cognitive diagnostic assessment for education : theory and applications*. Cambridge, United Kingdom : Cambridge University Press.

Loye, N. (2008). *Conditions d'élaboration de la matrice Q des modèles cognitifs et impact sur sa validité et sa fidélité*. Thèse de doctorat inédite, Université d'Ottawa, Ottawa, Ontario.

Macready, G. B. et Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of educational statistics, 2*(2), 99-120.

Ministère de l'Éducation, du Loisir et du Sport (MELS, 2006). *Échelles des niveaux de compétence. Enseignement secondaire, premier cycle*. Québec, Québec : Gouvernement du Québec.

Nichols, P. D., Chipman, S. F. et Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, New Jersey : Erlbaum.

Polya, G. (1945). *How to solve it*. Princeton, New Jersey : Princeton University Press.

Roussos, L., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A. et Templin, J. L. (2007). The fusion model skills diagnosis system. Dans J. P. Leightone et M. J. Gierl (dir.), *Cognitive diagnostic assessment for education : theory and applications*. New York, New York : Cambridge University Press.

Rupp, A. A. et Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and psychological measurement, 68*(1), 78-96.

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, Floride : Academic Press.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. Dans N. Frederiksen, R. Glaser, A. Lesgold et M. G. Shafto (dir.), *Diagnostic monitoring of skill and knowledge acquisition*. Hillsdale, New Jersey : Erlbaum.

Tatsuoka, K. K. (1983). Rule-space : an approach for dealing with misconceptions based on item response theory. *Journal of educational measurement, 20*(4), 345-354.

Tatsuoka, K. K. (2005). *Rule space method : cognitively diagnostical tool*. Communication présentée au National Council for Measurement in Education Training Program, Montréal, Québec.

Willse, J. (2009). *Retrofitting cognitive diagnostic models to large scale tests : problems with dimensionality*. Communication présentée au National Council for Measurement in Education, San Diego, Californie.

Annexe 1

Éléments de compétence dans la résolution d'un problème de mathématiques

0. Contrôler le passage d'une étape à l'autre

☐ 0.01 Démarrer (EV/EX)
☐ 0.02 Articuler les résultats des différents sous-problèmes et/ou cas possibles (IN/EX)
☐ 0.03 Détecter un raisonnement circulaire et l'éliminer de la solution (EV)
☐ 0.04 Reconnaître une impasse ou une invraisemblance (EV)
☐ 0.05 Identifier la source de l'impasse ou de l'invraisemblance, et y revenir (EV)
☐ 0.06 Anticiper la suite (EV)
☐ 0.07 Reconnaître l'atteinte de l'objectif ou d'un sous-objectif (EV)

1. Analyser le problème

☐ 1.01 Identifier/interpréter l'objectif à partir de l'énoncé (EX)
A : Langage naturel B : Symbolique C : Graphique
☐ 1.02 Identifier/interpréter les données, les hypothèses à partir de l'énoncé (EX)
A : Langage naturel B : Symbolique C : Graphique
☐ 1.03 Découper le système en en gardant l'essentiel (EV/EX)
☐ 1.04 Représenter à l'aide d'un graphique/diagramme/schéma (EX)
☐ 1.05 Identifier les principes, lois et/ou formules qui s'appliquent (EV/EX)
☐ 1.06 Identifier les variables (EX)
☐ 1.07 Identifier les paramètres (EX)
☐ 1.08 Identifier le ou les ***objets*** mathématiques sous-jacents (EV) : _____
_____
☐ 1.09 Reconnaître un problème bien défini (EV)
☐ 1.10 Traiter un problème mal défini (simplifications, ajout d'hypothèses) (EV/EX)
☐ 1.11 Mettre en équations (EX)

2. Élaborer le plan de résolution

☐ 2.01 Raisonner à partir d'un graphique/diagramme/schéma (EX/EV)
☐ 2.02 Raisonner à partir de cas particuliers (EV/IN)
☐ 2.03 Raisonner à partir d'un problème similaire (EV/IN)
☐ 2.04 Raisonner à partir d'une ***propriété d'un objet*** sous-jacent au problème (EV/IN) : _____
☐ 2.05 Explorer/expérimenter à l'aide de l'ordinateur (EV/IN/EX)
☐ 2.06 Identifier les ***méthodes de résolution*** applicables (EV)
☐ 2.07 Choisir la ou les ***méthodes de résolution*** en fonction de critères (précision, coût, etc.) (EV)
☐ 2.08 Structurer la ***résolution*** en décomposant en sous-problèmes (EV/IN)
☐ 2.09 Identifier les cas possibles (niveau macro) (EV)
☐ 2.10 Utiliser les règles d'inférence (niveau macro) (IN/EV)

3. Exécuter le plan

☐ 3.01 Identifier et démontrer une nouvelle propriété (EV/IN)
☐ 3.02 Utiliser une ***propriété*** associée à un ***objet*** mathématique (niveau micro) (IN/EV) : _____
☐ 3.03 Identifier les cas possibles (niveau micro) (EV/IN)
☐ 3.04 Utiliser les règles d'inférence (niveau micro) (IN/EV)
☐ 3.05 Effectuer des manipulations algébriques/analytiques (IN) : _____
_____
☐ 3.06 Écrire un algorithme (IN/EX)
☐ 3.07 Utiliser un algorithme (IN) : _____
☐ 3.08 Programmer (IN/EX)
☐ 3.09 Utiliser une fonction logicielle (IN/EX) : _____
☐ 3.10 Estimer un résultat (EV/IN)
☐ 3.11 Calculer un résultat avec précision (IN/EV)
☐ 3.12 Vérifier localement les résultats intermédiaires (EV/IN)
☐ 3.13 Appliquer une technique de calcul

4. Revenir sur la solution

☐ 4.01 Vérifier les propriétés générales (unités, ordre de grandeur, invariants, etc.) (EV)
☐ 4.02 Valider avec des cas particuliers (EV/IN)
☐ 4.03 Valider à partir d'un problème similaire (EV)
☐ 4.04 Expliquer les anomalies (EX/EV)
☐ 4.05 Généraliser les résultats obtenus (EV/IN)
☐ 4.06 Interpréter les résultats (EX/EV)

Annexe 2

Éléments de contenus visés par l'épreuve diagnostique

A000 Algèbre et fonctions
Manipulations arithm./algébriques
A101 Nombre décimal
A102 Nombre rationnel
A103 Nombre irrationnel
A104 Monôme
A105 Polynôme
A107 Expression rationnelle
A108 Addition/soustraction (et propriétés)
A109 Multiplication/division (et propriétés)
A110 Solution d'une équation
A111 Nombre « *e* »
A112 Logarithme
A113 Puissances (ou exposants)
A114 Factorisation
A115 Identités remarquables
Relations, fonctions
A201 Fonction
A202 Domaine de définition
A203 Codomaine (ensemble-image)
A204 Coordonnées cartésiennes
A205 Coordonnées à l'origine
A206 Croissance/décroissance
A207 Extremum (minimum, maximum)
A208 Composée de fonctions
A209 Égalité ou équation
A210 Inégalité ou inéquation
A211 Produit de fonctions
A212 Quotient de fonctions
A213 Racine d'une fonction
A216 Valeur d'une fonction en un point
A217 Distance
A218 Pente
Relations, fonctions particulières
A301 Fonction linéaire (droite)
A302 Fonction quadratique
A303 Fonction polynomiale (degré>2)
A304 Fonction rationnelle
A305 Fonction exponentielle
A306 Fonction logarithmique
A307 Fonction « racine »
A308 Fonction trigo 1 (sin, cos, tan)
A309 Fonction trigo 2 (sec, cosec, cotan)
A310 Fonction trig. inverses (arc sin, …)
A311 Fonction valeur absolue
A312 Conique
A314 Fonction « racine cubique »
T000 Trigonométrie
T101 Rapport trigo 1 (sin, cos, tan)
T102 Rapport trigo 2 (sec, cosec, cotan)
T103 Cercle trigonométrique
T104 Radian
T105 Identité trigonométrique
T106 Règle des sinus
T107 Règle des cosinus
T108 Sin (ou cos) d'une somme

G000 Géométrie
Figures/éléments géométriques
G101 Droites/segments remarquables
G104 Triangle
G105 Quadrilatère
G106 Polygone à plus de 4 côtés
G107 Cercle
Concepts métriques
G201 Périmètre/longueur/distance
G202 Aire
G203 Volume
G204 Angle
Relations spatiales
G301 Parallélisme
G302 Perpendicularité
G303 Translation
G304 Rotation
G305 Réflexion/symétrie
G306 Tangente
G307 Homothétie
Relations métriques
G401 Isométrie/congruence
G402 Similitude/proportionnalité
G403 Relations entre des angles
G404 Théorème de Pythagore
M000 Matrices
M101 Matrice
M102 Transposée
M103 Somme de matrices
M104 Produit de matrices
M105 Déterminant d'une matrice
M106 Solution d'un système $Ax = b$
M107 Inverse d'une matrice
M108 Rang d'une matrice
M109 Matrice identité
M110 Système d'équations
V000 Vecteurs
V101 Addition vectorielle
V102 Coordonnées polaires
V103 Norme vectorielle
V104 Produit scalaire
V105 Produit vectoriel
V106 Vecteurs
D000 Calcul différentiel
D101 Continuité
D102 Limite
D103 Infini
D104 Convergence (suite ? série ?)
D105 Dérivée
D106 Point critique
D107 Série de Taylor
D108 Règle de L'Hôpital
N000 Calcul intégral
N102 Intégration par fractions partielles
N103 Intégration par parties
N104 Intégration par substitution
N105 Primitive
N106 Intégrale

Chapitre 2

Utilisation du degré de certitude et du degré de réalisme dans un contexte d'évaluation diagnostique

Serge Boulé et Dany Laveault

Dans un contexte d'évaluation en appui à l'apprentissage, il peut être utile pour un étudiant de recevoir un feedback sur le réalisme de l'autoévaluation de ses réponses. Si le degré de réalisme varie en fonction des propriétés des items et des différences individuelles, son utilité dans ce contexte peut être réduite. Notre recherche a permis d'observer que, plus les étudiants sont performants, plus ils sont réalistes, et ce, quel que soit le genre. Également, plus l'item est facile, plus les étudiants sont réalistes. Les liens entre les niveaux taxonomiques des items, leur discrimination et le réalisme ne sont pas clairs et méritent d'être étudiés davantage.

1. INTRODUCTION

Cette étude s'inscrit dans le cadre d'une recherche appliquée qui porte sur l'utilisation du degré de certitude en docimologie. Lors de la passation de tests, on peut demander aux individus d'exprimer leur certitude quant à la qualité de leurs réponses. Il s'agit en quelque sorte d'ajouter une modalité de réponse à chaque item permettant ainsi à l'individu d'indiquer son degré de certitude pour chacune de ses réponses.

Henmon (1911) introduit le concept de certitude des étudiants envers leur réponse. Le degré de certitude d'un étudiant que les réponses qu'il donne sont correctes est un élément qui peut être ajouté au processus d'évaluation et qui tient compte des degrés variables de connaissance. Leclercq et Poumay (2004, p. 3) définissent les degrés de certitude comme [l]*'ensemble des jugements, des analyses, des régulations, conscientes ou non (mais qu'il importe de rendre explicites, observables et conscientes) effectués par l'apprenant sur ses propres performances (processus ou produits), dans des situations de PRÉ, PER ou POST performance*.

Il existe plusieurs façons d'exprimer le degré de certitude. Reach, Zerrouki, Leclercq et D'Ivernois (2005) proposent une échelle de probabilités d'exactitude qui pourrait varier par échelons de *pas sûr du tout* à *absolument sûr*. En fait, l'étudiant estime correctement ou non la qualité de ses réponses (bonne ou non), ce qui est adapté à l'évaluation des apprentissages.

Il y a des erreurs plus graves que d'autres, surtout celles pour lesquelles nous sommes sûrs de ne pas nous tromper. Il serait utile de savoir si un étudiant se doute qu'il se trompe lorsqu'il donne une mauvaise réponse à un item, surtout dans le cas où la mauvaise réponse pourrait avoir des conséquences en situation réelle. Hunt (1993) distingue d'ailleurs trois états de la connaissance : connaissance, ignorance et méconnaissance. C'est cette dernière catégorie qui est particulièrement critique pour toute régulation des apprentissages.

En comparant la certitude exprimée et la qualité des réponses, nous pouvons établir si l'individu est réaliste lorsqu'il juge de la qualité de ses réponses. La mesure du réalisme proposée s'obtient en vérifiant la concordance entre le degré de certitude exprimé et la qualité de la réponse obtenue. L'étudiant réaliste tend à anticiper une bonne réponse lorsque celle-ci est bonne et une réponse incorrecte lorsque celle-ci est incorrecte. L'utilisation du degré de certitude peut alors avoir une fonction formative à partir du moment où l'on donne un feedback à l'individu sur les éléments qui ont incité ce dernier à manquer de réalisme, ce que Hunt (1993) appelle la méconnaissance.

La certitude, et par conséquent le réalisme, sont-ils des traits ou des caractéristiques dépendant du contexte ? En contextes formatif et diagnostique, il est important de savoir si le réalisme de l'étudiant est indépendant du type de question posée et si des facteurs personnels l'influencent. Nous devons comprendre l'effet de ces facteurs pour que l'utilisation du degré de certitude soit exempte de biais et soit valide dans de tels contextes.

2. ÉTAT DE LA QUESTION

Il y a des situations de connaissance complète ou discrète, comme dans le cas où seule une réponse exacte serait possible. Ce sont principalement des questions de connaissance selon la taxonomie des objectifs cognitifs de Bloom. Il y a d'autres situations où l'on peut faire une déduction ou une approximation, par exemple lorsqu'il faut estimer la distance entre Vancouver et Halifax. Il s'agit alors d'une connaissance partielle où la réponse se situe sur une échelle continue. Prenons l'exemple d'un item qui porterait sur la capitale de l'Ontario. Les choix de réponses pourraient comprendre : Toronto, Ottawa, Montréal et Paris. L'étudiant qui répondrait Paris serait plus loin – au sens propre et figuré – de la bonne réponse que celui qui répondrait Ottawa, ce qui est une erreur commune et attendue des élèves ontariens. De plus, l'erreur serait encore nettement plus grave s'il choisissait Paris en indiquant un degré de certitude élevé.

Les recherches de Leclercq (1993) sur le degré de certitude ont ouvert de nouvelles perspectives en ce domaine puisqu'elles ajoutent une dimension formative. Cette dernière consiste en un feedback sous la forme d'un profil de réalisme. Leclercq postule que l'indice de réalisme obtenu fournit un feedback à l'étudiant et lui permet d'améliorer son estimation de ses compétences. En d'autres mots, transmettre à l'étudiant de l'information sur son degré de réalisme pourrait être formateur.

Henmon (1911) ainsi que Jonsson et Allwood (2003) ont montré que plusieurs variables affectent le degré de réalisme et qu'il est difficile de comprendre leur fonctionnement. S'il était démontré que le degré de réalisme pouvait varier en fonction des groupes d'individus, la fidélité et la validité des interprétations du réalisme des élèves par les enseignants en seraient considérablement affectées. Par conséquent, la qualité du feedback fourni à l'élève en souffrirait.

Il est également probable que les caractéristiques des items du test peuvent avoir une influence sur les degrés de certitude exprimés. Par exemple, un item ayant une amorce ou des leurres défectueux pourrait fausser le jugement d'un individu sur la qualité de sa réponse. Un item mal formulé pourrait ainsi inciter l'individu à douter de sa performance et par conséquent affecter le réalisme de son jugement par rapport à la qualité de sa réponse pour cet item en particulier.

En évaluation diagnostique, l'interprétation des résultats à un test pourrait varier en fonction des exigences cognitives de la tâche, et l'interprétation pourrait être différente selon la propension d'un individu à surestimer la qualité de ses réponses. Par exemple, un individu

pourrait bénéficier d'une rétroaction qui favoriserait une remédiation au niveau individuel. Si par contre il y a une surestimation alors que l'individu peut être considéré réaliste en général, il faut alors vérifier, lorsque c'est possible, si les membres du groupe ou de la classe sont réalistes par rapport à la même tâche. Si tel est le cas, il faut favoriser la régulation individuelle au plan de l'apprentissage. Dans le cas où tout le groupe manquerait de réalisme par rapport à une tâche, le manque de réalisme de l'individu pourrait bien être attribuable à des exigences cognitives démesurées ou à une ambiguïté de l'item. Dans le cas où tout le groupe se surestime, on doit favoriser la régulation collective et viser une régulation de l'apprentissage ou du degré de réalisme.

3. RECENSION DES ÉCRITS SUR LE SUJET

3.1. Effet selon le sexe

Il est probable que le degré de certitude exprimé par le répondant est lié à des caractéristiques personnelles comme le sexe. Swineford (1938) a montré que les garçons ont tendance à prendre plus de risques que les filles. Les garçons surestimeraient légèrement plus la probabilité de l'exactitude de leurs choix (Beyer, 2002 ; Pallier, 2003 ; Sieber, 1974 ; Stankov, 1998). Selon Beyer, ce phénomène serait particulièrement fréquent chez les étudiants ayant des attentes faibles et pourrait interagir avec le niveau de performance de l'étudiant.

Un rapport de l'Office de la qualité et de la responsabilité en éducation de l'Ontario montre qu'il y a un lien entre le rendement, la perception que les étudiants ont de leur niveau de réussite et la valorisation de la réussite dans ces matières (Office de la qualité et de la responsabilité en éducation, 1999). D'après ce rapport, les filles se disent plus souvent bonnes en lecture et en écriture que les garçons, qui, pour leur part, s'estiment plus souvent que les filles bons en mathématiques. À cause de l'impact possible du genre sur le degré de réalisme des étudiants, nous croyons important de tenir compte de ce facteur et d'examiner son effet sur le degré de certitude de manière à éviter de confondre certains effets lors des analyses statistiques.

3.2. Effet du niveau d'habileté

Fabre (1980) est d'avis que la certitude des étudiants dépend de leurs connaissances. Lorsque le niveau de performance est bas, l'étudiant peut apprécier sa performance incorrectement sur la foi d'indices non pertinents. Nous ne connaissons cependant pas la nature du lien entre

le niveau d'habileté et le degré de réalisme. Les étudiants peu performants dans une matière pourraient très bien exceller dans une autre. Ils pourraient être en mesure d'avouer leur ignorance et d'être réalistes quant à l'estimation de la qualité de leurs réponses dans un domaine, mais pas dans un autre.

3.3. Effet des propriétés des items

Pintrich (2002) décrit la connaissance de soi comme l'un des trois types de connaissance métacognitive. L'habileté à évaluer sa propre connaissance est un aspect de la connaissance de soi. De plus, le réalisme de l'étudiant entrerait dans la taxonomie révisée d'Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths et Wittrock (2001), sous la rubrique des dimensions de la connaissance qui correspondent à la connaissance métacognitive. Sur le plan cognitif, le degré de certitude exprimé pourrait être relié au niveau taxonomique des items.

La difficulté et la discrimination de l'item peuvent aussi avoir un rôle à jouer. Jacobs (1974) a montré que l'expression du degré de certitude pouvait varier en fonction de la difficulté de l'item. De Finetti (1965) a également souligné que les items difficiles entraînent un certain degré d'incertitude. Enfin, un item défectueux qui discriminerait peu à cause d'une ambiguïté dans sa formulation pourrait créer de l'incertitude. Nous croyons que dans l'interprétation des résultats sur le degré de réalisme des sujets, il faut prendre en considération de telles situations.

L'ensemble des recherches précédentes ne porte pas spécifiquement sur la variation du degré de certitude causée par les propriétés des items. Cependant, il semble y avoir suffisamment de raisons pour justifier la prise en considération de ces variables. Afin de mieux connaître les facteurs qui affectent le manque de réalisme chez les étudiants, il nous paraît important de déterminer si le degré de réalisme de l'étudiant est indépendant du test et quels sont les facteurs qui l'influencent.

À la lumière de ce qui précède, deux questions se posent. La surestimation des habiletés varie en fonction de différences individuelles comme le genre ou le niveau d'habileté. Comment et dans quelle mesure ces deux facteurs sont-ils liés à l'expression du degré de certitude et, par conséquent, au degré de réalisme de l'étudiant ? Un item qui serait difficile, ambigu ou qui s'adresserait à des habiletés cognitives supérieures créerait de l'incertitude. Comment et dans

quelle mesure l'expression du degré de certitude et, par conséquent, du réalisme des étudiants varie-t-elle en fonction du niveau taxonomique des items et de leurs propriétés métriques ?

4. MÉTHODE

4.1. Participants

L'échantillon de circonstance étudié est formé de 252 participants inscrits à une école d'éducation des adultes de l'Ontario. Ceux-ci se sont présentés à un test de classement en mathématiques. Après l'élimination des copies incomplètes, 152 copies sont retenues. Nous constatons que dans plusieurs cas, des données sont manquantes en raison de l'abandon ou parce que l'étudiant n'a pas répondu ou n'a pas exprimé son degré de certitude. Puisqu'il s'agissait d'une évaluation diagnostique plutôt que sommative, les enjeux perçus par les participants étaient relativement peu importants. Pour les besoins de cette recherche, nous ne pouvons pas traiter les données manquantes par imputation sans postuler des patrons de réponses et d'expressions du degré de certitude.

Nous avons retenu 83 copies chez les femmes et 69 copies chez les hommes. Puisque les scores totaux se distribuent normalement, les niveaux de performance ont été catégorisés en utilisant 27 % des scores les plus faibles et 27 % des scores supérieurs pour établir trois catégories : les niveaux faible, moyen et fort. Ces valeurs limites ont été choisies puisqu'elles correspondent aux deux points d'inflexion de la courbe normale. Nous pourrons ainsi décrire plus facilement notre échantillon.

4.2. Instruments

L'instrument de mesure a été développé en fonction des exigences en mathématiques du curriculum du secondaire du ministère de l'Éducation de l'Ontario, tout en respectant les bases méthodologiques utilisées en docimologie, telles que les ont stipulées Laveault et Grégoire (2002). Le test de type questionnaire à choix multiple (QCM) comprend 40 items avec quatre choix de réponses chacun. L'instrument de mesure a subi deux cycles de mise à l'essai. La cohérence interne du test est très bonne ($\alpha = 0{,}91$). Des renseignements supplémentaires sur les propriétés métriques de l'instrument sont données par Boulé (2007). En plus des réponses aux items, le participant doit exprimer son degré de certitude par rapport à ses réponses. Les items du test varient en fonction des niveaux taxonomiques. Nous retenons

les quatre premières catégories de la taxonomie des objectifs cognitifs de Bloom : 1) connaître, 2) comprendre, 3) appliquer, 4) analyser (Anderson et collab., 2001). Les items ne sont pas des cas typiques de chaque niveau taxonomique et ils n'ont pas été rédigés de manière à assurer une représentativité optimale des différents niveaux. Nous constatons qu'il est difficile d'en arriver à un consensus sur l'appartenance des items à un niveau taxonomique particulier et que certains items sont des cas limites entre deux niveaux ou encore font intervenir plus d'un niveau.

4.3. Procédure

Le test est administré dans un centre d'évaluation surveillé. Les tests ont tous été administrés au même endroit durant une période de quelques mois. Les risques de diffusion étaient minimes puisqu'il s'agit d'un test diagnostique pour des fins de classement et d'aiguillage. Il n'y a donc ni conséquences irrémédiables ni intérêt à tricher.

4.4. Considérations éthiques

Le projet de recherche a été soumis au Comité d'éthique à la recherche de l'Université d'Ottawa. Après évaluation, le comité a déterminé que cette recherche se situait sous le seuil minimal de risque pour les participants et que l'anonymat des participants était garanti. Le Comité d'éthique à la recherche de l'Université d'Ottawa a donc donné son aval.

4.5. Méthode d'analyse des résultats

Des analyses de types univarié et multivarié sont utilisées. Le but recherché est de détecter la présence ou l'absence de liens entre le degré de réalisme et les variables indépendantes liées aux caractéristiques des élèves ou aux propriétés métriques des items.

5. RÉSULTATS

Dans cette section, nous comparons les effets des niveaux de performance sur le degré de réalisme en fonction du sexe de manière à vérifier notre première hypothèse. La relation entre le coefficient de discrimination, le niveau de difficulté et le niveau taxonomique des items est aussi analysée afin de vérifier la seconde hypothèse. Pour des détails supplémentaires sur les méthodes d'analyse utilisées, on peut consulter l'étude de Boulé (2007).

Les données sont réparties en deux matrices. La première matrice ; *R*, est formée en fonction des réponses individuelles et permet les calculs liés au degré de réalisme des étudiants *Rs* selon Leclercq (1993) et à leur niveau de performance. Les étudiants sont aussi catégorisés selon le sexe. La deuxième matrice, *Q*, met en évidence les propriétés métriques des items, leur niveau taxonomique ainsi que le degré de réalisme associé à chaque item, ce que Gilles (2002) appelle l'indice de réalisation des prédictions *Rq*. Des calculs permettent d'obtenir des valeurs pour des variables construites telles que le degré de réalisme de l'étudiant, l'indice de réalisation des prédictions de l'item et le niveau de performance de l'étudiant.

5.1. Propriétés métriques des items

Le degré de difficulté moyen des items, calculé selon le pourcentage d'items réussis, est de 0,50. L'indice de discrimination moyen des items du test calculé selon la méthode de Kelley (1939) est de 0,58. Deux items ont retenu notre attention par leur niveau élevé de difficulté et leur très faible coefficient de discrimination. Nous remarquons qu'il s'agit d'items qui présentent une faible corrélation item/total et qui ne contribuent pas ou presque à la cohérence interne du test.

5.2. Effet des différences individuelles

Le tableau 2.1 présente les résultats d'une analyse de régression multiple où nous avons cherché à prédire le degré de réalisme à partir des variables indépendantes qui sont, en l'occurrence, le score total au test et le genre du sujet. Seul le score du sujet au test de mathématiques est retenu comme prédicteur significatif du réalisme pour obtenir un coefficient de détermination (R^2) significatif de 0,37 ($F = 87{,}22$; $p \leq 0{,}05$). Le score du sujet explique donc 37,2 % de la variabilité totale du réalisme exprimé par le sujet et est donc lié de façon significative à son degré de réalisme. La mesure d'association révèle un h^2 partiel de 0,62 qui constitue la corrélation maximale qu'il est possible d'obtenir entre le score et le degré de réalisme de l'étudiant. Le test de linéarité, pour sa part, indique un écart de la linéarité significatif ($F = 2{,}42$; $p \leq 0{,}05$) ce qui nous porte à conclure qu'il y a des contributions non linéaires dans les données recueillies.

Tableau 2.1
Modèle proposé par l'analyse de la régression pour les sujets

Modèle retenu	***R***	***R*²**	***F***	σ	**η² partiel**
Score du sujet	0,61	0,37	87,22	0,00	0,62

Modèle retenu et coefficients	**Coefficients bruts**		**Coefficients standardisés**		
	B	S_B	β	*T*	σ
Ordonnée à l'origine	5,07	0,89		5,70	0,00
Score du sujet	0,61	0,06	0,61	9,34	0,00

Nous avons effectué une analyse de la variance en plan factoriel de manière à savoir si le fait qu'il s'agit d'un garçon ou d'une fille peut avoir une importance dans certaines conditions. Le tableau 2.2 nous présente les résultats de l'analyse de la variance effectuée selon un plan factoriel à deux facteurs, incluant leur interaction. Les valeurs *F* pour le genre et la performance sont respectivement de 1,70 et 47,05. La valeur *F* pour le genre indique que la différence n'est pas significative. Le devis méthodologique ainsi que le nombre de sujets ne nous permettent pas d'avoir suffisamment de puissance avec 25,4 % de probabilité de rejeter l'hypothèse nulle lorsqu'elle est fausse. Il était donc peu probable d'en arriver à des résultats significatifs en fonction du genre des étudiants. Nous soupçonnons que la curvilinéarité du lien entre le niveau de performance des participants et leur degré de réalisme peut réduire la puissance de notre analyse pour ce qui est des effets du genre des étudiants et de l'interaction *rendement* × *sexe* sur le degré de réalisme. Dans le cas des niveaux de performance, la valeur *F* est élevée et nous pouvons conclure qu'il y a une différence significative. Nous notons un h^2 partiel assez élevé pour cette variable (39,7 %), ce qui indique qu'une bonne partie de la variance est expliquée par les différents niveaux de cette condition. Le R^2 global nous indique aussi que 41,4 % de la variance de l'indice de réalisme est expliquée par le modèle à deux critères de classification.

L'analyse de la variance en plan factoriel ne nous permet pas de rejeter l'hypothèse nulle selon laquelle il n'y a pas de différence significative entre les moyennes des hommes et des femmes. Par contre, nous rejetons l'hypothèse nulle concernant l'effet principal lié aux niveaux de performance de l'étudiant. Les résultats montrent que le degré de réalisme de l'étudiant varie en fonction de son niveau de performance. Par ailleurs, la valeur *F* pour l'interaction entre les deux conditions étudiées est de 1,14 et n'est pas significative. Comme nous l'avons expliqué antérieurement, il n'y aurait donc pas de preuve d'interaction

significative entre les deux variables indépendantes. Cela étant dit, la faible puissance (à 0,25) ne permet pas de conclure qu'il n'y a pas d'interaction significative.

Tableau 2.2
Sommaire de l'analyse de la variance

Variable dépendante : indice du réalisme							
Source	*SC*	*Dl*	*CM*	*F*	σ	η^2 **partiel**	**Puissance**
Sexe	16,51	1	16,51	1,70	0,19	0,01	0,25
Performance	913,19	2	456,59	47,05	0,00	0,40	1,00
Sexe × performance	22,09	2	11,05	1,14	0,32	0,02	0,25
Erreur	1 387,80	143	9,70				
Total	27 549,00	149					
$\alpha = 0,05$	$R^2 = 0,41$						

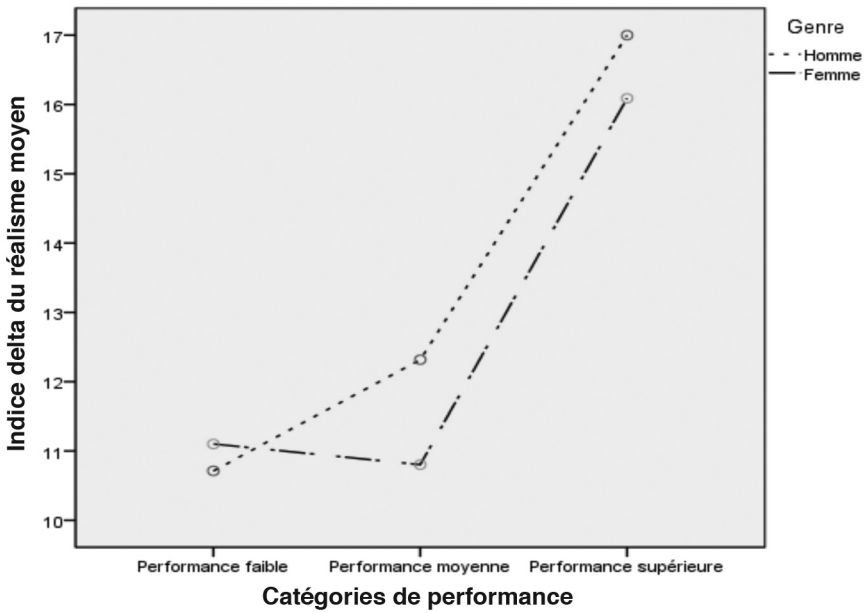


Figure 2.1
Degré de réalisme moyen des sujets selon le niveau de performance et le genre

La figure 2.1 nous permet de mieux visualiser la relation qui existe entre le degré de réalisme et le score. Elle permet de constater que plus les étudiants sont performants, plus ils ont un score de degré de réalisme élevé, et ce, quel que soit le sexe. Le lien entre le degré de

réalisme des étudiants et leur niveau de performance semble de prime abord se faire de la même façon chez les hommes que chez les femmes. Les deux fonctions indiquent des valeurs minimales de réalisme pour des scores δ (scores transformés de façon linéaire selon les paramètres : $\overline{x} = 13$; $\sigma = 4$) entre 10 et 12, soit pour les valeurs légèrement inférieures à la moyenne. Pour les valeurs très inférieures ou très supérieures à la moyenne, les valeurs du degré de réalisme ont tendance à être plus élevées, surtout pour les valeurs très au-dessus de la moyenne.

5.3. Effet des propriétés des items

Le modèle présenté au tableau 2.3 cherche à prédire l'indice de réalisation des prédictions de l'item (ou indice de réalisme par item) à partir du niveau de difficulté et du coefficient de discrimination de l'item. Ces deux variables indépendantes sont retenues dans l'équation de prédiction pour obtenir un coefficient de détermination (R^2) significatif de 0,65 ($F = 34{,}19$; $p \leq 0{,}05$). Ces variables expliquent environ les deux tiers de la variation du réalisme associée à un item en particulier. Des analyses présentées ultérieurement montreront que la contribution au modèle du coefficient de discrimination est difficile à expliquer et présente des faiblesses en raison de la distribution des données.

Comme le rapporte le tableau 2.3, le coefficient de discrimination contribue peu au R^2 (de 0,60 à 0,65) du modèle à deux variables indépendantes. La mesure d'association entre le niveau de difficulté et l'indice de réalisation des prédictions de l'item révèle un h^2 partiel de 0,90, ce qui laisse entrevoir une importante composante non linéaire dans la relation entre les deux variables.

Tableau 2.3
Modèle proposé par l'analyse de la régression pour les items

Modèle retenu	***R***	***R*²**	**Influence sur *R*²**	***F***	σ	**η² partiel**
Degré de difficulté et	0,78	0,60	0,60	57,76	0,00	0,90
coefficient de discrimination	0,81	0,65	0,05	34,19	0,00	0,99

Modèle retenu et coefficients	**Coefficients bruts**		**Coefficients standardisés**		
	B	S_B	β	*T*	σ
(Constante)	1,13	1,55		0,73	0,00
Niveau de difficulté *p*	0,68	0,12	0,68	6,30	0,00
Coefficient de discrimination *D*	0,24	0,12	0,24	2,20	0,00

La figure 2.2 présente la relation entre le niveau de difficulté des items et l'indice de réalisation des prédictions associé à ces derniers. Nous observons la présence d'un lien linéaire étroit entre les deux variables (R^2 = 0,60). L'indice de réalisation des prédictions augmente en fonction du niveau de difficulté de l'item. Trois valeurs au bas de la figure retiennent notre attention. Nous avons indiqué leur niveau de difficulté (*p*), leur coefficient de discrimination (*D*), leur niveau taxonomique (*T*) et des mesures de distances : Mahalanobis (*M*), Cook (*C*) et levier (*L*). Nous notons que la valeur représentant l'item le plus difficile peut être considérée comme une donnée extrême (*M* > 5).

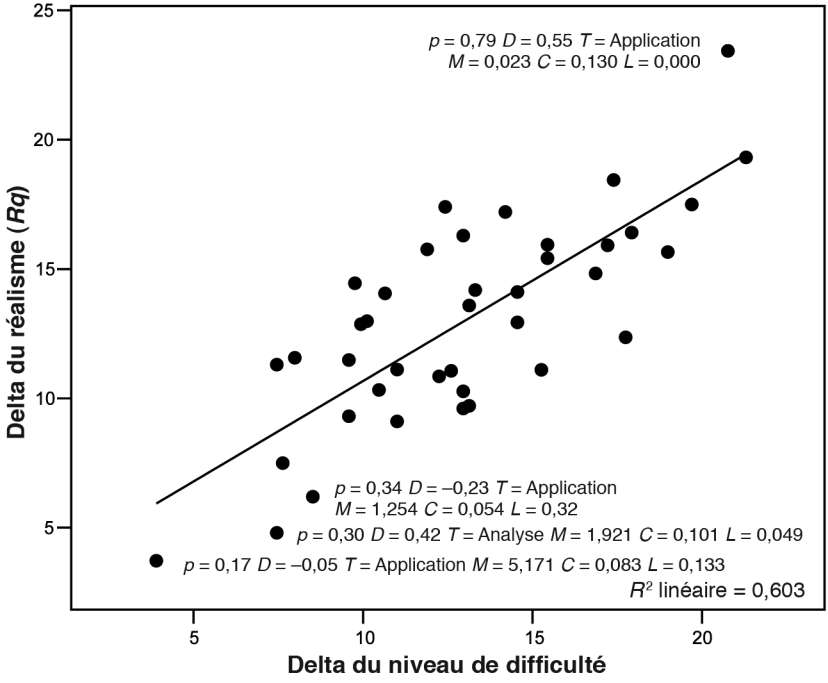


Figure 2.2
Lien entre le degré de réalisme et le niveau de difficulté

La figure 2.3 présente le lien entre le coefficient *delta* de discrimination de l'item et l'indice *delta* de réalisation des prédictions associé à celui-ci (R^2 = 0,27). Nous observons que la corrélation est fortement influencée par un petit nombre de valeurs extrêmes, les mêmes qui ont été observées à la figure 2.2. Les mesures de distances révèlent que les deux items qui discriminent le moins, en plus d'être considérés extrêmes (*M* = 13,86 et 8,18), ont suffisamment d'effet de levier (*L* = 0,36 et 0,21) pour influencer les résultats de l'analyse de régression.

Deux des trois items ont un coefficient de discrimination *D* très faible et sont du niveau taxonomique de type application. Les mesures de distances nous montrent que le retrait de ces items altérerait la pente de la droite de régression entre ces deux variables ainsi que la nature du lien linéaire observé. En supprimant les trois items en question, le R^2 linéaire passe de 0,27 à 0,06. Puisque l'instrument utilisé est essentiellement composé d'items qui discriminent bien, il est difficile de prévoir ce que serait l'impact d'items qui discrimineraient modérément ou peu. Ainsi, paradoxalement, un instrument constitué d'un plus grand nombre d'items ambigus aurait été nécessaire afin de mieux observer la relation entre discrimination et indice de réalisme par item. En l'absence d'un plus grand nombre d'items de ce genre, nos conclusions sur l'effet d'une faible discrimination de l'item sur le réalisme de l'item demeurent limitées, la corrélation entre ces deux variables étant affectée par une restriction de l'étendue des valeurs de discrimination.

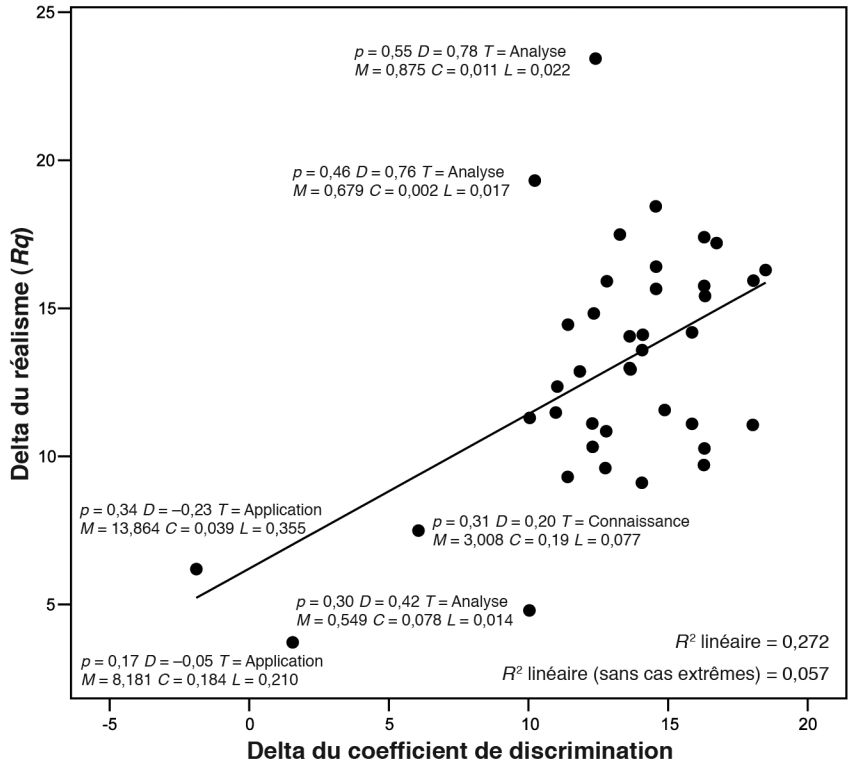


Figure 2.3
Lien entre le degré de réalisme et le coefficient de discrimination

L'analyse portant sur l'impact du niveau taxonomique sur le degré de réalisme ou l'indice de réalisation des prédictions soulève des problèmes. Certains niveaux taxonomiques des items incluent un petit nombre d'items et plusieurs cas limites sont difficiles à catégoriser dans un seul niveau taxonomique. Même si on s'attend à ce que les items de niveaux taxonomiques d'ordre supérieur entraînent une baisse du degré de certitude exprimé, la classification même de ces items est complexe et vient quelque peu fausser les résultats. Les items du niveau de l'application sont communs dans les instruments d'évaluation en mathématiques. L'instrument choisi contient une majorité d'items de ce niveau. Ceux qui ne l'étaient pas peuvent être considérés comme des cas limites, en ce sens qu'ils s'éloignent rarement des habiletés cognitives du niveau de l'application. Or, comme on dispose ainsi d'un échantillon d'items relativement petit et d'un petit nombre de niveaux taxonomiques différents, les conditions ne sont pas optimales pour une étude en profondeur des effets des différents niveaux taxonomiques sur le degré de réalisme des étudiants quant à leurs réponses aux items. Les tentatives de regroupement factoriel ont révélé certains patrons. Les résultats de ces analyses ne permettent cependant pas d'identifier une forme de classification taxonomique quelconque (Boulé, 2007).

En somme, les analyses effectuées permettent de constater qu'il y a des différences significatives entre les moyennes du degré de réalisme associé aux items dans le cas de la difficulté et de la discrimination des items. À la lumière de ces résultats, nous pouvons rejeter l'hypothèse nulle concernant l'effet principal lié à la difficulté des items. Il semblerait que le degré de réalisme exprimé par rapport à un item est lié aux propriétés de l'item. Nous ne pouvons pas rejeter l'hypothèse nulle pour l'effet principal lié aux niveaux taxonomiques des items. Les résultats montrent que le degré de réalisme de l'étudiant et son degré de certitude varient en fonction de la difficulté des items.

6. DISCUSSION

Cette recherche proposait de vérifier, dans un contexte d'évaluation diagnostique, l'effet des propriétés métriques et cognitives des items sur le degré de certitude exprimé par les étudiants et, par conséquent, sur le degré de réalisme associé à un item en particulier. Également, nous avions envisagé l'hypothèse que les différences individuelles pouvaient aussi avoir une influence sur le degré de réalisme des étudiants. Dans cette section, nous allons examiner les résultats obtenus et voir quelles sont les pistes de recherche à considérer.

Les écrits sur le sujet nous permettaient de pressentir que le degré de réalisme pouvait varier en fonction de certaines différences individuelles. Il était ainsi raisonnable de croire qu'un étudiant très performant serait plus réaliste dans son appréciation de la qualité de ses réponses. Nous pouvions aussi présumer que les hommes étaient plus téméraires que les femmes et qu'ils avaient une propension à surestimer leurs habiletés en mathématiques. Les résultats confirment notre hypothèse que les étudiants plus performants sont plus réalistes par rapport à l'estimation de la qualité de leurs réponses que ceux qui sont moins performants, ce qui est conforme aux observations de Beyer (2002), de Fabre (1980) et de Stankov (1998). Il semble important de fournir aux étudiants moins performants un feedback qui tienne compte du réalisme de leur réponse, en particulier dans toutes les situations pour lesquelles l'étudiant croyait avoir donné une réponse correcte avec un haut degré de certitude, alors que ce n'était pas le cas. La prise en compte du réalisme des sujets permet, au plan diagnostique, non pas de mieux pondérer les résultats, comme en évaluation sommative, mais plutôt de mieux cibler le feedback et la régulation des apprentissages.

Par ailleurs, nous ne pouvons pas rejeter l'hypothèse nulle sur l'effet principal, qui stipulait que le genre de l'individu pouvait être lié à son degré de réalisme, ce qui corrobore les constats de Henmon (1911) et de Jonsson et Allwood (2003). Il semblerait que les hommes et les femmes de notre échantillon expriment leur certitude par rapport à leurs réponses avec des degrés de réalisme relativement équivalents. On constate que plus les étudiants sont performants, plus ils ont un degré de réalisme élevé, et ce, quel que soit leur sexe. La quasi-absence de relation entre le genre de l'individu et son degré de réalisme peut aussi être attribuable à plusieurs autres facteurs. Les recherches, jusqu'à présent, ont été menées surtout avec des enfants. La situation chez les adultes n'est peut-être pas la même et mérite sans doute d'être explorée davantage. Ces résultats sèment suffisamment de doutes pour justifier d'autres analyses sur des échantillons de plus grande taille avec des sujets ayant un niveau de performance moyen.

Un autre aspect de notre recherche consistait à vérifier l'effet des propriétés métriques et cognitives des items sur le degré de certitude exprimé par les étudiants et, par conséquent, sur le degré de réalisme associé à un item en particulier. Le niveau de difficulté de l'item constitue la première variable des propriétés des items analysées. Les résultats appuient la prémisse que plus les items sont faciles, plus le degré de certitude et de réalisme de l'individu sera élevé. Ceci concorde d'ailleurs avec ce que l'on trouve dans la recension des écrits sur le sujet (de Finetti, 1965 ; Gilles, 2002 ; Jacobs, 1974). Cette conclusion n'est pas particulièrement surprenante, mais elle confirme néanmoins la force

potentielle de l'impact de la qualité métrique d'un item sur le degré de réalisme exprimé par rapport à cet item. Un item mal réussi et difficile qui aurait un indice faible de réalisation des prédictions est en quelque sorte un item pouvant être considéré comme défectueux. Comme Gilles (2002), nous concluons que le degré de certitude exprimé par rapport à un item constitue un autre élément pouvant s'ajouter aux propriétés métriques de l'item et qui peut servir à juger de sa qualité intrinsèque, de sa pertinence et de son utilité pratique.

Nous pouvons tirer la même conclusion en ce qui concerne le coefficient de discrimination de l'item, puisque le réalisme de l'étudiant semble varier en fonction des différents niveaux de discrimination de l'item. Plus un item discrimine bien, plus l'étudiant a une propension à être réaliste en ce qui concerne la qualité de sa réponse, un peu comme pour le niveau de difficulté d'un item : un item qui ne discrimine pas ou presque peut être défectueux de fait ou entraîner une confusion qui n'est pas voulue. Ces items sont à éviter dans à peu près tout contexte d'évaluation puisqu'ils n'apportent guère d'information utile. Après avoir analysé les données, il nous a semblé qu'aux fins de recherches ultérieures, il pourrait être intéressant d'utiliser un instrument qui est constitué d'items de degrés de discrimination variables. Nous pressentons que la relation entre le coefficient de discrimination de l'item et le degré de réalisme associé à un item mérite d'être étudiée et documentée en profondeur au moyen d'un ensemble d'items présentant une distribution plus étendue des coefficients de discrimination.

De leur côté, les niveaux taxonomiques posent plusieurs défis de taille. Avec un échantillon relativement petit de niveaux taxonomiques, les conditions n'étaient pas propices à une étude en profondeur des effets des différents niveaux taxonomiques sur le degré de réalisme des étudiants. Un compte rendu détaillé est cependant disponible (voir Boulé, 2007).

Enfin, l'étude diagnostique du réalisme peut contribuer à jeter un éclairage nouveau sur les performances différentielles de groupes d'étudiants. Des étudiants trop optimistes peuvent négliger de réviser leurs réponses ou prendre moins de temps pour remplir un questionnaire, entraînant des résultats différents. Cette notion est potentiellement prometteuse pour tenter d'expliquer les différences entre garçons et filles lorsqu'elles se produisent.

7. CONCLUSION

L'objectif de cette recherche consistait à vérifier l'utilité pratique et diagnostique du degré de certitude. Une modalité de réponse qui permettait à l'élève d'exprimer sa certitude par rapport à la qualité

de ses réponses fut ajoutée à un test, ce qui permettait d'estimer le degré de réalisme de l'élève. Les résultats permettent de constater que l'expression du degré de certitude ne varie pas en fonction du sexe de l'élève. Nous notons que plus le niveau de performance de l'élève est élevé, plus l'élève est réaliste par rapport à la qualité de ses réponses. Également, plus un item est facile, plus l'élève est réaliste.

Le devis utilisé dans cette recherche présente quelques difficultés qui en limitent la généralisation des résultats. D'une part, l'instrument est constitué d'items limités en nombre et qui ne se distribuent pas également parmi les catégories taxonomiques. D'autre part, le coefficient de discrimination moyen des items est particulièrement élevé et la distribution des coefficients de discrimination des items est fortement asymétrique. Ces deux conditions balisent l'interprétation des résultats. Des recherches ultérieures pourraient vérifier nos hypothèses en utilisant des instruments qui présentent des caractéristiques souhaitées en matière de niveaux taxonomiques et de coefficients de discrimination des items.

La contribution de cette recherche se situe davantage dans les questions qu'elle soulève que dans les réponses qu'elle apporte. La force de l'impact potentiel du coefficient de discrimination d'un item sur le degré de réalisme exprimé par rapport à cet item reste à vérifier à partir de séries de données appropriées. La prise en compte de la discrimination d'un item pourrait contribuer à une interprétation plus valide des résultats lors de l'utilisation des degrés de certitude.

On ne s'attendait pas à trouver un lien potentiellement curvilinéaire entre le niveau d'habileté et le degré de réalisme en fonction du genre. Les recherches et les analyses qui portent sur l'impact du genre sur le degré de réalisme et sur la surestimation postulent une relation linéaire entre ces variables. Nous considérons que la linéarité de ces relations doit être vérifiée lors d'analyses mettant ces variables en relation.

RÉFÉRENCES

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J. et Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. New York, New York: Longman.

Beyer, S. (2002). The effects of gender, dysphoria, and performance feedback on the accuracy of self-evaluations. *Sex roles, 47*(9-10), 453-464.

Boulé, S. (2007). *Utilisation du degré de certitude dans un contexte d'évaluation diagnostique critériée*. Thèse de maîtrise inédite, Université d'Ottawa, Ottawa.

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British journal of mathematical and statistical psychology, 18*, 87-123.

Fabre, J.-M. (1980). *Jugement et certitude*. Berne-Francfort, Suisse : Peter Lang.

Gilles, J.-L. (2002). *Qualité spectrale des tests standardisés universitaires*. Thèse de doctorat inédite, Université de Liège, Belgique.

Henmon, V. A. C. (1911). The relation of the time of a judgement to its accuracy. *Psychological review, 18*, 186-201.

Hunt, D. (1993). Human self-assessment : theory and application to learning and testing. Dans : D. Leclercq et J. Bruno (dir.), *Item banking : interactive testing and self-assessment*. Heidelberg, Allemagne : Springer Verlag.

Jacobs, S. S. (1974). Behavior on objective tests under theoretically adequate, inadequate and unspecified scoring rules. *Journal of educational measurement, 12*(1),19-29.

Jonsson, A. C. et Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and individual differences, 34*(4), 559-574.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of educational psychology, 30*(1), 17-24.

Laveault, D. et Grégoire, J. (2002). *Introduction aux théories des tests en psychologie et en sciences de l'éducation*. Bruxelles, Belgique : De Boeck.

Leclercq, D. (1993). Validity, reliability and acuity of self-assessment in educational testing. Dans D. Leclercq et J. Bruno (dir.), *Item banking : interactive testing and self-assessment*. Heidelberg, Allemagne : Springer Verlag.

Leclercq, D. et Poumay, M. (2004). Une définition opérationnelle de la métacognition et ses mises en œuvre. Communication présentée à la 21e conférence internationale de l'AIPU, Marrakech, Maroc.

Office de la qualité et de la responsabilité en éducation (1999). *Étude numéro 4*. Toronto, Ontario : Office de la qualité et de la responsabilité en éducation.

Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex roles, 48*(5-6), 265-276.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into practice, 41*(4), 219-225.

Reach, G., Zerrouki, A., Leclercq, D. et d'Ivernois, J.-F. (2005). Adjusting insulin doses : from knowledge to decision. *Patient education and counseling, 56*(1), 98-103.

Sieber, J. E. (1974). Effects of decision importance on ability to generate warranted subjective uncertainty. *Journal of personality and social psychology, 30*(5), 688-694.

Stankov, L. (1998). Calibration curves, scatterplots and the distinction between general knowledge and perceptual tasks. *Learning and individual differences, 10*(1), 29-50.

Swineford, F. (1938). Measurement of a personality trait. *Journal of educational psychology, 29*(4), 295-300.

Chapitre 3

Intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques et données de l'enquête TEIMS

Gilles Raîche, Diane Leduc, Martin Riopel et Claire Isabelle

À partir des trois rondes (1995, 1999, 2003) de l'enquête Tendances de l'enquête internationale sur les mathématiques et les sciences (TEIMS)*, nous voyons si ses données sous-tendent une intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques. Notre hypothèse suppose que le niveau d'intégration des pratiques augmente avec les rondes et que les structures factorielles sont modifiées. Nos résultats montrent au contraire que ces dernières sont constantes et qu'il y a stabilité du niveau d'intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques.*

Dans le cadre d'une recherche en persévérance et réussite scolaire[1], nous étudions l'intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques, notamment à partir des données d'enquêtes nationales et internationales, en particulier celle des *Tendances de l'enquête internationale sur les mathématiques et les sciences* (TEIMS). S'appuyant sur une analyse transversale des items des enquêtes, l'étude met en relation ces deux types de pratiques, ancrées dans un contexte d'approches par compétences, selon les quatre grandes dimensions (Raîche, Cantin et Lalonde,

1. Recherche subventionnée dans le cadre des Actions concertées du Fonds québécois de recherches sur la société et la culture (FQRSC) et par le ministère de l'Éducation, du Loisir et du Sport du Québec (MELS).

2005 ; Raîche, 2006) : apprentissage, authenticité, équité et intégration. Ces dimensions s'arriment aux nouvelles approches d'enseignement qui se mettent graduellement en place dans les institutions scolaires et sous-tendent une réflexion sur le cadre de référence des divers ministères de l'Éducation et la qualité de l'évaluation des apprentissages.

Notre but ici est de présenter la première phase du projet, qui consiste à analyser les items des questionnaires. Nous décrirons, dans un premier temps, la problématique de la recherche, avec son cadre théorique, ses objectifs et la méthodologie de la première phase ; dans un deuxième temps, nous décrirons brièvement l'enquête TEIMS en fournissant quelques explications sur les analyses de ses items. Nous aborderons au passage les liens entre ces résultats et nos hypothèses de recherche, qui supposent que le niveau d'intégration des pratiques augmente au fil des rondes du TEIMS et que les structures factorielles sont modifiées.

1. PROBLÉMATIQUE

1.1. Contexte de la recherche

Longtemps, il a été considéré que les pratiques d'évaluation des apprentissages devaient être dissociées des pratiques pédagogiques. L'élève apprenait et, par la suite, il était évalué. Il fallait à tout prix éviter d'enseigner en vue d'un test. Cette conception de l'évaluation des apprentissages est actuellement en transformation. Les nouvelles approches par compétences sous-tendent que les pratiques pédagogiques, à tout ordre d'enseignement, ne peuvent plus être séparées des pratiques d'évaluation des apprentissages qui les accompagnent toujours (Biggs, 1995 ; Wiggins, 1998 ; Stiggins, 2005). Pédagogie et évaluation vont ainsi de pair. De plus en plus, les pratiques pédagogiques doivent préparer les élèves au processus d'évaluation de leurs apprentissages et, simultanément, les pratiques d'évaluation de leurs apprentissages doivent contribuer au développement de ces mêmes apprentissages. Par leur intégration, chacune de ces pratiques augmente alors la portée de l'autre, contribuant ainsi non seulement à accroître la persévérance et à faciliter la réussite scolaire de l'élève, mais également à améliorer la qualité de l'enseignement.

Par ailleurs, les révisions des curriculums scolaires à travers le monde, sous diverses appellations, visent à ce que les apprentissages soient réalisés dans un contexte significatif, authentique et qui se rapproche ainsi de la réalité professionnelle ou quotidienne des apprenants. Dans cet état d'esprit, les études sur la formation des enseignants

et, plus récemment, sur l'évaluation des apprentissages témoignent de cette préoccupation d'intégration des pratiques d'évaluation aux pratiques pédagogiques. Cette préoccupation se retrouve aussi chez ceux et celles qui doivent faire vivre et fonctionner les systèmes éducatifs. Comme nous le verrons, les publications professionnelles et gouvernementales, au Québec comme ailleurs, sous diverses formes et à tous les niveaux d'enseignement, le montrent bien. Nous n'avons qu'à considérer les divers mémoires des associations et des instances syndicales aux consultations sur le système d'éducation, les avis ministériels, les politiques d'évaluation des apprentissages, les règlements, etc.

1.2. Cadre théorique

1.2.1. *Écrits professionnels et gouvernementaux*

L'intérêt pour l'intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques n'est toutefois pas apparu spontanément, et on le voit se profiler depuis plusieurs années. Ainsi, déjà en 1979, le ministère de l'Éducation du Québec[2], dans son énoncé de politique et son plan d'action, indique que par une mesure et une évaluation appropriée, il incombe à la commission scolaire et, plus particulièrement, aux enseignants, de suivre les élèves, de déterminer leurs aptitudes, leurs difficultés et de favoriser leurs apprentissages. On voit déjà poindre l'idée d'une intégration des pratiques d'évaluation des apprentissages et des pratiques pédagogiques, même si, à cette époque, la tendance est beaucoup plus de limiter l'évaluation à des fins certificatives, soit à la reconnaissance des études. Selon De Lorimier (1988), à ce moment, et même un peu avant, le ministère de l'Éducation du Québec considérait l'évaluation comme un acte pédagogique qui devrait s'insérer dans un processus de développement et de correction des apprentissages plutôt que dans un mécanisme de sélection et d'élimination des élèves. Il ne s'agit toutefois pas encore, dans la réalité quotidienne, d'intégrer formellement les pratiques d'évaluation et les pratiques pédagogiques en une seule entité.

La dernière politique d'évaluation des apprentissages, celle de 2003, du ministère de l'Éducation du Québec (Gouvernement du Québec, 2003) reflète beaucoup plus clairement les tendances modernes, ainsi que les principes précédemment mentionnés. Par

2. Ce ministère a changé de nom en 2005 pour le ministère de l'Éducation, du Loisir et du Sport (MELS).

exemple, les fonctions et le processus relatifs à l'évaluation énoncés dans cette politique (p. 29-35) soulignent clairement, entre autres, les éléments suivants :

- l'importance du passage du paradigme de l'enseignement à celui de l'apprentissage ;
- l'importance de la régulation de la démarche de l'élève à celle de l'enseignant ;
- l'intérêt accru pour l'aide à l'apprentissage et la reconnaissance des compétences, définies comme les deux fonctions principales de l'évaluation ;
- l'accent mis sur l'évaluation faite en vue d'une contribution à la réussite éducative.

Il nous apparaît pertinent de mentionner qu'en dehors de la province de Québec, selon une enquête du Conseil des ministres de l'Éducation du Canada réalisée en 1996, il semblerait que des enseignants ont exercé des pressions sur le gouvernement pour permettre une meilleure intégration des pratiques d'évaluation et des pratiques pédagogiques (Ontario secondary schools teachers federation, 2003a, 2003b, 2003c), affirmant que l'approche centrée sur les résultats nuit à cette intégration et les force à enseigner spécifiquement en vue des épreuves provinciales. Dans ce cas-ci, il semble donc que les intervenants des établissements scolaires sont les véritables initiateurs d'une volonté de changement.

1.2.2. *Travaux de recherche sur les pratiques d'évaluation des apprentissages*

Les recherches qui se sont intéressées aux effets sur le terrain de l'évolution de l'évaluation des apprentissages au Canada comme ailleurs, principalement dans le contexte des approches par compétences, renseignent peu sur les pratiques. Dans plusieurs cas, ces recherches visent plutôt à vérifier l'impact de pratiques qui s'intéressent à la réussite ou la motivation des élèves. Rarement, les pratiques sont considérées dans leur ensemble et encore moins en lien avec les pratiques pédagogiques.

Toutefois, deux études réalisées en Ontario font exception à la règle. La première, effectuée par Forgette-Giroux, Simon et Bercier-Larivière (1996), s'intéresse à la perception des enseignants quant aux pratiques d'évaluation des apprentissages en classe. À cette fin, des enseignants ont été interrogés à propos de leurs pratiques courantes d'évaluation des apprentissages. Selon cette étude, même si les enseignants estiment s'acquitter convenablement de cette tâche, ils

soulignent un décalage important entre celle-ci et celles qu'exige la réforme de l'éducation de leur province. Cette distance se fait sentir, entre autres, par rapport à l'évaluation de performances, de compétences, d'habiletés supérieures et d'attitudes. Les enseignants indiquent aussi que la grande faiblesse de l'évaluation des apprentissages provient de la difficulté de varier les modalités d'évaluation et de les adapter aux besoins individuels des élèves, et que le besoin de formation le plus important pour eux est l'intégration de l'évaluation aux activités d'apprentissages de l'élève, d'où l'intérêt de notre étude.

La seconde recherche, plus récente, a été réalisée par Forgette-Giroux, Simon, Turcotte, Ferne et Choi (2006). Elle représente un bon début, mais comporte toutefois plusieurs limites. Les chercheurs se sont aventurés un peu plus loin dans l'examen des pratiques d'évaluation et des pratiques pédagogiques, sans toutefois les étudier dans leur interaction. Ils ont surtout comparé les pratiques d'évaluation et les pratiques d'enseignement déclarées des enseignants de l'Ontario et du Québec qui ont participé à l'enquête *Progress in International Reading Literacy* Study (PIRLS) en 2001 et, au niveau pancanadien, au *Programme d'indicateurs du rendement scolaire* (PIRS) en 2002. Leurs résultats indiquent des différences dans les pratiques entre les trois populations, mais les causes de ces écarts ne sont pas expliquées.

L'étude se limite à la lecture et à l'écriture à partir des enquêtes du PIRLS et du PIRS, uniquement pour l'administration des questionnaires de 2001 et de 2002. Les données recueillies de l'enquête *Programme international pour le suivi des acquis des élèves* (PISA) et du TEIMS, qui s'intéressent aussi aux mathématiques et aux sciences, gagneraient à être utilisées aussi. Sans oublier que les données de toutes ces enquêtes sont disponibles depuis 1995 et qu'elles ont été administrées à plusieurs occasions. Il est donc possible de les utiliser dans le but d'étudier l'évolution des pratiques d'évaluation et des pratiques pédagogiques au fil de l'implantation des réformes de l'éducation dans ces deux provinces. Enfin, cette recherche, fait plus important au regard de notre propre recherche, ne s'est pas intéressée spécifiquement à l'étude de l'intégration des pratiques d'évaluation et des pratiques pédagogiques.

1.2.3. *Dimensions pour l'analyse dans un contexte d'approches par compétences*

Les écrits contemporains qui portent sur l'évaluation des apprentissages soulignent, occasionnellement, différents principes qui doivent être sous-jacents à l'évaluation des compétences. Tous ces principes ont toutefois la particularité de nécessiter une intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques. Comme

on l'a vu dans l'introduction, Raîche, Cantin et Lalonde (2005 ; Raîche, 2006) ont tenté d'identifier et de regrouper ces différents principes selon une structure constituée de quatre grandes dimensions : apprentissage, authenticité, équité et intégration. Certaines dimensions trouvent aussi un écho dans les principes que Tardif (2006) considère à la base de l'évaluation des compétences, mais ces derniers reposent sur une logique différente de la nôtre.

La première de ces dimensions renvoie au fait que dans une approche par compétences l'apprentissage de l'élève est au cœur de l'évaluation des apprentissages. Pour cette raison, l'évaluation doit avoir pour principe d'être personnelle, puisque c'est l'élève qui apprend, et non son groupe classe. Elle doit aussi être au service de l'apprentissage, être planifiée et en étroite collaboration avec le processus didactique. Ces quatre éléments reflètent bien les positions d'Angelo et Cross (1998), Biggs (1995) et Stiggins (2005) quant au fait que l'évaluation des apprentissages doit être centrée sur l'élève, sur la classe et sur l'implication de l'élève, ainsi que sur la mise en pratique, particulièrement au Québec, de l'évaluation formative et des compétences (Scallon, 1988, 2004).

La seconde dimension concerne l'authenticité des tâches d'évaluation (Wiggins, 1989). Évaluer des apprentissages dans une approche par compétences exige que les tâches d'évaluation soient en lien avec les actions qu'un élève doit effectuer dans la société pour qu'elles aient un sens pour lui et ses apprentissages. Elles doivent donc être réalistes, être significatives, c'est-à-dire être intéressantes et stimulantes pour les élèves, et avoir autant la qualité du rendement que sa justification comme préoccupation principale (Ndinga, 2010). Tardif (2006) souligne bien le caractère contextuel de l'évaluation des apprentissages dans une approche par compétences, ainsi que la *transférabilité* des apprentissages associés.

La troisième dimension est celle de l'équité, qui s'appuie notamment sur des notions de transparence et d'équivalence des évaluations. Dans le contexte des approches par compétences, l'apprentissage est tributaire de la compréhension des résultats d'évaluation, autant pour lui permettre d'apprendre que pour donner un sens à ces résultats. C'est la raison pour laquelle il faut aussi, défi important et encore difficile à réaliser, tendre à ce que les résultats soient comparables d'un groupe classe à un autre, d'un enseignant à un autre et d'un établissement d'enseignement à un autre. Il s'agit aussi de rendre intelligibles les résultats d'évaluation ; c'est probablement ce qui explique l'engouement pour l'utilisation de plus en plus répandue des échelles descriptives d'appréciation ou de vérification globales (*holistic scoring*, *scoring*

rubrics). Robin et Simon (2004), comme Tardif (2006), soulignent d'ailleurs l'adéquation de ces échelles dans un contexte d'évaluation des apprentissages pour soutenir l'apprentissage (*assessment for learning*).

Enfin, la quatrième dimension, abordée par Roegiers (2001), appelle des actions d'intégration selon divers éléments. En premier lieu, il s'agit d'enseigner dans le but de faire réussir les tâches d'évaluation. Par exemple, la réalisation d'exercices qui préparent concrètement les élèves à une tâche d'évaluation est une pratique qui intègre l'évaluation à l'enseignement. Ensuite, il s'agit de viser à ce que les modalités d'évaluation soient intégrées les unes aux autres. Nous pouvons penser ici à la complexité progressive des évaluations tout au long d'un trimestre ou à l'harmonisation des tâches d'évaluation formatives et sommatives. Enfin, l'intégration des évaluations et des apprentissages doit se faire dans le programme d'études, lui-même dans une approche programme (Dorais, 1990).

Le tableau 3.1 présente un instantané de ces dimensions sous-jacentes à l'intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques dans le contexte des approches par compétences.

Tableau 3.1
Dimensions de l'évaluation des apprentissages dans le contexte des approches par compétences (Raîche, 2006)

Dimensions	Éléments
Apprentissage	1. L'évaluation est au service de l'apprentissage. 2. L'évaluation doit être individuelle. 3. L'évaluation est en étroite collaboration avec le processus didactique. 4. L'évaluation est planifiée.
Authenticité	1. Les tâches d'évaluation sont authentiques. 2. Les tâches d'évaluation sont significatives. 3. La qualité et la justification du rendement sont au cœur des préoccupations.
Équité	1. La correction des tâches d'évaluation est transparente. 2. La correction des tâches d'évaluation est équivalente d'un moment à un autre.
Intégration	1. Les tâches d'évaluation sont intégrées à l'enseignement. 2. Les modalités d'évaluation sont intégrées les unes aux autres. 3. Les modalités d'évaluation sont intégrées au programme d'études.

Il est intéressant de noter que cette perception de l'évaluation des apprentissages intégrée aux pratiques pédagogiques se fait aussi sentir dans les études qui ne traitent pas directement de l'évaluation des apprentissages. À titre d'exemple, Gauthier et Lessard (2005) décrivent le rôle à venir de l'enseignant comme celui d'un professionnel de l'intervention pédagogique, un rôle où il devra mettre en relation trois éléments fondamentaux : la situation éducative, les savoirs de l'enseignant et le jugement. Cette mise en relation ne semble possible qu'avec l'intégration des pratiques d'évaluation des apprentissages et des pratiques pédagogiques.

1.3. Objectifs du projet de recherche

Plusieurs objectifs nous guident dans notre étude. D'abord, nous voulons identifier les structures quant aux pratiques d'évaluation des apprentissages. Il s'agit de découvrir, à l'aide d'analyses exploratoires, la dimensionnalité et les structures propres aux pratiques d'évaluation et aux pratiques pédagogiques. Quel est le contenu de ces pratiques ? Il est ici nécessaire de vérifier si les structures obtenues sont constantes au cours des diverses années d'administration des enquêtes et si elles sont équivalentes d'une province à l'autre. Ensuite, en examinant les relations qui existent entre les pratiques d'évaluation et les pratiques pédagogiques, nous tentons de déterminer si elles se modifient au cours des années et en fonction des curriculums. Puis, nous examinons plus particulièrement l'intégration des pratiques afin d'évaluer son évolution. Depuis quand est-elle présente dans les pratiques ? S'est-elle intensifiée au cours des ans ? Enfin, par cette recherche, nous essayons de trouver des pistes d'action pour préparer la relève du personnel éducatif. Pour ce faire, en fin de projet, des groupes de discussion, constitués de chercheurs, de collaborateurs et d'intervenants des milieux de pratiques, seront réunis pour jeter les bases d'un inventaire d'applications intégrant l'évaluation à l'enseignement en classe.

Notre recherche tente donc d'approfondir l'une des quatre dimensions précédemment mentionnées, soit l'intégration. Elle vise aussi à repousser les limites rencontrées par l'équipe de Forgette-Giroux et à utiliser notamment les politiques et les écrits ministériels pour explorer ce qu'il en est vraiment de l'intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques, toujours dans le contexte des approches par compétences. Plus concrètement, nous désirons obtenir un portrait général de la situation et analyser la structure et l'évolution de ces dimensions au fils des ans, compte tenu des modifications des curriculums.

2. MÉTHODOLOGIE

Dans ce qui suit, nous présentons uniquement la première phase de notre projet de recherche, visant le premier objectif, qui consiste plus particulièrement à vérifier si les données des enquêtes nous permettent d'identifier les structures des pratiques et de mesurer le degré d'intégration de ces mêmes pratiques. Il est à noter que cette phase sera, par la suite, suivie de la passation de questionnaires sur l'intégration auprès de professeurs du secondaire.

Rappelons que la méthodologie de cette première phase repose entièrement sur les données des enquêtes nationales et internationales. Les questionnaires administrés (ou qui le seront au cours de la réalisation de ce projet de recherche) aux parents, aux élèves, aux enseignants et aux directions par les grandes enquêtes nationales et internationales ont permis de recueillir des informations précieuses en ce qui a trait non seulement au rendement éducatif, mais aussi aux pratiques pédagogiques et d'évaluation des apprentissages, auprès de très grands échantillons. En principe, les données de toutes les enquêtes suivantes sont disponibles à la communauté de chercheurs, moyennant un engagement clair à protéger les items.

- *Tendances de l'enquête internationale sur les mathématiques et les sciences (TEIMS),* rondes 1995, 1999, 2003 et 2007 ;
- *Progress in International reading literacy study* (PIRLS), rondes 2001 et 2006 ;
- *Programme d'indicateurs du rendement scolaire* (PIRS), rondes 1996, 1997, 1998, 1999, 2001, 2002, 2004 et 2007 ;
- *Programme international pour le suivi des acquis des élèves* (PISA), rondes 1998, 1999, 2201, 2002 et 2004.

Toutefois, il s'avère difficile d'obtenir tous les items pour chacune de ces enquêtes. Ainsi, pour des raisons de confidentialité et malgré plusieurs tentatives, nous n'avons pas pu obtenir les données du PISA. Par conséquent, étant donné l'ampleur de cette enquête, le nombre d'items prévu pour effectuer nos analyses a été passablement réduit. Pour le PIRLS, le caractère spécifique de la lecture ne nous permettait pas, dans les items, de bien voir les structures des pratiques ni de généraliser nos résultats. Nous n'avons donc pas utilisé ces données pour faire des analyses plus approfondies. Quant au PIRS (rebaptisé depuis quelques années le Programme pancanadien d'évaluation [PCPE]), le caractère national de cette enquête, plutôt qu'international, pose quelques difficultés de comparaison. En effet, les items du PIRS, comme ceux des autres enquêtes, sont évidemment construits en fonction des programmes d'études afin de rendre compte des particularités des systèmes éducatifs, ce qui rend la comparaison des curriculums

inutile. Dans ce cas, nos analyses pourraient permettre de voir s'il y a intégration des pratiques au Canada, mais pas ailleurs dans le monde. Finalement, seules les données du TEIMS pouvaient être analysées en fonction de nos objectifs puisque nous avons eu accès à l'ensemble des items pour les rondes 1995, 1999, 2003 et 2007.

2.1. Particularités de l'enquête TEIMS

L'enquête TEIMS représente la continuité d'une longue série d'études sur le rendement scolaire menées par l'Association internationale pour l'évaluation du rendement scolaire (AIE ou, en anglais, IEA, pour *International Association for the Evaluation of Educational Achievement*), une coopérative internationale indépendante constituée d'institutions de recherche nationales et d'agences gouvernementales. Depuis sa fondation en 1959, l'AIE a mené plus d'une quinzaine d'études portant sur les acquis des élèves dans les domaines des mathématiques, des sciences, des langues, de l'éducation à la citoyenneté et de la lecture.

Les concepteurs de l'enquête TEIMS compilent les résultats d'élèves de 4e année (âgés de 9 à 10 ans), de 8e année (âgés de 13 à 14 ans) (Crocker, 2002) et de la dernière année du cours secondaire, afin de documenter les grandes tendances concernant la performance des élèves au fils des ans. Entre 1995 et 2007, de 41 à 67 États ont obtenu, tous les quatre ans, les résultats en mathématiques et en sciences auprès d'un demi-million d'élèves et les ont transmis à l'AIE.

2.2. Instrumentation

Afin de procurer des renseignements complémentaires sur les résultats des évaluations et de permettre aux États participants de retracer les changements dans les pratiques éducatives, les concepteurs de l'enquête demandent aux élèves, aux enseignants ainsi qu'aux directions des écoles de remplir un questionnaire à propos du contexte d'enseignement des mathématiques et des sciences. Combinées à des analyses des guides, des manuels des élèves et de tout le matériel de soutien pédagogique, les grandes tendances qui ressortent de ces données dressent un portrait des changements relatifs à l'implantation des pratiques et des politiques éducatives. Les résultats obtenus sont compilés dans une série de rapports qui procurent de l'information utile aux théoriciens et aux praticiens des États ayant participé à l'étude, sur l'enseignement des sciences et des mathématiques ainsi que sur le développement de nouvelles avenues pour l'amélioration des performances des élèves. Il est à souligner que dans ce chapitre, *les États* désigne tous les pays et les ressorts (telles les provinces) qui participent à l'enquête TEIMS.

2.3. Déroulement

Au point de départ, il a fallu classer les items du TEIMS en fonction des deux pratiques que sont l'évaluation et la pédagogie. Cette opération a été réalisée dans le but de faciliter le choix des items qui seront soumis aux analyses factorielles. Les items sélectionnés devaient se retrouver sans modifications du libellé de la question et des choix de réponses à l'intérieur de toutes les enquêtes réalisées en 1995, 1999, 2003 et 2007. Malheureusement, nous n'avons pas pu avoir accès à la traduction française des items et avons dû effectuer notre travail à partir de la version anglaise.

2.4. Méthode d'analyse

Afin de décrire la structure des pratiques d'évaluation et des pratiques pédagogiques, une analyse factorielle exploratoire a été réalisée. En premier lieu, le nombre de facteurs à extraire a été déterminé à partir de l'analyse des valeurs propres de la matrice des corrélations de Pearson entre les items retenus selon le test de l'éboulis de Cattell. Cette analyse du nombre de facteurs à retenir a été répétée pour les années d'administration suivantes : 1995, 1999 et 2003. Ensuite, les facteurs ont été obtenus selon une extraction par maximum de vraisemblance, en appliquant par la suite une rotation oblique par la méthode oblimin. Cette extraction a aussi été réalisée pour chacune des années d'administration du TEIMS.

Enfin, dans le but de vérifier l'intégration des pratiques d'évaluation aux pratiques pédagogiques, à condition que les structures factorielles nous indiquent vraiment l'existence de ces deux facteurs, la corrélation de Pearson entre les facteurs a été calculée. Plus cette corrélation est élevée, plus l'intégration des pratiques d'évaluation aux pratiques pédagogiques devrait être importante. Les coefficients de fidélité des diverses échelles de mesure associées à ces facteurs ont aussi été calculés.

3. ANALYSES

3.1. Classification des items

Comme première étape pour pouvoir effectuer nos analyses, nous devions classer les items du TEIMS en fonction des deux pratiques que sont l'évaluation et la pédagogie. Ceci constituait une étape décisive pour la suite de notre projet. Avant même que nous puissions rattacher chacun des items à une pratique, plusieurs écueils se présentaient.

D'abord, la traduction en français des items en modifiait passablement le sens. De plus, plusieurs items n'étaient pas traduits et les démarches pour obtenir la traduction française d'une grande quantité d'items se sont avérées trop complexes. Ensuite, les données du Québec et des autres provinces n'étaient pas toujours différentes les unes des autres. C'est pourquoi nos résultats ne pourront pas distinguer les structures factorielles selon la province où a été administré le TEIMS. Dans le même ordre d'idées, certains items ne sont pas administrés dans tous les pays. Il fallait donc faire le tri parmi tous les items pour séparer ceux qui sont communs à tous les pays de ceux qui ne le sont pas.

Enfin, s'est posée la question de la transformation des items d'une ronde à l'autre. Ainsi, comme le montre le tableau 3.2, certains items sont communs à toutes les rondes, d'autres sont propres à chaque ronde et d'autres enfin changent selon la ronde. Les deux premières colonnes illustrent les changements de code des items d'une ronde à l'autre.

Tableau 3.2
Items retenus pour les analyses factorielles

2003	1995/99	QUESTIONS
BSBMHWSG	BSBMSGRP	In your math lessons, how often do you work together in small groups?
BSBMHMDL	BSBMEVLF	In your math lessons, how often do you relate what you are learning in mathematics to your daily life?
BSBMHROH	BSBMHWDS	In your math lessons, how often do you review your homework?
BSBMHBHC	BSBMHWCL	In your math lessons, how often do you begin your homework in class?
BSBMHHQT	BSBMTEST	In your math lessons, how often do you have a quiz or test?
BSBMHCAL	BSBMCALC	In your math lessons, how often do you use calculators?
BSBSHDEI	BSBSDEMO	How often do you watch the teacher demonstrate an experiment or investigation in your science lessons?
BSBSHCEI	BSBSEXPR	How often do you conduct an experiment or investigation in your science lessons?
BSBSHWGO	BSBSSGRP	How often do you work in small groups on an experiment or investigation in your science lessons?
BSBSHBHC	BSBSHWCL	How often do you begin your homework in class in your science lessons?
BSBSHHQT	BSBSTEST	How often do you have a quiz or test in your science lessons?

Certains items sont propres à certains pays et n'ont donc pas été administrés au Canada. Comme ils ne contenaient pas toutes les données nécessaires à notre étude, nous avons dû les rejeter. Enfin, pour disposer de suffisamment d'items, nous avons dû accepter de conserver certains items même s'ils avaient subi de légères modifications. En dernier, il ne reste que onze items, modifiés ou non, communs aux trois premières rondes. On trouvera leur version anglaise au tableau 3.2.

3.2. Analyse factorielle exploratoire

Dans un premier temps, l'analyse de la dimensionnalité de la structure factorielle a été réalisée à partir de l'application du test de l'éboulis de Cattell aux valeurs propres de la matrice des coefficients de corrélation de Pearson. Les figures 3.1 à 3.3 illustrent le graphique des éboulis séparément pour les années d'administration du TEIMS : 1995, 1999 et 2003. En premier lieu, on remarquera la similarité de la courbe des éboulis entre ces trois figures. Il semble donc que le nombre de facteurs à retenir soit le même quelle que soit l'année d'administration du TEIMS. Ensuite, la courbe affiche un changement de pente abrupt à partir de la deuxième valeur propre : il semblerait alors que deux facteurs permettraient de représenter correctement la structure factorielle.

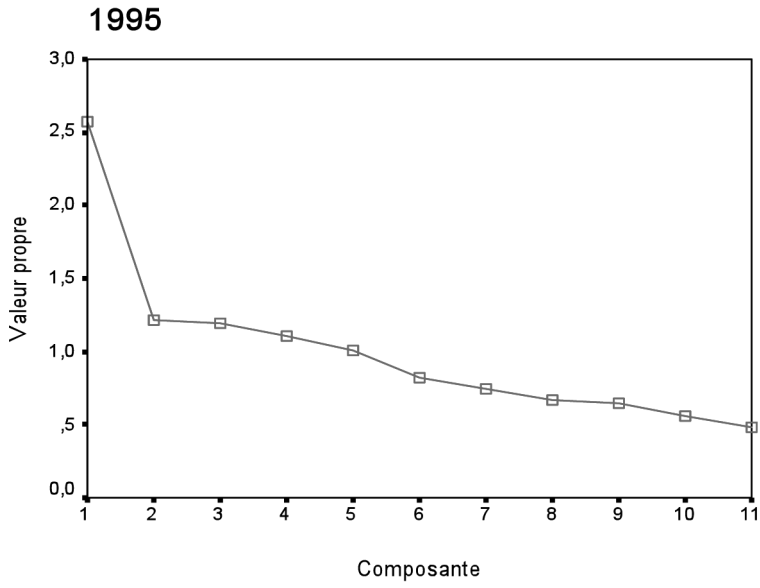


Figure 3.1
Dimensionnalité pour la ronde 1995

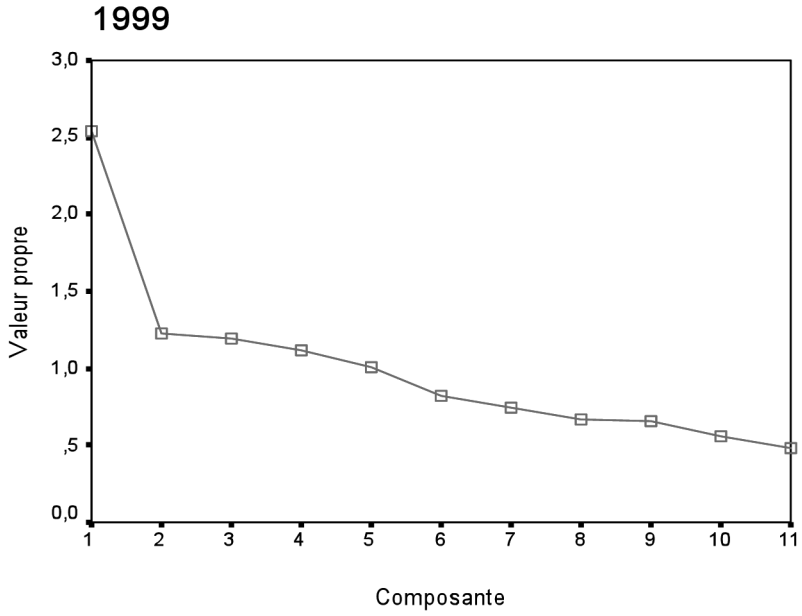


Figure 3.2
Dimensionnalité pour la ronde 1999

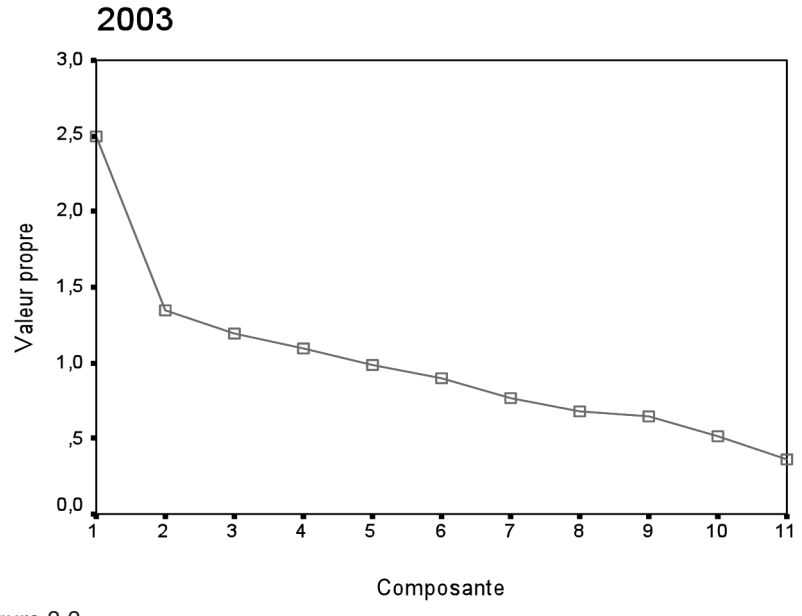


Figure 3.3
Dimensionnalité pour la ronde 2003

Le test de l'éboulis de Cattell ayant permis d'émettre l'hypothèse de la présence de deux facteurs quelle que soit l'année d'administration du TEIMS, une analyse factorielle exploratoire avec rotation oblique est ensuite réalisée pour chacune des années d'administration du questionnaire. Les résultats sont présentés au tableau 3.3. On y remarque, comme pour le test de l'éboulis, que la structure factorielle est strictement équivalente quelle que soit l'année d'administration du questionnaire. Pour l'analyse, on a retenu seulement les coefficients de saturation supérieurs ou égaux à 0,40 pour identifier la nature des facteurs. Toutefois, dans les structures factorielles finales, on retiendra tous les items.

On remarquera que quelle que soit l'année d'administration du questionnaire, on retrouve deux regroupements d'items significatifs qui semblent montrer des pratiques qui sont plutôt d'ordre pédagogique pour le premier facteur. En ce qui a trait au second facteur, il semblerait qu'il soit associé à des pratiques évaluatives. On retrouverait ainsi les dimensions que nous avions retenues dès le début ; une structure bifactorielle semble alors de mise.

Puisque nous n'avons pu obtenir, tel qu'espéré, une structure factorielle avec rotation oblique qui reflète adéquatement les pratiques pédagogiques et les pratiques évaluatives, nous poursuivons l'analyse en calculant les corrélations de Pearson entre les deux facteurs pour chacune des années d'administration du TEIMS. Au tableau 3.4, on remarque que cette corrélation est égale à environ 0,50 quelle que soit l'année d'administration. Il semble donc qu'il y ait une corrélation importante et positive entre les pratiques pédagogiques et les pratiques évaluatives. Ainsi, plus les pratiques pédagogiques sont importantes, plus les pratiques évaluatives sont elles aussi importantes.

Enfin, dans le même tableau sont présentés les coefficients de fidélité associés à ces deux facteurs pour chacune des années d'administration du questionnaire. C'est le premier facteur, soit celui qui est associé à la pédagogie, qui affiche le coefficient de fidélité le plus élevé (entre 0,65 et 0,73). Avec si peu d'items (seulement 11), il s'agit d'une valeur fort acceptable. En ce qui a trait au facteur associé aux pratiques d'évaluation, le coefficient de fidélité varie entre 0,43 et 0,49. Avec 11 items, il s'agit toutefois d'une valeur un peu faible. Il serait alors utile d'avoir d'autres items pour améliorer la précision de l'échelle des pratiques d'évaluation.

Tableau 3.3
Corrélation entre les facteurs évaluation et pédagogie des onze items finaux[1]

Items	1995		1999		2003	
	Péd	Éval	Péd	Éval	Péd	Éval
In your math lessons, how often do you work together in small groups?	0,13	0,35	0,13	0,35	0,12	0,22
In your math lessons, how often do you relate what you are learning in mathematics to your daily life?	0,20	0,39	0,21	0,39	0,15	0,26
In your math lessons, how often do you review your homework?	0,22	0,30	0,21	0,30	0,08	0,15
In your math lessons, how often do you begin your homework in class?	0,14	0,31	0,14	0,30	0,14	0,22
In your math lessons, how often do you have a quiz or test?	0,20	**0,40**	0,20	**0,40**	0,18	**0,62**
In your math lessons, how often do you use calculators?	0,10	0,23	0,11	0,23	0,10	0,12
How often do you watch the teacher demonstrate an experiment or investigation in your science lessons?	**0,63**	0,39	**0,61**	0,39	**0,53**	0,36
How often do you conduct an experiment or investigation in your science lessons?	**0,82**	0,33	**0,83**	0,33	**0,89**	0,38
How often do you work in small groups on an experiment or investigation in your science lessons?	**0,47**	0,40	**0,46**	0,39	**0,70**	0,33
How often do you begin your homework in class in your science lessons?	0,26	**0,41**	0,25	**0,40**	0,21	**0,36**
How often do you have a quiz or test in your science lessons?	0,30	**0,47**	0,29	**0,47**	0.21	**0,61**

1 Les coefficients de saturation en caractères gras sont ceux qui représentent le mieux le facteur. Seules les valeurs supérieures ou égales à 0,40 sont retenues.

Tableau 3.4
Corrélation entre les facteurs évaluation et pédagogie

Ronde	Corrélation entre les facteurs évaluation et pédagogie	Fidélité des échelles pédagogie (évaluation)
2003 (N = 4104)	0,48	0,73 (0,49)
1999 (N = 14907)	0,50	0,65 (0,43)
1995 (N = 14951)	0,51	0,66 (0,44)

4. DISCUSSION

L'intégration des pratiques évaluatives et des pratiques pédagogiques retient l'attention depuis au moins deux décennies, et elle est à l'ordre du jour des politiques récentes en matière d'évaluation. De plus, les enquêtes nationales et internationales fournissent un nombre considérable de données qui peuvent être utiles pour les chercheurs. Cependant, à la lumière de nos résultats, force est de constater que l'important échantillon que propose l'enquête TEIMS n'est pas aussi riche que nous l'espérions pour étudier les pratiques évaluatives et pédagogiques. En réalité, trop de facteurs étaient à prendre en considération dans notre analyse : la traduction en français des items, les items communs à plusieurs pays, les changements d'items d'une ronde à l'autre. Notre échantillon s'est rétréci et est devenu peu représentatif pour vérifier l'intégration de l'évaluation dans les pratiques d'enseignement.

Nos résultats montrent également une structure à deux facteurs regroupant une dimension pédagogique et une dimension évaluative et une corrélation acceptable entre les deux. Toutefois, l'échantillon de 11 items est trop petit pour nous permettre de tirer des conclusions quant à l'intégration de ces deux dimensions. Il faut souligner aussi que ces dimensions ne sont pas nécessairement le reflet des pratiques à proprement parler. Il y aurait donc lieu de poursuivre notre étude avec d'autres données pour vérifier si les items du TEIMS, ou provenant d'autres enquêtes, nous permettraient de trouver davantage de liens entre les pratiques évaluatives et pédagogiques.

Il serait également utile de porter éventuellement un jugement qualitatif sur l'implication de ces échelles dans la pratique et si possible d'effectuer des rapprochements avec les résultats de recherches antérieures. Dans ce jugement, il faudra souligner le peu d'informations que ces items nous donnent, malgré leur nombre élevé, et leur peu de représentativité du modèle que nous avons développé en termes de pratiques d'évaluation.

5. CONCLUSION

En arrière-plan de cette recherche se trouve le souci d'utiliser et d'analyser les nombreuses données d'enquêtes nationales et internationales pour étudier l'intégration des pratiques évaluatives et pédagogiques. Pour diverses raisons et suite à une démarche préliminaire, nous n'avons retenu que les données du TEIMS pour effectuer nos analyses. Tout au long de cette première étape nous avons effectué différents types d'analyses pour tenter d'identifier les structures de ces pratiques et mesurer leur niveau d'intégration les unes par rapport aux autres. Nos résultats montrent que nous retrouvons, dans les données du TEIMS, les dimensions évaluatives et pédagogiques telles que nous les avions envisagées et, par conséquent, une structure à deux facteurs.

Les résultats obtenus ne nous permettent toutefois pas de pouvoir utiliser ces échelles de mesure issues du TEIMS dans des applications pratiques en recherche, en évaluation de programme ou en soutien aux pratiques d'évaluation des enseignants. Il serait peut-être utile de répéter cette démarche d'analyse avec les données d'autres enquêtes internationales ou nationales telles que le PISA, le PIRS ou le PIRLS, et de comparer les résultats avec ceux d'autres pays pour vérifier la stabilité de la structure obtenue à partir des données canadiennes. Toutefois, nous pensons qu'il est plus profitable pour notre recherche de développer des échelles sans nous appuyer sur les données des enquêtes internationales. C'est la raison pour laquelle nos travaux futurs seront dirigés principalement vers l'élaboration d'échelles de mesure des pratiques d'évaluation en classe en lien avec les dimensions théoriques de ces pratiques, dimensions à partir desquelles nous avons entrepris cette recherche.

RÉFÉRENCES

Biggs, J. (1995). Assessing for learning : some dimensions underlying new approaches to educational assessment. *Alberta journal of educational research, 41*(1), 1-17.

Crocker, R. K. (2002). *Résultats d'apprentissage : analyse critique du domaine au Canada*. Rapport présenté au Conseil des statistiques canadiennes de l'éducation. Ottawa, Ontario : Conseil des statistiques canadiennes de l'éducation.

Dorais, S. (1990). Réflexion en six temps sur l'approche-programme. *Pédagogie collégiale, 4*(1), 37-41.

De Lorimier, J. (1988). *Propositions et politiques sur l'école : principales interventions des dix dernières années*. Québec, Québec : Conseil supérieur de l'éducation.

Forgette-Giroux, R., Simon, M. et Bercier-Larivière, M. (1996). Les prtiques d'évaluation des apprentissages en salle de classe : perception des enseignantes et des enseignants. *Revue canadienne de l'éducation*, *21*(4), 384-395.

Forgette-Giroux, R., Simon, M., Turcotte, C., Ferne, T. et Choi H. (2006). *Examen des pratiques d'enseignement et d'évaluation du personnel enseignant anglophone et francophone de l'Ontario et du personnel enseignant francophone du Québec à la lumière de la PIRLS et du PIRS*. Toronto, Ontario : Colloque CSCE-CRSH 2006.

Gauthier, C. et Tardif, M. (2005). *La pédagogie. Théories et pratiques de l'Antiquité à nos jours* (2e édition). Montréal, Québec : Gaëtan Morin.

Gouvernement du Québec (2003). *Politique d'évaluation des apprentissages*. Québec, Québec : ministère de l'Éducation.

Ndinga, P. (2010). *Authenticité des tâches d'évaluation : état des lieux*. Communication présentée au 78e Congrès de l'Association francophone pour le savoir (ACFAS), Montréal, Québec.

Ontario secondary schools teachers federation (2003a). *Accountability and the grade 10 literacy test : background on accountability and testing issues*. Toronto, Ontario : Ontario secondary schools teachers federation.

Ontario secondary schools teachers federation (2003b). *Accountability for all : a new partnership vision*. Toronto, Ontario : Ontario secondary schools teachers federation.

Ontario secondary schools teachers federation (2003c). *Response to the 11 Government's proposed secondary school reform : Introduction and overview*. Toronto, Ontario : Ontario secondary schools teachers federation.

Raîche, G. (2006). *L'évaluation dans un contexte d'approches par compétences*. Communication présentée dans le cadre des acivités du CEFRES, Montréal, Québec.

Raîche, G., Cantin, A. et Lalonde, M.-F. (2005). *Enseigner l'instrumentation en évaluation des apprentissages à de futurs enseignants à l'enseignement supérieur*. Communication présentée au congrès annuel de l'Association pour le développement de la mesure et de l'évaluation en éducation (ADMÉÉ), Québec, Québec.

Robin, T. et Simon, M. (2004). What's still wrong with rubrics : focusing on the consistency of performance criteria across scale levels. *Practical assessment, research and evaluation, 9*(2).

Roegiers, X. (2001). *Une pédagogie de l'intégration. Compétences et intégration des acquis dans l'enseignement*. Bruxelles, Belgique : de Boeck.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Saint-Laurent, Québec : Éditions du Renouveau pédagogique.

Scallon, G. (1988). *L'évaluation formative des apprentissages. Tome 1 : la réflexion*. Québec, Québec : Les Presses de l'Université Laval.

Simon, M. et Forget-Giroux, R. (2001). A rubric for scoring postsecondary academic skills. *Practical assessment research and evaluation, 7*(18).

Stiggins, R. J. (2005). *Student-involved assessment for learning*. San Francisco, Californie : Jossey-Bass.

Tardif, J. (2006). *L'évaluation des compétences. Documenter le parcours de développement*. Montréal, Québec : Chenelière Éducation.

Wiggins, G. (1998). *Educative assessment : designing assessments to inform and improve student performance.* San Francisco, Californie : Jossey-Bass.

Wiggins, G. (1989). A true test : toward more authentic and equitable assessment. *Phi delta kappan, 70*(9).

Chapitre 4

Validité des situations de compétence

Élaboration d'une grille d'analyse

Micheline-Joanne Durand et Isabelle Trépanier

L'approche par compétences se distingue des approches traditionnelles, tant du point de vue du rôle de l'évaluation que de l'instrumentation associée. Les auteurs s'entendent pour affirmer que les compétences se manifestent dans des situations d'un niveau de complexité supérieur exigeant la mobilisation d'un ensemble articulé de ressources. Pour montrer la validité d'une situation, le processus général d'élaboration d'un modèle suggéré par Silvern (1972 : voir Legendre, 2005) a été suivi afin d'établir une grille d'analyse descriptive (prototype). Ce sont les résultats représentant les caractéristiques d'une bonne grille d'analyse et d'une situation pertinente que cette recherche de développement a tenté de décrire.

1. INTRODUCTION

De nombreux systèmes éducatifs ont pris le virage de l'approche par compétences au cours de la dernière décennie. Cette approche exige des changements en profondeur, tant dans les approches pédagogiques qu'évaluatives. L'objet des évaluations a en effet changé : il ne s'agit plus uniquement de mémoriser des connaissances, mais de développer des compétences. Puisque toute nouvelle compétence nécessite des savoirs, des savoir-faire et des savoir-être, le concept de situation demeure une pierre angulaire du processus d'inférence en évaluation des compétences. Les auteurs affirment d'ailleurs que les compétences se manifestent dans des situations d'un niveau de complexité supérieur exigeant la mobilisation d'un ensemble articulé de ressources. Dans ce

contexte, on tentera de concevoir des situations d'évaluation qui favorisent la manifestation maximale des apprentissages qu'on veut évaluer (Laurier, Tousignant et Morissette, 2005, p. 63).

Dans un premier temps, on déterminera quelle est la nature des concepts liés à l'évaluation des compétences en situation, en consultant la littérature spécialisée. Le concept de situation et les concepts associés ont ainsi été utilisés dans l'analyse de contenu effectuée au moyen du processus de l'anasynthèse de Silvern (1972 : voir Legendre, 2005). Ce processus a été élaboré précisément pour passer d'un ensemble de données à l'élaboration d'un modèle. Il a été alors possible d'établir la nature, et les caractéristiques d'une situation d'apprentissage et d'évaluation (SAÉ). Dans un deuxième temps, ces caractéristiques ont déterminé les critères qui ont été développés à l'intérieur d'une grille comprenant une échelle descriptive (*rubric*). Troisièmement, cette grille a permis de porter un jugement sur cinq situations complexes en mathématiques, puis des correctifs nécessaires ont été apportés à la grille. Finalement, de nouvelles situations ont été analysées par huit experts à l'aide de cette grille améliorée. Les commentaires formulés par ceux-ci font l'objet de discussions sur la validité des situations d'apprentissage et d'évaluation utilisées dans différentes disciplines, dans l'enseignement tant primaire que secondaire au Québec. Puisqu'une démarche d'évaluation est une démarche d'inférence, la question centrale demeure la pertinence d'une situation permettant de faire émerger des compétences et de recueillir des traces pouvant aider l'enseignant dans le processus consistant à porter un jugement le plus fiable possible.

2. CADRE CONCEPTUEL : ÉVALUATION DES COMPÉTENCES EN SITUATION AUTHENTIQUE

Bien que le point de départ de cette recherche soit la notion de situation d'apprentissage et d'évaluation, une brève recension des écrits indique que cette expression ne semble pas univoque. On retrouve, dans les écrits portant sur l'approche par compétences et la didactique, différents termes rapprochés, notamment : tâche complexe, situation-problème, situation didactique, situation a-didactique, activité, situation complexe, situation de compétences, etc. Le ministère de l'Éducation, du Loisir et du Sport (MELS) (2006) distingue, dans le *Cadre de référence pour l'enseignement secondaire*, les situations d'apprentissage et d'évaluation et les situations d'évaluation (SÉ). Même si ces termes sont peu utilisés dans la littérature spécialisée, tous les auteurs semblent s'entendre pour souligner que c'est dans l'action complexe que l'élève a l'occasion de développer ses compétences et que c'est

toujours dans l'action que l'on pourra inférer des compétences chez l'apprenant. Notre cible sera donc élargie aux situations pertinentes permettant d'évaluer des compétences, que nous appellerons des *situations de compétences*.

2.1. Écrits relatifs à la didactique des mathématiques

Nous nous attarderons, dans un premier temps, sur le concept de situation traité dans certains écrits en didactique des mathématiques. Pallascio (2005) définit la situation-problème comme une situation réaliste, se produisant en début d'apprentissage dans le but d'introduire de nouvelles connaissances. L'élève est alors amené à chercher une solution adéquate en adoptant différentes stratégies, par opposition aux problèmes d'application, dans lesquels il est plutôt amené à appliquer des connaissances déjà acquises. Brousseau (2000) a construit un modèle important en didactique des mathématiques, aujourd'hui utilisé dans d'autres disciplines. Il définit d'abord une situation comme un contexte dans lequel des utilisations particulières d'une connaissance mathématique sont considérées comme formant un système. Il définit aussi une tâche comme une action acceptée déterminée à l'intérieur d'une situation. Selon l'auteur, les situations pourraient être subdivisées en deux sous-groupes (voir la figure 4.1).

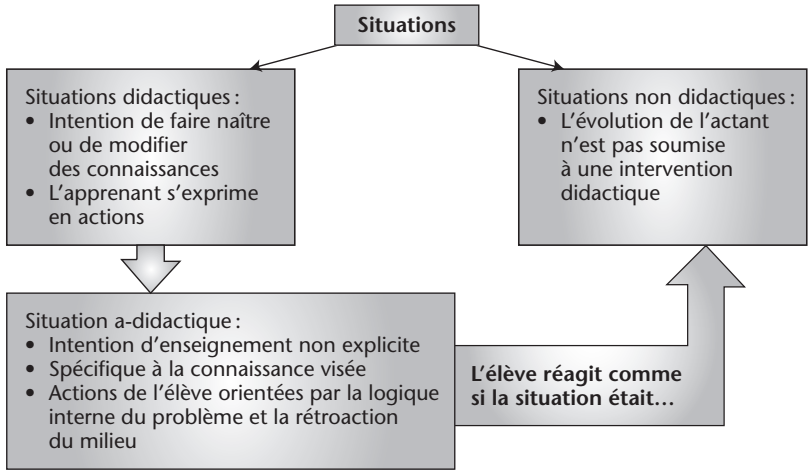


Figure 4.1
Types de situations d'après Brousseau (2000)

La logique dominante est ici celle de la découverte dans une perspective constructiviste. Les situations a-didactiques sont construites de manière à offrir une rétroaction à l'apprenant, permettant, dans un monde idéal, d'invalider les procédures erronées et de valider les procédures adéquates. En didactique des mathématiques, de telles situations supposent que les élèves mobilisent des ressources afin de produire quelque chose, ce qui nous rapproche de l'idée de situation permettant d'évaluer des compétences. La question de l'évaluation reste peu exploitée dans les différents écrits consultés.

2.2. Écrits généraux sur l'évaluation en situation

Louis (1999), reprenant les propos de Wiggins (1993) et Hart (1994), aborde le concept d'évaluation en situation authentique ainsi :

> *lorsqu'elle présente à l'élève des tâches : qui expriment des situations tirées de la vie normale ; qui sont signifiantes et motivantes pour l'élève ; et qui permettent de comprendre ou de résoudre un problème fréquemment rencontré dans la vie extrascolaire* (p. 78).

Louis mentionne également que l'évaluation en situation authentique suppose la mobilisation de ressources sous forme de savoirs, de savoir-faire et d'attitudes dans un contexte potentiel réaliste. Si les tâches supposent une intégration de connaissances acquises, alors ce sont des tâches complexes. Pour Louis, la tâche comporte une mise en situation offrant une contextualisation, les actions principales que devront entreprendre les élèves (tâches), les consignes ou précisions, les caractéristiques du résultat attendu, la définition d'un auditoire-cible (ou d'un contexte d'application) ainsi que l'instrument d'évaluation (grille d'appréciation, comportant généralement de trois à cinq niveaux).

Selon Legendre (2005), une situation pédagogique serait un contexte dans lequel se déroule un processus d'enseignement-apprentissage et une situation d'apprentissage représenterait spécifiquement une *situation pédagogique planifiée par le Sujet et qui suscite effectivement son apprentissage* (p. 1238).

Pour Scallon (2004), les situations pour l'évaluation s'inscrivent dans un continuum de complexité : situations de connaissances, d'habiletés, de stratégies et enfin de compétences. La distinction entre ces deux dernières est liée à la complexité de la situation ressentie par l'élève (nombre de ressources à mobiliser, problèmes plus ou moins bien définis). Pour cet auteur, une situation peut être soit un problème au sens strict du terme, soit une tâche ou un projet complexe. Il définit une situation-problème comme étant *toute tâche complexe, tout projet qui pose à l'élève des défis, dont celui de mobiliser des ressources* (p. 112).

Selon lui, pour permettre aux élèves de s'exercer et aux enseignants d'inférer les compétences développées, les situations de compétences présentent les caractéristiques suivantes : permettre une production attendue, être réalistes ou authentiques, demander la mobilisation d'un grand nombre de ressources (complexité) qui sont précisées et, selon les possibilités, présenter des problèmes mal définis (donc plus ouverts). Pour ce qui est des situations d'évaluation, l'élève peut les réaliser de manière autonome.

Roegiers (2003) rappelle que, dans la logique des compétences, les épreuves servant à évaluer les élèves sont constituées de situations se rapprochant de la réalité avec un certain niveau de complexité. Ces situations répondent à plusieurs conditions, les trois principales étant : correspondre à la compétence à évaluer, être significatives pour l'élève et véhiculer des valeurs positives. De plus, une épreuve d'évaluation se planifie en fonction de la compétence à évaluer et présente de nouvelles situations. Les critères d'évaluation sont utilisés au moins trois fois avant de porter un jugement (permettant de vérifier si l'élève réussit au moins deux fois sur trois). Finalement, les supports et consignes sont clairement établis pour les élèves, les indicateurs sont spécifiés et la grille d'évaluation est élaborée. Cet auteur évoque également le concept de situation-problème dans un contexte présentant un obstacle à franchir, dans une intention didactique. Une situation-problème nécessite de la part de l'élève un processus de résolution et suppose une production. Roegiers ne distingue pas les concepts de tâche et de situation, qui semblent équivalents à son sens, mais rappelle la distinction faite par Tardif (voir Roegiers, 2004) entre tâche-source et tâche-cible. Dans les deux cas, il s'agit d'une situation-problème, mais le but est fort différent. Dans le premier cas, on souhaite la construction d'un savoir alors que, dans le second, on veut provoquer une intégration. Dans la présente recherche, c'est la situation-cible que nous retiendrons pour l'évaluation des compétences.

À propos de la complexité d'une situation, Gerard (2007) distingue une situation complexe d'une situation compliquée. Dans le premier cas, la situation combine des éléments que l'élève connaît, qu'il a appris, mais dont le contexte est nouveau, tandis qu'une situation compliquée propose des savoirs et des savoir-faire nouveaux, encore peu maîtrisés.

Par opposition, une situation qui serait compliquée demanderait à l'élève de faire appel à des ressources encore peu maîtrisées par lui. En ce sens, la situation a-didactique de Brousseau serait une situation à la fois compliquée et complexe, puisque l'élève est amené, en adoptant différentes stratégies, à confronter ses conceptions *a priori* et à les modifier afin de développer de nouvelles connaissances.

Pour le ministère de l'Éducation, du Loisir et du Sport (2006), une situation d'apprentissage et d'évaluation est composée d'une problématique contextualisée ainsi que d'une ou plusieurs tâches et activités associées aux connaissances. Les activités d'apprentissage visent l'acquisition ou la restructuration de connaissances et peuvent avoir lieu à différents moments d'une situation d'apprentissage et d'évaluation. De telles activités peuvent être évaluées par différents moyens. Les tâches complexes, quant à elles, *amènent l'élève à prendre conscience des ressources dont il dispose, à choisir celles qui sont pertinentes et à les utiliser de manière efficace dans un contexte donné* (p. 11). Pour qu'une tâche soit considérée comme complexe, elle doit faire appel à la compétence, stimuler des productions de la part des élèves, permettre des démarches ou productions personnelles, présenter un problème nouveau pour l'apprenant, permettre l'évaluation d'au moins une compétence selon les critères du Programme de formation de l'école québécoise (PFÉQ) et être adaptée, selon les critères, au niveau et à la période de l'année correspondante.

Hall et Burke (2003) soulignent l'importance d'activités authentiques pour l'apprentissage et l'évaluation. Pour ces auteurs, l'évaluation formative d'activités authentiques permet à l'enseignant de porter un jugement sur les apprentissages de l'élève, intégrant ainsi l'évaluation à l'apprentissage. Ils mentionnent également les travaux de Nuttall (1987 : voir Hall et Burke, 2003), qui décrit les tâches valides dans un tel processus : les tâches concrètes à l'intérieur de l'expérience de l'individu, présentées clairement, perçues comme relevant de préoccupations de l'apprenant (un défi motivant à sa mesure). Selon ces auteurs, ce type de tâche offre la possibilité de porter un jugement plus éclairé : les tâches valorisent le processus de l'élève, permettant ainsi d'observer le raisonnement et les stratégies utilisés. Pour ce faire, il est primordial, selon les auteurs, que les intentions et les critères d'évaluation soient explicites pour l'élève et que ce dernier ait un rôle à jouer dans le dialogue évaluatif.

Pour Durand et Chouinard (2006), les différentes situations avec lesquelles les enseignants peuvent travailler varient en complexité. C'est ainsi que la situation d'apprentissage représente *les dispositifs favorisant à la fois la réalisation des apprentissages et la fonction régulatrice de l'évaluation* (p. 128). Ces situations, faites d'une série d'activités, ont pour double objectif le développement des compétences et la régulation des apprentissages. En effet, la situation d'apprentissage est structurée en trois phases : activités de préparation, activités de réalisation et activités d'intégration, qui permettent *de réaliser des apprentissages complexes et authentiques qui sont liés aux compétences disciplinaires et transversales qu'il doit maîtriser* (p. 129). Selon ces auteurs, les situations

d'apprentissage et d'évaluation possèdent les caractéristiques suivantes : signifiantes, authentiques, contextualisées, complexes, elles suscitent la mobilisation de nombreuses ressources, supposent une production unique, se réalisent en coopération, se déroulent à moyen terme et font partie d'une famille de situations d'apprentissage.

La synthèse des différents concepts liés à la situation, selon les auteurs consultés, met en évidence la structuration des composantes et leurs relations en fonction de la pertinence de cet outil pédagogique. Celle-ci est illustrée dans le tableau 4.1, selon le niveau de pertinence pour l'évaluation des compétences.

Puisque, dans le cadre du renouveau pédagogique, la notion de situation d'apprentissage et d'évaluation est utilisée, c'est le terme qui sera retenu ici pour signifier toute situation contextualisée permettant d'inférer des compétences. Suite au cadre conceptuel régissant les situations de compétences et à l'analyse, c'est-à-dire l'identification et la cueillette des données pertinentes et la synthèse, l'objectif est maintenant de mettre en évidence des caractéristiques communes qui constituent les éléments d'une grille descriptive analytique permettant de juger de la validité de différentes situations d'apprentissage et d'évaluation.

3. MÉTHODOLOGIE

Le but premier étant de produire un objet de recherche, la grille d'analyse sert de modèle et fournit un procédé nouveau (Legendre, 2005), c'est-à-dire un référentiel dans le choix ou la conception de situations d'apprentissage et d'évaluation validées. Quelques méthodologies se prêtent à ce type de démarche, dont les méthodes de l'analyse de la valeur pédagogique (AVP) conçues par Roque, Langevin et Riopel (1998) et le modèle de recherche de développement technologique de Nonnon (1993). Bien qu'ils fournissent un cadre qui s'applique à un prototype en phase de conception et de mise au point en plusieurs étapes, nous considérons que le processus d'anasynthèse de Silvern (1972 : voir Legendre, 2005) correspond davantage à notre démarche de développement. La démarche suivante a été effectuée à partir de la problématique de départ : 1) analyse du concept ; 2) synthèse et élaboration d'un prototype ; 3) simulation ; 4) modèle. Dans notre cas, la simulation s'est réalisée en deux étapes. La figure 4.2 illustre le processus d'anasynthèse de Silvern.

Tableau 4.1
Pertinence des outils pédagogiques sous l'appellation « situation pour l'évaluation des compétences », selon les auteurs consultés

COMPLEXITÉ	Sources / Pertinence (compétences)	MELS (2006) ; Durand et Chouinard (2006)	Scallon (2004)	Roegiers (2004)	Louis (1999)	Hall et Burke (2003)	Brousseau (2000) ; Pallascio (2005)
	Niveau de pertinence minimal (connaissances)	Activités de connaissances	Situation de connaissances Situation d'habiletés	Situation de classe (activité)			Exercice
	Niveau de pertinence moyen : habiletés-stratégies et/ou approche (construction des connaissances)	Tâche complexe	Situation de stratégies	Situation-problème d'apprentissage	Tâche complexe	Tâche	Situation a-didactique, situation-problème
	Niveau de pertinence maximal (compétences)	SAÉ + SÉ	Situation de compétences (tâche complexe)	Situation (tâche) cible	Situation d'évaluation	Activité authentique pour l'apprentissage et l'évaluation	

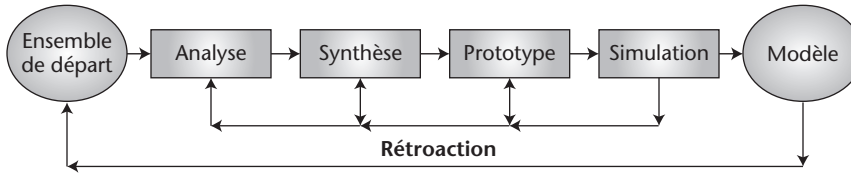


Figure 4.2
Processus d'anasynthèse de Silvern (1972 : voir Legendre, 2005)

3.1. Les étapes de l'anasynthèse

Dans un premier temps et à la suite de la recension des écrits, les données pertinentes ont été identifiées et compilées. Chacun des énoncés se rapportant au concept-clé a été classé selon la nature quadripartite du message, soit : information formelle, axiologique, praxique et explicative. Le concept de situation a été analysé en fonction de sa nature, de sa visée et de ses caractéristiques selon la technique de l'analyse de contenu. Cette distinction avait pour but de faciliter la synthèse des caractéristiques d'une situation donnée.

Dans un deuxième temps, une structuration des composantes et de leurs relations a permis de réaliser une synthèse en mettant en évidence les caractéristiques essentielles et celles qui sont secondaires. Chacune de ces caractéristiques a ensuite été représentée sous forme de critères.

Dans un troisième temps, un prototype a été élaboré sous la forme d'une grille descriptive (*rubric*) représentant ainsi la meilleure synthèse possible.

Dans un quatrième temps, le prototype a été appliqué à quelques situations en enseignement des mathématiques au primaire et au secondaire (simulation 1).

Dans un cinquième temps, une première rétroaction a été effectuée afin d'apporter des correctifs à la suite de l'utilisation de la grille et des commentaires des enseignants ; des limites ont été fixées. Ensuite, un prototype amélioré a été développé.

Dans un sixième temps, la grille améliorée a été validée par huit experts qui l'ont utilisée pour analyser des situations d'apprentissage et d'évaluation dans différentes disciplines et de niveaux d'enseignement primaire et secondaire (simulation 2). Ils ont formulé des commentaires généraux et ont critiqué les différents aspects de la grille : les critères, les échelons, les descripteurs et la clarté des énoncés. Ces

commentaires ont été analysés en relevant les points positifs, les points négatifs et les recommandations. Finalement, une dernière grille a été construite qui servira de modèle pour analyser la validité des situations d'apprentissage et d'évaluation.

3.2. Première modélisation : élaboration d'une grille analytique pour les situations d'évaluation et d'apprentissage

Tout d'abord, un consensus se dégage sur le fait qu'une situation donnée permet d'inférer des compétences seulement si elle présente une certaine complexité. Puisque la complexité est reliée, entre autres, à la mobilisation des ressources, cette dimension fait partie de la grille analytique. Dans le cas de SAÉ exceptionnelles, le transfert est possible, ajoutant à la complexité de la situation. Une situation complexe suppose aussi un contexte qui suscite l'engagement des élèves. Pour ce faire, la situation doit avoir un sens pour l'élève (signifiance), que ce soit d'après une situation authentique ou une situation réaliste. Elle est, idéalement, motivante pour l'élève tout en offrant un contexte qui représente un certain sens. Ces deux premières dimensions essentielles devraient se retrouver dans une situation d'évaluation et d'apprentissage.

Ensuite, comme le rappellent Durand et Chouinard (2006), une situation d'évaluation et d'apprentissage repose sur des compétences déterminées, tant disciplinaires que transversales, qui sont réellement développées et qui se manifestent selon une démarche d'apprentissage, et non seulement à une seule occasion. Il suffit donc que la situation permette effectivement l'évaluation des compétences et soit conforme au Programme de formation de l'école québécoise. Ainsi, on retrouve, dans la grille, une dimension de conformité au programme. Une situation d'évaluation et d'apprentissage exceptionnelle repose sur une intention éducative (en vue de conscientiser, de prendre position, par exemple), sur des compétences disciplinaires et transversales, et dévoile explicitement les ressources à mobiliser, comme le rappelle Scallon (2004). De plus, pour qu'une situation d'évaluation et d'apprentissage soit pertinente, elle doit présenter des activités ou tâches répondant à l'intention de départ pour les trois phases (préparation, réalisation et intégration de l'apprentissage).

Enfin, puisque le but ultime d'une situation d'évaluation et d'apprentissage est l'évaluation des compétences, celle-ci doit en faire partie et être bien instrumentée. L'évaluation est planifiée en présentant explicitement différents outils d'évaluation adaptés aux tâches. Les critères d'évaluation comportent idéalement une grille descriptive

ou un outil exposant la façon dont les critères ciblés sont utilisés dans la tâche, au moyen d'indices observables. On retrouve aussi, dans un souci de cohérence, des endroits où les traces des élèves pourront être observées. Enfin, pour l'évaluation de certains critères, la production attendue est idéalement personnelle et répond à l'intention exprimée.

Les choix effectués pour la grille d'analyse se veulent un exercice de logique. Les critères adoptés seront ordonnés selon un ordre de priorité qui nous semble pertinent. Tout d'abord, pour permettre d'inférer des compétences, il faut absolument que la situation offre un certain niveau de complexité et que l'élève s'engage effectivement à relever le défi. C'est ce qui explique que la complexité et la signifiance sont regroupées, puisqu'elles sont des conditions nécessaires dans le contexte d'inférence de compétences. Ensuite, si une telle situation ne répondait pas ou presque aux exigences particulières du programme de formation ou du référentiel de compétences, elle pourrait difficilement être utilisée. C'est la raison pour laquelle on trouve ensuite la dimension de conformité au programme. Puis, pour que la situation soit pertinente, un maximum de tâches qui la composent doivent être en lien avec l'intention annoncée. Par contre, les éléments relatifs aux outils d'évaluation sont très importants, mais ils n'invalident pas la situation, qui pourrait quand même permettre l'émergence et l'évaluation des compétences. Toutefois, en laissant l'évaluateur construire lui-même ses propres outils, la place laissée au jugement de chacun pourrait occasionner un problème de fiabilité. Par conséquent, il est nettement préférable que tous les outils soient inclus, bien détaillés, incluant des traces d'élèves prévues pour l'évaluation.

3.3. Mise à l'épreuve de la grille

Afin de procéder à une mise à l'épreuve de la grille descriptive analytique, quatre situations d'évaluation et d'apprentissage et une situation d'évaluation proposées à des élèves de différents niveaux en mathématiques ont été examinées. Pour chacune de ces situations, la grille d'analyse a été remplie en indiquant le niveau atteint par chaque critère, suivi d'un bref commentaire mentionnant les points forts et les points faibles, puis un jugement global a été rendu. Le tableau 4.2 illustre la grille d'analyse utilisée pour la simulation 1.

Tableau 2
Grille descriptive analytique pour l'évaluation de situations d'apprentissage et d'évaluation (simulation 1)

La situation…		Manifestations observables			
		A : SAÉ exceptionnelle	**B : SAÉ pertinente**	**C : SAÉ à retravailler**	**D : SAÉ à refaire**
Critères d'évaluation	COMPLEXITÉ des tâches proposées	La situation propose un défi qui exige la mobilisation d'un ensemble de ressources permettant le transfert.	La situation propose un défi qui exige la mobilisation d'un ensemble de ressources.	La situation propose une série d'activités interreliées à une thématique.	La situation propose une série d'activités.
	SIGNIFIANCE de la situation	La situation propose un contexte réaliste ou authentique très motivant, suscitant l'engagement de l'élève.	La situation propose un contexte réaliste ou authentique motivant.	La situation est partiellement contextualisée et peu motivante.	La situation est décontextualisée.
	CONFORMITÉ du contenu	La planification de l'apprentissage comporte explicitement tous les éléments suivants : DGF (intention éducative), CT, CD, ressources internes (savoirs, savoir-faire et savoir-être), ressources externes.	La planification de la situation comporte explicitement la plupart des éléments suivants : intention pédagogique, CT, CD, ressources internes (savoirs, savoir-faire et savoir-être), ressources externes.	La planification de la situation identifie les éléments suivants : CT, CD, ressources internes (savoirs, savoir-faire et savoir-être), ressources externes.	La planification de la situation identifie certains des éléments suivants : CT, CD, ressources internes (savoirs, savoir-faire et savoir-être), ressources externes.

	PERTINENCE de la structure	Toutes les activités sont en lien avec l'intention, et ce, pour chacune des phases : préparation, réalisation et intégration.	Les activités sont en lien avec l'intention, et ce, pour chacune des phases : préparation, réalisation, intégration.	Le lien entre certaines activités et l'intention est absent.	On ne retrouve que des activités de préparation.
Critères d'évaluation	VALIDITÉ de l'évaluation des compétences	La situation propose des outils d'évaluation pertinents, variés pour les compétences ciblées.	La situation prévoit des outils d'évaluation pertinents pour l'évaluation des compétences transversales et disciplinaires prévues.	La situation prévoit des outils d'évaluation pour l'évaluation des compétences prévues.	La situation ne prévoit pas d'outils d'évaluation ou prévoit des outils d'évaluation ne permettant pas l'évaluation des compétences visées.
		Les critères d'évaluation ciblés sont reformulés en éléments observables et les traces des élèves sont prévues.	Les critères d'évaluation ciblés sont présents et les traces des élèves sont prévues.	Certains critères d'évaluation ciblés sont présents et les traces des élèves sont prévues.	Les traces des élèves sont prévues.
		La production attendue est identifiée, unique et en lien avec l'intention éducative.	La production attendue est identifiée, unique et en lien avec l'intention pédagogique.	La production attendue est unique.	Une production est attendue dans le cadre de la situation.

4. PRÉSENTATION DES RÉSULTATS APRÈS LA PREMIÈRE SIMULATION

Globalement, une situation d'évaluation et d'apprentissage qui se classerait au niveau A (exceptionnelle) pour chacune des dimensions de la grille serait considérée exemplaire. Pour qu'une situation soit pertinente, elle devrait atteindre au moins le niveau B pour chaque critère, c'est-à-dire ne comporter aucun élément important à retravailler pour inférer des compétences du programme de formation. Une situation d'évaluation et d'apprentissage qui se situerait au niveau C (à retravailler) pour la conformité du contenu, la pertinence de la structure ou la validité de l'évaluation des compétences serait à retravailler seulement si elle est suffisamment complexe pour inférer des compétences et comporte un minimum de signifiance. Dans le cas d'une situation d'évaluation, comme celle utilisée dans le cadre de notre analyse, la signifiance ne serait pas un critère exigé. Une situation d'évaluation et d'apprentissage qui se verrait attribuer le niveau D ou C pour la complexité et la signifiance serait à refaire complètement.

4.1. Analyse des situations d'évaluation et d'apprentissage

4.1.1. *Situation d'évaluation et d'apprentissage 1 : Les bienfaits du dieu soleil, 3e année primaire*

La grille analytique a été, dans le cas de cette situation d'évaluation et d'apprentissage, très facile à appliquer puisque les critères sont presque tous situés au plus haut niveau ; il n'y a pas eu d'hésitation. Les libellés étaient parfaitement clairs. Par contre, pour le second aspect de la validité, les critères d'évaluation ne pouvaient être tous présents puisqu'ils n'étaient pas prévus pour la compétence transversale. De plus, le terme « présence » peut porter à confusion entre la présence en termes d'outils, d'indices observables ou de planification des traces. Il faudra voir dans les autres situations d'évaluation et d'apprentissage si cet aspect est plus clair ou s'il faudra le préciser.

4.1.2. *Situation d'évaluation et d'apprentissage 2 : Un débat organisé, 1re année du 1er cycle du secondaire*

Cette situation serait pratiquement à écarter : elle se classe dans la catégorie inférieure et serait probablement impossible à récupérer. Le fait que les trois premiers critères ne sont pas satisfaisants en fait une situation d'évaluation et d'apprentissage qui permettra difficilement une démarche d'inférence des compétences, étant limitée à certains aspects

seulement de la compétence *Raisonner à l'aide de concepts et de processus mathématiques* au primaire ou *Déployer un raisonnement mathématique* au secondaire.

4.1.3. *Situation d'évaluation et d'apprentissage 3 : Un peu d'ordre ! 2e année du 1er cycle du secondaire*

Dans une perspective plus globale, cette situation d'évaluation et d'apprentissage ne serait pas à conserver, même si certaines idées pourraient inspirer une autre situation d'évaluation et d'apprentissage, puisqu'elle ne comporte pas les conditions nécessaires pour une démarche d'inférence de compétences.

4.1.4. *Situation d'évaluation et d'apprentissage 4 : Le piratage… combien ça coûte ? 1re année du 2e cycle du secondaire (3e secondaire)*

Cette situation d'évaluation et d'apprentissage comporte certaines composantes pertinentes et d'autres exceptionnelles ou à retravailler. Globalement, il serait difficile de recommander cette situation d'évaluation et d'apprentissage telle quelle, étant donné certaines lacunes, notamment l'absence d'évaluation de la compétence transversale et le manque d'activités pour chacune des phases. Il faudrait que certains éléments soient complétés et que sa structure soit retravaillée, mais l'idée générale demeure pertinente.

Globalement, si un enseignant souhaitait utiliser cette situation d'évaluation comme pratique en classe, il pourrait effectivement évaluer les compétences disciplinaires des élèves, mais des adaptations mineures seraient à envisager.

4.2. Discussion : aspects du prototype à retravailler

Bien que le prototype mis au point se soit avéré un outil d'analyse intéressant, permettant de mettre en valeur les points forts et les points à retravailler des situations d'évaluation et d'apprentissage et situation d'évaluation analysées, des correctifs se sont avérés nécessaires concernant la terminologie utilisée et certains aspects plus fondamentaux. Quelques éléments ont parfois été difficiles à juger. Mentionnons qu'en général, les niveaux supérieur (A) et inférieur (D) ont été relativement faciles à utiliser, mais que des réajustements se sont avérés nécessaires. Le tableau 4.3 présente la synthèse des réajustements à faire au prototype, par rapport aux situations d'évaluation et d'apprentissage analysées.

Tableau 4.3
Rétroactions à apporter au prototype suite à la simulation 1

APPRÉCIATION GLOBALE	ASPECTS DU PROTOTYPE À RETRAVAILLER
SAÉ 1 B – pertinente, mais l'ajout d'outils d'évaluation serait nécessaire	VALIDITÉ : Modifier le terme « présence » des critères ; il serait suffisant de mentionner qu'ils sont ciblés.
SAÉ 2 D – à refaire	C1 – COMPLEXITÉ : Les niveaux C et D sont à modifier, pour inclure le cas où une seule activité est proposée.
	C5 – VALIDITÉ : Revoir la gradation du 1er aspect en fonction des compétences disciplinaires et transversales. Au niveau C, seules les compétences disciplinaires seront mentionnées.
SAÉ 3 D – à refaire, mais certains aspects pourraient resservir dans une nouvelle SAÉ, comme les problèmes de réinvestissement	C3 – CONFORMITÉ : Ajouter un élément de cohésion entre les activités et la planification, et ce pour les niveaux A et B.
	– Au niveau B, enlever l'expression *la plupart des éléments…* étant donné que les éléments mentionnés ensuite devraient tous se retrouver dans la planification pour l'attribution de cet échelon.
	C5 – VALIDITÉ : Retravailler le 2e aspect de la validité pour inclure la cohérence critères-traces-activités proposées. Les activités doivent bel et bien permettre l'évaluation des critères annoncés.
SAÉ 4 C – avec un bon potentiel ; à retravailler	C2 – SIGNIFIANCE : Il est difficile de juger de la motivation et du niveau d'engagement des élèves sans d'autres traces ou commentaires d'enseignants. On pourrait peut-être ajuster les niveaux A et B par le potentiel d'intérêt des thèmes auprès de la clientèle ciblée.
	C3 – CONFORMITÉ : Puisque les savoir-être ne sont pratiquement jamais détaillés dans les SAÉ observées, il pourrait être intéressant de retirer la liste détaillée des ressources internes pour ne laisser que ce dernier terme.
	C4 – PERTINENCE : Inclure dans le libellé la possibilité qu'une seule activité soit proposée.
	C5 – VALIDITÉ : Au 1er aspect, spécifier en C qu'il s'agit de compétences disciplinaires.
SE 5 Grille non adaptée	Dans le cas de la SÉ qui constitue une épreuve ministérielle, il serait intéressant de créer une grille adaptée, dans laquelle il faudrait modifier la signifiance, la conformité et la validité selon la fonction et les caractéristiques des SÉ. Une telle grille pourrait faire l'objet d'un travail ultérieur.

Par contre, certains aspects peuvent fausser l'analyse. Celle-ci, d'ailleurs, n'est pas exhaustive en raison de l'absence de copies types des élèves, qui auraient été précieuses pour juger de la validité de l'évaluation et de la motivation des élèves, jusqu'à un certain point. Des commentaires d'enseignants qui ont testé ces situations d'évaluation et d'apprentissage auraient également pu fournir des informations pertinentes. Certains critères dépendent de la clientèle ciblée ; par exemple, la complexité peut varier selon le niveau de maîtrise des élèves à l'intérieur d'un groupe-classe. Des informations complémentaires telles que les antécédents des élèves et leur niveau de familiarité avec le sujet et avec les concepts auraient été utiles dans le cadre de cette analyse. Puis, le critère de validité comportait trois éléments qui pourraient théoriquement mener à trois niveaux différents, ce qui serait plus difficile à interpréter pour porter un jugement valide et fiable. Ces trois aspects pourraient peut-être être regroupés ou scindés pour en former de nouveaux. Enfin, le petit nombre de situations utilisées pour la mise à l'épreuve de la grille représente également une limite importante à la validation la grille dans le cadre de cette première simulation.

D'autres critères auraient pu être utilisés. La souplesse et l'adaptabilité de la situation, les possibilités de différenciation et de régulation, l'efficacité de l'organisation auraient pu figurer dans la grille, mais nous pourrions les retenir comme critères de perfectionnement éventuels de notre analyse.

5. PRÉSENTATION DES RÉSULTATS APRÈS LA DEUXIÈME SIMULATION

5.1. Prototype amélioré

La grille améliorée se veut une deuxième tentative dans la démarche de validation. Cette grille, construite selon les observations et l'évaluation faites à partir des quatre situations d'évaluation et d'apprentissage analysées précédemment, tient compte des différents commentaires déjà formulés à leur sujet. L'analyse de nouvelles situations d'évaluation et d'apprentissage par des experts permettra de porter un autre regard sur le prototype et de le bonifier pour en arriver au modèle définitif.

Huit experts se sont prononcés sur la pertinence de la grille améliorée en l'utilisant pour analyser des situations d'évaluation et d'apprentissage de différentes disciplines enseignées au primaire ou au secondaire. Cinquante-deux commentaires ont été analysés, soit une moyenne de six commentaires par expert. Ces commentaires

représentaient un point positif (dans 73 % des cas) ou un point faible (dans 15 % des cas), ou proposaient des améliorations sous forme de recommandations (dans 12 % des cas). Ces commentaires sont regroupés selon qu'ils concernent : 1) la grille en général ; 2) les critères d'évaluation ; 3) les échelons ; 4) les descripteurs ; 5) la formulation générale des énoncés.

5.1.1. *La grille en général (22 commentaires)*

Un certain consensus se rapportant à l'objectif général se dégage de l'analyse. La grille permet tout d'abord d'identifier les éléments qui sont plus ou moins développés et, ainsi, d'apporter des corrections et des ajustements à la situation d'évaluation et d'apprentissage analysée. Elle permet aussi de dresser un bon portrait global en donnant une vue d'ensemble des éléments recherchés et en validant les caractéristiques d'une bonne situation d'évaluation et d'apprentissage. De plus, la grille semble très bien construite ; elle est concise tout en étant complète, ce qui permet d'effectuer efficacement et rapidement l'analyse de la situation d'apprentissage et d'évaluation. Finalement, ce serait une bonne grille de régulation en cours de tâche. Certains experts considèrent par contre que la grille n'indique pas si certains éléments précis sont manquants et qu'elle ne permet pas un approfondissement pertinent en ne suscitant pas suffisamment de régulation de la part de l'enseignant. Cet aspect sera repris dans la recommandation des critères d'évaluation.

5.1.2. *Les critères d'évaluation de la grille (14 commentaires)*

Une majorité d'experts s'entendent pour constater que les critères sont bien définis, qu'ils ont tous une dimension clairement identifiée et qu'ils sont variés. Les critères d'évaluation sont jugés pertinents, car ils se rapportent aux caractéristiques les plus importantes d'une situation d'évaluation et d'apprentissage, à savoir : la signifiance, la contextualisation et la complexité. Quant à la conformité du contenu et aux trois phases de l'action en classe, on les retrouve comme il se doit. Finalement, on remarque qu'une attention particulière a été portée aux outils d'évaluation, aux outils de consignation des résultats, de même qu'à la production finale attendue. Un expert a trouvé certains critères difficiles à comprendre, et des recommandations ont été faites afin d'ajouter un critère portant sur la régulation de l'enseignement et de reformuler le dernier critère.

5.1.3. *Les échelons de la grille (3 commentaires)*

Peu d'experts se sont prononcés sur la pertinence des échelons. Deux commentaires sont positifs, à savoir que les échelons permettent de situer très facilement la situation d'évaluation et d'apprentissage selon les quatre degrés donnés et que cette gradation est cohérente et juste. Un expert déplore le fait que le seuil de réussite n'est pas clairement établi. Pourtant, l'échelon C (situation d'évaluation et d'apprentissage à retravailler) correspond bien à un travail partiellement réussi.

5.1.4. *Les descripteurs de la grille (7 commentaires)*

On retrouve ici l'unanimité quant à la pertinence des descripteurs. Facilement identifiables, les descripteurs correspondent bien aux caractéristiques de la situation d'évaluation et d'apprentissage. De plus, l'observation du niveau supérieur des descripteurs (A) fournit des pistes d'adaptation de la situation, car elle donne de l'information sur les améliorations à apporter. En effet, on souligne que les descripteurs permettent de voir les éléments qui seraient à ajouter pour rendre les situations d'évaluation et d'apprentissage plus pertinentes, voire exceptionnelles. Une recommandation est formulée par un des experts concernant les outils d'autoévaluation et les grilles qui devraient être utilisées dans les tâches proposées.

5.1.5. *La formulation des énoncés de la grille (6 commentaires)*

En général, on trouve que la tâche de l'enseignant est facilitée par la formulation précise des énoncés. Les descripteurs sont bien précis et l'évaluation des compétences est bien détaillée, de façon claire et nette. Par contre, il semble que les descripteurs du dernier critère soient moins clairs pour l'un des experts, donc plus difficiles à utiliser.

5.2. Discussion

Il semble qu'au départ, les experts n'aient pas tous envisagé les caractéristiques d'une situation d'évaluation et d'apprentissage de la même façon. C'est pour cette raison que l'on retrouve surtout des commentaires positifs qui font l'unanimité et quelques commentaires négatifs ou recommandations de la part d'un seul expert. Globalement, la grille décrit bien ce qu'elle est censée analyser et le nombre d'échelons est pertinent. L'échelon A indique que l'enseignant peut utiliser la situation d'évaluation et d'apprentissage telle quelle : elle comprend tous les éléments nécessaires et ceux-ci sont jugés adéquats. De plus, cet échelon permet de saisir les changements à apporter afin d'améliorer

l'un ou l'autre des critères. Ainsi, la grille peut servir en même temps d'outil de rétroaction efficace. Nous croyons qu'il n'est pas nécessaire de proposer des mécanismes de régulation dans la situation d'évaluation et d'apprentissage, car celle-ci se fait souvent oralement et spontanément lorsque l'enseignant est en interaction avec ses élèves. Par contre, cette indication pourrait être ajoutée au critère de la planification de l'évaluation. Ce dernier sera, de plus, reformulé afin de préciser davantage certains éléments qui paraissaient moins clairs aux yeux de quelques experts.

L'analyse poussée de la documentation spécialisée nous a permis de trouver les caractéristiques importantes d'une SAÉ, et celles-ci se sont révélées judicieuses selon les experts. En effet, les descripteurs ont permis de juger de la qualité d'une situation d'évaluation et d'apprentissage, tout en confirmant la pertinence de la grille d'analyse. Plusieurs auteurs (Arter et Chappuis, 2007 ; Durand et Chouinard, 2006 ; Laurier et colla., 2005 ; Roegiers, 2004 ; Scallon, 2004 ; Wiggins, 1998) mettent en évidence les qualités d'une bonne grille descriptive, notamment : la pertinence des critères, la qualité des descripteurs (l'univocité, l'observabilité et la qualité des observations présentées), les distinctions tranchées entre les échelons (ils sont distincts, indépendants les uns des autres et ne se recoupent jamais) ainsi que la cohérence (concernant la structure et la constance de la progression de même que l'organisation de l'information). C'est ce qui a été observé par les experts et qui sera utilisé dans l'élaboration du modèle, étape finale du processus d'anasynthèse de Silvern.

6. CONCLUSION

Ce projet de recherche avait pour but la construction et l'expérimentation d'une grille d'analyse descriptive d'un outil privilégié d'évaluation dans le cadre de l'approche par compétences : la situation d'évaluation et d'apprentissage. Bien que son appellation ne fasse pas consensus, cet instrument avait été sélectionné parce qu'il représente le point de départ d'une évaluation des compétences, une sorte de condition nécessaire (mais non suffisante) pour une démarche d'inférences des compétences.

Le prototype construit à partir des résultats des études réalisées a permis l'évaluation des situations d'évaluation et d'apprentissage, mais certains éléments ont dû être retravaillés dans le cadre d'une grille modifiée. Par contre, à cause des différences de vues et de caractéristiques entre les situations d'évaluation et d'apprentissage et les situations d'évaluation, notre grille s'est avérée moins efficace pour le

traitement des situations d'évaluation et n'a pas été réutilisée dans ce contexte. Une nouvelle grille mieux adaptée aux caractéristiques d'une situation d'évaluation pourrait être plus pertinente.

L'utilisation de notre grille a permis de tirer certaines conclusions quant aux situations traitées, ce qui a rendu possible la rédaction d'une deuxième version du prototype, qui, cette fois, a été validée par huit experts indépendants. Les commentaires de ceux-ci ont permis de consolider notre grille et de produire un modèle. Ce dernier pourrait être utilisé à plus grande échelle et ainsi profiter aux différents acteurs du milieu scolaire qui élaborent ou utilisent des situations d'apprentissage et d'évaluation. Dans ce cas-ci, la méthodologie de Silvern a été d'une grande utilité pour produire le modèle. Une des limites de ce modèle repose sur le manque de précision de chacun des critères, soit les caractéristiques d'une situation d'apprentissage et d'évaluation. Nous avons aussi indiqué qu'il serait pertinent d'analyser des copies types d'élèves afin de vérifier les manifestations possibles de la compétence et de placer la situation d'apprentissage et d'évaluation dans une famille de situations pour observer la progression des élèves dans leur apprentissage. Selon Scallon (2004), un des avantages de l'utilisation des échelles descriptives réside dans la plus grande justesse des évaluations. On pourrait, dès lors, utiliser notre modèle de grille d'analyse pour que différents experts évaluent un même lot de situations d'apprentissage et d'évaluation et, ainsi, avoir un jugement qui traduirait son haut niveau de concordance. Ces différents aspects font l'objet d'une nouvelle recherche subventionnée par le Fonds québécois de recherche sur la culture et la société (FQRSC), dont les résultats seront disponibles en 2013.

RÉFÉRENCES

Arter, J. A. et Chappuis, J. (2006). *Creating and recognizing quality rubrics*. Portland, Oregon : Educational Testing Service.

Brousseau, G. (2000). *Éducation et didactique des mathématiques*. Communication au Congrès d'Aguas Calientes, Mexique. Article paru en espagnol dans la revue mexicaine *Educación matemática*, *12*, 5-39.

Durand, M.-J. et Chouinard, R. (2006). *L'évaluation des apprentissages, de la planification de la démarche à la communication des résultats*. Montréal, Québec : Hurtubise HMH.

Gerard, F.-M. (2007). La complexité d'une évaluation des compétences à travers des situations complexes : nécessités théoriques et exigences du terrain. Dans M. Ettayebi, R. Opertti et P. Jonnaert (dir.), *Logique de compétences et développement curriculaire : débats, perspectives et alternative pour les systèmes éducatifs*. Paris, France : L'Harmattan.

Hall, K. et Burke, W. M. (2003). *Making formative assessment work*. Londres, Royaume-Uni : Open University Press, McGraw-Hill Education.

Hart, D. (1994). *Authentic assessment : a handbook for educators*. New York, New York : Addison-Wesley.

Laurier, M. D., Tousignant, R. et Morissette, D. (2005). *Les principes de la mesure et de l'évaluation des apprentissages* (3e édition). Montréal, Québec : Gaëtan Morin.

Legendre, R. (2005). *Dictionnaire actuel de l'éducation* (3e édition). Montréal, Québec : Guérin.

Louis, R. (1999). *L'évaluation des apprentissages en classe : théorie et pratique*. Laval, Québec : Études Vivantes.

Ministère de l'Éducation, du Loisir et du Sport (2006). *L'évaluation des apprentissages au secondaire. Cadre de référence*. Direction générale de la formation des jeunes. Québec, Québec : Gouvernement du Québec.

Nonnon, P. (1993). *Proposition d'un modèle de recherche de développement technologique en éducation. Regards sur la robotique pédagogique*. Paris, France : Publications du service de technologie de l'éducation de l'Université de Liège et de l'Institut national de recherche pédagogique. Technologies nouvelles et éducation.

Pallascio, R. (2005). Les situations-problèmes : un concept du nouveau programme de mathématique. *Vie pédagogique*, *136*, 32-35.

Rocque, S., Langevin, J. et Riopel, N. (1998). L'analyse de la valeur pédagogique au Canada. *La valeur des produits, procédés et services, 76*, 6-11.

Roegiers, X. (2004). *L'école et l'évaluation. Des situations pour évaluer les compétences des élèves*. Bruxelles, Belgique : De Boeck.

Roegiers, X. (2003). *Des situations pour intégrer les acquis scolaires*. Bruxelles, Belgique : De Boeck Université.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Saint-Laurent, Québec : Éditions du Renouveau pédagogique inc.

Wiggins, G. (1993). *Educative assessment : Designing assessments to inform and improve student performance*. San Francisco, Californie : Jossey-Bass.

Partie 2

LE JUGEMENT ET L'ARGUMENTATION DE LA VALIDITÉ EN ÉVALUATION

Chapitre 5

Analyse de l'argumentation de la validité des inférences d'évaluation dans les politiques institutionnelles d'évaluation des apprentissages des établissements d'enseignement collégial québécois

Karine Paquette-Côté et Gilles Raîche

La présente recherche de nature exploratoire constitue une première tentative de validation de la structure d'argumentation interprétative de Kane (2006) par l'application de cette structure à l'analyse de politiques institutionnelles d'évaluation des apprentissages (PIEA) du réseau collégial québécois. Une analyse de contenu de politiques institutionnelles d'évaluation des apprentissages (PIEA) a été réalisée à partir de la structure d'argumentation interprétative de Kane (2006). Cette recherche a permis de formuler des hypothèses quant à l'exhaustivité, à l'exclusivité et à la pertinence des catégories du modèle de Kane (2006) dans le contexte de l'évaluation des apprentissages à l'enseignement collégial. Elle a aussi permis de proposer l'ajout de deux arguments supplémentaires à la structure d'argumentation initiale de façon à chercher à assurer la crédibilité de l'inférence d'évaluation aux yeux des acteurs concernés.

1. INTRODUCTION

Dans l'enseignement supérieur, les apprentissages devraient être authentiques et significatifs puisqu'ils sont en rapport direct avec les actes et la profession future ou actuelle des étudiants (Gipps, 1994 ; Scallon, 2004 ; Wiggins, 1993). Au plan de l'évaluation,

qu'il s'agisse de celle des apprentissages traditionnels ou de celle des compétences, il est aujourd'hui reconnu que le processus d'évaluation est toujours limité quant à son degré d'authenticité face à la tâche réelle (Raîche, 2008). Le processus d'évaluation demeure un exercice de simulation durant lequel la performance de l'étudiant se distingue nécessairement de sa performance en contexte réel. À ce propos, Mucchielli (1971, p. 30-31) rappelle qu'un test est une mise à l'épreuve expérimentale qui correspond à une situation simulée. La correspondance entre la performance en contexte réel et le résultat obtenu à un test fait principalement appel à la notion de validité.

Traditionnellement, la validité d'une évaluation a souvent été définie comme la capacité d'un test à bien mesurer ce qu'il est censé mesurer. En 1951, Cureton définissait la validité comme ayant deux dimensions, la pertinence et la fiabilité (*relevance* et *reliability*), et comme étant la corrélation entre le score obtenu au test et le *vrai* score critérié (*« true » criterion score*) (p. 623). En 1966, l'American Psychological Association propose trois types de validité : la validité de construit, la validité de contenu et la validité critériée, cette dernière étant prédictive ou concomitante (Cronbach, 1971). Cette conception de la validité ayant été développée pour la mesure en psychologie, Cronbach (1971) souligne le besoin de définir la validité en fonction de son utilisation et de son interprétation en éducation. La validité est surtout associée à la validité d'un test, mais on accorde plus d'importance à l'interprétation. Cronbach (1971, p. 447) mentionne d'ailleurs que parce que chaque interprétation a son propre degré de validité, on ne peut jamais arriver à la conclusion qu'un test en particulier est valide[1]. Il ajoute que tous les aspects et tous les détails d'une procédure de mesure peuvent influencer la performance et donc ce qui est mesuré (p. 449). Cronbach place la validité dans un contexte de validation, laquelle correspond au processus visant à évaluer la précision des prédictions ou des inférences réalisées sur la base du résultat obtenu à un test. En 1989, Messick définit la validité comme jugement évaluatif intégré du degré auquel la preuve empirique et le rationnel théorique soutiennent la justesse et la pertinence des inférences et des actions basées sur les résultats d'une évaluation[2]. Cette définition suppose une conception de la validité en tant qu'argument, telle que l'a introduite Cronbach (1980), qui décrit la validation du jugement évaluatif comme un processus rhétorique dans lequel l'évaluateur doit justifier son juge-

1. « *Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test "is valid"* » (Cronbach, 1971, p. 447).
2. « *Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the* adequacy *and* appropriateness *of inferences and actions based on test scores or other modes of assessment* » (Messick, 1989, p. 13).

ment par la présentation d'arguments réalistes fondés sur des preuves empiriques. Deux modèles principaux ont été développés en parallèle dans les années 1990 et 2000 en accord avec cette conception de la validité : l'un est plutôt centré sur la preuve empirique et l'autre, sur le rationnel théorique. Le premier est une approche de la conception de l'évaluation basée sur les faits probants, le modèle ECD (*evidence-centred design*) développé par Mislevy et ses collaborateurs (Mislevy, Almond et Lukas, 2004 ; Mislevy et Haertel, 2006a, 2006b ; Mislevy, Steinberg, Almond et Lukas, 2006). Le second est une structure d'argumentation interprétative développée par Kane (1992, 2004, 2006), qui peut aussi être considérée comme un modèle méthodologique pouvant servir à la validation des arguments interprétatifs en évaluation des apprentissages. Le modèle de Kane et celui de Mislevy et de ses collaborateurs, ainsi que celui de Toulmin (1964) et les travaux de Messick (1989) qui les ont inspirés, sont d'ailleurs aujourd'hui des références incontournables en ce qui concerne le concept de la validité en mesure et évaluation (Lissitz, 2009). Toutefois, la mise à l'épreuve empirique de ces modèles théoriques reste encore à faire.

La recherche à l'origine du présent chapitre visait à examiner la question suivante : comment peut-on assurer la validité de l'interprétation des résultats de l'évaluation des apprentissages des étudiants à l'enseignement supérieur ? Cette exploration a été réalisée par une application de la structure d'argumentation interprétative de Kane (2006) à l'analyse du contenu des politiques institutionnelles d'évaluation des apprentissages (PIEA) qui orientent les pratiques en évaluation des apprentissages à l'enseignement supérieur et plus particulièrement à l'enseignement collégial.

2. CONTEXTE THÉORIQUE

La Commission d'évaluation de l'enseignement collégial (CEEC) fut créée en 1993 pour contribuer au développement de la qualité, de la crédibilité et de la reconnaissance de la formation dispensée dans les établissements d'enseignement collégial québécois. La Commission d'évaluation de l'enseignement collégial québécois devait préciser les mécanismes d'évaluation des pratiques institutionnelles tant au plan de l'évaluation des apprentissages que de celle des programmes d'études. Dès l'automne 1994, chaque établissement d'enseignement collégial du Québec était dans l'obligation d'adopter les mesures suivantes (CEEC, 1994a, p. 4) :

– *définir et appliquer une politique institutionnelle d'évaluation des apprentissages (PIEA) ;*

- *définir et appliquer une politique institutionnelle d'évaluation des programmes d'études (PIEP);*
- *soumettre à la Commission, pour des fins d'évaluation, sa politique institutionnelle d'évaluation des apprentissages (PIEA) et sa politique institutionnelle d'évaluation des programmes d'études (PIEP);*
- *collaborer avec la Commission à l'évaluation de la mise en œuvre des programmes d'étude qu'il dispense.*

Dans le domaine de l'évaluation des apprentissages, ce sont les politiques institutionnelles d'évaluation des apprentissages (PIEA) qui permettent à la Commission d'évaluation de l'enseignement collégial de porter un jugement sur la qualité de l'évaluation des apprentissages des étudiants dans les établissements d'enseignement collégial. Une politique institutionnelle d'évaluation des apprentissages est :

> [*u*]*n instrument qui permet à un collège d'orienter, d'encadrer et de soutenir les activités reliées à l'évaluation des apprentissages. Elle établit les objectifs poursuivis par le collège, les principes et les valeurs qui orientent les actions, ainsi que les responsabilités de tous les groupes concernés* (Conseil des collèges, 1992, p. 38).

L'évaluation laisse une marge d'autonomie aux enseignants, mais elle relève de la responsabilité institutionnelle, c'est-à-dire que l'établissement est responsable de soutenir et d'appuyer les enseignants dans l'exercice de leurs fonctions d'évaluateurs. Les moyens que doivent mettre en place les établissements à cette fin doivent être spécifiés par eux dans les politiques institutionnelles d'évaluation des apprentissages. Chaque établissement d'enseignement collégial doit donc établir et appliquer sa propre politique institutionnelle d'évaluation des apprentissages, permettre à la Commission d'évaluation de l'enseignement collégial de procéder à leur évaluation et appliquer les modifications qu'elle recommande. Selon la Commission d'évaluation de l'enseignement collégial (1994b, p. 11), une politique institutionnelle d'évaluation des apprentissages doit comporter les composantes essentielles suivantes : les finalités et les objectifs, les moyens, le partage des responsabilités et, enfin, les moyens et les critères de l'autoévaluation de l'application de la politique. Cette dernière composante recouvre les processus et les actions prévus par l'établissement d'enseignement collégial pour procéder à l'évaluation de sa propre politique institutionnelle d'évaluation des apprentissages. La Commission d'évaluation de l'enseignement collégial (*ibid.*) suggère que cette autoévaluation soit réalisée en tenant compte des critères suivants : *la conformité de l'application avec le texte de la politique, l'efficacité de cette application pour garantir la qualité de l'évaluation des apprentissages et l'équivalence de l'évaluation des apprentissages pour contribuer à en assurer l'équité.* Selon le deuxième critère, les établissements doivent veiller à ce que leur

politique institutionnelle d'évaluation des apprentissages permette de garantir la qualité de l'évaluation des apprentissages. Différentes études faites sur l'évaluation des apprentissages fournissent des moyens de respecter ce critère, notamment en ce qui concerne la notion de validité de l'évaluation.

La fonction première de l'évaluation est de porter un jugement, lequel doit être crédible et cohérent avec les inférences issues de la procédure de mesure. Ces inférences sont produites à partir d'évidences elles-mêmes produites à partir d'une procédure de mesure (Cronbach, 1971). Cronbach, Linn, Brennan et Haertel (1997, p. 376) rappellent que la mesure consiste à faire l'évaluation d'un échantillon de performances. Ce qui nous intéresse dans l'évaluation en éducation, c'est le niveau de compétence d'un étudiant dans un domaine de compétences et pas seulement dans un échantillon restreint de performances. Cronbach et collab. précisent que le domaine de généralisation, c'est-à-dire la question de savoir jusqu'à quel point et à quel ensemble d'observations les résultats de la mesure peuvent être généralisés, dépend des utilisations et des interprétations qui seront faites de ces résultats.

Les preuves empiriques ne peuvent à elles seules permettre la validation du jugement évaluatif. Comme on l'a vu dans l'introduction, pour Messick (1989), la validité est un jugement évaluatif intégré du degré auquel la preuve empirique et le rationnel théorique supportent la justesse des inférences et des actions fondées sur les résultats d'une évaluation. Les évidences ou preuves empiriques doivent faire partie intégrante d'un raisonnement logique (Cronbach, 1980). Cronbach décrit la validation du jugement évaluatif comme un processus rhétorique dans lequel l'évaluateur doit justifier son jugement par la présentation d'arguments réalistes fondés sur des preuves empiriques. En 1988, il propose de fonder la validation de l'interprétation et de l'utilisation des résultats des évaluations sur la logique de l'argument évaluatif (Cronbach, 1988). S'appuyant sur cette proposition et sur la structure de l'argumentation élaborée par Toulmin (1964), Kane (2006) propose une structure d'argumentation interprétative, qui peut aussi être considérée comme un modèle méthodologique, pouvant servir à la validation des arguments interprétatifs en évaluation des apprentissages.

La figure 5.1 illustre la structure d'argumentation interprétative pour l'interprétation d'un trait développée par Kane (2006).

Selon Kane, un *trait* est une disposition à agir ou à performer d'une certaine façon en réponse à certains types de stimuli ou de tâches, dans des situations données. La signification d'un trait est

donnée par le domaine dans lequel il est défini, mais l'interprétation d'un trait suppose aussi, implicitement du moins, que certains attributs sous-jacents ou latents expliquent les régularités observées dans la performance (Loevinger, 1957 : voir Kane, 2006, p. 30). On peut s'attendre à ce que les personnes qui présentent un trait fort performent bien dans des tâches ou des situations liées à ce trait. Kane (2006) indique que dans la plupart des cas où une terminologie du trait est utilisée, l'attribut mesuré suppose une combinaison de composantes pratiques, mais les fonctions de l'attribut sont considérées comme s'il n'y avait qu'un seul trait unidimensionnel qui opérerait de la même façon en tant que déterminant de la réussite à tous les items d'un test (Henrysson, 1971, p. 146 : voir Kane, 2006, p. 32).

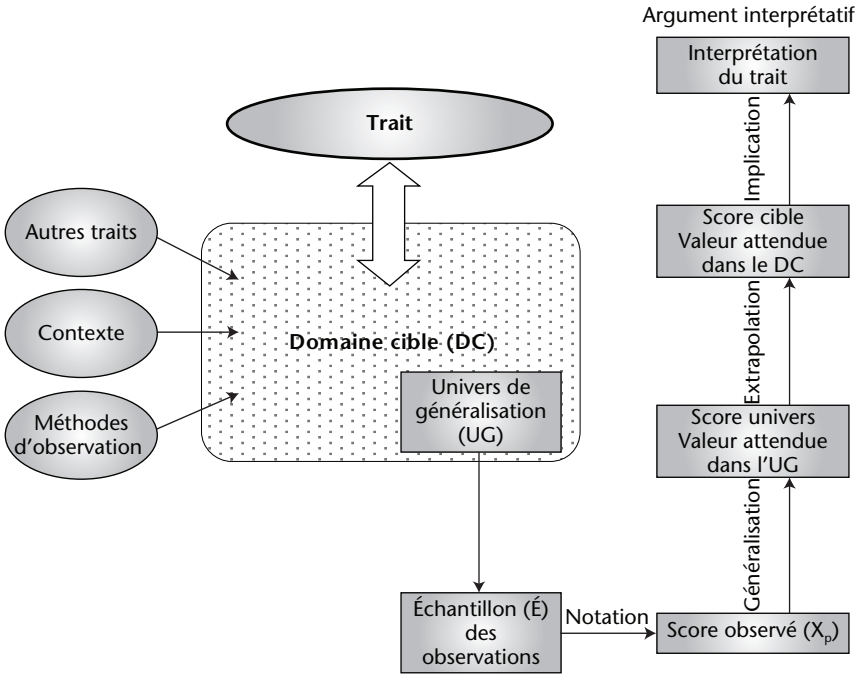


Figure 5.1
Procédure de mesure et d'argumentation interprétative pour l'interprétation d'un trait (adapté de Kane, 2006, p. 33)

On peut rapprocher le terme *trait* du terme *compétence*, qui est plus couramment utilisé en évaluation des apprentissages dans les établissements d'enseignement québécois. En effet, la notion de compétence peut elle aussi être considérée comme unidimensionnelle tout en sous-tendant une combinaison de composantes pratiques, et elle

suppose elle aussi la régularité de la performance *dans une famille de situations comparables qui exigent leur intégration* (Scallon, 2004, p. 204). De plus, la notion de compétence est en rapport direct avec les actes et la profession future ou actuelle des étudiants (Raîche, 2008) ; sa signification est donc déterminée par le domaine auquel elle est associée, tout comme c'est le cas pour le trait.

Selon les termes de Kane (2006), le *domaine cible* est un ensemble d'observations liées au trait. Il s'agit de l'ensemble des manifestations du trait, dans les divers contextes dans lesquels il est impliqué, ainsi que des liens qui l'unissent à d'autres traits et d'autres domaines. La décision d'inclure certaines observations, plutôt que d'autres, dans un domaine cible dépend généralement de l'expérience ou de suppositions à propos des processus impliqués dans les observations. Ainsi, par exemple, l'analyse d'un article scientifique peut être une tâche faisant partie du domaine cible de la méthodologie de la recherche parce qu'on présume qu'elle requiert les mêmes habiletés ou des habiletés connexes ou des performances constitutives (*component performances*). Le fait qu'il peut être difficile d'inclure certaines tâches dans une évaluation ne signifie pas qu'elles ne font pas partie du domaine cible d'un trait. L'objectif n'est pas de définir un domaine bien ordonné ou un domaine qui soit facile à évaluer, mais d'identifier la gamme des observations associées à la compétence ou au trait qui nous intéresse. Le *score cible* correspond à la valeur de la performance attendue d'une personne dans le domaine cible, c'est-à-dire à la performance attendue d'une personne à l'ensemble des observations pouvant être associées au trait qui nous intéresse. Le score cible résulte donc de l'interprétation du score obtenu par une personne en lien avec le domaine cible. En évaluation des apprentissages, le score cible serait donc une estimation de la compétence de l'étudiant dans le domaine, résultant de l'interprétation du ou des scores obtenus par un étudiant aux tests ou aux tâches d'évaluation. Le score en lui-même n'est pas différent, mais son interprétation est étendue au domaine. La correspondance entre le domaine cible et la mesure du trait est une notion importante dans le développement de la procédure d'évaluation et dans la validation de l'interprétation des résultats de l'évaluation.

Dans la plupart des cas, en évaluation des apprentissages, il est quasi impossible de former un échantillon aléatoire ou représentatif des observations à partir du domaine cible et de généraliser le *score observé* d'une personne dans cet échantillon (ex. : résultat obtenu à un test) au score attendu d'une personne dans le domaine cible (score cible). La gamme des observations contenues dans l'évaluation est par conséquent plus restreinte que celle du domaine cible. Par conséquent, les observations contenues dans les mesures des

traits sont généralement tirées d'un sous-ensemble du domaine cible, souvent un très petit sous-ensemble (Fitzpatrick et Morrison, 1971 : voir Kane, 2006 ; Kane, 1982 : voir Kane, 2006). Kane (2006) nomme ce sous-ensemble l'*univers de généralisation*, soit un échantillon des observations du domaine cible à partir duquel sont tirées les observations qui constitueront la mesure (par exemple, les items d'un examen, les critères d'évaluation d'un stage, etc.). En évaluation des apprentissages, l'univers de généralisation représente donc le programme d'études ou, de façon plus restreinte, le cours. Le score attendu d'une personne dans l'univers de généralisation est le *score univers*. C'est l'interprétation du score par rapport à l'univers de généralisation, donc, en évaluation des apprentissages, l'estimation de la compétence de l'étudiant dans le cours ou dans le programme d'études. Alors que le domaine cible pour la compétence d'une personne en méthodologie de la recherche devrait inclure, par exemple, l'examen critique et l'application de la méthodologie dans une grande variété de documents (ex. : journaux, manuels, magazines, articles de revues scientifiques, rapports de recherche), de réponses (ex. : répondre à des questions spécifiques, faire un résumé écrit ou oral, expérimenter une méthode, produire un rapport de recherche) et de contextes (ex. : en classe, à la maison, à la bibliothèque, en milieu professionnel), l'univers de généralisation pour une mesure de la compétence en méthodologie de la recherche peut parfois être limité aux réponses à des questions objectives portant sur la lecture d'un article scientifique au cours d'un examen ou à l'application d'une méthode dans la réalisation d'un travail de session. Ces performances constituent effectivement des exemples d'observations liées à la méthodologie de la recherche, mais elles forment une très petite partie du domaine cible de la compétence en méthodologie de la recherche. Dans de telles circonstances, le score observé, dans ce cas-ci le résultat obtenu à l'examen ou pour le travail de session, peut bien représenter le score univers, soit la performance attendue de la personne en situation d'examen ou pour la réalisation d'un travail de session. Toutefois, on peut se demander s'il reflète avec justesse le score cible, soit la performance attendue de la personne en méthodologie de la recherche, sachant que le domaine cible, la méthodologie de la recherche, inclut une diversité beaucoup plus grande d'observations que celles de l'univers de généralisation. C'est ce questionnement, la représentativité du score observé par rapport au score cible, qui est à la base de la validation de l'argument interprétatif pour l'interprétation d'un trait.

L'interprétation de la performance observée en termes de score cible requiert une chaîne de raisonnements, de la performance à la mesure au score observé, du score observé au score univers, et du score

univers au score cible. De plus, l'interprétation d'un trait a plusieurs implications, dont les relations avec d'autres variables, l'impact des interventions sur ce trait, les exceptions à l'interprétation et l'étendue ou l'importance des différences entre les groupes (Cronbach, 1971, p. 448 ; Kane, 2006, p. 32). La plupart des interprétations ont des implications qui vont au-delà du domaine cible et qui doivent être validées. Le tableau 5.1, donne une vue d'ensemble de l'argument interprétatif pour l'interprétation d'un trait tel que décrit par Kane (2006).

Tableau 5.1
L'argument interprétatif pour l'interprétation d'un trait (adapté de Kane, 2006, p. 34)

I1 : Notation (*scoring*) : de la performance à la mesure au score observé	
A1.1	Les règles de passation et de notation sont appropriées.
A1.2	Les règles de passation et de notation sont appliquées tel que spécifié.
A1.3	La passation et la notation sont exemptes de biais.
A1.4	Les données s'adaptent à tous les modèles d'échelles de mesure employés dans la notation.
I2 : Généralisation : du score observé au score univers	
A2.1	L'échantillon des observations est représentatif de l'univers de généralisation.
A2.2	L'échantillon des observations est suffisamment grand pour contrôler l'erreur aléatoire.
I3 : Extrapolation : du score univers au score cible	
A3.1	Le score univers est lié au score cible.
A3.2	Il n'y a pas d'erreur systématique qui soit susceptible de miner l'extrapolation.
I4 : Implication : du score cible à la description verbale de l'interprétation du trait	
A4.1	Les implications associées au trait sont appropriées.
A4.2	Les propriétés des scores observés soutiennent les implications associées à l'étiquette du trait.

L'argument interprétatif représenté à droite de la figure 5.1 comprend quatre niveaux d'inférence pour l'interprétation des résultats d'un test en termes de trait : la notation, la généralisation, l'extrapolation et l'implication (I1 à I4 dans le tableau 5.1). Lors de l'inférence de notation (I1), la performance observée dans l'échantillon des observations qui constituent la mesure est notée en se rapportant à un score observé (un score brut ou un score sur une échelle). Par l'inférence de généralisation (I2), le score observé est généralisé au score univers, c'est-à-dire que le score obtenu à la mesure est interprété dans un contexte plus large qui est l'univers de généralisation. Par l'inférence

d'extrapolation (I3), le score univers est extrapolé au score cible, c'est-à-dire que le score est interprété comme étant représentatif non plus de la performance dans un sous-ensemble d'observations, mais du domaine en entier. Finalement, par l'inférence d'implication (I4), les implications associées au trait sont ajoutées à l'interprétation du score dans le domaine de façon à décrire ses relations avec d'autres traits et d'autres domaines, et à faire certaines mises en garde quant aux exceptions ou aux diverses interprétations pouvant être produites concernant le trait en question ou le score. Ces quatre niveaux d'inférence (I1 à I4 dans le tableau 5.1) sont accompagnés des arguments (A1.1 à A4.2 dans le tableau 5.1) qui garantissent la validité de chaque inférence et, par le fait même, la validité de l'interprétation du trait. Les quatre niveaux d'inférence accompagnés de leurs arguments de validité sont représentés dans la figure 5.2 et expliqués ensuite en détail, de l'inférence de notation à l'inférence d'implication.

L'inférence de notation assigne un score à la performance de chaque personne en utilisant une règle de notation, laquelle fournit la garantie (*warrant*) de la validité de l'inférence de notation. On assume alors que les critères de notation sont appropriés et qu'ils sont appliqués tel que prévu, que le processus de passation et de notation est exempt de biais et que tous les modèles statistiques (échelle de mesure, calcul des scores) employés dans la notation sont appropriés.

L'inférence de généralisation étend l'interprétation du score observé à partir de l'évaluation d'un échantillon d'observations, qui constitue la mesure, jusqu'à la valeur attendue dans l'univers de généralisation, soit le score univers. La valeur du score est la même, mais son interprétation s'étend d'une affirmation sur un ensemble spécifique d'observations à une affirmation sur la performance attendue dans l'univers de généralisation. La garantie de cette inférence découle de la théorie de l'échantillonnage statistique et dépend du fait que l'on assume la représentativité de l'échantillon des observations et la justesse de la taille de l'échantillon pour contrôler l'erreur d'échantillonnage. Ainsi, pour que la généralisation de l'interprétation du score observé au score univers soit considérée comme valide, on doit s'assurer que les items sélectionnés pour constituer la mesure sont en nombre suffisant pour permettre de contrôler l'erreur aléatoire et qu'ils représentent bien l'ensemble des items constituant l'univers de généralisation.

L'extrapolation de l'univers de généralisation au domaine cible étend l'interprétation du score univers au score cible. Dans l'inférence d'extrapolation, on assume que le score univers est lié (rationnellement ou empiriquement) au score cible et que l'extrapolation est

relativement exempte d'erreurs systématiques ou d'erreurs aléatoires. Encore une fois, le score ne change pas, mais l'interprétation du score s'étend de l'univers de généralisation au domaine cible.

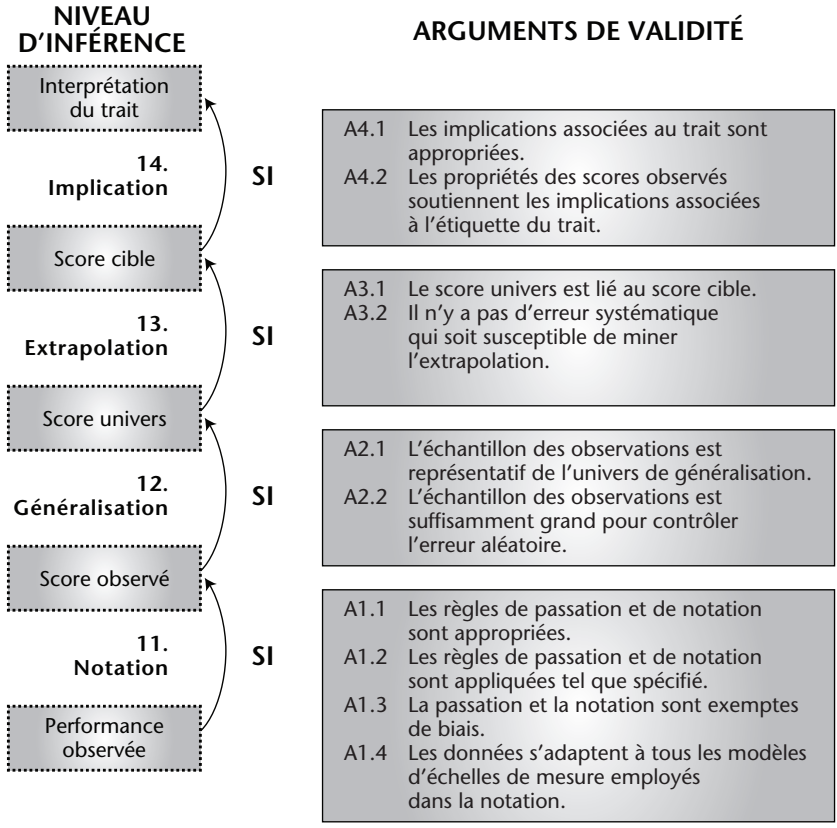


Figure 5.2
Les quatre niveaux d'inférence de l'interprétation d'un trait et les arguments garantissant leur validité

Les implications comportent les extrapolations de l'interprétation pour y inclure toute affirmation ou suggestion associée au trait. Les justifications des implications du trait fournissent l'autorisation pour toute inférence pouvant être formulée ou contenue dans la description du trait, dans l'étiquette du trait et dans les utilisations faites des résultats de l'évaluation. Par exemple, en adoptant une étiquette de trait déjà existante, un terme du langage courant associé à un sens communément accepté, le concepteur du test adopte implicitement dans l'interprétation proposée la signification étendue du terme,

ou alors il se voit obligé de s'opposer à toute inférence non justifiée pouvant être faite sur la base de l'étiquette du trait. Ou encore, s'il est attendu par exemple que le trait demeure stable dans le temps, les résultats empiriques qui laissent croire qu'il en est effectivement ainsi devraient appuyer l'interprétation qui en est donnée ; dans le cas contraire, l'implication devrait spécifier que la stabilité du trait dans le temps n'a pas été démontrée et qu'elle n'est pas assurée.

Pour que l'argument interprétatif décrit au tableau 5.1 et représenté dans la figure 5.2 soit convaincant, toutes les inférences prises individuellement doivent être convaincantes. Une faille dans n'importe laquelle d'entre elles invalide l'argument interprétatif en entier, et ce, même si les preuves empiriques des autres inférences ont bien été établies (Crooks et collab., 1996 : voir Kane, 2006, p. 34).

Selon de Ketele et Gerard (2005), la maîtrise de chacun des critères d'une épreuve bien construite ne garantit pas la maîtrise de la compétence visée. Il s'agit là d'une préoccupation fondamentale de la démarche de validation de l'inférence d'évaluation. En effet, le modèle de Kane (2006) propose une démarche méthodologique permettant de s'assurer de la validité de chacune des inférences passant par la notation, la généralisation, l'extrapolation et l'implication du trait mesuré. Cette démarche s'apparente au concept de validité globale élaboré par Auger (2003) et qu'il définit ainsi :

> *Évaluation de la contribution de l'ensemble des validités spécifiques, de l'opérationnalisation des critères de scientificité, des protocoles mis en place, fondée sur des évidences empiriques et sur un rationnel théorique, en vue d'assurer l'adéquation et la justesse des inférences et des actions à partir d'informations recueillies ou de scores au test.*

C'est la validité démontrée de chacune des inférences qui rend valide l'argument interprétatif en entier, c'est-à-dire qu'elle permet d'établir avec rigueur le lien entre le score observé et le trait ciblé.

La structure d'argumentation décrite par Kane (2006) devrait théoriquement permettre de soutenir la validation des inférences d'évaluation. Comme une politique institutionnelle d'évaluation des apprentissages est un outil devant permettre d'orienter, d'encadrer et de soutenir les pratiques d'évaluation des apprentissages dans les établissements d'enseignement collégial, il faut s'attendre à ce qu'elle contienne les éléments permettant la validation des inférences d'évaluation. Elle devrait donc contenir les éléments présentés dans le modèle de Kane telles, par exemple, des spécifications sur les règles de notation appropriées, sur la réduction des biais associés à la notation, sur la représentativité des observations, etc. L'analyse des politiques institutionnelles d'évaluation des apprentissages selon la structure

d'argumentation décrite par Kane pourrait représenter un pas de plus pour améliorer la qualité de l'évaluation des apprentissages des étudiants dans les établissements d'enseignement.

Le rapprochement des politiques institutionnelles d'évaluation des apprentissages et du modèle théorique de Kane (2006), deux types d'assurances de qualité de l'évaluation des apprentissages, l'une pratique et l'autre théorique, permet de préciser l'objectif principal du projet, qui est de valider la structure d'argumentation interprétative de Kane par l'application de cette structure à l'analyse de politiques institutionnelles d'évaluation des apprentissages du réseau collégial québécois. Un objectif secondaire consiste à identifier des moyens que peuvent mettre en place les établissements pour chercher à assurer la validité des inférences d'évaluation des apprentissages des étudiants. Ce chapitre est consacré au premier objectif[3].

3. MÉTHODOLOGIE

3.1. Corpus d'analyse

Le corpus d'analyse est constitué de cinq politiques institutionnelles d'évaluation des apprentissages. Le tableau 5.2 décrit les établissements d'enseignement collégial dont la politique institutionnelle d'évaluation des apprentissages a été analysée.

Pour des raisons d'homogénéité, seules les politiques institutionnelles d'évaluation des apprentissages s'appliquant à la formation ordinaire, par opposition à la formation continue, ont été considérées. Les collèges privés non subventionnés n'ont pas été pris en compte pour les analyses, car leur contexte éducationnel diffère de celui des collèges subventionnés par le nombre d'étudiants généralement inférieur et par leur système de gestion et d'organisation. Sur les 64 collèges privés et publics, anglophones et francophones, dispensant la formation ordinaire et figurant sur la liste du ministère de l'Éducation, du Loisir et du Sport, 41 avaient publié leurs politiques institutionnelles d'évaluation des apprentissages sur leur site Internet au moment de la collecte des données en 2008. Les cinq politiques institutionnelles d'évaluation des apprentissages qui forment le corpus d'analyse ont été sélectionnées à partir de ces dernières de façon à représenter les collèges privés et publics, anglophones et francophones, de moyennes et grandes populations de la province de Québec.

3. Pour les résultats associés au second objectif, voir le mémoire de recherche à l'origine de ce chapitre (Paquette-Côté, 2009).

Tableau 5.2
Description des établissements d'enseignement collégial dont les politiques institutionnelles d'évaluation des apprentissages (PIEA) constituent le corpus d'analyse

PIEA	Population de la municipalité (en 2006)[a]	Public / Privé	Francophone / Anglophone	Nombre d'étudiants (en 2006)[b]
01	Moins de 50 000	Public	Francophone	Environ 2 500
02	Moins de 50 000	Public	Francophone	Moins de 1 500
03	Plus de 1 000 000	Public	Francophone	Environ 2 500
04	Plus de 1 000 000	Privé	Francophone	Moins de 1 500
05	Plus de 1 000 000	Public	Anglophone	Plus de 5 000

a Statistique Canada (<http://www.statcan.gc.ca>).
b Statistiques sur l'éducation (ministère de l'Éducation, du Loisir et du Sport, 2007).

Les résultats obtenus sont donc ceux qui se rapportent aux établissements d'enseignement collégial dont la politique institutionnelle d'évaluation des apprentissages a été analysée. Toutefois, puisque la Commission d'évaluation de l'enseignement collégial propose un cadre général de référence pour l'évaluation des politiques institutionnelles d'évaluation des apprentissages (Commission d'évaluation de l'enseignement collégial, 1994), les politiques institutionnelles d'évaluation des apprentissages des établissements d'enseignement collégial québécois sont relativement homogènes. De ce fait, l'interprétation des résultats de la présente recherche peut s'appliquer à des établissements d'enseignement collégial dont les politiques institutionnelles d'évaluation des apprentissages n'a pas été analysée.

3.2. Instrumentation

L'instrumentation de la recherche se compose d'une grille d'analyse de contenu et d'un outil de modélisation schématique. Une grille d'analyse de contenu a été construite à partir de la structure argumentaire développée par Kane (2006) (voir le tableau 5.1), qui permet d'assurer la validité de l'argumentation interprétative pour l'interprétation d'un trait. Cette grille d'analyse est constituée de cinq colonnes. La première colonne indique la politique institutionnelle d'évaluation des apprentissages analysée. La deuxième colonne exprime l'unité de l'analyse, soit un énoncé d'une phrase ou un ensemble de phrases exprimant une seule idée. Dans la troisième colonne, on identifie pour chaque extrait l'argument ou les arguments de validité du modèle de Kane (2006) qui lui sont associés. L'unité d'analyse pouvait aussi être codée *autre* lorsque l'énoncé se rapportait à la validité de l'inférence d'évaluation, mais qu'il ne correspondait à aucun argument du modèle de Kane

en particulier. La quatrième colonne permettait de coder un moyen d'application ou d'évaluation de l'argument identifié dans l'énoncé. Dans la cinquième colonne on trouve les interprétations de l'analyste, pouvant être, selon le cas : le moyen d'application ou d'évaluation de l'argument, une interprétation de l'énoncé ou de l'argument de validité auquel il référait, ou une interprétation du modèle de Kane. Lorsque chaque énoncé a été codé dans la grille d'analyse, les moyens d'application ou d'évaluation identifiés dans les politiques institutionnelles d'évaluation des apprentissages analysées sont modélisés au moyen du logiciel *MotPlus* du Centre de recherche LICEF de la Télé-université (Paquette, Lundgren-Cayrol et Léonard, 2008). *MotPlus* est un éditeur graphique qui *vise à soutenir une méthode de représentation graphique des connaissances développée à l'intention des concepteurs pédagogiques* (Rivard, 2007).

3.3. Déroulement

La méthodologie utilisée dans la recherche est axée sur l'analyse de contenu des politiques institutionnelles d'évaluation des apprentissages. La première étape de l'analyse consistait à analyser le contenu manifeste des politiques institutionnelles d'évaluation des apprentissages. Pour les fins de l'analyse de contenu, les textes ont été divisés en énoncés distincts, chaque énoncé devenant une unité d'analyse. Un énoncé est formé d'une phrase ou d'un ensemble de phrases exprimant une seule idée. À ce stade, n'étaient conservés que les éléments les plus significatifs des textes étudiés (Lessard-Hébert, Goyette et Boutin, 1996 ; Van der Maren, 1995). Le processus de catégorisation ou de classification suivait un modèle fermé (L'Écuyer, 1987), c'est-à-dire que les catégories étaient prédéterminées par la structure argumentaire permettant la validation de l'interprétation d'un trait décrite par Kane (2006). Un processus de modélisation a ensuite été enclenché, permettant de classer selon une représentation schématique les moyens pouvant être mis en place par les acteurs des établissements d'enseignement collégial pour chercher à assurer la validité des inférences d'évaluation des apprentissages des étudiants.

À l'étape de la modélisation, le processus de catégorisation prend la forme d'un modèle mixte (L'Écuyer, 1987, p. 59) permettant d'effectuer des regroupements, des subdivisions ou des ajouts au modèle de Kane (2006). La structure de base est celle de la représentation du modèle de Kane (voir la figure 5.2). Chaque argument de validité (A1.1 à A4.2) a été décliné en sous-modèle consistant en une représentation graphique de l'ensemble des moyens qui lui étaient associés. Chaque moyen identifié lors de l'analyse de contenu était ajouté à la

représentation graphique de l'argument correspondant. Les moyens d'application ou d'évaluation étaient classés et présentés en fonction des acteurs (direction d'établissement, professeurs, étudiants, etc.) responsables de leur application ou de leur évaluation.

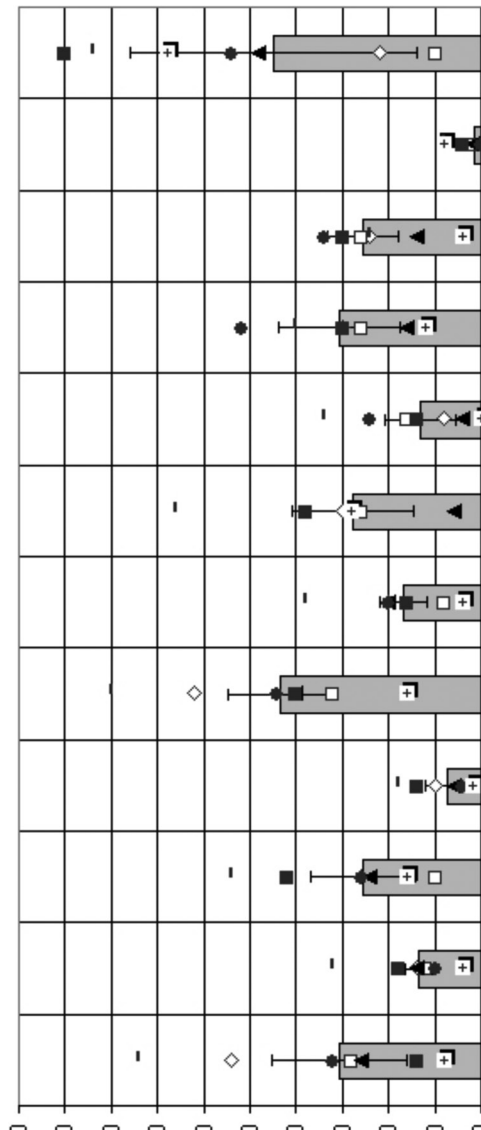
3.4. Méthode d'analyse et d'interprétation des données

Deux analyses ont été réalisées à partir des données recueillies : une analyse quantitative de statistiques descriptives et une analyse qualitative de synthèse. Une analyse quantitative de statistiques descriptives (fréquences absolues et relatives, moyennes, écarts types) a été effectuée à partir des données obtenues pour évaluer l'exhaustivité, l'exclusivité et la pertinence des catégories du modèle de Kane (2006) dans le contexte de l'évaluation des apprentissages en enseignement collégial, permettant ainsi de respecter l'objectif principal du projet de recherche. Cet objectif consistait à valider la structure d'argumentation interprétative de Kane par l'application de cette structure à l'analyse de politiques institutionnelles d'évaluation des apprentissages du réseau collégial québécois. Le second objectif visait à identifier des moyens que peuvent mettre en place les institutions pour chercher à assurer la validité des inférences d'évaluation des apprentissages des étudiants. Une analyse qualitative de synthèse a été réalisée pour faire ressortir les résultats les plus significatifs de chaque argument de la structure d'argumentation de la validité des inférences d'évaluation au regard des moyens relevés lors de l'analyse de contenu. Dans ce chapitre, on ne traite que des résultats liés au premier objectif[4].

4. RÉSULTATS

Un total de 424 énoncés tirés des cinq politiques institutionnelles d'évaluation des apprentissages retenues ont été analysés. La figure 5.3 présente la répartition des énoncés classés dans chaque catégorie pour chaque politique institutionnelle d'évaluation des apprentissages ainsi que les moyennes et écarts types. Rappelons qu'un même énoncé pouvait être classé dans plus d'une catégorie d'analyse. Le graphique présente aussi le nombre total de moyens identifiés pour chaque catégorie. Parmi ceux-ci, pour chaque catégorie du modèle, le nombre de moyens spécifiques correspond au nombre de moyens qui se rapportent uniquement à cette catégorie.

4. Les résultats associés au second objectif peuvent être consultés dans le mémoire de recherche à l'origine de cet article (Paquette-Côté, 2009).



	A1.1	A1.2	A1.3	A1.4	A2.1	A2.2	A3.1	A3.2	A4.1	A4.2	A5.1	A5.2
□ Moyenne	15,40	6,80	12,80	3,60	21,80	8,40	13,80	6,60	15,40	12,80	0,80	22,40
Écart type	7,30	1,48	5,67	2,41	5,59	2,61	6,57	3,85	6,58	3,77	0,84	15,55
◇ PIEA01	27	7	13	5	31	10	15	4	15	12	0	11
□ PIEA02	14	6	5	1	16	4	13	8	13	13	1	5
▲ PIEA03	13	7	12	3	20	10	3	2	8	7	1	24
● PIEA04	16	5	13	2	22	10	19	12	26	17	0	27
■ PIEA05	7	9	21	7	20	8	19	7	15	15	2	45
– Nb de moyens	37	16	27	9	40	19	33	17	20	12	4	42
+ Nb de moyens spécifiques	4	2	8	1	8	2	14	0	6	2	4	34

Figure 5.3
Répartition des énoncés en fréquences et nombre de moyens identifiés selon les catégories du modèle d'argumentation de la validité des inférences d'évaluation

La première ligne du tableau sous le graphique présente la moyenne du nombre d'énoncés catégorisés dans chaque argument de la structure d'argumentation de la validité des inférences d'évaluation, toutes politiques confondues. La deuxième ligne présente l'écart type associé à chacune de ces moyennes, c'est-à-dire la dispersion autour de la moyenne du nombre d'énoncés catégorisés dans chaque argument pour chaque politique analysée. Un écart type faible signifie que, pour chaque politique analysée, le nombre d'énoncés catégorisés dans l'argument identifié se situe près de la moyenne. On peut alors conclure que le nombre d'énoncés de l'argument correspondant est sensiblement homogène pour l'ensemble des politiques analysées. Dans le cas contraire, c'est-à-dire lorsque l'écart type est élevé, lorsque les données sont plus dispersées, on peut conclure que les politiques analysées diffèrent quant au nombre d'énoncés d'un argument donné. Les lignes 3 à 7 (PIEA01 à PIEA05) présentent, pour chaque politique analysée, le nombre d'énoncés catégorisés dans chaque argument de la structure d'argumentation de la validité des inférences d'évaluation. Rappelons qu'un même énoncé pouvait être classé dans plus d'une catégorie d'analyse. Les données des lignes 1 à 7 seront examinées en réponse au premier objectif de cette recherche portant sur la validation de la structure d'argumentation interprétative de Kane (2006).

La ligne 8, avant-dernière ligne du tableau, présente, pour chaque politique analysée, le nombre total de moyens d'application ou d'évaluation identifiés pour chaque argument de la structure d'argumentation de la validité des inférences d'évaluation. Un même moyen ayant plus d'une fonction pouvait être rattaché à plus d'un argument. La ligne 9, dernière ligne du tableau, présente, pour chaque politique analysée, le nombre de moyens d'application ou d'évaluation spécifiques à chaque argument. Il s'agit de moyens ne s'appliquant qu'à l'argument identifié et à aucun autre argument de la structure d'argumentation. Ces moyens sont identifiés dans le mémoire (Paquette-Côté, 2009) à l'origine du présent texte en réponse au second objectif de cette recherche.

Une analyse quantitative de statistiques descriptives (fréquences absolues et relatives, moyennes, écarts types) a été effectuée à partir des résultats de l'analyse de contenu pour évaluer l'exhaustivité, l'exclusivité et la pertinence des catégories du modèle de Kane (2006) dans le contexte de l'évaluation des apprentissages en enseignement collégial.

4.1. Exhaustivité des catégories du modèle de Kane (2006)

Les catégories doivent être exhaustives, en ce sens qu'elles doivent pouvoir permettre de classer tous les éléments analysés (L'Écuyer, 1987). L'analyse de la fréquence des énoncés classés et non classés dans les catégories du modèle de Kane (2006) permet de faire une évaluation de leur exhaustivité par rapport au matériel analysé. Comme on l'a vu, le processus de catégorisation à l'étape de la modélisation correspond à un modèle mixte (L'Écuyer, 1987, p. 59) permettant d'effectuer des regroupements, des subdivisions ou des ajouts au modèle de Kane si celui-ci ne couvre pas l'ensemble du contenu analysé. Le tableau 5.3 présente les résultats de la réduction des données dans le processus de catégorisation.

Tableau 5.3
Réduction des données dans le processus de catégorisation

	Processus de catégorisation	Fréquence (Nb d'énoncés)	Fréquence relative
1	Identification des énoncés significatifs	424	100 %
2	Catégorisation dans le modèle de Kane (2006) (A1.1 à A4.2)	268	63,21 %
	Inclassables	156	*36,79 %*
3	Catégorisation dans les deux catégories émergentes (A5.1 et A5.2)	116	27,36 %
	Inclassables	40	*9,43 %*

En tout, 268 énoncés (63,21 %) ont été classés dans les catégories du modèle de Kane (2006) tandis que 156 énoncés (36,79 %) n'ont pas pu être classés dans ces catégories. De l'analyse de ces derniers ont émergé deux nouvelles catégories, deux arguments ajoutés au modèle initial : *A5.1 L'ensemble des arguments est respecté de façon à ce que chacune des inférences considérées de façon individuelle soit convaincante* et *A5.2. Les processus d'apprentissage et d'évaluation sont connus des acteurs*. Ces deux arguments (A5.1 et A5.2) renvoient à la crédibilité de l'inférence d'évaluation. Après l'analyse des 156 énoncés en fonction des deux catégories ajoutées, seulement 40 énoncés restent inclassables dans les catégories du modèle, ce qui fait passer le pourcentage d'énoncés non considérés par le modèle de 36,79 % à 9,43 %. Ces 40 énoncés restants sont soit des énoncés généraux sur l'évaluation des apprentissages, qui n'apportent rien de particulier à l'argumentation de la validité des inférences d'évaluation, soit des pratiques non cohérentes avec l'argumentation de la validité de l'inférence d'évaluation.

4.2. Exclusivité des catégories du modèle de Kane (2006)

L'exclusivité des catégories signifie qu'un même élément ne peut être classé dans deux catégories différentes. La nécessité de l'exclusivité est toutefois contestée puisqu'un même énoncé peut parfois renfermer plus d'un sens (L'Écuyer, 1987). C'est d'ailleurs ce que révèlent les résultats de la recherche par rapport aux catégories du modèle de Kane (2006). Le tableau 5.4 illustre la répartition des énoncés selon le nombre de catégories dans lesquelles ils ont été classés.

Tableau 5.4
Répartition des énoncés selon le nombre de catégories dans lesquelles ils ont été classés

	Nombre de catégories										Total
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Fréq. (nb d'énoncés)	40	239	74	27	13	2	25	1	1	2	424
Fréq. rel.	9,43	**56,37**	17,45	6,37	3,07	0,47	5,89	0,24	0,24	0,47	100

La majorité (56,37 %) des énoncés analysés ont été classés dans une seule catégorie du modèle ; 17,45 % des énoncés ont été classés dans deux catégories ; 16,75 % des énoncés ont été classés dans plus de deux catégories. Enfin, 9,43 % des énoncés n'ont été classés dans aucune catégorie du modèle pour les raisons évoquées précédemment.

4.3. Pertinence des catégories du modèle de Kane (2006)

Les catégories du modèle semblent pertinentes puisqu'un certain nombre d'énoncés se rapportent à chacune d'elles (voir la figure 5.3). On pourrait toutefois remettre en question la pertinence de deux catégories, soit l'argument A5.1, dans lequel seulement quatre énoncés ont été classés, et l'argument A1.4, dans lequel seulement 18 énoncés ont été classés. Ainsi, si un moins grand nombre d'énoncés ont été catégorisés comme faisant référence à ces arguments, il est possible que ces arguments soient moins pertinents dans le contexte de l'évaluation des apprentissages dans les établissements d'enseignement collégial ou dans les politiques institutionnelles d'évaluation des apprentissages. L'assertion du nombre n'est toutefois pas suffisante pour remettre en question la pertinence des arguments du modèle. Il est possible que certains arguments du modèle soient pertinents pour assurer l'argumentation de la validité des inférences d'évaluation, mais que les pratiques liées à ceux-ci soient moins connues ou moins répandues

que d'autres. Dans ce cas, il est possible qu'elles soient moins bien représentées dans les politiques institutionnelles d'évaluation des apprentissages.

4.4. Adaptation de la structure d'argumentation de Kane (2006)

L'analyse des politiques institutionnelles d'évaluation des apprentissages du réseau collégial québécois en fonction de la structure d'argumentation interprétative de Kane (2006) révèle que l'ensemble de la structure s'applique à l'évaluation des apprentissages à l'enseignement collégial. Toutefois, la structure d'argumentation de Kane ne couvre pas l'ensemble de la pratique de l'évaluation des apprentissages à l'enseignement collégial. C'est pourquoi on suggère d'ajouter un niveau comprenant deux arguments supplémentaires au modèle initial. Il s'agit du niveau *crédibilité,* qui représente la condition pour que l'inférence puisse être considérée valide et qu'elle soit acceptée par les acteurs concernés. Ces derniers représentent aussi bien les personnes de l'établissement engagées dans la production ou l'interprétation de l'inférence d'évaluation (étudiants, professeurs et autres membres de l'établissement) que des personnes extérieures à l'établissement qui, elles aussi, interprètent les résultats de l'évaluation (employeurs, membres d'autres établissements d'enseignement, etc.). Le fait que l'ensemble des arguments est respecté de façon à ce que chacune des inférences considérées de façon individuelle soit convaincante (A5.1) et le fait que les processus d'apprentissage et d'évaluation sont connus des acteurs (A5.2) renforcent la crédibilité de l'inférence d'évaluation auprès des acteurs concernés. La figure 5.4 présente le modèle qui repose sur les résultats de la présente recherche : une adaptation de la structure d'argumentation de Kane, suggérée pour l'interprétation des inférences d'évaluation dans le contexte de l'évaluation des apprentissages au collégial.

5. DISCUSSION

Les résultats obtenus ont permis de constater que le modèle de Kane (2006) peut effectivement s'appliquer au contexte de l'évaluation des apprentissages en enseignement collégial. Toutefois, puisque la structure d'argumentation de Kane (2006) ne semble pas couvrir l'ensemble de la pratique de l'évaluation des apprentissages dans l'enseignement collégial, deux arguments ont été ajoutés à la structure d'argumentation initiale de façon à chercher à assurer la crédibilité de l'inférence d'évaluation aux yeux des acteurs concernés.

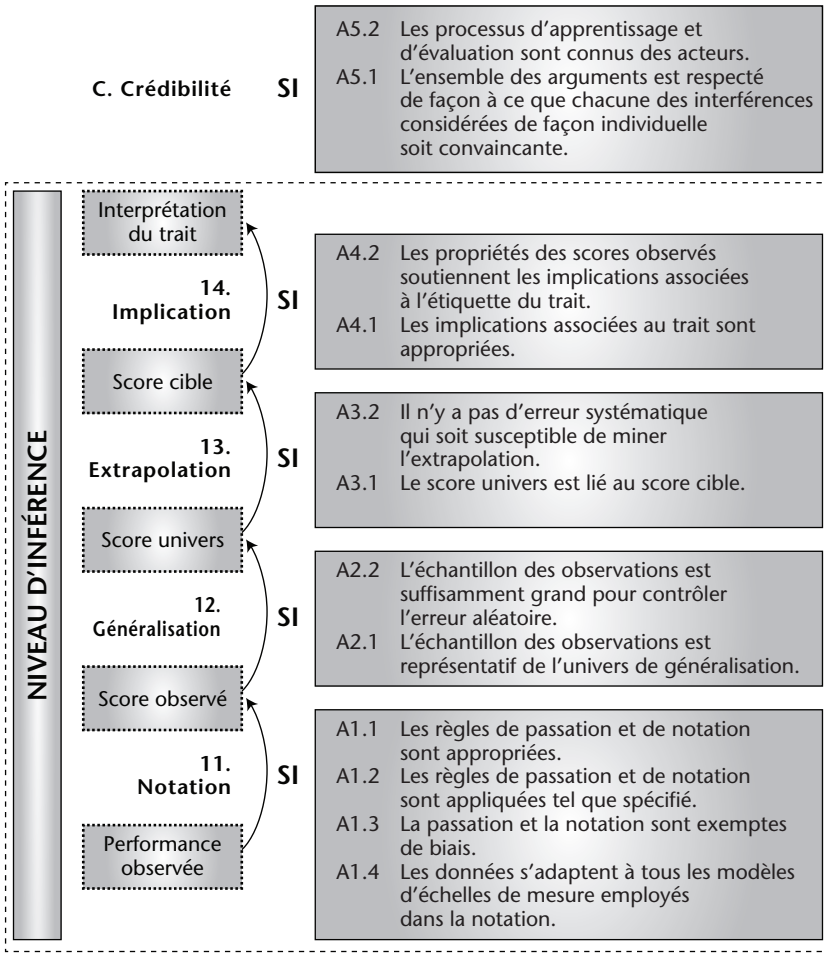


Figure 5.4
Structure d'argumentation de la validité des inférences d'évaluation des apprentissages

Pour ce qui est de l'exhaustivité des catégories du modèle de Kane (2006), la nécessité d'ajouter des catégories indique que les arguments de validité du modèle initial ne couvrent pas l'ensemble du phénomène de l'évaluation des apprentissages dans les établissements d'enseignement collégial. Le niveau *crédibilité* (voir la figure 5.4) a été ajouté à la structure d'argumentation de Kane pour représenter de façon plus complète la pratique de l'évaluation des apprentissages dans l'enseignement collégial. Ce niveau indique que l'inférence d'évaluation ne peut être jugée valide que si elle est acceptée par les acteurs concernés.

Selon Kane (2006), pour que l'argument interprétatif soit convaincant, chacune des inférences considérées de façon individuelle doit être convaincante. Un argument a donc été ajouté au modèle initial de Kane : *A5.1 L'ensemble des arguments est respecté de façon à ce que chacune des inférences considérées de façon individuelle soit convaincante*. De plus, pour que les acteurs acceptent comme valide l'inférence d'évaluation, ils doivent connaître les processus utilisés lors de l'évaluation des apprentissages. Un second argument a été ajouté en ce sens : *A5.2 Les processus d'apprentissage et d'évaluation sont connus des acteurs*.

En ce qui concerne l'exclusivité des catégories du modèle de Kane (2006), les catégories du modèle ne sont pas totalement exclusives, mais on pourrait difficilement s'attendre à ce qu'elles le soient étant donné qu'une même pratique dans un établissement d'enseignement peut avoir plusieurs fonctions et ainsi répondre à plusieurs critères de validité.

D'autre part, les analyses effectuées n'ont pas permis de juger de la pertinence des catégories du modèle de Kane (2006). Le critère du nombre d'énoncés catégorisés comme se rapportant à chaque argument n'est pas suffisant pour prononcer un jugement sur la pertinence de chacun d'eux. Il est possible que les politiques institutionnelles d'évaluation des apprentissages ne reflètent pas l'apport ou l'importance des pratiques pour assurer la validité de l'inférence d'évaluation des apprentissages. On pourrait donc considérer dans le cadre de recherches ultérieures de faire l'analyse de contenu de divers outils pédagogiques, tels que des plans de cours et des plans d'évaluation par exemple, ainsi que de procéder à des entretiens avec le personnel des établissements d'enseignement et avec des étudiants pour identifier les pratiques en vigueur dans les établissements d'enseignement. On pourrait également tenir compte de l'évaluation de l'efficacité perçue ou réelle de chaque pratique pour assurer la validité de l'inférence d'évaluation des apprentissages.

En examinant les moyennes et les écarts types du nombre d'énoncés classés dans chaque catégorie du modèle, on trouve une grande variabilité entre les différentes politiques institutionnelles d'évaluation des apprentissages. Cet écart entre les politiques institutionnelles d'évaluation des apprentissages concernant l'importance accordée à chaque argument du modèle peut refléter les valeurs des collèges et ce qui les distingue au plan de l'évaluation des apprentissages. L'objectif de cette recherche n'était pas de vérifier si les politiques institutionnelles d'évaluation des apprentissages contenaient tous les éléments permettant d'argumenter quant à la validité des inférences d'évaluation, ni même d'effectuer une comparaison entre

les différentes politiques institutionnelles d'évaluation des apprentissages. Toutefois, si le modèle suggéré et les hypothèses dégagées étaient eux-mêmes validés par les acteurs du milieu collégial et par l'analyse des pratiques, le modèle et les lignes directrices identifiés pourraient éventuellement servir à l'évaluation et à l'élaboration des politiques institutionnelles d'évaluation des apprentissages dans une perspective d'argumentation de la validité des inférences d'évaluation des apprentissages.

6. CONCLUSION

Cette recherche de nature exploratoire constitue une première tentative de validation de la structure d'argumentation interprétative de Kane (2006) par l'application de cette structure à l'analyse de politiques institutionnelles d'évaluation des apprentissages (PIEA) du réseau collégial québécois. Cette recherche poursuivait deux objectifs : 1) valider la structure d'argumentation interprétative de Kane par l'application de cette structure à l'analyse de politiques institutionnelles d'évaluation des apprentissages du réseau collégial québécois ; 2) identifier des moyens que peuvent mettre en place les institutions pour chercher à assurer la validité des inférences d'évaluation des apprentissages des étudiants.

La méthodologie retenue pour atteindre ces objectifs a été l'analyse de contenu. Cinq politiques institutionnelles d'évaluation des apprentissages, divisées en 424 énoncés distincts, ont été analysées au moyen d'une grille d'analyse de contenu construite à partir de la structure argumentaire développée par Kane (2006). Une modélisation schématique des moyens identifiés dans les politiques institutionnelles d'évaluation des apprentissages analysées a ensuite été réalisée, permettant d'effectuer des regroupements, des subdivisions ou des ajouts à la structure d'argumentation interprétative de Kane.

Les résultats obtenus ont permis de constater que le modèle de Kane (2006) peut effectivement s'appliquer au contexte de l'évaluation des apprentissages dans l'enseignement collégial. De plus, les analyses effectuées ont permis de formuler des hypothèses quant à l'exhaustivité, à l'exclusivité et à la pertinence des catégories du modèle de Kane dans ce contexte. Elles ont ainsi permis de proposer l'ajout de deux arguments à la structure d'argumentation initiale de façon à chercher à assurer la crédibilité de l'inférence d'évaluation aux yeux des acteurs concernés.

Par la même occasion, cette recherche a permis d'identifier des moyens, ou des lignes directrices, permettant de chercher à assurer l'argumentation de la validité des inférences d'évaluation en évaluation des apprentissages dans l'enseignement supérieur. Des moyens particuliers d'application et d'évaluation restent encore à trouver pour étoffer les lignes directrices identifiées dans cette recherche. Cela pourrait ainsi contribuer à faire en sorte que les politiques institutionnelles d'évaluation des apprentissages deviennent de véritables outils en évaluation des apprentissages, soutenant les pratiques et servant aux établissements d'enseignement, aux enseignants, aux professionnels et aux étudiants de façon à assurer la validité de l'interprétation des résultats de l'évaluation des apprentissages des étudiants de l'enseignement supérieur.

RÉFÉRENCES

Auger, R. (2003). *Clarification conceptuelle et proposition d'opérationnalisation de quelques critères de scientificité de la recherche en éducation : le cas de la saturation et de la complétude*. Communication présentée à l'Association pour la recherche qualitative, Trois-Rivières, Québec. Montréal, Québec : Université du Québec à Montréal.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. Dans R. L. Brennan (dir.), *Educational measurement* (4e édition). Westport, Connecticut : Praeger.

Commission d'évaluation de l'enseignement collégial (1994a). *La Commission d'évaluation de l'enseignement collégial : sa mission et ses orientations*. Québec, Québec : ministère de l'Éducation du Québec.

Commission d'évaluation de l'enseignement collégial (1994b). *L'évaluation des politiques institutionnelles d'évaluation des apprentissages – Cadre de référence*. Québec, Québec : ministère de l'Éducation du Québec.

Conseil des collèges (1992). *L'enseignement collégial : des priorités pour un renouveau de la formation*. Québec, Québec : Gouvernement du Québec.

Cronbach, L. J. (1988). Five perspectives on validity argument. Dans H. Wainer et H. I. Braun (dir.), *Test validity*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Cronbach, L. J. (1980). Validity on parole : how can we go straight. Dans W. B. Schrader (dir.), *New directions for testing and measurement : measuring achievement : progress over a decade*. San Francisco, Californie : Jossey-Bass.

Cronbach, L. J. (1971). Test validation. Dans R. L. Thorndike (dir.), *Educational measurement* (4e édition). Washington, District of Columbia : American Council on Education.

Cronbach, L. J., Linn, R. L., Brennan, R. L. et Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and psychological measurement*, *57*(3), 373-399.

Cureton, E. E. (1951). *Validity*. Dans E. F. Lindquist (dir.), *Educational measurement*. Washington, District of Columbia : The American Council on Education.

De Ketele, J.-M. et Gérard, F.-M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences. *Mesure et évaluation en éducation*, *28*(3), 1-26.

Gipps, C. V. (1994). *Beyond testing : towards a theory of educational assessment*. Washington, District of Columbia : The Falmer Press.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement : interdisciplinary research and perspective, 2*(3), 135–170.

Kane, M. T. (2006). Validation. Dans R. L. Brennan (dir.), *Educational measurement* (4e édition). Westport, Connecticut : Praeger.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin, 112*(3), 527-535.

L'Écuyer, R. (1987). L'analyse de contenu : notion et étapes. Dans J.-P. Deslauriers (dir.), *Les méthodes de la recherche qualitative*. Québec, Québec : Presses de l'Université du Québec.

Lessard-Hébert, M., Goyette, G. et Boutin, G. (1996). *La recherche qualitative, fondements et pratiques*. Montréal, Québec : Éditions Nouvelles.

Lissitz, R. W. (dir.) (2009). *The concept of validity : revisions, new directions, and applications*. Charlotte, Caroline du Nord : Information Age.

Messick, S. (1989). Validity. Dans R. L. Linn (dir.), *Educational measurement* (3e édition). New York, New York : American Council on Education et Macmillan.

Ministère de l'Éducation, du Loisir et du Sport (2007). *Tableau 07. L'effectif scolaire des établissements d'enseignement collégial selon la région administrative, l'établissement, le type de formation, le service d'enseignement et le réseau d'enseignement (2004 ou 2005 ou 2006). Statistiques détaillées sur l'éducation*. Québec, Québec : ministère de l'Éducation, du Loisir et du Sport.

Mislevy, R. J., Almond, R. G. et Lukas, J. F. (2004). *A brief introduction to evidence-centered design*. CSE Report 632. Los Angeles, Californie : The National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE), University of California.

Mislevy, R. et Haertel, G. (2006a). *Implications of evidence-centered design for educational testing*. Draft PADI technical report 17. Menlo Park, Californie : SRI International.

Mislevy, R. J. et Haertel, G. D. (2006b). Implications for evidence-centered design for educational testing. *Educational measurement : issues and practice*, *25*(4), 6-20.

Mislevy, R. J., Steinberg, L. S., Almond, R. G. et Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. Dans D. M. Williamson, R. J. Mislevy et I. I. Bejar (dir.), *Automated scoring of complex tasks in computer-based testing*. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Mucchielli, R. (1971). *L'examen psychotechnique, connaissance du problème. Applications pratiques*. Paris, France : ESF.

Paquette, G., Lundgren-Cayrol, K. et Léonard, M. (2008). The MOT+ visual language for knowledge-based instructional design. Dans L. Botturi et T. Stubs (dir.), *Handbook on virtual instructional design languages – theories and practices*. Hershey, Pennsylvanie : Informing Science Reference.

Paquette-Côté, K. (2009). *Analyse de l'argumentation de la validité des inférences d'évaluation dans les politiques institutionnelles d'évaluation des apprentissages des établissements d'enseignement collégial québécois*. Mémoire de maîtrise inédit. Montréal, Québec : Université du Québec à Montréal.

Raîche, G. (2008). *L'évaluation des apprentissages à l'enseignement supérieur : vers une vision intégrative de l'évaluation des apprentissages* (2e édition). Montréal, Québec : Université du Québec à Montréal.

Rivard, J. (2007). *Logiciel MotPlus. Éditeur de modèles de connaissances. Manuel de l'utilisateur. Version du logiciel : 1.6.5*. Québec, Québec : LICEF, Télé-Université.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Saint-Laurent, Québec : Éditions du Renouveau pédagogique.

Toulmin, S. E. (1964). *The uses of argument*. Cambridge, Royaume-Uni : Cambridge University Press.

Van der Maren, J.-M. (1995). *Méthodes de recherche pour l'éducation*. Montréal, Québec : Les Presses de l'Université de Montréal.

Wiggins, G. P. (1993). *Assessing student performance : exploring the purpose and limits of testing*. San Francisco, Californie : Jossey-Bass.

Chapitre 6

Validité du jugement professionnel des enseignants du primaire dans un contexte d'approche par compétences

Pascal Ndinga

Le présent chapitre propose un modèle visant à assurer la validité du jugement professionnel des enseignants du primaire dans un contexte de l'approche par compétences. Ce modèle s'inspire d'une analyse des principales références des encadrements du ministère de l'Éducation, du Loisir et du Sport du Québec portant sur la valeur accordée au jugement professionnel des enseignants. Le modèle argumentatif de l'interprétation de la validité de Kane a aussi servi de référence. Ses indicateurs apparentés aux pratiques usuelles des enseignants du primaire sont utilisés. L'approche de l'élaboration d'un modèle a été privilégiée. Cette application confirme la validité du jugement professionnel.

1. INTRODUCTION

L'importance de la notion de validité n'est plus à démontrer dans le domaine de la psychométrie, comme le montre Angoff (1988), cité par Laveault et Grégoire (1997, p. 189) : *En psychométrie, la validité a toujours été considérée comme le concept le plus fondamental et le plus important*. Ce statut découle du fait que la crédibilité et la valeur des informations fournies par un test, quel qu'il soit, sont tributaires de la preuve de la validité de celles-ci. Ainsi, il incombe au constructeur et à l'utilisateur de tests de procéder à

une cueillette minutieuse des éléments prouvant la précision de l'information fournie par rapport au but visé par l'outil. Il s'agit de répondre à la question : les données obtenues au terme de la démarche évaluative reflètent-elles réellement le niveau de développement des compétences des individus en cause ? Pour répondre à cette question, il importe de connaître les procédés et les mécanismes ayant servi à l'élaboration des outils utilisés, à la collecte des données et par-dessus tout, à leur interprétation.

L'implantation du renouveau pédagogique au Québec a mis en lumière le jugement professionnel des enseignants. En effet, bien que le jugement professionnel ait toujours été un élément pris en compte dans les différentes étapes du processus évaluatif (Durand, 2008 ; McMillan, 2000 : voir Durand et Chouinard, 2006 ; ministère de l'Éducation, du Loisir et du Sport, 2006), la place prépondérante qu'il occupe dorénavant en évaluation constitue la véritable innovation de la réforme. Cette innovation a suscité de nombreuses réactions et critiques, dont les plus vives traduisent urtout l'incertitude face à la qualité des jugements. De plus, le caractère éminemment subjectif (Lafortune, 2008 ; ministère de l'Éducation, du Loisir et du Sport, 2006 ; Scallon, 2004) du jugement contribue à entretenir le doute sur la validité des décisions découlant de cette démarche. À ce jour, très peu de recherches ont été menées sur l'évaluation fondée principalement sur le jugement.

Dans ce contexte, comment vérifier la validité du jugement professionnel des enseignants ? Le présent chapitre tente de répondre à cette question en proposant un modèle qui se veut pratique. Dans les lignes qui suivent, nous décrivons sa structure et son articulation. Mais d'abord, une étude documentaire sur le sujet permet de faire le point sur cette question. On aborde également la définition des concepts cruciaux, en l'occurrence celui de la validité et celui du jugement, accompagnée des critères de qualité rattachés à ces deux concepts. La section suivante présente le modèle proposé.

2. CONTEXTE THÉORIQUE

2.1. Définition des concepts

2.1.1. *La validité*

Bertrand et Blais (2004, p. 240) ont formulé une définition holistique, mais surtout opérationnelle du concept de validité : *La validité consiste en un jugement basé sur des preuves empiriques et une argumentation de nature théorique qui vise à justifier l'interprétation des scores obtenus à la*

suite de l'administration d'un test dans un contexte donné. Ces auteurs insistent sur le caractère procédural de la démarche de validation des résultats, car, selon eux, la démarche se résume en un processus d'accumulation de preuves, lesquelles font l'objet d'un jugement rigoureux. Il s'agit donc d'une démonstration par la preuve ou d'une argumentation logique et solide. Il apparaît également dans cette définition que la validité d'un jugement est intimement liée à celle de l'argumentaire c'est-à-dire la preuve qui le sous-tend. L'argumentaire est constitué des différentes étapes du processus de collecte des informations qui seront ultimement interprétées. Autrement dit, non seulement l'interprétation des informations doit être rigoureuse, mais il faut également observer la même rigueur dans chacune des étapes du processus visant à recueillir ces données. La validité est donc un jugement devant être porté sur l'argumentaire, construit pour démontrer le bien-fondé de l'interprétation des résultats observés dans une situation d'évaluation donnée, et ce, tout au long de la démarche évaluative.

Par ailleurs, la définition de Bertrand et Blais (2004) s'apparente, à bien des égards, à celle de Kane (2006), qui conçoit la validité comme un processus argumentatif constitué de quatre étapes ou échelons : la notation (*scoring*), la généralisation, l'extrapolation et l'implication (Paquette-Côté, 2008). Le modèle que nous proposons porte essentiellement sur le premier niveau du modèle de Kane, soit la notation. Car, en évaluation des compétences, *la validité met en jeu deux composantes principales de la démarche d'évaluation : les situations d'évaluation (tâches) et les outils de jugement* (Scallon, 2004, p. 261). Dans le modèle de Kane, les pratiques évaluatives des enseignants du primaire sont surtout de l'ordre de la notation. Toutefois, l'emploi des outils de jugement et la prescription par le ministère de l'Éducation, du Loisir et du Sport de la pratique du travail en équipe de cycle ou en équipe-école ont entraîné implicitement l'expérimentation des autres niveaux du processus argumentatif du modèle de Kane. Par exemple, la suffisance des observations et leur représentativité par rapport à l'objet d'évaluation assurent un certain niveau de contrôle des biais. Cette tâche est partagée entre l'enseignant et son équipe de cycle, qui doivent prescrire les règles et modalités appropriées visant à garantir la crédibilité du jugement qui en découlera. Ces différentes actions correspondent au deuxième échelon du modèle de Kane : la généralisation.

2.1.2. *Le jugement*

Le jugement consiste à faire une analyse et une synthèse des données recueillies sur les apprentissages de l'élève (MELS, 2006). Pour Lafortune (2008), le jugement est une démarche qui consiste à se

prononcer sur le développement de compétences. Le jugement professionnel, quant à lui, se fonde sur des informations recueillies dans des situations complexes. Cette complexité suppose une rigueur accrue afin d'assurer un jugement valable, car par définition, un jugement comporte toujours une part de subjectivité (Lafortune, 2008). Selon Gerard (2002 : voir Durand, 2008, p. 63), *l'objectivité de l'évaluation est impossible, car la subjectivité est inévitablement présente et nécessaire ; évaluer, c'est donner du sens au résultat observé, donc une opération subjective*. Dans le même ordre d'idées, Durand (2008, p. 66) affirme que *le jugement est un acte professionnel qui ne peut revêtir un caractère de totale objectivité. La posture de l'enseignant évaluateur est déterminante*.

Devant l'inéluctable subjectivité dans le jugement professionnel, le ministère de l'Éducation, du Loisir et du Sport (2006, p. 6) a prescrit une série d'actions visant à assurer la crédibilité des jugements portés sur les apprentissages des élèves. Parmi ces actions, il est suggéré que l'enseignant travaille en collaboration avec ses collègues ou avec d'autres professionnels. L'individualisme dans ce domaine n'est pas conseillé. Cette référence aux collègues concerne implicitement l'équipe de cycle ou l'équipe-école puisque le renouveau pédagogique prescrit ces modalités de travail. Le ministère de l'Éducation, du Loisir et du Sport soutient aussi qu'il est essentiel que les enseignants prennent tous les moyens pour éliminer le plus possible la subjectivité lorsqu'ils ont à porter des jugements. Car, en plus de pouvoir demander une justification ou même une révision du jugement de l'enseignant, les parents, les élèves et la direction de l'école peuvent le remettre en question.

Ainsi, la validité d'un jugement est tributaire d'un ensemble d'actions caractérisées par la rigueur. À cet égard, le ministère de l'Éducation, du Loisir et du Sport (2006) affirme que la condition fondamentale pour une évaluation de qualité est que celle-ci doit allier rigueur et transparence. La rigueur implique, de la part de l'enseignant, de définir l'objet de l'évaluation, de préciser les cibles des apprentissages qui doivent être vérifiés, et de choisir et utiliser divers méthodes et outils appropriés dans le contexte de l'évaluation. Durand et Chouinard (2006, p. 301) formulent les mêmes recommandations afin d'en arriver à un jugement éclairé et fondé. Quant à la transparence en évaluation, elle exprime surtout le souci constant que doit manifester l'enseignant pour informer les élèves des attentes à leur égard, relativement aux apprentissages, aux critères et aux exigences appliqués. Il s'agit aussi d'expliquer tant aux élèves qu'aux parents les jugements et décisions concernant les apprentissages de l'élève. Cette justification constitue l'essence même de l'argumentaire évoqué précédemment. C'est la preuve du bien-fondé du jugement qui en découle.

En somme, compte tenu du caractère hautement subjectif du jugement, il est crucial que cette tâche ne soit pas réservée à une seule personne. Plusieurs facteurs fortifient cette position. Parmi ceux-là, il y a la nature complexe des objets d'évaluation dans l'approche par compétences, le potentiel de remise en question des jugements émis par les enseignants, tant par les élèves et les parents que par la direction des établissements scolaires, et, par-dessus tout, l'usage des outils de jugement pour l'évaluation des compétences. La collaboration apparaît donc comme une condition essentielle pour établir un jugement professionnel éclairé, argumenté, car il se développe nécessairement dans une interaction avec les pairs et dans une pratique réflexive soutenue (Lafortune, 2008). Cette auteure soutient par ailleurs que l'interaction suppose une incertitude et une remise en question continue, des facteurs que le professionnel devra prendre en compte en faisant preuve de tolérance et d'ouverture d'esprit.

Rappelons que le renouveau pédagogique en cours au Québec comporte une prescription fondamentale qui constitue en soi un cadre privilégié des interactions entre les enseignants. Il s'agit de l'équipe de cycle ou de l'équipe-école. Les enseignants partageant les mêmes préoccupations peuvent ainsi échanger et évoluer ensemble dans l'exercice de leur métier. Pour Durand et Chouinard (2006, p. 303), il est plutôt *préférable que les décisions de fin de cycle se prennent en équipe de cycle afin qu'elles soient le plus justes possible*. Dans cette perspective, l'enseignant doit agir en amont (c'est-à-dire en cours de cycle) en alimentant l'équipe de cycle avec les informations nécessaires en vue d'en arriver à un jugement éclairé. En effet, c'est en recueillant suffisamment d'informations pertinentes de façon régulière et en utilisant une instrumentation rigoureuse, que les enseignants se mettent en mesure de porter un jugement bien documenté (Durand, 2008).

Ces critères inhérents au jugement correspondent à ceux qu'on retrouve à l'étape de la notation, une des étapes cruciales du processus conduisant au jugement argumenté, selon le modèle de Kane (2006). Or, la notation comporte des sources de biais qui doivent être contrôlées afin d'assurer la validité du jugement professionnel des enseignants. Durand et Chouinard (2006, p. 298) ont en effet répertorié sept sources de biais de la notation susceptibles d'affecter le jugement des enseignants. Il s'agit notamment de :

1. la place de la production de l'élève dans la pile : effet de l'ordre ;
2. l'effet de halo ;
3. les attentes et les critères émergents ;
4. l'effet de contamination ;
5. la personnalité du correcteur, sa posture ;

6. l'effet de contraste ;
7. la fatigue du correcteur.

1. La place de la production de l'élève dans la pile

La place de la production de l'élève dans la pile de copies peut influencer son évaluation. Il est donc fortement recommandé de corriger un critère à la fois pour toutes les copies, de les changer de place pour le second critère et ainsi de suite. Mais l'efficacité de cette procédure reste tributaire de la probité, du sens des responsabilités, de l'éthique et de la compétence professionnelle du **seul** enseignant évaluateur. L'équipe de cycle pourrait se répartir cette charge tout en renforçant la crédibilité du jugement qui en découle. Nous proposons que chaque enseignant de l'équipe de cycle ou de l'équipe-école se charge de la correction d'un seul critère à la fois pour l'ensemble des élèves du cycle dans des tâches complexes. L'équipe participerait alors, de façon concertée, à l'évaluation des apprentissages.

2. L'effet de halo

Il s'agit de l'effet sur l'évaluation de la connaissance par l'évaluateur de certains attributs de l'élève évalué. Par exemple, le fait de savoir qu'un élève a de façon récurrente de mauvais résultats (ou à l'inverse, de très bons résultats) pourrait déteindre sur le jugement du correcteur. C'est pourquoi il est recommandé de corriger anonymement les copies. Dans ce contexte, on peut raisonnablement penser qu'une correction par l'équipe de cycle ou l'équipe-école réduirait considérablement la portée de l'effet de halo. La crédibilité du jugement serait davantage renforcée si chaque membre de l'équipe ne corrigeait qu'un seul critère d'une tâche complexe.

3. Les attentes et les critères émergents

Les attentes et critères émergents concernent le biais important occasionné par la tendance que peut avoir l'enseignant évaluateur à modifier sa façon de corriger en cours de correction, créant ainsi de nouveaux critères. La façon de contrôler ce type de biais est de définir préalablement les critères et de les communiquer aux élèves, quitte à les modifier en concertation avec eux s'ils s'avèrent peu clairs. Ce faisant, la notion de transparence (critère fondamental d'un jugement de qualité) serait assurée. Dans un contexte d'évaluation par l'équipe de cycle, ce problème devrait être moins apparent puisque les critères seraient préalablement définis.

4. L'effet de contamination

S'apparentant à l'effet de halo, l'effet de contamination se rapporte au fait que l'enseignant est influencé par les notes antérieures de l'élève. La réputation de l'élève, bonne ou mauvaise, en matière de résultats scolaires peut influencer les évaluations que l'enseignant fera de ses productions. Ici aussi, l'effet de contamination sera sensiblement limité dans un contexte de correction par les membres de l'équipe de cycle ou de l'équipe-école.

5. La personnalité du correcteur, sa posture

Durand et Chouinard (2006, p. 300) ont trouvé quatre postures (contrôleur, entraîneur-coach, conseiller ou consultant) que peut adopter un enseignant évaluateur et qui sont susceptibles d'entraîner une variation du jugement d'un groupe à l'autre appartenant au même niveau. À l'instar de tout groupe d'individus, il y a des enseignants qui sont plus sévères que d'autres et d'autres encore qui sont plutôt indécis. Ces derniers ont tendance à accorder à tous les élèves d'un groupe donné une série de notes variant très peu ou pas du tout (quasi-uniformité de la note accordée à un groupe d'individus). Durand et Chouinard suggèrent, entre autres, le recours à l'équipe de cycle pour surmonter cette difficulté liée à la personnalité du correcteur. C'est aussi le fondement de la présente proposition. En effet, nous proposons d'aller plus loin encore. Il s'agit en fait de procéder à une correction partagée des tâches complexes, avec un seul critère par enseignant. Le facteur de la personnalité du correcteur n'entrerait plus en ligne de compte.

6. L'effet de contraste

Pour ce qui de l'effet de contraste, il correspond au cas où l'enseignant estimerait qu'une des copies satisfait parfaitement aux attentes pour devenir, à tort ou à raison, la copie de référence pour toutes les autres. Tout comme dans le cas des biais relatifs aux critères émergents, le correcteur a tendance à s'éloigner des critères initiaux et à en adopter de nouveaux, représentés par la *copie modèle*. Une fois de plus, dans un contexte de partage du processus d'évaluation en équipe de cycle ou en équipe-école, l'effet de contraste s'en trouverait minimisé, voire annihilé. Durand et Chouinard (2006) proposent comme solution de définir les critères au préalable et d'en informer les élèves. Si cette procédure assure la transparence quant au processus d'évaluation, rien n'est moins sûr en ce qui concerne le contrôle du biais relatif à l'effet de contraste.

7. La fatigue du correcteur

Dans une correction de productions complexes pour un grand groupe, il est possible que l'attention et la précision du correcteur tendent à diminuer après la correction de quelques copies. C'est l'effet de la fatigue. Durand et Chouinard (2006) suggèrent alors de recourir aussi aux traces antérieures laissées par l'élève et aux anecdotes que l'enseignant évaluateur aura recueillies sur le travail de l'élève avant la présente production. Cette façon de procéder permet d'atténuer l'effet de fatigue du correcteur. Elle constitue une évaluation complète qui porte autant sur la production que sur les stratégies adoptées par l'élève. Dans le contexte d'évaluation partagée avec l'équipe de cycle ou l'équipe-école, l'enseignant recueillera les informations sur le travail de l'élève. Les productions complexes seront évaluées en collaboration avec les membres de l'équipe de cycle. Ainsi, l'évaluation d'une production réalisée par l'équipe de cycle sera complétée par les informations recueillies par l'enseignant attitré pendant les apprentissages ou en cours de cycle.

Il convient de relever ici la différence entre la notation telle que la conçoivent Durand et Chouinard (2006) et la notation selon Kane (2006). Pour les premiers, il s'agit de l'acte de correction des productions complexes des élèves, tandis que pour Kane, la notation englobe tout ce qui relève de l'administration d'un test. On s'assure aussi du contrôle de l'erreur aléatoire en choisissant un échantillon suffisamment grand des observations (Paquette-Côté, 2008). Ainsi, la validité à l'étape de la notation dans le modèle de Kane demande d'assurer la pertinence des règles de passation et de notation du test, l'application rigoureuse de ces règles, l'absence de biais tant dans la notation qu'au cours de la passation et enfin, l'adaptation des données découlant de la notation aux types d'échelles utilisés. Elle inclut donc la notation au sens que lui donnent Durand et Chouinard. Assurer la validité du jugement professionnel des enseignants revient à minimiser le plus possible, à défaut de les éliminer complètement, les biais du processus de notation. Comme il a été dit plus haut, la validité est un processus d'accumulation de preuves visant à assurer la qualité du jugement.

Pour sa part, Scallon (2004) a répertorié une série de critères permettant d'apprécier la validité des démarches d'évaluation des habiletés complexes ou des compétences. À l'aide de ces critères, on examine tour à tour chaque situation-problème ou tâche, l'ensemble des situations-problèmes, l'outil de jugement, les effets du procédé d'évaluation et l'application des principes de justice et d'équité. Ce faisant, on vérifie divers types de validité, notamment la validité de construit, la validité de contenu, la validité d'apparence et la validité de

conséquence. À cet égard, Scallon (2004) aborde la question de la validité d'un point de vue plutôt analytique, par opposition à l'approche globale préconisée par Bertrand et Blais (2004). Examinons les arguments avancés par Scallon concernant les différents types de validité qu'il a ainsi énumérés.

L'examen de la situation-problème consiste à vérifier que celle-ci a suscité la mobilisation de ressources de même nature pour y répondre. Cela passe par l'observation des traces laissées par l'élève, son cheminement. La validité de contenu se vérifie par le rapport entre la situation d'évaluation et la situation d'apprentissage, c'est-à-dire la similitude entre les situations sources et les situations cibles, étant entendu que ces dernières visent l'évaluation. Cela correspond au 2e niveau du modèle de Kane (2006). À cette étape, on s'assure de la représentativité de l'échantillon de l'univers de généralisation à partir d'un échantillon des observations.

Pour ce qui est de la validité d'apparence, elle est déterminée par la perception qu'ont les sujets des situations d'évaluation. Selon Scallon (2004, p. 263), la question fondamentale inhérente à la validité d'apparence réside dans le fait qu'on doit toujours préciser sa portée. Il s'agit de déterminer si celle-ci se limite aux situations d'évaluation ou englobe toute la démarche nécessitant un jugement de plusieurs personnes. Scallon ajoute (p. 265) : *La validité des grilles d'évaluation ou des échelles d'appréciation (analytiques ou globales) est étroitement liée à la validité des situations-problèmes, conçues pour inférer des habiletés complexes ou des compétences.* La validité de conséquence, quant à elle, concerne les retombées attendues de l'évaluation. À cet égard, Scallon (*ibid.*) déclare :

> *Certaines pratiques d'évaluation ont des conséquences positives sur la motivation des élèves, mais aussi sur l'application des programmes d'études. Vérifier la capacité d'utiliser savoirs et savoir-faire, par exemple, ne peut qu'inciter les enseignants à mieux préparer leurs élèves en ce sens.*

Ainsi, tant dans la perspective analytique (Scallon, 2004) que dans la perspective holistique (Bertrand et Blais, 2004), la validité demeure un processus de jugement. Il est raisonnable de penser que la pratique d'une évaluation collégiale améliorerait la validité des évaluations des tâches complexes. Une recommandation de Scallon (2004, p. 263) va dans ce sens : *Cette appréciation ne peut relever d'une seule personne. La démarche devrait être entreprise au sein de groupes d'enseignants partageant les mêmes préoccupations d'évaluation ou par des conseillers pédagogiques ou plusieurs personnes responsables de la conception des situations d'évaluation.*

Les enseignants de l'équipe de cycle agiraient comme un jury pour porter un jugement sur une production complexe. Ce jugement serait nettement moins subjectif s'il était prononcé par un jury. C'est précisément ce qui est attendu de l'équipe de cycle ou de l'équipe-école dans le modèle que nous proposons et dont la structure est analysée dans la section suivante.

3. APPROCHE THÉORIQUE

L'approche théorique adoptée dans ce chapitre est celle du développement d'un modèle. Willet (1992, p. 33) définit le modèle comme étant *une description et une représentation schématique, systématique et consciemment simplifiée d'une partie du réel, faites au moyen de signes, de symboles, de formes géométriques ou graphiques et de mots*. La limite de cette définition réside dans le fait que seule la structure ou la forme du modèle est considérée. À cet égard, la proposition de Mucchielli (2004, p. 153) précise cette définition du modèle puisqu'elle s'attache à ses fonctionnalités : *Le modèle fournit une représentation simplifiée d'un type de phénomène particulier en vue de faciliter la compréhension*. Ainsi, en clair, un modèle est un instrument visant à expliciter un phénomène plutôt complexe. Notre proposition conjugue ces deux aspects de la définition du modèle qui consiste à préciser tant sa forme que ses fonctionnalités.

4. DÉVELOPPEMENT DU MODÈLE

4.1. Description du modèle

Le modèle proposé détermine les différents points d'ancrage du jugement professionnel des enseignants en faisant appel à un processus évaluatif. Ce jugement étant toujours en filigrane dans les différentes étapes du processus évaluatif (McMillan, 2000 : voir Durand et Chouinard, 2006 ; ministère de l'Éducation, du Loisir et du Sport, 2006 ; Durand, 2008), les éléments de validité associés à chacune des étapes de ce processus sont aussi mis en lumière. Dans l'évaluation par compétences, le processus évaluatif comporte plusieurs étapes dont certaines sont caractéristiques de l'enseignant ou lui sont même exclusives, et d'autres sont partagées avec son équipe de cycle ou équipe-école.

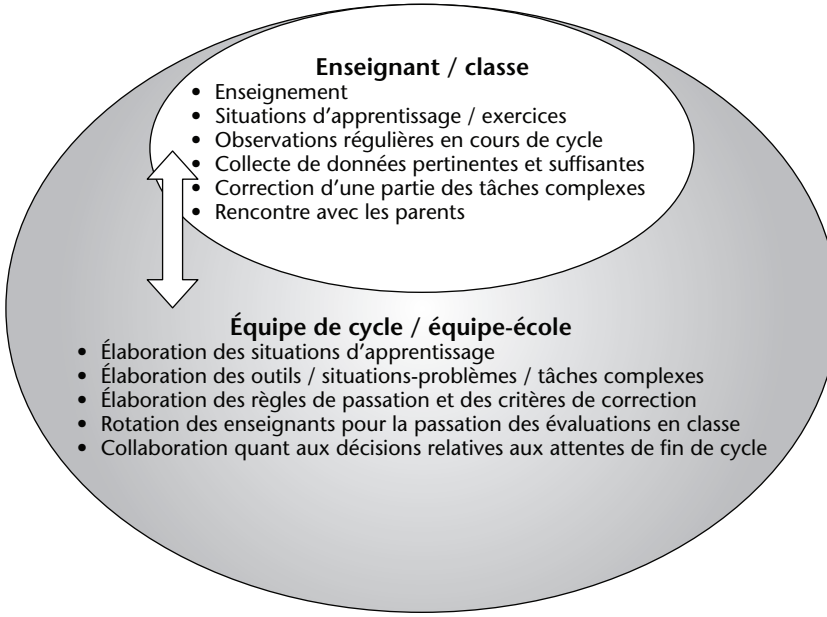


Figure 6.1
Interfaces de rôles distinctifs de l'enseignant et de l'équipe de cycle

4.1.1. *Rôle de l'enseignant*

L'enseignant est le premier acteur de l'évaluation des apprentissages des élèves. Le gouvernement du Québec (ministère de l'Éducation, du Loisir et du Sport, 2006) lui reconnaît à cet égard les compétences de plein droit quant à son jugement professionnel. L'exercice de ce jugement se manifeste dans les actes collectifs ou individuels qu'il pose tout au long du processus évaluatif. Ainsi, il lui revient d'assurer l'enseignement des notions ou des contenus prévus dans le programme de formation. Pour ce faire, il choisit les méthodes pédagogiques. Cependant, ce choix ne s'effectue pas au hasard. Il résulte d'une démarche articulée que l'enseignant entreprend en s'appuyant sur plusieurs ressources. Parmi celles-ci se trouve l'équipe de cycle. Elle répertorie les méthodes pédagogiques et les stratégies adaptées à chaque situation d'apprentissage et laisse à l'enseignant attitré le soin de choisir celle qui sera effectivement appliquée en classe. La sélection des méthodes pédagogiques adaptées à chaque situation d'apprentissage par l'équipe de cycle constitue une action qui assure une certaine validité dans la mesure où les membres de l'équipe agissent en concertation. Cette collégialité est considérée comme étant nécessaire, car elle contribue à parfaire la

formation continue des enseignants. Cet état des choses peut être le reflet d'un partage par les enseignants de valeurs essentielles, comme le montre Lusignan (2008, p. 30) : *Des enseignants qui se respectent et se font confiance sur le plan pédagogique peuvent aisément se donner des objectifs communs de formation continue et contribuer à se former mutuellement. Ils sont alors moins compétitifs et travaillent davantage en collégialité.*

L'enseignant oriente les élèves dans les situations d'apprentissage et les exercices qui s'y rattachent. Pour ce faire, il met en œuvre ses qualifications de pédagogue. Il interagit avec les élèves, les accompagne en situation d'apprentissage ; il intervient au besoin, régule la méthode pédagogique afin de tirer le meilleur parti des apprentissages. C'est aussi à l'enseignant que revient la tâche d'effectuer les observations régulières en cours de cycle. Il consigne les traces des apprentissages réalisées par l'apprenant. Ces traces seront utiles lors de la prise de décision en fin cycle ; elles compléteront l'analyse faite par l'équipe de cycle quant à la progression de l'élève en fonction des attentes. Dans cette même veine, l'enseignant collige toutes les informations pertinentes et nécessaires en vue des délibérations de l'équipe de cycle.

Certes, dans l'exercice de ces différentes actions, l'enseignant paraît agir seul. Ainsi, on peut s'interroger sur le niveau de validité des données recueillies. Rappelons-nous cependant que l'enseignant agit dans le prolongement des initiatives concertées de l'équipe de cycle ou de l'équipe-école. En effet, il est suggéré que l'enseignant ne corrige qu'une partie des tâches complexes, celles qui concernent un seul critère par exemple. Ce faisant, il devra cependant corriger toutes les copies des élèves du cycle, y compris celles des élèves des classes des autres membres de l'équipe de cycle. L'enseignant n'est donc plus seul à apprécier les productions portant sur les tâches complexes. Cette façon de procéder assure la validité apparente (Scallon, 2004) des données ainsi recueillies.

Une autre tâche de l'enseignant, et non la moindre, consiste à communiquer avec l'élève, la direction de l'école et les parents. Il s'agit de les informer de ce qui se passe en classe quant au déroulement des apprentissages, au développement des compétences. Rappelons que les parents, la direction de l'école de même que l'élève peuvent remettre en question le jugement de l'enseignant en matière d'évaluation des compétences (MELS, 2006). Dans ce contexte, la concertation des membres de l'équipe de cycle au cours des différentes étapes de la démarche évaluative semble être un gage d'assurance pour l'enseignant, dans la mesure où sa décision est l'aboutissement d'un processus collégial fondé et raisonné. L'enseignant n'est pas seul à avoir porté un

jugement sur le produit résultant d'une tâche complexe. Il peut alors se présenter en toute confiance devant ses interlocuteurs et montrer le bien-fondé de sa décision.

4.1.2. *Rôle de l'équipe de cycle*

L'équipe de cycle est constituée d'enseignants qui partagent les mêmes préoccupations par rapport à une discipline et qui ont des élèves appartenant à un même cycle d'enseignement. Ainsi, il y aurait trois équipes de cycle pour chaque école primaire qui compte trois cycles, soit de la première à la sixième année. Quant à l'équipe-école, elle regroupe les enseignants d'une même école qui décident de se concerter, et ce, au-delà de la délimitation des cycles. Le modèle proposé dans ce chapitre situe le rôle de l'équipe de cycle ou de l'équipe-école sur plusieurs plans, notamment celui de l'élaboration des situations d'apprentissage, de l'élaboration des outils ou des situations-problèmes (tâches complexes), de l'élaboration des règles de passation et des critères de correction, de la rotation des enseignants pour la passation des évaluations en classe, de la correction d'un seul critère par chaque enseignant de l'équipe de cycle ou de l'équipe-école, et ce, pour l'ensemble des sujets du cycle et de la collaboration quant aux décisions relatives aux attentes de fin de cycle, après validation de l'enseignant attitré.

Le principe de la collégialité des membres de l'équipe de cycle fait de l'élaboration des situations d'apprentissage un des éléments clés de la validité du jugement dans le contexte d'évaluation par compétences. Les membres de l'équipe de cycle ont ainsi la possibilité d'améliorer les situations d'apprentissage par des échanges et des discussions. Ces situations d'apprentissage constitueront une réserve dans laquelle les enseignants pourront choisir celle qui sera appliquée. Cette participation collective à l'élaboration des situations d'apprentissage assure sa pertinence et la cohérence des tâches par rapport au contenu du programme.

L'élaboration des situations-problèmes permet aux membres de l'équipe de cycle d'assurer la validité de contenu. En effet, à l'instar des situations d'apprentissage, les situations-problèmes font l'objet d'une étroite collaboration entre les membres de l'équipe de cycle. À cet égard, plusieurs scénarios sont possibles. Par exemple, les membres de l'équipe peuvent proposer que chacun élabore d'abord une ou deux situations-problèmes qui seront ensuite examinées et discutées en équipe. À cette étape du processus, consacrée à la construction des tâches, les discussions portent surtout sur la cohérence des situations-problèmes par rapport au contenu du programme. La nature de la

situation-problème est aussi analysée, c'est-à-dire que l'on s'assure que son niveau de complexité est approprié par rapport aux attentes et au contenu prescriptif du programme.

Après la construction des situations-problèmes, l'équipe doit aussi s'atteler à la tâche de rédaction des instructions ou des règles d'administration de celles-ci. Cette démarche vise essentiellement à prescrire les conditions optimales d'utilisation et d'application des situations-problèmes. Cela concerne autant la durée que le contenu d'apprentissage requis avant l'évaluation. La prescription des règles d'administration des situations-problèmes a aussi un effet de standardisation dans la mesure où un autre enseignant que celui qui est attitré à une classe donnée peut assurer l'exercice avec très peu d'incidence (voire aucune) quant aux sujets visés. Il en va de même pour les suppléants qui remplacent occasionnellement le titulaire. Au plan scientifique, ce contrôle *a priori* des biais de méthodes assure la comparabilité interclasse des résultats des élèves d'un même cycle.

Mais, avant d'en arriver à cette comparaison, il faut aussi avoir respecté les mêmes critères de corrections des productions issues des situations-problèmes ou tâches complexes. Ces critères sont élaborés en même temps que se forment les situations-problèmes. Il s'agit d'un canevas retraçant les grandes lignes du processus, qui laisse aussi place à une certaine liberté de la part du correcteur quant au niveau d'appréciation. Car, en fin de compte, il s'agit d'un jugement, et tout jugement porte sa part de subjectivité.

La standardisation des critères s'accompagne d'un autre élément important : la rotation des enseignants pour la passation des évaluations en classe. Il est en effet proposé que les enseignants de l'équipe de cycle procèdent à une rotation en vue de la passation des évaluations en classe. Cela permet de minimiser l'effet de l'administrateur unique. Imaginons le scénario où c'est le même enseignant qui administre les situations-problèmes à ses propres élèves. La tentation peut s'avérer forte pour lui de fournir des explications susceptibles de donner des indices de réponse à ses élèves en situation d'incompréhension ou de faiblesse face à une tâche complexe donnée. La rotation des enseignants dans les classes du cycle permet à chacun d'eux de s'en tenir aux critères et aux conditions établis pour chaque situation-problème. De ce fait, on contrôle la source de biais de méthode liés à l'administrateur (à défaut de la neutraliser ou de l'éliminer complètement), ce qui contribue à améliorer la validité des informations recueillies.

La correction d'un seul critère par enseignant, et ce, pour l'ensemble des élèves d'un même cycle, constitue un autre levier pour assurer la validité du jugement. En effet, même avec des critères

communs bien définis, il est possible que certains enseignants se montrent plus ou moins indulgents que d'autres dans l'application de ces critères. Ce phénomène, Durand et Chouinard (2006) l'appellent *la posture ou personnalité du correcteur*. Imaginons ce phénomène dans une situation où chaque enseignant utilise les mêmes critères que les autres membres de l'équipe-cycle, mais où chacun corrige tous les critères de la situation-problème. Les écarts entre les élèves du même cycle seraient directement attribuables à la différence de rigueur des correcteurs quant à l'application des critères de correction. En ne faisant corriger qu'un seul critère par chaque enseignant du cycle, le biais introduit par la différence possible des niveaux de rigueur des correcteurs à appliquer les critères de correction se trouve éliminé. Ainsi, au lieu de favoriser certains élèves du cycle au détriment des autres, l'écart entre les enseignants et les niveaux de rigueur des correcteurs sera distribué sur l'ensemble des copies. Les différences entre les résultats obtenus refléteraient davantage des différences réelles entre les résultats observés qu'un défaut du processus de correction. La validité des résultats serait donc assurée.

Avant l'étape ultime de communication des résultats à qui de droit, les membres de l'équipe de cycle doivent collaborer en vue de porter un jugement quant au niveau de développement des compétences de chaque élève. Ils délibèrent en interprétant les performances observées sur la base des attentes de fin de cycle. Durant ces délibérations, le titulaire de la classe de l'élève dont la performance est examinée agit comme personne-ressource pour donner à l'équipe de cycle les informations pertinentes tirées de ses observations personnelles. Il est le témoin privilégié des réalisations de l'élève en cours de cycle. L'ensemble des données recueillies par le titulaire, y compris la performance de l'élève dans les situations-problèmes administrées par un membre de l'équipe de cycle, constitue le fondement de l'argumentaire du jugement global porté sur le niveau de développement des compétences de l'apprenant. Ainsi, ce jugement est l'expression d'un consensus de l'équipe de cycle. Dans ce contexte, la rencontre avec les parents, qui consiste fondamentalement à renseigner ou à éclairer ceux-ci sur le travail de l'élève, constitue aussi une occasion pour l'enseignant d'expliquer ou de justifier le jugement découlant de ce processus. La figure 6.2 illustre les interactions possibles dans le cas d'une équipe de cycle comprenant quatre membres. On peut l'envisager avec trois membres ou davantage. Mais il ne faut pas perdre de vue qu'à l'instar de tout travail d'équipe, l'efficacité de l'équipe est inversement proportionnelle au nombre de participants.

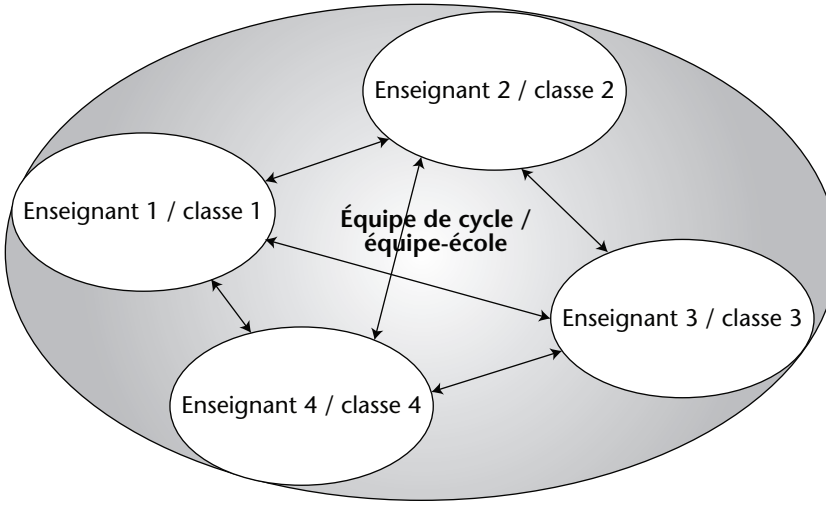


Figure 6.2
Interactions entre quatre enseignants de l'équipe de cycle

4.2. Considérations éthiques

Au plan éthique, il importe de signaler que l'enseignant attitré à une classe donnée demeure le seul interlocuteur des élèves, des parents et de la direction de l'école pour tout ce qui concerne la communication des résultats. Le rapport entre enseignant et élèves doit être absolument préservé. On doit donc s'assurer que le travail de collaboration en équipe de cycle ne sera pas perçu par les élèves comme une occasion de divulguer à tous les membres de cette équipe leur performance individuelle. Cela vaut aussi pour les parents d'élèves. Ainsi, on pourrait justifier la collaboration des membres de l'équipe de cycle par l'importance de s'assurer de la pertinence des situations d'apprentissage, la qualité des tâches complexes proposées, la pertinence des critères d'évaluation de ces tâches et les interprétations qui s'y rattachent. Ces arguments demeurent valables même si la décision prise en fin de cycle résulte d'une concertation entre les membres de l'équipe de cycle.

5. CONCLUSION

Le modèle proposé dans ce chapitre vise à assurer la validité du jugement des enseignants du primaire dans le contexte de l'approche par compétences. Pour ce faire, il expose l'opérationnalisation du travail en

équipe de cycle. Les membres de l'équipe de cycle y agiraient comme un jury pour porter un jugement reposant sur des arguments. En plus de la validité du jugement, on assure par le fait même l'équité envers les élèves. Le modèle définit aussi les rôles et les fonctions de l'enseignant en tant que responsable d'une classe dans un cycle donné, d'une part, et en tant que membre d'une équipe de cycle, d'autre part.

Ce modèle mise grandement sur la collaboration optimale des participants membres de l'équipe de cycle. Ceux-ci doivent faire corps et travailler comme une entité pour assurer le bon fonctionnement du modèle. Ainsi, chaque membre de l'équipe de cycle doit s'engager pleinement et faire preuve de tolérance et d'ouverture d'esprit. Le jugement professionnel de l'enseignant membre de l'équipe de cycle dépend de ses qualités personnelles ; l'enseignant est constamment en proie à l'incertitude et il se remet en question (Lafortune, 2008). Assurer la validité du jugement professionnel des enseignants dans le contexte actuel de l'approche par compétences relève de la responsabilité individuelle et collective.

Toutefois, la poursuite de cet objectif peut s'avérer difficile, car les membres de l'équipe sont des individus aux personnalités et caractères variés. La formation des maîtres ne prépare pas non plus suffisamment les futurs enseignants à une collaboration aussi poussée que celle qui est préconisée dans ce modèle. Pourtant, nous pensons qu'il s'agit d'une voie toute tracée dans le contexte du renouveau pédagogique en cours au Québec et de l'évaluation des compétences (tâches complexes). Cette difficulté pourrait être surmontée avec le concours de la direction de l'école, qui créerait une dynamique en instaurant un climat favorable à la collaboration entre les membres de l'équipe de cycle. Par exemple, la direction pourrait s'assurer de la disponibilité des professionnels qui soutiendraient et encourageraient l'équipe de cycle dans son travail.

L'expérimentation de ce modèle constituera le premier jalon vers la mise en œuvre de futures pistes de recherche. Cette mise en pratique prendrait la forme d'un projet pilote permettant d'enrichir le modèle et d'ouvrir des pistes de recherche. Par exemple, le projet pilote pourrait servir de sujet de mémoire de maîtrise. Ensuite, on pourrait envisager d'étudier l'effet de cette application sur la perception qu'ont les enseignants de leur charge de travail, les changements des pratiques pédagogiques, la motivation des apprenants, etc.

RÉFÉRENCES

Angoff, W. H. (1988). Validity : an evolving concept. Dans H. Wainer et H.I. Braun (dir.), *Test validity*. Hillsdale, New Jersey : Lawrence Erlbaum.

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

Durand, M.-J. (2008). La démarche d'évaluation dans une approche basée sur le jugement professionnel. *Vie pédagogique, 148*, p. 63-67.

Durand, M.-J. et Chouinard, R. (2006). *L'évaluation des apprentissages. De la planification de la démarche à la communication des résultats*. Montréal, Québec : Hurtubise HMH.

Gerard, F.-M. (2002). L'indispensable subjectivité de l'évaluation. *Antipode, 156*, avril, p. 26-34.

Kane, M. T. (2006). Validation. Dans R. L. Brennan (dir.), *Educational measurement* (4e édition). Westport, Connecticut : Praeger.

Lafortune, L. (2008). *L'éthique et le jugement professionnel dans le référentiel de compétences professionnelles pour l'accompagnement d'un changement prescrit*. Communication présentée à l'Association pour le développement des méthodes d'évaluation en éducation (ADMÉE).

Laveault, D. et Grégoire, J. (1997). *Introduction aux théories des tests en sciences humaines*. Bruxelle, Belgique : De Boeck Université.

Lusignan, G. (2008). De la formation initiale à la formation continue : un continuum à établir. *Vie pédagogique, 147*, p. 27-32.

McMillan, J. H. (2000). *Basic assessment concepts for teachers and school administration*. Paper presented at the annual meeting of the American Educational Research Association, La Nouvelle-Orléans, Louisiane.

Ministère de l'Éducation, du Loisir et du Sport (2006). *La valeur accordée au jugement professionnel des enseignants. Question et éléments de réponse – Principales références dans les encadrements ministériels*. Québec, Québec : Gouvernement du Québec.

Mucchielli, A. (2004). *Dictionnaire des méthodes qualitatives en sciences humaines et sociales*. Paris, France : Armand Colin.

Paquette-Côté, K. (2008). *Analyse de l'argumentation de la validité des inférences d'évaluation dans les politiques institutionnelles d'évaluation des apprentissages des établissements d'enseignement collégial québécois*. Communication présentée à l'Association pour le développement des méthodes d'évaluation en éducation (ADMÉE).

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Saint-Laurent, Québec : Éditions du Renouveau pédagogique.

Willet, G. (1992). *La communication modélisée*. Ottawa, Ontario : Éditions du Renouveau pédagogique.

Chapitre 7

Estimation du degré d'accord entre des experts lors du calibrage d'un test de concordance de script avec le modèle à facettes de Rasch

Jean-Guy Blais, Bernard Charlin, Julie Grondin,
Carole Lambert, Nathalie Loye et Robert Gagnon

Ce chapitre présente une démarche reposant sur le modèle à facettes de Rasch pour déterminer le degré d'accord entre des experts quant au classement d'une série de 30 items d'un test de concordance de script ayant comme objectif l'évaluation de la compétence en raisonnement clinique dans le domaine de la radio-oncologie du poumon.

1. INTRODUCTION

Plusieurs situations d'évaluation en éducation demandent l'assistance de personnes identifiées comme des juges, des experts ou des correcteurs et qui ont comme tâche de classer des objets, des observations, des performances ou des éléments d'un ensemble dans des catégories qualitatives déterminées et le plus souvent ordonnées. Ainsi en est-il de l'appréciation d'une vaste gamme de performances et de compétences complexes dans des situations qui ne peuvent être totalement standardisées avec des résultats dichotomiques, c'est-à-dire des bonnes et des mauvaises réponses, mais qui nécessitent plutôt des appréciations nuancées représentant différents niveaux du construit visé. Il est donc tout à fait possible que ces juges, experts ou correcteurs ne soient pas du même avis et que le classement réalisé ne soit pas le même d'une personne à l'autre.

Sans être parfait, le degré de l'accord entre les personnes doit cependant être tout de même assez élevé, car la validité du classement et des décisions qui peuvent en découler dépend aussi de cet accord. Dans ces situations, la validité se retrouve alors au confluent des caractéristiques de l'outil d'évaluation, la plupart du temps une grille d'appréciation avec des critères et des échelles ordonnées pour consigner la cote, et du degré de convergence des jugements des personnes concernées. La possibilité que les jugements de toutes les personnes appelées à effectuer la tâche de classement divergent, quelquefois légèrement mais peut-être aussi fortement, illustre l'importance de disposer d'une procédure éprouvée permettant d'identifier les personnes qui produisent les classements les plus éloignés de ceux de la majorité et de déterminer l'impact du retrait des données les concernant. Dans le cadre de ce chapitre, nous allons donc présenter une procédure reposant sur le modèle de mesure à facettes de Rasch, tel que décrit par Linacre (1989, 1994), pour déterminer le degré d'accord des classements réalisés par un groupe d'experts avant et après le retrait de certaines données qui apparaissent problématiques par rapport à la majorité des experts. Les données analysées sont constituées des cotes attribuées par des radio-oncologues à 30 items d'un test de concordance de script ayant comme objectif l'évaluation de la compétence en raisonnement clinique dans le domaine de la radio-oncologie du poumon.

La procédure de notation d'un test de concordance de script s'effectue en deux étapes. Dans un premier temps, des experts, dans ce cas-ci des radio-oncologues, attribuent une cote à chacun des items du test. Dans un deuxième temps, la distribution des cotes attribuées est utilisée pour la notation de résidents dans la spécialité et d'étudiants en médecine. L'article présente donc des résultats de la première étape, centrés sur l'étude du degré d'accord entre les spécialistes qui doivent attribuer une cote aux items d'un test de concordance de script. Cette étude trouve sa pertinence dans le fait que la validité de la procédure de notation des étudiants, lorsque ceux-ci sont soumis à un test de concordance de script, est directement dépendante du degré d'homogénéité ou d'hétérogénéité des cotes attribuées par les spécialistes.

2. CONTEXTE THÉORIQUE

2.1. L'accord entre des observateurs, des experts, des correcteurs ou des juges

La subjectivité des jugements humains fait l'objet d'une réflexion approfondie dans les domaines de l'éducation, de la médecine ou de la psychologie depuis moins d'un siècle. À partir des années 1940, de

nombreux travaux en médecine, en particulier en radiologie ou en psychologie, ont fait naître l'idée que l'observateur peut être une source importante d'erreurs de mesure. Landis et Koch (1975) ont proposé une étude documentaire riche de plus de soixante-dix références sur ce sujet. De nombreux exemples y sont donnés montrant que durant les années 1940 à 1960, les chercheurs de plusieurs domaines commençaient à concevoir l'idée de l'existence d'erreurs intra-observateur et inter-observateurs.

En médecine, l'interprétation de radioscopies des poumons a ainsi donné lieu à de nombreux travaux à une époque où la guérison de la tuberculose dépendait du diagnostic précoce et juste du médecin (Birkelo, Chamberlain et Phelps, 1947 ; Cochrane et Garland, 1952 : voir Landis et Koch, 1975, p. 101 ; Fletcher et Oldham, 1949 ; Yerushalmy, 1947 ; Yerushalmy, Garland, Harkness, Hinshaw, Miller, Shipman et Zwerling, 1951). À partir des années 1970, de nombreuses publications sur l'étude de l'accord et de la fidélité entre des observateurs, des experts, des correcteurs ou des juges voient le jour en psychologie, en médecine, en counseling et en éducation. Plus récemment, dans ce dernier domaine, on peut recenser plusieurs études sur l'accord entre des correcteurs de productions écrites d'étudiants (par exemple, Blais, 1998 ; Boosma, Van Duijn et Snijders, 2001 ; Braun, 1988 ; Congdon et McQueen, 2000 ; Klein et Taub, 2005 ; Longford, 1994 ; Sudweeks, Reeve et Bradshaw, 2005 ; Wolfe et Kao, 1996). La diversité et l'instabilité des jugements humains sur les éléments plus subjectifs des textes produits dans ce contexte soulèvent un certain nombre de questions, qui sont celles qui surgissent dans toutes les situations de ce type. Ainsi, la cote qu'attribuera chacun des correcteurs à certains critères peut être influencée par le moment de la journée, par le temps alloué à chaque correction, par la précision des consignes, par la fatigue due à un grand nombre de corrections, par son expérience, par sa formation, par la difficulté des items ou des tâches à corriger, par le format de l'échelle d'appréciation (nombre de points et étiquettes), par des problèmes personnels, par un environnement inadéquat, etc. (Congdon et McQueen, 2000 ; Lipkins, Jones et Halkitis, 1996 ; Lunz, Stahl et Wright, 1994 ; Murphy et Davidshofer, 1988 ; Wolfe et Kao, 1996).

Le jugement porté pouvant donc varier d'une personne à l'autre et d'une occasion à une autre, Raymond et Houston (1990) ont montré que l'éthique exige l'étude systématique de l'effet possible de l'observateur, de l'expert, du correcteur ou du juge et de ses caractéristiques afin de mieux le comprendre et de mieux le contrôler. Il convient donc d'utiliser un modèle de mesure approprié pour ce faire : un modèle qui permettra de tenir compte de certaines des caractéristiques des

personnes qui pourraient influencer le résultat d'une tâche de classement ou de correction. C'est ce que le modèle à facettes de Rasch permet de faire relativement aisément avec des devis complets ou incomplets (c'est-à-dire avec des valeurs manquantes).

2.2. Le modèle à facettes de Rasch pour l'estimation du degré d'accord

Les modèles de la famille du modèle de Rasch sont des modèles dits de *trait latent* et ils supposent donc que le résultat d'une personne peut être prédit ou expliqué par certaines caractéristiques qui ne sont pas directement observables. Différents modèles peuvent être utilisés selon le contexte et les particularités des données recueillies. Linacre (1989, 1994) a proposé d'utiliser une version relativement simple du modèle de Rasch, un modèle à facettes, pour apprécier le degré d'accord entre des personnes appelées à porter un jugement. Ce modèle permet d'observer l'importance de différentes facettes d'une situation de jugement à l'aide de paramètres uniques indépendants des autres paramètres du modèle. Comme tous les modèles de Rasch, il permet aussi de construire une échelle de mesure à intervalles égaux commune aux différentes facettes et de disposer de statistiques permettant de déterminer la qualité de l'ajustement des données au modèle. Les écarts entre les données réelles et les valeurs attendues, étant donné le modèle choisi, permettent de déceler au cas par cas chacun des juges ou des sujets, objets ou performances présentant un comportement particulier s'éloignant des conditions de mesure considérées adéquates.

L'application de ce modèle nécessite un réseau d'items, de sujets et de juges qui se chevauchent minimalement et où plusieurs combinaisons sont possibles. Dans un devis idéal, tous les juges réalisent toutes les tâches (c'est le cas dans la recherche présentée dans ce chapitre). Mais dans la plupart des situations réelles, ce devis idéal est trop onéreux et les juges partagent seulement une partie des tâches d'évaluation. Heureusement, les analyses avec les modèles de Rasch peuvent aussi être réalisées avec des devis incomplets, c'est-à-dire avec des données manquantes, et ainsi s'appliquer à des situations plus réalistes.

Le modèle à facettes de Rasch place les sujets, les critères (ou les items) et les juges (des experts dans ce cas) sur une même échelle de mesure. Voici un exemple du modèle pour une situation où il y a des sujets, des juges, des critères à apprécier et une échelle de cotes ordonnées pour le classement :

$$P_{nijk} = \frac{e^{B_n - D_i - C_j - F_k}}{1 + e^{B_n - D_i - C_j - F_k}} \quad (1)$$

Si on l'exprime selon le logarithme du rapport de cotes (*odds-ratio*), on obtient la relation ci-dessous :

$$\text{Log}\left[P_{nijk} \Big/ \left(1 - P_{nijk}\right) \right] = B_n - D_i - C_j - F_k \quad (2)$$

Dans ce modèle, P_{nijk} représente la probabilité que la performance du sujet n pour le critère i se voie attribuer la cote k par le juge j ; B_n est la position du sujet n ; D_i est la position du critère i ; C_j est la position du juge j ; F_k est la difficulté associée au fait de passer de la cote $k - 1$ à la cote k.

Accompagnant ce modèle, on retrouve deux statistiques permettant de tester l'ajustement des données. Le calcul de la première statistique, u_i, est basé sur la moyenne des carrés des différences entre la cote du sujet et l'espérance de cette cote telle que la propose le modèle. Le calcul de la seconde statistique, v_i, est basé sur cette même moyenne, pondérée cette fois afin de diminuer l'influence des grands résidus. Ces deux statistiques peuvent être standardisées pour en faciliter l'interprétation ; dans les écrits sur le modèle de Rasch, elles prennent alors le nom de statistiques d'*infit* et d'*outfit* (pour leur calcul, voir entre autres Wright et Masters, 1982 ; Wu, Adams et Wilson, 1988). Ces indices statistiques permettent d'identifier les données qui ne répondent pas aux conditions de base du modèle, d'examiner les caractéristiques de ces données et de vérifier de quelle façon elles contreviennent à la mesure. Ils permettent donc d'identifier, par exemple, les personnes qui produisent des classements problématiques ou les tâches pour lesquelles les classements sont trop divergents par rapport à la majorité des personnes impliquées.

Smith (2004) avance qu'il est possible de développer avec ces statistiques standardisées une valeur de référence possédant un taux d'erreur de type 1 similaire pour chacune des queues de la distribution, de même que pour plusieurs conditions différentes. Les indices d'ajustement standardisés posséderaient un taux d'erreur de type 1 stable avec différentes tailles d'échantillon. Par conséquent, les indices standardisés constituent un meilleur choix pour vérifier la qualité de l'ajustement entre les données et le modèle. C'est pourquoi ils ont été retenus pour l'étude de l'adéquation des données au modèle.

2.3. Le raisonnement clinique et le test de concordance de script

Le raisonnement clinique est le processus de réflexion et de résolution de problème en action lorsqu'un clinicien rencontre un patient. Il s'agit d'un processus complexe qui est également tributaire de l'attitude des personnes, de leurs connaissances et de leurs habiletés. Plusieurs outils ont été proposés pour recueillir des données afin de réaliser l'évaluation de cette compétence. Les items à réponse choisie, les simulations, les raisonnements d'experts ont tour à tour été mis à contribution à cette fin, mais ils se sont révélés plutôt insatisfaisants (Newble, Hoare et Baxter, 1982 ; Swanson, Norcini et Grosso, 1987 ; van der Vleuten et Newble, 1996). Des chercheurs ont donc proposé une approche différente pour évaluer le raisonnement clinique, approche qui repose sur le concept de *script* (Charlin, Roy, Brailovsky, Goulet et van der Vleuten, 2000). Ce concept, relié à celui de *schéma*, renvoie à des séquences d'événements qui surviennent fréquemment dans un ordre particulier (Fayol et Monteil, 1988). Il repose sur l'hypothèse que lorsque l'être humain est confronté à une situation nouvelle, certaines caractéristiques de cette situation activent un réseau de connaissances et d'expériences antérieures pour classer et interpréter l'information disponible. Par exemple, durant leur formation, les étudiants en médecine développent, par souci d'efficacité et pour gagner du temps, ce qu'on appelle des *scripts de maladie*, qui contiennent des symptômes et des signes ainsi que les liens potentiels entre les deux. Ces scripts sont activés pour établir des diagnostics et, de toute évidence, ils devraient s'affiner avec les années et l'expérience.

Le test de concordance de script a ainsi été développé pour obtenir des données sur la compétence du raisonnement clinique suivant l'idée du script de maladie. Il contient des items qui se présentent sous la forme de vignettes et qui ont quatre parties distinctes. D'abord, les candidats prennent connaissance du problème d'un patient fictif ayant des caractéristiques particulières et se font proposer une décision qui peut être diagnostique, thérapeutique ou investigatrice, selon ce qui est visé par l'item. Ensuite, une information supplémentaire est fournie et les candidats doivent décider si cette information les amène à changer la décision proposée. Les items sont donc semblables à des items à réponse choisie, mais il n'y a pas de bonne ou de mauvaise réponse univoque, car la situation est incomplète et contient des éléments vaguement définis. Plusieurs réponses sont possibles : certaines sont plus appropriées, certaines, un peu moins appropriées et d'autres, carrément inappropriées. Les candidats doivent répondre d'après une échelle comportant des catégories ordonnées ayant deux pôles. Il s'agit d'une échelle qui s'apparente donc à ce que les chercheurs ont l'habitude d'appeler une échelle de réponse de type *Likert* et qui situe en

quelque sorte l'effet sur le candidat de la nouvelle information au sujet de l'hypothèse suggérée (une plus ou moins grande confiance dans cette hypothèse après une réévaluation de la situation).

3. MÉTHODOLOGIE

3.1. Les sujets

La technique du test de concordance de script requiert dans un premier temps la sélection d'un groupe d'experts qui permettra de produire une distribution de réponses à chacun des items en vue d'établir une procédure de notation. Dans le contexte du test de concordance de script, un expert est défini comme étant un praticien expérimenté dont la présence dans un jury est légitime, considérant le niveau de formation des candidats qui auront à passer le test. Le critère d'acceptation dans l'étude était d'avoir terminé la totalité du programme de formation en radio-oncologie. Les domaines d'expertise choisis pour cette étude étaient ceux dont la prévalence est la plus forte chez les radio-oncologues. Le but était de s'assurer qu'un nombre suffisant de radio-oncologues (≥ 10 à 15) soient considérés comme spécialistes dans chacun des domaines étudiés (> 10 ou plus de 50 cas traités par année). Dans cette étude, 62 spécialistes en radio-oncologie ont été invités à participer à l'étude et 47 d'entre eux ont accepté de passer un test de concordance de script de 90 items.

3.2. L'instrument

Le test de concordance de script utilisé comprenait 90 items pour le diagnostic oncologique de problèmes reliés au poumon, au sein et au système urologique (trois sections de 30 items). Dans la présente étude, seules les données provenant des items sur les problèmes reliés au poumon sont analysées. La figure 7.1 ci-dessous présente l'exemple d'une vignette avec trois items qui a été utilisée dans la recherche de Lambert en 2005. On y retrouve les quatre parties d'un item d'un test de concordance de script : 1) une présentation de la situation : *Homme de 56 ans avec néoplasie pulmonaire…* ; 2) une hypothèse : *Un traitement de chimio et…* ; 3) une nouvelle information suite à un examen : *Un ganglion sus-claviculaire droit de 2,5 cm* ; 4) une échelle pour coter l'hypothèse en tenant compte de la nouvelle information : *contre-indiqué, moins indiqué, ça ne change rien, un peu plus indiqué, beaucoup plus indiqué*. D'autres étiquettes pour l'échelle de réponses sont possibles

selon que la tâche demandée est du domaine de la prédiction ou du conseil, mais le format général d'un item d'un test de concordance de script est sensiblement le même.

1. Vignette clinique

Homme de 56 ans avec néoplasie pulmonaire non à petites cellules T3 N2 au lobe inférieur du poumon droit

	Si vous pensiez à…	**Et que le patient rapporte, que vous trouvez à l'examen clinique ou au bilan d'extension**	**Ce traitement devient…**
A.	Un traitement de chimio et de radiothérapie à dose radicale	Un ganglion sus-claviculaire droit de 2,5 cm	–2 –1 0 +1 +2
B.	Un traitement de chimio et de radiothérapie à dose radicale	Un épanchement pleural malin à droite	–2 –1 0 +1 +2
C.	Un traitement de chimio et de radiothérapie à dose radicale	Un syndrome de la veine cave supérieure	–2 –1 0 +1 +2

Échelle : –2 contre-indiqué, –1 moins indiqué, 0 ça ne change rien, +1 un peu plus indiqué, +2 beaucoup plus indiqué

Figure 7.1
Un exemple d'item d'un test de concordance de script (Lambert, 2005)

Les spécialistes en radio-oncologie ont également précisé leur nombre d'années d'expérience ainsi que leur milieu de pratique (universitaire ou non). Ils ont également répondu aux trois questions suivantes : Combien avez-vous traité de patient(e)s atteints du cancer 1) du sein, 2) de la prostate ou 3) du poumon, au cours de la dernière année ? L'échelle de réponses correspondait dans ces trois cas à : Aucun(e), moins de 10, entre 10 et 50, plus de 50.

3.3. Déroulement

La tâche demandée aux 47 spécialistes en radio-oncologie était de classer les items dans l'une au l'autre des catégories de réponses qui accompagnent les items. La procédure a été réalisée en présence d'un assistant de recherche et selon ses consignes. Le test a été passé sous forme papier/crayon et la passation des 90 items a pris de 60 à 75 minutes. Le classement réalisé par les radio-oncologues a ensuite

été utilisé pour construire les scores de résidents dans la spécialité et d'étudiants en médecine qui avaient accepté de passer le test. L'objectif final étant la production d'un test qui discrimine les candidats selon le niveau de formation. Cette dernière étape ne fait pas l'objet de la présente étude ; le lecteur est invité à consulter l'étude de Lambert (2005) pour prendre connaissances des résultats de cette partie de la recherche.

3.4. Considérations éthiques

L'étude présentée constitue une nouvelle analyse de données avec un modèle de mesure différent du modèle mis en œuvre dans la recherche originale de Lambert (2005). Nonobstant ce fait, il est important de mentionner que l'étude originale avait obtenu l'aval du comité d'éthique de la Faculté de médecine de l'Université de Montréal et que tous les participants avaient donné leur consentement en remplissant un formulaire à cet effet. Les réponses ont été traitées de façon anonyme et elles sont demeurées strictement confidentielles. La participation à l'étude était volontaire, et aucune rémunération n'a été consentie aux participants.

3.5. Méthode d'analyse des résultats

Le modèle de Rasch appliqué comporte deux facettes principales, qui sont respectivement les experts et les items, et deux facettes secondaires, qui sont le nombre d'années d'expérience et le nombre de cas traités durant les deux dernières années. Les items du test ne proposant pas tous les mêmes échelles de réponses, la structure des échelles a été considérée comme pouvant varier d'un item à l'autre. C'est une structure de crédit partiel qui a été imposée pour ces échelles (Mead, 2008, p. 19-20 ; Wright et Masters, 1982). Le modèle pour les deux facettes principales avec une structure de crédit partiel et n experts, i items et k catégories pour le classement est :

$$\mathrm{Log}\left[P_{nijk} \Big/ \left(1 - P_{nijk}\right) \right] = B_n - D_i - F_k \qquad (2)$$

Le modèle qui inclut les deux facettes secondaires et k catégories pour le classement correspond à :

$$\mathrm{Log}\left[P_{nijk} \Big/ \left(1 - P_{nijk}\right) \right] = B_n - D_i - C_j - E_l - F_{ik} \qquad (3)$$

Dans l'étude, il y a $n = 47$ experts, $i = 30$ items, $k = 5$ catégories pour le classement et $j = l = 4$ catégories pour chacune des facettes secondaires.

Plusieurs analyses ont été réalisées avec les deux modèles (deux ou quatre facettes) en retirant de façon itérative, en partie ou en totalité, les données des experts dont les classements ne s'ajustaient pas adéquatement au modèle tel que signalé par les valeurs des indices d'ajustement standardisés *infit* et *outfit*. Deux logiciels ont été utilisés pour effectuer les analyses : *Winsteps* et *Facets*. Une contrainte de contexte a été imposée pour ces analyses. Il s'agit du fait que le nombre d'experts retirés devait être minimal tout en produisant un classement relativement homogène. La raison de ce choix était que l'exercice demandé aux experts est plutôt exigeant en termes de temps à y consacrer et que ceux-ci se prêtant volontairement à l'exercice, il n'est pas toujours facile de justifier leur élimination *a posteriori* en raison d'un désaccord avec leurs collègues. La participation volontaire à des expériences semblables à l'avenir pourrait être compromise si les réponses d'un nombre trop élevé de personnes se voyaient exclues des analyses. Il s'agit donc d'une décision pragmatique, mais qui met aussi en lumière une autre difficulté inhérente à ce type d'opération, soit celle de déterminer le nombre minimal d'experts nécessaire pour obtenir un degré d'homogénéité raisonnable dans des domaines d'investigation où le nombre d'experts disponibles est moins élevé. Dans l'étude réalisée, la participation de 47 experts permet de retirer un certain nombre d'entre eux sans trop nuire à la validité de la démarche, mais dans le cas où le nombre d'experts est plus réduit, moins de 15 selon l'étude de Gagnon, Charlin, Coletti, Sauvé et van der Vleuten (2005), cette opération devient plus délicate.

4. RÉSULTATS

Évidemment, il serait trop fastidieux de présenter en détail dans ce texte tous les tableaux et figures rendant compte des diverses analyses réalisées. Il faut faire des choix de sorte que le lecteur puisse saisir l'utilité de la démarche en fonction des objectifs visés. La première analyse est celle qui utilise les données complètes et, par souci de parcimonie, les deux facettes principales. Le tableau 7.1 (p. 150) présente certains des résultats de cette analyse obtenus à partir du logiciel *Facets* (Linacre, 1989). La première colonne contient le code d'identification de l'expert (un numéro par exemple), la deuxième, sa position par rapport aux autres sur une échelle en logit (pour *logistic unit*), et la troisième indique l'erreur type standard de l'estimation de la deuxième colonne. Les deux colonnes suivantes contiennent respectivement la

valeur des statistiques d'ajustement *infit* et *outfit* standardisées. Dans cette présentation, les données sont ordonnées en fonction de la valeur de ces statistiques qui peuvent être positives ou négatives. Une valeur trop élevée dans l'une ou l'autre des directions pour une des deux statistiques indique un problème d'ajustement des données au modèle de mesure. Pour les analyses, nous avons fixé l'intervalle à [–2,5 ; +2,5] pour envisager la présence d'un problème d'ajustement. Évidemment, l'intervalle pourrait être différent, plus grand ou plus petit selon le cas, mais notre expérience avec la modélisation de Rasch dans le cas où le nombre de sujets et d'items est relativement petit (Blais, Grondin, Loye et Raîche, 2010 ; Blais et Grondin, soumis), de même que les suggestions de Lawton, Bhakta, Chamberlain et Tennant (2004), confirment que ce choix est raisonnable. D'autres tableaux pourraient être présentés pour décrire d'une manière semblable les caractéristiques des items, mais comme ces tableaux sont aussi nombreux que le nombre d'analyses réalisées et que les items ne sont pas l'objet direct de l'analyse, ils ne seront pas présentés dans ce chapitre.

Les présentations plus visuelles des figures 7.2 et 7.3 (p. 151-152) permettent tout de même de mieux percevoir la position des experts et celle des items sur l'échelle de logit [–2 ; +2], et de se faire rapidement une idée des positions relatives des valeurs des deux statistiques d'ajustement *infit* et *outfit*. À l'examen de la figure 7.1, on constate ainsi qu'il y a une concentration des experts à l'intérieur de l'intervalle [–1 ; +1] logit et que les experts 30 et 31 se situent à l'écart de la majorité. Cela n'indique pas *a priori* que ces experts sont plus ou moins en accord avec leurs collègues, même si c'est un bon indice à cet effet. Il sera toutefois possible de le déterminer avec plus de précision en étudiant les valeurs des statistiques d'ajustement. De plus, on observe que les items se situent aussi presque tous, sauf les items P2, P12 et P20, à l'intérieur de l'intervalle [–1 ; +1] logit. La figure 7.3 permet de mieux percevoir l'ensemble des valeurs des statistiques d'ajustement pour les experts et d'identifier ceux pour qui au moins une de ces valeurs se situe à l'extérieur de l'intervalle choisi de [–2,5 ; +2,5] pour l'une ou l'autre des statistiques d'ajustement. Ainsi, ce sont les classements des experts 27, 30 et 46 qui ressortent comme s'éloignant le plus du classement de leurs collègues lorsque le modèle avec les deux facettes principales est appliqué.

Tableau 7.1
Résultats de l'analyse avec les deux facettes principales et des données complètes

Expert	Mesure	Erreur type	*Infit*	*Outfit*
E46	0,61	0,22	2,60	4.60
E27	0,94	0,22	3,50	2,80
E30	1,24	0,23	3,10	2,40
E22	–0,34	0,25	2,20	1,80
E36	0,46	0,22	2,10	1,90
E20	0,22	0,23	1,70	0,70
E43	–0,16	0,24	1,50	0,20
E45	0,27	0,22	1,60	1,10
E07	–0,53	0,25	0,70	1,20
E31	1,60	0,25	1,70	1,50
E33	0,06	0,23	1,30	0,40
E21	–0,16	0,24	1,10	0,80
E15	–0,34	0,25	0,80	–0,10
E35	–0,05	0,24	0,70	0,10
E16	0,16	0,23	0,40	0,30
E10	–0,22	0,24	0,20	0,10
E08	–0,28	0,25	0,00	0,20
E02	0,27	0,22	0,20	0,10
E39	–0,53	0,25	0,00	0,10
E18	–0,05	0,24	0,10	–0,10
E40	0,06	0,23	–0,20	–0,20
E19	0,00	0,23	0,00	–0,30
E11	0,32	0,22	–0,40	–0,50
E29	0,27	0,22	–0,50	–0,50
E05	–0,92	0,26	–0,70	–0,30
E23	–0,66	0,26	–0,30	–0,50
E09	–0,11	0,24	0,10	–0,70
E32	–0,16	0,24	–0,80	–0,40
E04	–0,22	0,24	–0,80	–0,60
E12	0,16	0,23	–1,00	–0,90
E47	–0,47	0,25	–0,90	–0,60
E37	–0,22	0.24	–0.90	–0,90
E17	0,06	0,23	–1,20	–0,80
E06	–0,28	0,25	–1,10	–0,90
E44	–0,22	0,24	–1,20	–1,00
E01	–0,47	0,25	–1.20	–1,20
E28	–0,34	0,25	–0.60	–1,30
E24	–0,11	0,24	–1,60	–0,80
E25	–0,05	0,24	–1,40	–1,40
E34	–0,05	0,24	–1,70	–1,10
E03	–0,40	0,25	–1,50	–1,40
E42	0,16	0,23	–1,90	–1,20
E26	–0,28	0,25	–1,50	–1,50
E38	–0,05	0,24	–1,80	–1,10
E14	–0,28	0,25	–1,60	–1,50
E13	–0,11	0,24	–1,40	–1,60
E41	–0,40	0,25	–1,50	–1,50

Mes.	Experts	Items	Mes.
2			2
		P12	
	E31		
		P20	
1	E27	P17	1
		P11	
	E46	P13 P15 P24 P4	
	E11 E36	P19	
	E02 E12 E16 E20 E29 E42 E45	P23 P30	
0	E17 E18 E19 E25 E33 E34 E35 E38 E40	P5 P6 P7	0
	E04 E06 E08 E09 E10 E13 E14 E21 E24 E26 E32 E37 E43 E44	P10 P28 P29 P3 P9	
	E01 E03 E15 E22 E28 E41 E47	P1 P14 P16 P22 P25 P27	
	E07 E23 E39	P21 P8	
		P18 P26	
–1	E05		–1
–2		P2	–2

Figure 7.2
Position relative des experts et des items pour l'analyse avec les deux facettes principales et des données complètes

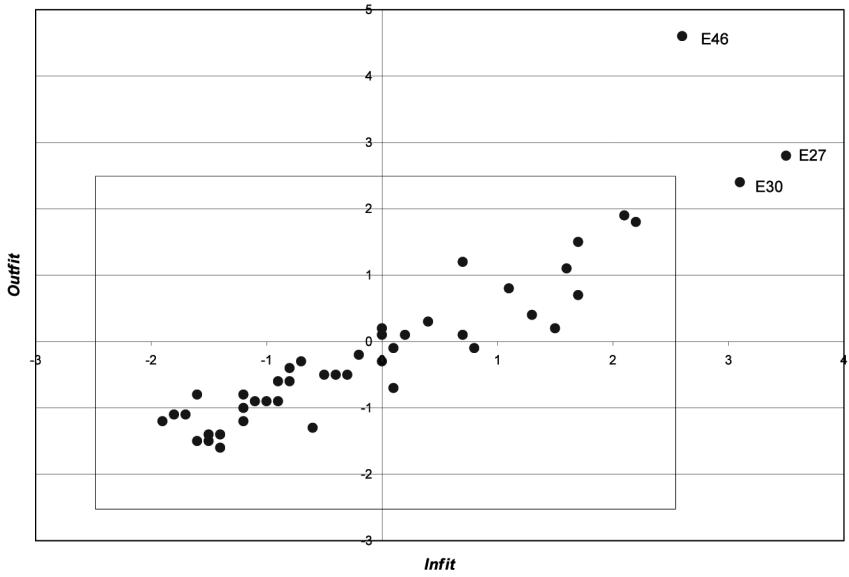


Figure 7.3
Valeurs des statistiques *infit* et *outfit* pour l'analyse avec les deux facettes principales et les données complètes

La seconde étape de la démarche pourrait donc être de retirer les données associées à ces experts et de reprendre l'analyse des données amputées de ces trois vecteurs. Cependant, une autre voie est également possible. En effet, comme le logiciel utilisé fournit un tableau qui identifie les classements qui produisent les résidus les plus élevés, il est possible de retirer uniquement certaines valeurs de certains experts plutôt que l'ensemble des réponses d'un ou plusieurs experts. Un exemple de ce que fournit le logiciel pour l'application du modèle avec les deux facettes principales et les données complètes est présenté au tableau 7.2 (p. 154). De plus, la figure 7.4 (p. 155) illustre les valeurs des statistiques d'ajustement obtenues lorsque ces données incomplètes sont modélisées. On observe que deux points se retrouvent à l'extérieur du double intervalle [–2,5 ; +2,5]. Évidemment, en examinant uniquement des tableaux de statistiques d'ajustement avant et après le retrait de vecteurs complets ou de données isolées, il est difficile de décider quelle est la stratégie la plus appropriée, d'autant plus qu'un raffinement continuel est toujours possible. En effet, le fait de retirer les données correspondant à un expert, partiellement ou complètement, finit le plus souvent par singulariser d'autres données en vertu des valeurs des statistiques d'ajustement observées. Il faut donc utiliser une autre information pour rendre la comparaison efficace et faciliter

la décision quant à l'ensemble de données à conserver en vue de la mise au point du système de notation des résidents et des étudiants, qui est l'objectif ultime de toute l'opération de cette première étape de calibrage d'un test de concordance de script. Heureusement, deux statistiques sont disponibles pour faire cette comparaison ; l'indice de séparation *G* et l'indice de fidélité de la séparation *R* (Bond et Fox, 2001, p. 206-207). Ces deux indices sont complémentaires et apportent la même information, soit le degré d'homogénéité ou d'hétérogénéité de l'échelle de mesure créée avec le modèle. Ils permettent donc de déterminer la capacité de la modélisation à différencier les personnes ou les items. L'indice de séparation *G* peut avoir des valeurs entre 0 et, en théorie, l'infini. Quant à lui, l'indice de fidélité de la séparation *R* peut avoir des valeurs entre 0 et 1. De plus, à partir de l'indice de séparation *G*, il est possible de calculer le nombre de strates de scores qu'il est pertinent de considérer à partir des données modélisées. Le nombre de strates est ainsi égal à ([4*G* + 1]/3) (Wright et Masters, 1982, p. 92).

Dans une situation où c'est l'échelle de mesure des scores à attribuer aux résidents et aux étudiants qui nous intéresse, les valeurs recherchées des indices *G* et *R* devraient être élevées, soit près des maximums possibles, car cela indiquerait que l'échelle de mesure permet de différencier efficacement les individus entre eux. On pourrait ainsi avoir une échelle avec plusieurs strates de scores possibles. Une échelle de mesure qui créerait une seule strate effective de scores serait inutile dans la majorité des applications, car cela indiquerait qu'il est difficile de distinguer les scores entre les individus. Cependant, l'objectif de la première étape du calibrage d'un test de concordance de script est à l'opposé de ce raisonnement, car ce qui est recherché, c'est l'homogénéité des cotes données et non leur hétérogénéité. Ainsi, plus les valeurs de ces deux indices seront près de la valeur 0, plus le classement des experts sera homogène. Cette situation indiquera la présence d'une seule strate et un degré d'accord élevé pour ce classement.

Le tableau 7.3 et la figure 7.5 (p. 156-157) présentent les valeurs de ces deux statistiques pour plusieurs ensembles de données analysés avec le modèle incluant les deux facettes principales et le modèle intégrant les deux facettes secondaires. Les analyses ont été faites avec les données complètes, en enlevant des données isolées et en retirant des vecteurs complets de données pour certains experts (trois, quatre et cinq experts).

Au tableau 7.3, on observe ainsi qu'avec l'ensemble 1 de données complet, les valeurs de l'indice *G* sont respectivement égales à 1,63 et 1,45 pour les modèles avec deux et quatre facettes. Celles de l'indice *R* sont respectivement égales à 0,73 et 0,68. Le nombre de strates

Tableau 7.2
Réponses avec des valeurs résiduelles élevées pour l'analyse avec les deux facettes principales et des données complètes

Résidu	Résidu standardisé	Expert	Item
–2,70	–5,10	E46	P20
1,80	4,10	E36	P02
2,80	3,60	E43	P14
2,80	3,50	E09	P14
2,60	3,20	E20	P29
2,60	3,20	E45	P29
2,60	3,10	E02	P14
–2,70	–3,10	E22	P15
–2,30	–3,10	E46	P15
2,50	3,10	E45	P01
–2,60	–3,00	E22	P13
2,40	2,90	E36	P29
2,40	2,90	E36	P14
–2,20	–2,90	E46	P13
2,40	2,90	E18	P03
2,40	2,90	E20	P22
–2,40	–2,80	E30	P22
2,30	2,80	E33	P03
–2,40	–2,80	E31	P10
–2,30	–2,70	E27	P03
2,30	2,70	E16	P03
–2,30	–2,60	E30	P01
–2,20	–2,60	E45	P24
–2,20	–2,50	E31	P23

possibles en appliquant ces deux modèles serait donc d'environ 2,5 et 2,25, respectivement. Il y a ainsi un gain lorsqu'on utilise quatre facettes plutôt que deux, mais il est minime et, idéalement, nous recherchons la présence d'une seule strate pour les experts. Ces premiers résultats incitent donc à poursuivre les analyses en retirant soit des vecteurs complets de données en fonction des valeurs des statistiques d'ajustement, soit des données isolées en fonction de valeurs résiduelles élevées. Cette dernière stratégie ne semble pas être la meilleure pour les données complètes de départ. En effet, avec deux et quatre facettes, l'indice *G* a comme valeurs 2,06 et 1,83 et l'indice *R,* les valeurs 0,81 et 0,77. Ces valeurs sont plus élevées que celles qui ont été obtenues précédemment ce qui augmente le nombre de strates possibles à 3,08 et 2,77. Avec les ensembles de données 5 et 6, on observe une diminution des valeurs de ces deux indices lorsque les vecteurs de données de trois experts (27, 30 et 46) sont retirés complètement, mais le gain le plus important survient avec les ensembles 7 et 8 lorsqu'on retire les

données pour un quatrième expert (31). On obtient alors des valeurs pour l'indice *G* de 0,75 et 0,68 et de 0,36 et 0,31 pour l'indice *R* avec les modèles à deux et quatre facettes. Le nombre de strates possibles diminue ainsi à 1,33 et 1,24. Finalement, en retirant les données d'un cinquième expert (36), le nombre de strates possibles s'approche considérablement de la valeur idéale recherchée pour le modèle intégrant les quatre facettes, avec une valeur *G* de 0,56 pour 1,08 strate possible. Considérant ces résultats et la contrainte fonctionnelle d'essayer de retirer un minimum de données complètes, il est possible de dire que l'objectif de la première étape a été globalement atteint et qu'un ensemble de données adéquat a été constitué. On peut donc considérer qu'une procédure de notation reposant sur les classements effectués par les 42 experts restants sera suffisamment valide pour être appliquée aux résidents et aux étudiants. La démarche itérative accompagnant la première étape du calibrage pourrait donc se terminer ici et servir de base à une étude s'attachant à la deuxième étape, soit celle de la construction des scores d'un test de concordance de script et de leur capacité à discriminer des candidats se trouvant à différents niveaux de formation.

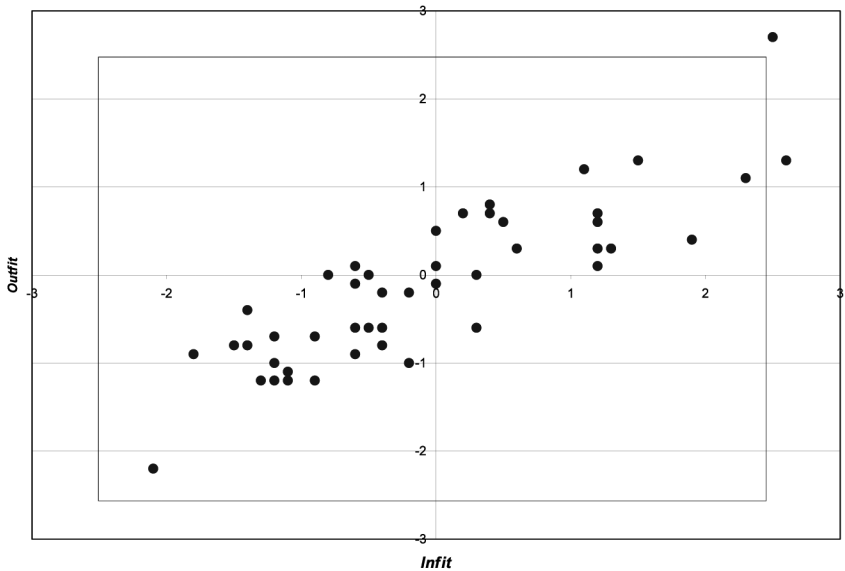


Figure 7.4
Valeurs des statistiques *infit* et *outfit* pour l'analyse avec les deux facettes principales et les données avec des résidus élevés converties en valeurs manquantes

Tableau 7.3
Valeurs de séparation et de fidélité pour les différents ensembles de données analysés et modèles utilisés

Analyses	Ensemble	Séparation	Fidélité
1	2 facettes; données complètes	1,63	0,73
2	4 facettes; données complètes	1,45	0,68
3	2 facettes; certaines réponses enlevées	2,06	0,81
4	4 facettes; certaines réponses enlevées	1,83	0,77
5	2 facettes; réponses de trois experts enlevées	1,40	0,66
6	4 facettes; réponses de trois experts enlevées	1,33	0,64
7	2 facettes; réponses de quatre experts enlevées	0,75	0,36
8	4 facettes; réponses de quatre experts enlevées	0,68	0,31
9	2 facettes; réponses de cinq experts enlevées	0,67	0,31
10	4 facettes; réponses de cinq experts enlevées	0,56	0,24

5. DISCUSSION

Un premier élément de discussion suite aux analyses réalisées porte sur le nombre minimal d'experts nécessaires pour assurer une certaine robustesse au processus et une validité aux scores qui pourraient être attribués aux candidats. Dans le cas qui nous concerne, le nombre élevé d'experts disponibles (47) a permis d'envisager plusieurs scénarios de retrait de données et de produire une modélisation révélant une homogénéité suffisamment élevée avec le retrait des données de cinq experts. Lorsque le nombre d'experts est moins élevé, il est possible qu'un degré d'homogénéité aussi intéressant, par exemple avec une seule strate, soit plus difficile à obtenir, même si Gagnon et collab. (2005) ont montré qu'un nombre de 15 experts pouvait être suffisant à certaines conditions pour obtenir des distributions de scores cohérentes en vue de la notation. Il semble tout de même qu'un nombre un peu plus élevé que 15 devrait être recherché pour donner une certaine latitude quant à la possibilité d'explorer le retrait des données d'experts trop divergents dans les classements réalisés. De plus, comme il existe peu de recherches qui ont été faites sur cet aspect du processus de validation d'un test de concordance de script, il serait souhaitable de poursuivre les investigations dans cette direction afin de mieux déterminer les conditions qui permettent d'en établir la robustesse en vue de la notation des candidats.

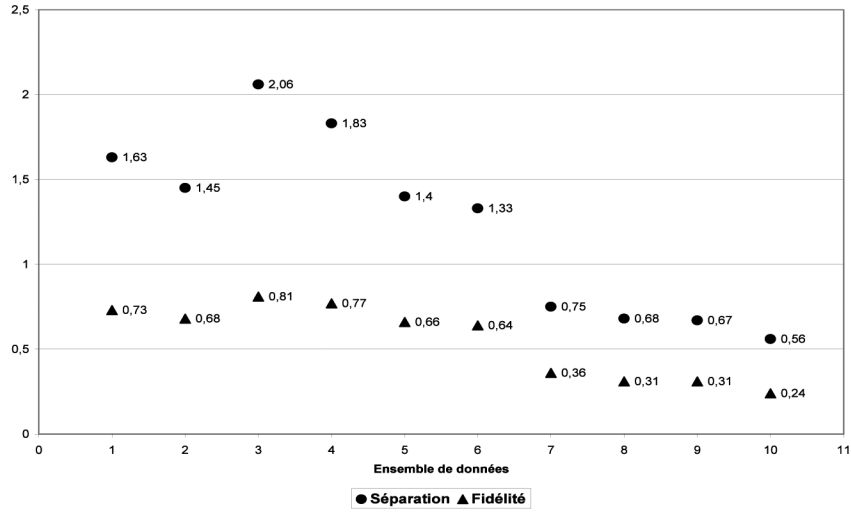


Figure 7.5
Valeurs de séparation et de fidélité pour les différents ensembles de données analysés et les modèles appliqués

Un deuxième élément de discussion, en rapport direct avec le premier, concerne la stratégie de retrait de données jusqu'à l'obtention de valeurs satisfaisantes, tant du point de vue des statistiques d'ajustement *infit* et *outfit* que de celui de l'indice de séparation *G,* et qui peut être envisagée selon un retrait complet ou un retrait partiel des données d'un expert. Nous avons privilégié la première approche dans les analyses finales et obtenu un résultat satisfaisant en terme d'homogénéité des jugements des experts, mais il n'est pas dit que dans un autre contexte le retrait partiel ne serait pas plus avantageux, d'autant plus qu'il évite d'exclure des experts et que le modèle de Rasch se prête bien à des analyses avec des données manquantes. D'autres analyses dans cette direction seraient sûrement appropriées, surtout dans un contexte où les experts sont peu nombreux.

Un troisième élément de discussion porte sur l'impact de la présence de facettes secondaires sur la modélisation. Il pourrait être tentant à première vue d'intégrer un plus grand nombre de facettes secondaires dans les analyses, mais, comme l'illustrent les résultats du tableau 7.3, le gain réalisé n'est pas toujours substantiel. Dans les analyses, les catégories des facettes secondaires se distinguaient très peu les uns des autres sur l'échelle commune en logits (ces résultats n'ont pas été présentés) et ne s'ajustaient pas toujours très bien au modèle. En effet, la présence de quatre catégories pour chacune des

facettes secondaires suppose aussi la présence d'un nombre suffisamment élevé d'experts dans chacune des catégories pour que les résultats possèdent une quelconque robustesse ou stabilité, d'autant plus que les analyses n'ont même pas tenu compte des interactions possibles entre les deux facettes secondaires, soit le nombre d'années d'expérience et le nombre de cas des deux dernières années. Le lien entre le nombre d'experts et le nombre de facettes secondaires considérées est important et pourrait avoir un impact non négligeable sur la stabilité des résultats. Il serait probablement plus sage de limiter le nombre de facettes au minimum et de les utiliser seulement lorsque leur contribution a été démontrée par ailleurs.

Finalement, un dernier élément de discussion concerne une situation qui ne se trouvait pas dans les analyses. Il est possible que certains problèmes d'ajustement soient attribuables à des items mal formulés, peu clairs ou ambigus. De la même manière qu'il est possible d'identifier des experts pour lesquels les classements ne s'accordent pas au modèle de mesure, les indices d'ajustement *infit* et *outfit* permettent d'identifier des items qui apparaissent problématiques par rapport à l'adéquation des données au modèle de mesure. Dans notre recherche, peu importe l'ensemble de données analysé, aucun des 30 items n'a produit d'indices dont les valeurs se situaient à l'extérieur de l'intervalle [–2,00 ; +2,00]. Cette situation ne nous a pas menés à considérer des analyses où certains items auraient pu être exclus de façon itérative en concomitance avec le retrait d'experts.

6. CONCLUSION

Le test de concordance de script est un outil relativement récent (Charlin et collab., 2000) qui vise à mesurer le processus de réflexion et de résolution de problèmes en action lorsqu'un clinicien rencontre un patient, c'est-à-dire le raisonnement clinique. Le processus de notation à partir d'un test de concordance de script s'effectue en deux étapes : dans un premier temps, des experts attribuent une cote à chacun des items du test ; puis, dans un deuxième temps, la distribution des cotes attribuées est utilisée pour la notation des candidats. Dès le départ, il faut donc déterminer quelle est la meilleure distribution des cotes qui serviront à noter les candidats ; une étude de concordance des jugements des experts peut contribuer à préciser les caractéristiques de cette distribution. L'approche décrite exige donc de faire des choix et de prendre des décisions au sujet des objectifs et du contexte d'utilisation d'un test de concordance de script. Étant donné le nombre d'experts disponibles et les enjeux plus ou moins critiques qu'implique la notation des étudiants, il ressort que ces décisions ne peuvent probable-

ment pas être standardisées pour tous les tests de concordance de script et exigent à chaque fois une étude détaillée justifiant la pertinence de la distribution retenue pour le processus de notation. De plus, d'autres recherches sont nécessaires pour étudier la question du nombre d'experts et celle de l'influence de la distribution des réponses des experts pour assurer une stabilité au processus de notation. Ces recherches doivent être menées sur deux fronts, soit en utilisant des données existantes et en procédant par rééchantillonnage, soit en procédant par simulation pour mettre à l'épreuve la robustesse des procédures.

RÉFÉRENCES

Birkelo, C. C., Chamberlain, W. E. et Phelps, P. S. (1947). Tuberculosis case finding : a comparison of effectiveness of various roentgenographic and photofluoroscopic methods. *Journal of The American Medical Association*, *133*, 359-366.

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*. Beijing, Chine : Foreign Language Teaching and Research Press.

Blais, J.-G. et Grondin, J. (soumis). The influence of labels associated with anchor points of Likert-type response scales in survey questionnaires. *Journal of applied measurement*.

Blais, J.-G., Grondin, J., Loye, N. et Raîche, G. (2010). Rasch model's contribution to the study of items and item response scales formulation in opinion/perception questionnaires. Dans N. Brown, B. Duckor, K. Draneyet, M. Wilson (dir.), *Advances in Rasch measurement : volume two*. Maple Grove, Minnesota : Jam Press.

Bond, T. G. et Fox, C. M. (2001). *Applying the Rasch model : fundamental measurement in the human sciences*. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Boomsma, A., Van Duijn, M. A. J. et Snijders, T. A. B. (2001*). Essays on item response theory*. New York, New York : Springer.

Braun, H. I. (1988). Understanding scoring reliability : experiments in calibrating essay readers. *Journal of educational statistics, 13*(1), 1-18.

Charlin, B., Roy, L., Brailovsky, C., Goulet, F. et van der Vleuten, C. (2000). The script concordance test : a tool to assess the reflective clinician. *Teaching and learning in medicine*, *12*, 189-195.

Charlin, B., Tardif, J. et Boshuizen, H. P. A. (2000). A theory of clinical knowledge organization. *Academic medicine*, *75*, 182-190.

Cochrane, A. L. et Garland, L. H. (1952). Observer error in the interpretation of chest films. *Lancet, 2*(6733), 505-509.

Congdon, J. P. et McQueen, J. (2000). The stability of rater severity in large scale assessment programs. *Journal of educational measurement, 37*(2), 163-178.

Fayol, M. et Monteil, J. M. (1988). The notion of script : from general to developmental and social psychology. *European bulletin of cognitive psychology, 8*, 335-361.

Fletcher, C. M. et Oldham, P. D. (1949). The problem of consistent radiological diagnosis in coalminers' pneumoconiosis. *British journal of industrial medicine, 6*, 168-183.

Gagnon, R., Charlin, B., Coletti, M., Sauvé, E. et van der Vleuten, C. (2005). Assessment in the context of uncertainty : how many members are needed on the panel of reference of a script concordance test ? *Medical education, 39*, 284-291.

Klein, J. et Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing writing, 10*, 134-148.

Lambert, C. (2005). *Le test de concordance de scripts : un outil pour évaluer le raisonnement clinique des résidents en radio-oncologie*. Mémoire de maîtrise inédit. Montréal, Québec : Université de Montréal.

Landis, J. R. et Koch, G. G. (1975). A review of statistical methods in the analysis of data arising from observer reliability studies, parts I and II. *Statistica neerlandica, 3*, 101-123 et 151-161.

Lawton, G., Bhakta, B. B., Chamberlain, M. A. et Tennant, A. (2004). The Behçet's disease activity index. *Rheumatology, 43*(1), 73-78.

Linacre, J. M. (1994). Constructing measurement with a many-facet Rasch model. Dans M. Wilson (dir.), *Objective measurement, theory into practice, volume 2*. Norwood, New Jersey : Ablex.

Linacre, J. M. (1989). *Many-facet, Rasch measurement*. Chicago, Illinois : Mesa Press.

Lipkins, R. H., Jones, J. P. et Halkitis, P. N. (1996). *On the trial of the deviant rater : influence of item effects and method on consistency*. Paper presented at the annual meeting of the American Educational Research Association, New York, New York.

Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of educational and behavioral statistics, 19*, 171-200.

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

Mead, R. J. (2008). *A Rasch primer : the measurement theory of Georg Rasch*. Psychometrics services memorandum 2008-001. Maple Grove, Minnesota : Data Recognition Corporation.

Murphy, K. R. et Davidshofer, C. O. (1988). *Psychological testing : principles and applications*. Englewood Cliffs, New Jersey : Prentice Hall.

Newble, D. I., Hoare, J. et Baxter, A. (1982). Patient management problems : issues of validity. *Medical education, 16*, 137-142.

Raymond, M. R. et Houston, W. M. (1990). *Detecting and correcting for rater effects in performance assessment*. Iowa City, Iowa : American College Testing Program.

Smith, R. M. (2004). Fit analysis in latent trait measurement models. Dans E. V. Smith Jr. et R. M. Smith (dir.), *Introduction to Rasch measurement : theory, models and applications*. Maple Grove, Minnesota : JAM Press.

Sudweeks, R. R., Reeve, R. et Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing writing*, *9*, 239-261.

Swanson, D. B., Norcini, J. J. et Grosso, L. J. (1987). Assessment of clinical competences : written and computer-based simulations. *Assessment and evaluation in higher education*, *12*, 220-246.

Van der Vleuten, C. P. M. et Newble, D. (1996). Methods of assessment in certification. Dans D. Newble, B. Jolly et R. Wakeford (dir.), *The certification and recertification of doctors*. Cambridge, Royaume-Uni : Cambridge University Press.

Wolfe, E. W. et Kao, C. W. (1996). *Expert/novice differences in the focus and procedures used by essay scorers*. Paper presented at the annual meeting of the American Educational Research Association, New York, New York.

Wright, B. D. et Masters, G. N. (1982). *Rating scale analysis : Rasch measurement*. Chicago, Illinois : Mesa Press.

Wu, M. L., Adams, R. J. et Wilson, M. R. (1988). *ACER ConQquest, generalised item response modeling software*. Camberwell, Australie : Australian Council for Education Research.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis with special reference to X-ray techniques. *Publication health report, 62,* 1432-1449.

Yerushalmy, J., Garland, L. H., Harkness, J. T., Hinshaw, H. C., Miller, E. R., Shipman, S. J. et Zwerling, H. B. (1951). Evaluation of the role of serial chest roentgenograms in estimating the progress of disease in patients with pulmonary tuberculosis. *American review of tuberculosis, 61*, 225-248.

LISTE DES CONTRIBUTEURS

Jean-Guy BLAIS
Université de Montréal, Québec, Canada
jean-guy.blais@umontreal.ca

Serge BOULÉ
Centre canadien de leadership en évaluation
sboule@lecle.com

Carole BURNEY-VINCENT
École Polytechnique de Montréal, Québec, Canada
carole.burney-vincent@polymtl.ca

France CARON
Université de Montréal, Québec, Canada
france.caron@umontreal.ca

Bernard CHARLIN
Université de Montréal, Québec, Canada
bernard.charlin@umontreal.ca

Micheline-Joanne DURAND
Université de Montréal, Québec, Canada
mj.durand@umontreal.ca

Michel GAGNON
École Polytechnique de Montréal, Québec, Canada
michel.gagnon@polymtl.ca

Robert GAGNON
Université de Montréal, Québec, Canada
robert.gagnon@umontreal.ca

Julie GRONDIN
Université du Québec à Rimouski, Québec, Canada
julie_grondin@uqar.qc.ca

Claire ISABELLE
Université d'Ottawa, Ontario, Canada
claire.isaBelle@uottawa.ca

Carole LAMBERT
Université de Montréal, Québec, Canada
carole.lambert.1@umontreal.ca

Dany LAVEAULT
Université d'Ottawa, Ontario, Canada
dlaveaul@uottawa.ca

Diane LEDUC
Université du Québec à Montréal, Québec, Canada
leduc.diane@uqam.ca

Nathalie LOYE
Université de Montréal, Québec, Canada
nathalie.loye@umontreal.ca

David MAGIS
Université de Liège, Belgique
david.magis@ulg.ac.be

Pascal NDINGA
Université du Québec à Montréal, Québec, Canada
ndinga.pascal@uqam.ca

Karine PAQUETTE-CÔTÉ
Université du Québec à Montréal, Québec, Canada
paquette-cote.karine@courrier.uqam.ca

Jenny PINEAULT
Université de Montréal, Québec, Canada
jenny.pineault@umontreal.ca

Gilles RAÎCHE
Université du Québec à Montréal, Québec, Canada
raiche.gilles@uqam.ca

Martin RIOPEL
Université du Québec à Montréal, Québec, Canada
riopel.martin@uqam.ca

Michèle TESSIER-BAILLARGEON
Université de Montréal, Québec, Canada
michele.tessier-baillargeon@umontreal.ca

Isabelle TRÉPANIER
Commission scolaire de Montréal, Québec, Canada
trepanieri@csdm.qc.ca

RÉSUMÉS EN ANGLAIS

PARTIE 1
L'ÉVALUATION DES APPRENTISSAGES ET LES PRATIQUES PÉDAGOGIQUES

CHAPITRE 1
Validité du diagnostic issu d'un mariage entre didactique et mesure sur un test existant
Nathalie Loye, France Caron, Jenny Pineault, Michèle Tessier-Baillargeon, Carole Burney-Vincent et Michel Gagnon

Diagnosis of cognitive processes used to answer items in a test is based on the cognitive structure underlying the items, identified by experts and formalized in a Q matrix. The DINA model includes two indicators of the Q matrix quality and, thereby, of the validity of the diagnosis. This study presents results related to the validity of the diagnosis with a Q matrix based on classification of items according to an ontology of mathematics and subject skills based on didactics. The data comes from the mathematics test given in June 2008 to students newly admitted to the École Polytechnique de Montréal.

CHAPITRE 2
Utilisation du degré de certitude et du degré de réalisme dans un contexte d'évaluation diagnostique
Serge Boulé et Dany Laveault

In the context of assessment for learning, it could be useful for students to receive feedback on the degree of realism regarding the self-assessment of their answers. If, however, the degree of realism varies according to item properties and individual differences, its usefulness in this context may be limited. Our results show that, independently of gender, the more students perform, the more realistic they are. Students are more realistic with less difficult items. The relationships between taxonomy levels, discrimination coefficient of items and realism are not clear and should be studied further.

CHAPITRE 6

Validité du jugement professionnel des enseignants du primaire dans un contexte d'approche par compétences

Pascal Ndinga

This chapter suggests a model that aims to ensure elementary school teachers' soundness of professional judgement within a competency-based approach environment. This model is the result of analyzing key references within the Ministère de l'Éducation, du Loisir et du Sport of Quebec's official parameters regarding the value attributed to teachers' professional judgement. Kane's argumentative model of interpreting soundness was also used as a reference. The indicators related to elementary school teachers' everyday practices are raised. The approach for developing a model was used. This application ensures the soundness of professional judgment.

CHAPITRE 7

Estimation du degré d'accord entre des experts lors du calibrage d'un test de concordance de script avec le modèle à facettes de Rasch

Jean-Guy Blais, Bernard Charlin, Julie Grondin, Carole Lambert, Nathalie Loye et Robert Gagnon

This chapter illustrates a strategy based on Rasch's facets model to estimate experts' agreement for the task of classifying 30 items from a script concordance test developed to assess clinical reasoning competency in the field of lung radio-oncology.

Collection Mesure et Evaluation

La notion de validité a évolué depuis ses premières définitions vers 1950, de sorte qu'on la considère aujourd'hui non plus comme une caractéristique intrinsèque de la mesure, mais plutôt en relation avec l'utilisation et l'interprétation du score associé à la mesure. Les résultats des évaluations en éducation possèdent rarement une interprétation signifiante en eux-mêmes. Le score prend son sens dans le cadre de référence utilisé pour l'interprétation et dans les inférences d'évaluation.

Cet ouvrage est le produit de la collaboration de chercheurs et d'intervenants en éducation à l'occasion d'un colloque organisé en mai 2009 à Ottawa au Canada lors du 77e congrès annuel de l'Association francophone pour le savoir (ACFAS). Ce colloque visait à porter un regard critique sur les mécanismes pour assurer la validité de l'interprétation de la mesure en éducation afin d'en dégager des tendances, des solutions et de nourrir les pratiques pédagogiques ainsi que le développement de la recherche dans le domaine de la mesure et de l'évaluation en éducation. Ce colloque était organisé selon trois axes. Le premier, théorique, s'attardait notamment au développement des construits, aux avancées théoriques et aux modèles de mesure. Le deuxième, technique et technologique, portait sur l'instrumentation, les méthodes de sélection des critères d'évaluation et des items d'un test, les méthodes d'évaluation de la représentativité des items, les modèles de réponse à l'item et l'intégration des technologies dans la production du jugement. Ce sont ces deux axes qui ont fait l'objet du premier volume. Le troisième axe, pratique, qui s'intéressait, entre autres, aux domaines et aux contextes d'application, aux exemples d'application dans le système d'éducation et à la formation des professionnels de l'évaluation, est traité dans ce second volume.

GILLES RAÎCHE est professeur en mesure et évaluation au Département d'éducation et pédagogie à l'Université du Québec à Montréal. Il est aussi rédacteur en chef de la Revue des sciences de l'éducation.

KARINE PAQUETTE-CÔTÉ est spécialiste en sciences de l'éducation à l'unité d'enseignement et de recherche Éducation de la TÉLUQ et doctorante en éducation à l'Université du Québec à Montréal.

DAVID MAGIS est chargé de recherches, subventionné par le Fonds national de la recherche scientifique (FNRS) et rattaché à l'Université de Liège en Belgique.

ONT COLLABORÉ À CET OUVRAGE

Jean-Guy Blais
Serge Boulé
Carole Burney-Vincent
France Caron
Bernard Charlin
Micheline-Joanne Durand
Michel Gagnon
Robert Gagnon
Julie Grondin
Claire Isabelle
Carole Lambert
Dany Laveault
Diane Leduc
Nathalie Loye
Pascal Ndinga
Karine Paquette-Côté
Jenny Pineault
Gilles Raîche
Martin Riopel
Michèle Tessier-Baillargeon
Isabelle Trépanier

www.puq.ca

9 782760 526877
ISBN 978-2-7605-2687-7