Mesure et Évaluation

Sous la direction de
GILLES RAÎCHE, PASCAL NDINGA et HÉLÈNE MEUNIER

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

VOLUME 3 – Aspects pratiques

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 3 – Aspects pratiques

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

DES MÉCANISMES POUR ASSURER LA VALIDITÉ DE L'INTERPRÉTATION DE LA MESURE EN ÉDUCATION, Volume 2 – L'ÉVALUATION
Sous la direction de Gilles Raîche, Karine Paquette-Côté et David Magis
Avec la collaboration de Diane Leduc et d'Hélène Meunier
ISBN-978-2-7605-2687-7, 178 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 : 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, avenue de Milcamps 119, 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, PASCAL NDINGA et HÉLÈNE MEUNIER

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

VOLUME 3 – Aspects pratiques

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.
Volume 3, aspects pratiques

(Collection Mesure et évaluation)

Comprend des réf. bibliogr.
Comprend du texte en anglais.

ISBN 978-2-7605-3593-0

1. Tests et mesures en éducation – Évaluation. 2. Tests et mesures en éducation – Validité. 3. Tests et mesures en éducation – Interprétation des résultats. I. Raîche, Gilles, 1956- . II. Ndinga, Pascal, 1958- . III. Meunier, Hélène, 1957- . IV. Collection : Collection Mesure et évaluation.

LB3050.5.D473 2012 371.2601'3 C2012-941924-9

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 – 1er trimestre 2013 – 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

Gilles Raîche, Pascal Ndinga et Hélène Meunier

1. INTRODUCTION

La notion de validité a évolué depuis ses premières définitions vers 1950 ; ainsi, 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. Au départ, la validité était plutôt associée aux tests ; aussi a-t-on encore tendance à faussement parler 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 cette époque, soit celle de la validité de critère (*criterion-related validity*).

La première édition par Lindquist (1950) d'un des classiques en mesure et évaluation en éducation, soit *Educational measurement*, reflète bien cette situation, car la majeure partie du chapitre sur la validité (Cureton, 1950) traite de mesures critériées et de puissance de prédiction. Dans la seconde édition de cet ouvrage, Cronbach (1971), ne serait-ce que par le titre de son texte (*test validation*), insiste encore beaucoup sur le fait que la notion de validité est associée au test. Cependant, il souligne 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 à l'avant-scène pendant plusieurs années. Par exemple, la notion de validité de contenu sera très importante

dans le contexte des opérations de définition du domaine en éducation pour planifier les tâches d'évaluation des apprentissages. Oubliée quelque peu avec l'introduction des approches par compétences en éducation, l'importance de la validité de contenu pourrait ressurgir pour soutenir la mise en œuvre des programmes élaborés selon ces approches. Pour sa part, la notion de validité de concept est, et sera encore, primordiale lorsque vient le moment de vérifier ce que mesure réellement un instrument de mesure.

C'est encore Cronbach (1988) qui semble jeter un nouvel éclairage sur la notion de validité en introduisant 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) raffinent considérablement ce processus de validation conçu en tant que jugement évaluatif intégré du degré avec lequel les évidences empiriques et les justifications théoriques soutiennent 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 organisées, fondées sur la structure de l'argumentation, pour soutenir ce jugement évaluatif. Ces avancées autour de la notion de validité permettent de constater que les résultats des évaluations en éducation possèdent rarement une interprétation signifiante en eux-mêmes. Le score obtenu à un test prend son sens dans le cadre de référence utilisé pour l'interprétation et dans les inférences d'évaluation.

Cet ouvrage a été produit à la suite d'un colloque organisé en mai 2010 à Montréal (Québec, Canada) dans le cadre du 78e congrès annuel de l'Association francophone pour le savoir (ACFAS), qui conviait les chercheurs et les intervenants en éducation à faire le point sur les avancées accomplies ainsi que les défis qui se posent dans le domaine de la mesure et de l'évaluation en éducation afin de formuler des propositions et des stratégies susceptibles de répondre aux préoccupations actuelles et de rendre les jugements d'évaluation plus valides. Quatre grands axes composaient ce colloque. Le premier axe était celui de la mesure, qui exposait les avancées et les défis relatifs à l'élaboration des stratégies édumétriques et psychométriques. Le deuxième concernait l'acte d'évaluer qui englobe tous les aspects évaluatifs, soit la planification, l'organisation, l'interprétation et le jugement, pour se terminer par la communication de ce jugement. Le troisième abordait l'incontournable et omniprésente question de la technologie, les progrès qui la caractérisent et les défis qu'elle pose, tant comme outil pédagogique que comme objet d'apprentissage et

d'enseignement. Le quatrième axe traitait de la question générale des pratiques en classe ou sur le lieu de formation, du transfert et des applications pratiques des progrès réalisés à ce jour ainsi que des défis qu'elles soulèvent.

Le présent ouvrage, soit le volume 3, est consacré à l'axe pratique de la mesure et de l'évaluation, réunissant les sujets abordés lors des communications présentées à l'occasion de ce colloque. Les axes de la mesure et de l'évaluation ont été traités respectivement dans le premier et le deuxième volume de cette série.

2. CONTENU DE L'OUVRAGE

Cet ouvrage est divisé en sept chapitres. Le premier chapitre est consacré aux moyens identifiés dans les politiques institutionnelles d'évaluation des apprentissages (PIEA) pour assurer l'argumentation de la validité des inférences d'évaluation en évaluation des apprentissages au collégial. Issu du mémoire de maîtrise de son premier auteur, ce texte fait ressortir l'importance, malheureusement négligée, de la validité apparente de l'interprétation des résultats d'évaluation, soit sa crédibilité.

Le deuxième chapitre permet de comparer des outils technologiques servant à faire l'analyse de données à l'aide des modélisations issues du modèle de Rasch : RUMM2020 et Winsteps. Plus précisément, l'ergonomie de l'interface, les méthodes d'estimation des paramètres, l'adéquation des données aux modélisations et les informations produites par ces logiciels sont analysées.

Dans le troisième chapitre est abordée la mesure de l'aptitude physique générale lors des épreuves de sélection pour les études supérieures en éducation physique et sport au Maroc et en Algérie. La validité conceptuelle de plusieurs épreuves physiques censées mesurer un facteur unique d'aptitude physique générale des étudiants se destinant à une formation supérieure en éducation physique et sport est étudiée. Ce chapitre décrit ainsi une application de l'analyse factorielle exploratoire au domaine de l'éducation physique.

Une recension des écrits de la simulation comme technique d'enseignement et d'évaluation en sciences infirmières est l'objet du quatrième chapitre. Différents modèles de simulation sont présentés avec les avantages et les limites de chacun d'eux. Selon les conclusions obtenues, ces modèles de simulation ajoutent une plus grande valeur à l'enseignement et permettent d'obtenir des données complémentaires sur le degré d'atteinte des compétences.

Le cinquième chapitre présente le contexte de l'évaluation des apprentissages en arts plastiques et en danse contemporaine dans le réseau collégial québécois. De réforme en réforme, les auteurs donnent un aperçu de l'évolution des pratiques et des conceptions de l'évaluation des apprentissages. Tout en tenant compte des politiques et des nouvelles manières de faire, les auteurs soulignent que les professeurs sont de plus en plus appelés à varier leurs modalités d'évaluation des apprentissages pour s'assurer que les compétences à créer, à interpréter et à apprécier sont bien atteintes. Ces professeurs sont aussi appelés à davantage intégrer la fonction formative de l'évaluation des apprentissages dans leur enseignement.

L'état des lieux au regard de l'authenticité des tâches d'évaluation en milieu scolaire forme le sixième chapitre. Les écrits consultés et l'analyse qui en a découlé ont permis à l'auteur de constater l'état embryonnaire de cette question, tant sur le plan de la pratique que sur celui de la recherche. Ces écrits se situeraient encore à un stade dit de caractérisation et la recherche en évaluation devra transcender cette simple caractérisation par l'élaboration et la validation d'outils d'évaluation grâce auxquels il sera possible d'entreprendre des études corrélationnelles et expérimentales.

Enfin, le septième chapitre traite de l'impact de la méthode d'estimation du niveau d'habileté et du choix des premiers items sur l'efficacité de l'administration adaptative du test de classement en anglais, langue seconde, au collégial. Au moment où l'on est à élaborer au Québec une version adaptative informatisée de ce test, il est utile de se préoccuper de ses conditions d'administration. L'une de ces conditions, comme dans tout test adaptatif, est la formalisation de la règle de début du test : un seul item, plusieurs items, comment choisir les items les plus informatifs, etc.

3. CONCLUSION

En ayant comme objectif de porter un regard critique sur les mécanismes pour assurer la validité de l'interprétation de la mesure en éducation, on peut se rendre compte de la diversité des sujets qui se rapportent au concept de validité de la mesure en évaluation et à son application pratique.

Par exemple, au regard de la mesure, ce troisième volume aborde l'évaluation assistée par ordinateur et les tests adaptatifs, l'utilisation des logiciels permettant d'estimer les paramètres d'items et de personnes dans le contexte des modèles de Rasch, de même que l'application de l'analyse factorielle exploratoire à la validation conceptuelle.

Ce volume s'intéresse aussi à l'évaluation tant dans son application formelle que dans un cadre plus théorique. C'est ainsi que sont abordés les politiques institutionnelles d'évaluation des apprentissages, les modèles de simulation pour soutenir l'évaluation des apprentissages, la notion d'authenticité des tâches d'évaluation et, enfin, l'évolution des pratiques d'évaluation des apprentissages en arts plastiques et en danse.

Pour terminer, il est à noter que cet ouvrage utilise une adaptation, pour le Canada francophone, des règles de publication de l'APA (Raîche et Noël-Gaudreault, 2009).

RÉFÉRENCES

Cronbach, L. J. (1971). Test validation. Dans R. L. Thorndike (dir.), *Educational measurement* (2e éd.). Washington, District de Columbia : American council on education.

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

Cureton, E. E. (1950). Validity. Dans E. F. Cureton (dir.), *Educational measurement* (1re éd.). Washington, District de Columbia : American council on education.

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

Lindquist, E. F. (1950). *Educational measurement* (1re éd.). Washington, District de Columbia : American council on education.

Messick, S. (1989). Validity. Dans R. L. Linn (dir.), *Educational measurement* (3e éd.). Washington, District de Columbia : American council on education.

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.

Chapitre 1

Moyens relevés dans les politiques institutionnelles d'évaluation des apprentissages (PIEA) pour assurer la validité des inférences en évaluation des apprentissages au collégial

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 son application à l'analyse des politiques institutionnelles d'évaluation des apprentissages (PIEA) du réseau collégial québécois. L'un de ses objectifs consistait à relever des moyens que les institutions peuvent utiliser pour assurer l'argumentation de la validité des inférences au regard des apprentissages des étudiants. Une analyse de contenu des politiques institutionnelles d'évaluation des apprentissages et une modélisation schématique ont été réalisées à partir de la structure de Kane, menant à l'élaboration de lignes directrices permettant de chercher à assurer l'argumentation de la validité des inférences d'évaluation en évaluation des apprentissages au collégial.

1. INTRODUCTION

Depuis 1994, avec la création de la Commission d'évaluation de l'enseignement collégial (CÉEC), le système d'enseignement collégial du Québec s'est doté de mesures visant à garder un certain contrôle sur la qualité de la formation et de l'évaluation au sein du réseau collégial. Une de ces mesures est l'obligation, pour chaque établissement, de rédiger et de mettre en application une politique

institutionnelle d'évaluation des apprentissages (PIEA). 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, on s'attend à ce qu'elle contienne les éléments permettant la validation des inférences d'évaluation.

Traditionnellement, la validité en évaluation a été définie comme la capacité d'un test à bien mesurer ce qu'il prétend mesurer. En 1951, Cureton définissait la validité comme ayant deux dimensions, la pertinence et la fiabilité (*relevance and reliability*), et comme étant la corrélation entre le score obtenu au test et le score vrai critérié (*true criterion score*) (Cureton, 1951, 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 souligne le besoin de définir la validité en fonction de son utilisation et de son interprétation en éducation. La validité est encore principalement associée à la validité d'un test, mais plus d'importance est accordée à l'interprétation. Cronbach (p. 447) mentionne d'ailleurs que, puisque chaque interprétation a son propre degré de validité, on ne peut jamais arriver à la simple conclusion qu'un test particulier est valide[1]. Il ajoute ensuite que tous les aspects et tous les détails d'une procédure de mesure peuvent influencer la performance et, par conséquent, ce qui est mesuré (p. 449). Il 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 étant le jugement évaluatif intégré du degré auquel la preuve empirique et le rationnel théorique supportent la justesse et la pertinence des inférences et des actions basées sur les résultats d'une évaluation[2]. Cette définition implique une conception de la validité en tant qu'argument, telle que l'a introduite Cronbach (1980) en décrivant 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. Deux principaux modèles ont été développés en parallèle dans les années 1990 et 2000 en accord avec cette conception de la validité : l'un plus particulièrement centré sur la

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).

preuve empirique et l'autre, sur le rationnel théorique. Le premier est une approche de conception de l'évaluation basée sur la preuve ; c'est 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 relativement au 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 dont il est ici question poursuivait l'objectif de valider la structure d'argumentation interprétative de Kane (2006) par l'application de cette structure à l'analyse de politiques institutionnelles d'évaluation des apprentissages du réseau collégial québécois (Paquette-Côté, 2010 ; Paquette-Côté et Raîche, 2009, 2011). Un objectif secondaire était de relever dans le contenu de ces politiques des moyens que peuvent mettre en place les institutions pour chercher à assurer la validité des inférences d'évaluation au regard des apprentissages des étudiants. Le présent article se centre sur les résultats obtenus dans la poursuite de ce second objectif. Au moment de la rédaction du mémoire de recherche à l'origine de cette communication (Paquette-Côté, 2010), les études sur les applications de ces modèles récents ne faisaient que débuter.

2. CADRE THÉORIQUE

La structure d'argumentation interprétative proposée par Kane (2006) est présentée au tableau 1.1 et décrite de façon plus détaillée par la suite.

L'argument interprétatif comprend quatre niveaux d'inférence pour l'interprétation des résultats de l'évaluation : la notation, la généralisation, l'extrapolation et l'implication (I1 à I4 dans le tableau 1.1). 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 de mesure). 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. L'univers de généralisation est un échantillon des observations du domaine à partir duquel sont tirées les observations qui constitueront la mesure (par exemple les items d'un examen ou

Tableau 1.1.
L'argument interprétatif pour l'interprétation d'un trait
(adapté de Kane, 2006, p. 34)

I1 :	Notation (*scoring*) : à partir 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 telles que spécifiées.
	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 : à partir 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 : à partir du score univers au score cible
	A3.1. Le score univers est lié au score cible.
	A3.2. Il n'y a pas d'erreurs systématiques qui sont susceptibles de miner l'extrapolation.
I4 :	Implication : à partir 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 supportent les implications associées avec l'étiquette du trait.

les critères d'évaluation d'un stage). En évaluation des apprentissages, l'univers de généralisation représente donc le programme d'études ou, de façon plus restreinte, le cours. Par l'inférence d'extrapolation (I3), le score univers est extrapolé au score cible, c'est-à-dire que le score est interprété non plus comme représentatif de la performance à un sous-ensemble d'observations, mais comme étant représentatif du domaine cible en entier. Le domaine cible couvre l'ensemble des manifestations du trait, dans les divers contextes dans lesquels ce trait est impliqué, ainsi que les liens qui l'unissent à d'autres traits et à d'autres domaines. L'interprétation d'un trait a son lot d'implications, lesquelles peuvent inclure des 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 comportent des implications

qui dépassent le domaine cible et qui nécessitent d'être validées. Par l'inférence d'implication (I4), les implications associées au trait sont jointes à l'interprétation du score dans le domaine de façon à décrire ses relations avec d'autres traits et d'autres domaines, à faire certaines mises en garde quant aux exceptions ou aux diverses interprétations pouvant être produites au regard du trait en question ou du score. Ces quatre niveaux d'inférence (I1 à I4 dans le tableau 1.1) sont accompagnés des arguments (A1.1 à A4.2 dans le tableau 1.1) qui garantissent la validité de chaque inférence et, par le fait même, la validité de l'interprétation du trait. La structure d'argumentation décrite par Kane (2006) devrait théoriquement permettre de soutenir la validation des inférences d'évaluation.

3. MÉTHODOLOGIE

Dans un effort de mise à l'épreuve empirique de la structure d'argumentation interprétative proposée par Kane (2006), cette recherche menée en 2008-2009 consistait à appliquer cette structure à l'analyse de politiques institutionnelles d'évaluation des apprentissages du réseau collégial québécois (Paquette-Côté, 2010 ; Paquette-Côté et Raîche, 2009, 2011). Deux objectifs étaient alors poursuivis : 1) valider la structure d'argumentation interprétative de Kane (2006) 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 au regard des apprentissages des étudiants. Le présent article est centré sur le deuxième objectif. Les résultats associés au premier objectif ont fait l'objet d'une publication antérieure (Paquette-Côté et Raîche, 2011).

3.1. Corpus d'analyse

Le corpus d'analyse de la présente recherche est constitué de cinq politiques institutionnelles d'évaluation des apprentissages. Le tableau 1.2 présente la description des 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 ne sont pas pris en compte par les analyses, car leur contexte éducationnel diffère de celui des collèges subventionnés en ce qui a trait à leur nombre d'étudiants,

généralement inférieur, et à leur système de gestion et d'organisation, qui est particulier. Sur les 64 collèges privés et publics, anglophones et francophones, offrant la formation ordinaire et figurant sur la liste fournie par le ministère de l'Éducation, du Loisir et du Sport en 2009, 41 présentaient leur politique institutionnelle d'évaluation des apprentissages sur leur site Web au moment de la constitution du corpus d'analyse. Les cinq politiques institutionnelles d'évaluation des apprentissages retenues pour les analyses ont été sélectionnées à partir de celles-ci et en veillant à inclure des collèges privés et publics, anglophones et francophones, de moyennes et grandes populations de la province de Québec.

Tableau 1.2.
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 2500
02	Moins de 50 000	Public	Francophone	Moins de 1500
03	Plus de 1 000 000	Public	Francophone	Environ 2500
04	Plus de 1 000 000	Privé	Francophone	Moins de 1500
05	Plus de 1 000 000	Public	Anglophone	Plus de 5000

[a] Données issues de Statistique Canada (http://www.statcan.gc.ca).
[b] Données issues des *Statistiques détaillées sur l'éducation* (ministère de l'Éducation, du Loisir et du Sport, 2007).

Les résultats de la présente recherche sont limités 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 met à la disposition de la population et des collèges un cadre de référence sur l'évaluation des politiques institutionnelles d'évaluation des apprentissages (Commission d'évaluation de l'enseignement collégial, 1994), le contenu des politiques institutionnelles d'évaluation des apprentissages des établissements d'enseignement collégial québécois est relativement homogène. De ce fait, il est attendu que les interprétations issues des résultats de la présente recherche puissent s'appliquer à des établissements d'enseignement collégial dont la politique institutionnelle d'évaluation des apprentissages n'a pas été analysée.

3.2. Méthodes d'analyse

La méthodologie utilisée dans la réalisation de cette recherche est centrée sur l'analyse de contenu de politiques institutionnelles d'évaluation des apprentissages. Une première étape consistait à analyser le contenu manifeste des politiques institutionnelles d'évaluation des apprentissages au moyen d'une grille d'analyse construite à partir de la structure argumentaire pour la validation de l'interprétation d'un trait décrite par Kane (2006). Une deuxième étape de l'analyse consistait à modéliser les informations recueillies à l'aide du logiciel MotPlus développé par le Centre de recherche LICEF de la Télé-université (Rivard, Banville, Gareau, Léonard, Mihaila, Paquette et Rosca, 2006). La modélisation a permis de classer, par le recours à 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 au regard des apprentissages des étudiants. Au départ, la structure de catégorisation correspondait en tout point à la structure argumentaire de Kane (2006) telle que présentée au tableau 1.1. À l'étape de la modélisation, un modèle mixte (L'Écuyer, 1987, p. 59) a été adopté, permettant d'effectuer des regroupements, des subdivisions ou des ajouts au modèle de Kane (2006). Chaque argument de validité (A1.1 à A4.2) a été développé en sous-modèle consistant en une représentation graphique de l'ensemble des moyens qui lui sont associés. Chaque moyen identifié à l'étape de l'analyse de contenu a été ajouté à la représentation graphique de l'argument auquel il correspond. Les moyens d'application et/ou d'évaluation pour chaque argument de validité ont finalement été classés et présentés en fonction des acteurs responsables de leur application et/ou de leur évaluation.

4. RÉSULTATS DES ANALYSES DES POLITIQUES INSTITUTIONNELLES D'ÉVALUATION DES APPRENTISSAGES

L'un des objectifs poursuivis par cette recherche était d'identifier des moyens que peuvent mettre en place les établissements d'enseignement pour chercher à assurer la validité des inférences d'évaluation au regard des apprentissages des étudiants. Le présent chapitre offre une synthèse des résultats les plus significatifs pour chaque argument du modèle d'argumentation de la validité des inférences d'évaluation des apprentissages. Les moyens sont formulés sur la base du contenu des politiques institutionnelles d'évaluation des apprentissages et de façon à représenter tous les niveaux de la modélisation schématique.

4.1. Inférence de notation (I1)

L'inférence de notation (figure 1.1) est réalisée à travers la passation et la notation des activités d'évaluation de façon à pouvoir faire le passage de l'interprétation de la performance observée au score observé. La performance est alors interprétée en termes de score associé à un échantillon d'observation constituant la mesure.

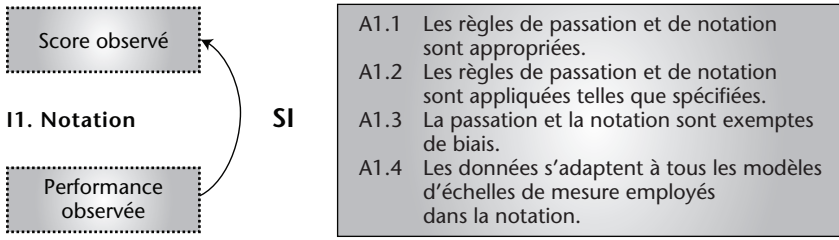


Figure 1.1.
Inférence de notation (I1)

4.1.1. *Les règles de passation et de notation sont appropriées (A1.1)*

Les moyens identifiés dans les politiques institutionnelles d'évaluation des apprentissages relativement à l'argument A1.1 concernent la planification des activités d'évaluation pour chercher à s'assurer que la passation et la notation sont appropriées.

Un premier moyen pour chercher à appliquer cet argument est la tenue d'activités de concertation entre les professeurs d'un département. Ensemble, les professeurs d'un département constituent un groupe d'experts du domaine cible. Leur concertation au sujet des objectifs, des contenus, des modes d'évaluation, des exigences et des seuils de réussite des cours et des programmes qui les concernent est un moyen de s'assurer que les règles de passation et de notation sont appropriées. La concertation des professeurs permet aussi de rechercher l'équivalence des objectifs d'apprentissage, des procédés et des critères d'évaluation dans un même cours donné par plusieurs professeurs. C'est donc un moyen de s'assurer que la correction des activités d'évaluation sommative est réalisée à partir de critères établis par l'ensemble des professeurs qui donnent le même cours et selon une pondération équivalente.

En établissant et en diffusant des règles départementales d'évaluation des apprentissages, le département peut chercher à s'assurer que les règles de passation et de notation sont appropriées, mais aussi qu'elles

tiennent compte des particularités de chaque programme d'études tout en garantissant une certaine uniformité des modalités d'évaluation au sein d'un même département contribuant à la recherche d'équivalence. Ces règles sont approuvées au sein du département et par la direction des études. L'élaboration de plans-cadres de cours auxquels les professeurs peuvent se référer lors de la planification de leurs cours et de leurs activités d'évaluation est un moyen pour le département de répondre à ces mêmes objectifs. Le professeur doit, quant à lui, rédiger un plan de cours conforme à la PIEA, aux règles départementales d'évaluation des apprentissages, au programme et au plan de cours cadre dans lequel il établit les modalités d'évaluation : la nature des épreuves (travaux, examens ou autres) servant à l'établissement de la note finale, leur pondération, leur durée et la date prévue. Le plan de cours et ses modifications durant la session sont approuvés par le département et les modalités d'analyse et d'approbation des plans de cours sont elles-mêmes approuvées par la direction des études. Toutes les étapes d'approbation constituent des consultations de différents groupes experts qui contribuent à la validité de l'argument interprétatif.

Enfin, l'organisation d'activités de perfectionnement, tout en offrant des services d'encadrement aux professeurs, permet de les tenir informés des derniers développements du domaine de la mesure et de l'évaluation et de leur donner les outils et les méthodes nécessaires pour s'assurer que les règles de passation et de notation soient appropriées.

4.1.2. *Les règles de passation et de notation sont appliquées telles que spécifiées (A1.2)*

L'établissement de règles de passation et de notation appropriées contribue à assurer la validité de l'inférence de notation, mais encore faut-il que ces règles soient appliquées telles qu'indiquées.

Concernant cet argument, la concertation entre les professeurs du département au sujet des objectifs, des contenus, des modes d'évaluation, des exigences et des seuils de réussite des cours et des programmes qui les concernent est aussi importante. Si les règles de passation et de notation sont discutées lors des activités de concertation et que les professeurs sont impliqués, s'ils ont participé à leur élaboration, on peut supposer qu'ils auront plus tendance à les appliquer telles que spécifiées.

Les professeurs devraient aussi s'assurer de connaître et d'appliquer la politique institutionnelle d'évaluation des apprentissages et les règles départementales d'évaluation des apprentissages lors de la passation et de la notation des activités d'évaluation dont ils ont

eux-mêmes la responsabilité. L'évaluation est alors fondée sur l'atteinte des objectifs d'apprentissage annoncés dans le plan de cours ou sur l'atteinte des éléments de la compétence, selon les standards de réussite prescrits ou définis par le département.

Les services d'encadrement et de perfectionnement contribuent aussi à s'assurer que les professeurs connaissent et appliquent la politique institutionnelle d'évaluation des apprentissages et les règles départementales d'évaluation des apprentissages.

Enfin, la direction des études peut, en plus de décrire et de diffuser les procédures administratives requises pour la passation et la notation des activités d'évaluation, évaluer les pratiques d'évaluation au sein du collège et rédiger un rapport écrit sur le processus d'évaluation des apprentissages qu'elle remettrait au conseil d'administration du collège à la fin de chaque session ou de chaque année. Ce rapport ferait, entre autres, état du degré d'application des règles de passation et de notation.

4.1.3. *La passation et la notation sont exemptes de biais (A1.3)*

L'inférence de notation peut être affectée par certaines sources de biais liées aux caractéristiques de l'évaluateur ou de l'étudiant ou à des caractéristiques environnementales. Le contenu des politiques institutionnelles d'évaluation des apprentissages renferme un certain nombre de moyens permettant de réduire ces sources de biais lors de la passation ou de la notation.

Les étudiants qui suivent un même cours devraient être évalués à partir des mêmes objectifs, des mêmes critères et selon une même pondération, quel que soit le professeur qui donne le cours. La concertation entre les professeurs d'un même département et entre les départements au sujet des modes et des instruments d'évaluation et à propos du degré d'exigence et des seuils de réussite est un moyen de rechercher l'équivalence des évaluations pour réduire les risques de biais liés à l'évaluateur. Il est également suggéré d'avoir un seul et même plan de cours pour un cours donné à plusieurs groupes d'étudiants, à la même session et dans un même secteur, ce qui permet de s'assurer que le contenu enseigné et les activités d'évaluation sont les mêmes pour tous, donc que l'évaluation est plus équitable. Dans l'une des politiques institutionnelles d'évaluation des apprentissages analysées, il est même suggéré d'avoir un seul et même examen final pour un cours donné à plusieurs groupes d'étudiants, à la même session et dans

un même secteur. En plus de réduire les risques de biais liés à l'évaluateur, cela facilite la reconnaissance des acquis lors de l'évaluation des équivalences.

Un autre moyen de réduire les risques de biais lors de la notation est l'établissement par le professeur de critères et de standards clairs préalablement à la passation et à la notation. En se référant lors de la notation aux critères et aux standards préétablis, le professeur réduit les risques de glissement lors de la notation et assure ainsi une évaluation plus équitable.

Faire porter l'inférence d'évaluation sur un seul score ou une seule activité d'évaluation pourrait être une source de biais liée aux conditions d'évaluation, à l'environnement ou à l'état physique ou psychologique de la personne évaluée au moment de la passation. Un moyen de réduire ces sources de biais peut être de répartir l'évaluation sur plusieurs activités en cours de session. La direction des études peut aussi structurer le calendrier des évaluations pour l'ensemble du collège afin de s'assurer qu'aucun étudiant n'a plus de deux évaluations par jour et que les conditions environnementales d'évaluation sont adéquates (grandeur et ameublement des locaux, lumière, contrôle des bruits environnants…).

Le professeur, notamment lors de l'évaluation de travaux d'équipe, devra veiller à s'assurer que la notation porte sur l'évaluation des apprentissages individuels et qu'elle n'est pas biaisée par la performance des autres étudiants.

L'étudiant a lui-même une responsabilité au regard de la réduction des risques de biais de notation en s'assurant de connaître et d'appliquer la PIEA et les règles départementales d'évaluation des apprentissages. Par exemple, un étudiant qui contrevient aux règles d'évaluation des apprentissages en trichant ou en plagiant lors d'une évaluation introduit un biais réduisant la validité de l'inférence de notation puisque celle-ci ne se rapporte pas à sa propre performance. À ce propos, le professeur peut réduire les risques de tricherie en s'assurant de la confidentialité des instruments d'évaluation avant leur administration aux étudiants. Un autre exemple de responsabilité de l'étudiant concerne le respect des normes de présentation lors de la rédaction de ses travaux écrits, ce qui peut réduire les risques que la notation de ses réalisations soit biaisée par leur présentation matérielle. Quant au professeur, la responsabilité lui revient de fournir aux étudiants les consignes de présentation des travaux.

4.1.4. *Les données s'adaptent à tous les modèles d'échelles de mesure employés dans la notation (A1.4)*

Les politiques institutionnelles d'évaluation des apprentissages n'abordent pratiquement pas la question des échelles de mesure. Une seule politique institutionnelle d'évaluation des apprentissages y fait directement référence dans deux énoncés. C'est à partir de ceux-ci qu'a pu être formulé un moyen qui se rapporte spécifiquement à cet argument : celui d'établir clairement l'échelle de notation employée et la façon dont les cotes sont interprétées numériquement.

Les autres énoncés ne sont qu'indirectement liés à l'argument ; c'est pourquoi les moyens tirés des politiques institutionnelles d'évaluation des apprentissages concernant cet argument sont d'ordre plus général. Ils concernent le développement des instruments de mesure et l'établissement des seuils de réussite sans traiter spécifiquement des préoccupations liées aux données de l'évaluation et aux échelles de mesure.

4.2. Inférence de généralisation (I2)

L'inférence de généralisation (figure 1.2) est réalisée à travers l'échantillon des observations. C'est la constitution de l'échantillon des observations qui assure la validité de l'inférence de généralisation qui, à son tour, permet le passage de l'interprétation du score observé au score univers. Le score est alors interprété non plus seulement en lien avec la mesure, mais en tenant compte également de sa représentativité d'un sous-ensemble plus large d'observations que constituent le cours et le programme d'études.

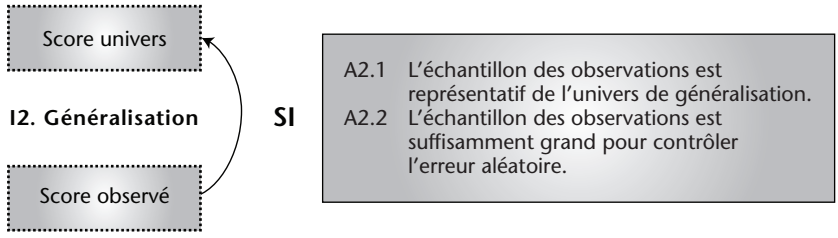


Figure 1.2.
Inférence de généralisation (I2)

4.2.1. *L'échantillon des observations est représentatif de l'univers de généralisation (A2.1)*

C'est à l'argument A2.1 du modèle original de Kane (2006) que le plus grand nombre d'énoncés analysés se réfèrent et c'est dans cet argument que le plus grand nombre de moyens d'application ou d'évaluation ont été identifiés, sans toutefois avoir le plus grand nombre de moyens propres à l'argument. Il s'agit donc d'une catégorie d'ordre plus général. Cet argument renvoie à la représentativité de l'échantillon des observations par rapport à l'univers de généralisation, donc à la représentativité des items ou des observations qui forment le cours et la mesure par rapport au programme d'études.

Le fait de planifier les activités d'apprentissage et d'évaluation en conformité avec les plans-cadres ou avec les objectifs et les standards de chacun des cours est un moyen pour le professeur de chercher à s'assurer que l'échantillon des observations représente bien l'univers de généralisation. La planification de l'évaluation permet de s'assurer que les contenus, les objectifs et les critères d'évaluation sont clairement présentés et établis dès le départ. Cette planification peut être faite parallèlement avec la rédaction du plan de cours qui présente : la compétence visée par le cours et l'apport du cours à la réalisation des objectifs du programme ; les objectifs de formation du cours présentés en énoncés de compétence et en standards ainsi que les objectifs du cours ; les contenus couverts par le cours ; les activités d'enseignement et d'apprentissage ; les activités d'évaluation sommative des apprentissages et leurs modalités, c'est-à-dire la nature des épreuves (travaux, examens ou autres) servant à l'établissement de la note finale, leur pondération, leur durée et la date prévue des évaluations. La conception des activités d'évaluation et le développement des instruments de mesure sont fondés sur les objectifs du cours ou sur les compétences à développer. Raîche (2008, p. 7) suggère d'ailleurs, outre la rédaction d'un plan de cours, la rédaction d'un plan d'évaluation qui présente les objectifs et les modalités d'évaluation plutôt que les activités d'enseignement et les objectifs d'apprentissage.

La répartition de l'évaluation sur plusieurs activités d'évaluation en cours de session permet aussi d'assurer une meilleure représentativité de l'univers de généralisation en augmentant le nombre d'observations et en diminuant les risques de biais liés aux conditions d'évaluation, à l'environnement ou à l'état physique ou psychologique de la personne évaluée au moment de l'évaluation.

Le département peut s'assurer de la concertation des professeurs sur les objectifs, les contenus, les modes d'évaluation, les exigences et les seuils de réussite des cours et des programmes qui les concernent.

Ainsi, les professeurs peuvent se concerter pour s'assurer que les items ou les observations qui forment la mesure sont représentatifs du cours et du programme. En plus de fonder le contenu de l'évaluation sur le jugement d'une série d'experts, cela permet la recherche d'équivalence des objectifs d'apprentissage, des procédés et des critères d'évaluation dans un même cours donné par plusieurs professeurs et, même dans ce cas, cela permet l'élaboration d'instruments d'évaluation communs.

Le département peut élaborer des règles d'évaluation propres au programme qui précisent comment la maîtrise de la compétence est évaluée pour obtenir la note de passage et élaborer un plan-cadre pour chacun des cours dont il est responsable. Le plan-cadre présente les contenus essentiels du cours, ce qui peut guider les professeurs dans la détermination des contenus devant faire l'objet de l'évaluation. De plus, en s'informant sur les besoins de formation des professeurs et en organisant, en collaboration avec le service de formation continue et la direction des études, des services d'encadrement et de perfectionnement, le département peut chercher à s'assurer que les professeurs disposent des méthodes et des outils nécessaires au développement d'instruments de mesure qui permettent de générer des résultats représentatifs de la compétence de l'étudiant dans le cours et le programme d'études.

Le comité de programme, quant à lui, est responsable de la détermination du contenu des cours en lien avec les objectifs et les standards du programme, du domaine et des exigences du collège et du Ministère. Il peut procéder, en collaboration avec le département et les professeurs, à l'évaluation des instruments et des méthodes d'évaluation afin de s'assurer qu'ils sont conformes à la politique institutionnelle d'évaluation des apprentissages, aux règles départementales d'évaluation des apprentissages, au programme et aux plans-cadres de cours.

4.2.2. *L'échantillon des observations est suffisamment grand pour contrôler l'erreur aléatoire (A2.2)*

Dans la planification de l'évaluation, un moyen pour le professeur de s'assurer que l'échantillon des observations est suffisamment grand pour contrôler l'erreur aléatoire est d'attribuer aux activités d'évaluation une pondération suffisamment importante pour qu'elle soit significative de la maîtrise globale des éléments de compétence, de la ou des compétences visées par le cours. Le département peut aussi prescrire cette pondération dans le plan-cadre du cours. Le fait de répartir l'évaluation sur plusieurs activités d'évaluation en cours de session permet aussi de contrôler l'erreur aléatoire en augmentant le nombre

d'observations et en diminuant les risques de biais liés aux conditions d'évaluation, à l'environnement ou à l'état physique ou psychologique de la personne évaluée au moment de l'évaluation.

4.3. Inférence d'extrapolation (I3)

C'est au niveau de l'inférence d'extrapolation (figure 1.3) qu'entrent en considération les questions relatives à la définition du programme d'études. Le score est alors interprété en tenant compte du programme d'études et du domaine cible. Les liens doivent ainsi être clairement établis entre le programme d'études et le domaine, car c'est par l'extrapolation qu'on peut émettre le jugement affirmant que l'étudiant évalué possède ou ne possède pas les compétences du domaine cible.

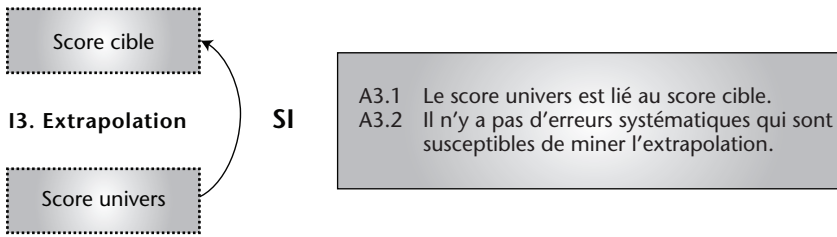


Figure 1.3.
Inférence d'extrapolation (I3)

4.3.1. *Le score univers est lié au score cible (A3.1)*

Pour l'argument A3.1, le résultat est interprété selon sa représentativité par rapport au domaine cible. Les moyens retenus pour appliquer ou évaluer cet argument de validité permettent donc d'établir un lien entre le programme et le domaine.

Le professeur planifie les activités d'apprentissage et d'évaluation en conformité avec la politique institutionnelle d'évaluation des apprentissages, les règles départementales d'évaluation des apprentissages, le programme et le plan-cadre du cours. Un service de soutien pédagogique assure un rôle d'assistance aux professeurs, et particulièrement aux nouveaux professeurs, dans la préparation et l'organisation des cours en accord avec les objectifs, les standards et les politiques du programme et du département. Le professeur indique dans le plan de cours la place du cours dans le programme d'études. Ce faisant, il établit le lien entre le cours, le programme et le domaine. En fournissant une bibliographie et des lectures complémentaires dans le

plan de cours, le professeur établit le lien entre le cours et le domaine cible et incite les étudiants à faire de même. Le professeur montre ainsi que son cours est constitué d'un sous-ensemble d'observations tirées du domaine cible et que d'autres éléments du domaine peuvent être explorés par l'étudiant lui-même, bien que ceux-ci n'aient pas été inclus dans l'univers de généralisation (ensemble des observations abordées dans le programme et plus particulièrement dans le cours et susceptibles de faire partie de l'évaluation). Lors de la planification de l'évaluation, le professeur centre l'évaluation sur les objectifs du programme.

Le département et le comité de programme sont responsables de l'élaboration des objectifs du programme. Le comité de programme élabore les critères spécifiques des compétences professionnelles du domaine et développe les critères et les standards d'évaluation pour le programme. Il assure la concordance entre les éléments de la politique institutionnelle d'évaluation des apprentissages, le développement de programme et les plans de cours. Il cherche aussi à favoriser l'harmonisation des pratiques et l'équivalence des évaluations pour tous les cours du programme. Chaque cours du programme constitue en quelque sorte un univers de généralisation duquel sont tirés les échantillons d'observations qui constituent chaque évaluation. L'harmonisation des pratiques et l'équivalence des évaluations favorisent l'établissement de liens entre les cours, donc entre différents univers de généralisation, et offrent une meilleure représentativité du domaine cible. Le département s'assure que les cours sont intégrés au programme et qu'ils contribuent à l'atteinte des objectifs du programme. Les liens entre un cours et les autres cours du programme sont présentés dans le plan-cadre de façon à bien établir les liens entre les différents univers de généralisation et ainsi mieux représenter le domaine cible.

L'évaluation des équivalences et des substitutions revient à vérifier l'équivalence de deux univers de généralisation au sein d'un même domaine cible ou parfois de domaines différents. Dans ces cas, la direction des études doit évaluer si le score univers est lié au score cible sans toutefois être en mesure d'évaluer la validité des arguments liés à la notation et à la généralisation, qui sont liées aux pratiques d'évaluation réalisées dans le cours servant d'équivalence ou de substitution et qui ne sont donc plus disponibles pour permettre une vérification de la validité des inférences réalisées. Un moyen suggéré dans les politiques institutionnelles d'évaluation des apprentissages pour évaluer les équivalences et les substitutions est d'élaborer et d'administrer un examen démontrant la maîtrise par l'étudiant de la compétence et des éléments de compétence prévus au cours faisant l'objet de la demande.

C'est ici, au niveau de l'extrapolation, que l'épreuve synthèse de programme prend toute son importance; elle permet d'attester l'intégration des apprentissages réalisés dans l'ensemble du programme. L'épreuve synthèse permet donc d'établir le lien entre les apprentissages de l'étudiant et le domaine cible. Le comité de programme élabore l'épreuve synthèse sur la base des compétences, des objectifs et des standards du programme d'études et non en fonction du contenu spécifique de chacun des cours du programme. L'épreuve synthèse tient compte des objectifs et des standards déterminés dans le programme, du profil de sortie de l'étudiant et de la recherche d'harmonisation et d'équivalence.

Avant de recommander au Ministère de décerner le diplôme à un étudiant, et ainsi attester la compétence de cet étudiant dans le domaine cible, les instances du collège vérifient si toutes les conditions de validation de l'argument interprétatif établies par le Ministère et par le collège ont été remplies.

4.3.2 *Il n'y a pas d'erreurs systématiques qui sont susceptibles de miner l'extrapolation (A3.2)*

Pour appliquer l'argument A3.2, il faut chercher à identifier et à contrôler les sources d'erreurs systématiques les plus sérieuses qui pourraient biaiser l'inférence d'extrapolation, c'est-à-dire l'interprétation du résultat relativement à la compétence dans le domaine cible.

Au moment de concevoir les modalités d'évaluation, lors du choix de la méthode d'évaluation (questions objectives, questions à développement, observations des performances, etc.), des règles de notation et des conditions d'observation, la concertation des professeurs au sujet des modes et des instruments d'évaluation permet d'évaluer, d'identifier et de contrôler les sources d'erreurs systématiques les plus sérieuses qui pourraient être susceptibles de miner l'extrapolation de l'interprétation de l'univers de généralisation au domaine cible.

Les politiques institutionnelles d'évaluation des apprentissages font très peu référence aux propriétés de la mesure et à la construction des instruments de mesure. La source d'erreur la plus souvent mentionnée est le niveau de langue parlée et écrite. En effet, l'évaluation des apprentissages d'un étudiant peut être biaisée s'il ne possède pas les compétences linguistiques requises par la méthode d'évaluation. L'étudiant a sa part de responsabilité à cet égard et il devrait chercher à s'assurer que sa performance ne sera pas biaisée par certaines compétences auxiliaires, telles que ses compétences linguistiques. Il peut donc se renseigner sur les exigences linguistiques selon lesquelles il

sera évalué et développer, durant ses études collégiales, les compétences linguistiques requises à son niveau d'études. Le professeur, quant à lui, peut détecter les sources d'erreurs systématiques et adresser l'étudiant qui éprouve des difficultés vers une ressource d'aide appropriée. S'il s'agit des compétences linguistiques, il peut par exemple diriger l'étudiant vers le centre d'aide en français (ou en anglais selon le cas) du collège.

4.4. Inférence d'implication (I4)

L'inférence d'implication (figure 1.4) consiste à interpréter le score en lien avec les implications associées avec la compétence et le domaine cible. Il s'agit de considérer dans l'interprétation du score les contextes dans lesquels la ou les compétences évaluées sont impliquées, les contextes scolaires et professionnels, les contextes de mesure et du quotidien. C'est dans l'inférence d'implication que sont aussi considérées les exceptions et les conditions particulières pouvant influencer l'interprétation.

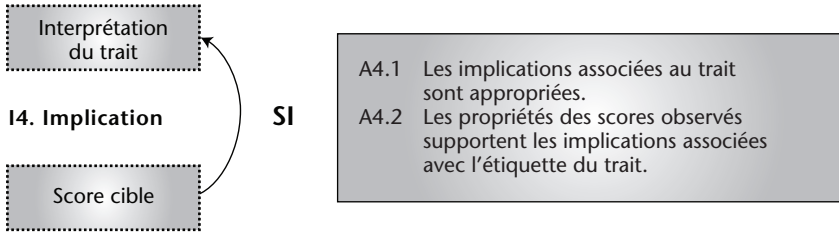


Figure 1.4.
Inférence d'implication (I4)

4.4.1. *Les implications associées au trait sont appropriées (A4.1)*

Les implications regroupent les extrapolations de l'interprétation pour y inclure toutes affirmations ou suggestions associées au trait ou à la compétence cible. Les implications concernent donc les inférences réalisées du résultat au domaine cible en passant par les liens établis entre le programme d'études et le domaine. C'est pourquoi le comité de programme doit définir le profil de sortie de l'étudiant inscrit au programme en lien avec le domaine cible. Le conseil d'administration, quant à lui, établit la liste des conditions que doit respecter l'étudiant pour être diplômé de son programme, par exemple avoir atteint tous les objectifs du programme et réussi l'épreuve synthèse et les épreuves ministérielles. L'établissement, en recommandant au Ministère de

décerner un diplôme à un étudiant, affirme que ce dernier, en ayant satisfait à chacune de ces conditions, devrait être compétent dans le domaine. Les conditions mentionnées concernant la sanction des études et la remise des diplômes constituent des implications associées aux traits ou aux compétences du domaine pour lequel l'étudiant est diplômé. Ainsi, les interprétations associées aux compétences maîtrisées sont bien justifiées par rapport au domaine cible.

Les implications concernent aussi toutes les exceptions appliquées au score et à son interprétation, c'est-à-dire toutes conditions particulières pouvant affecter le résultat qui ne sont pas nécessairement liées à la compétence dans le domaine cible. Il s'agit, par exemple, des exigences liées à la présence aux cours ou aux activités d'évaluation, à la présentation matérielle des travaux ou à la qualité de la langue. Tous ces exemples font référence à des compétences auxiliaires qui, selon l'établissement, doivent nécessairement être maîtrisées pour que la compétence dans le domaine cible soit entièrement atteinte. La direction des études détermine ces exceptions, le département les spécifie dans ses règlements et le professeur en informe les étudiants et les indique dans son plan de cours.

4.4.2. *Les propriétés des scores observés supportent les implications associées avec l'étiquette du trait (A4.2)*

L'explication des choix liés à la pondération contribue à s'assurer que les propriétés des scores observés supportent les implications associées au trait ou à la compétence cible. L'interprétation du score sera différente selon qu'on choisit une pondération plus importante de l'évaluation finale ou une répartition équivalente de la pondération des différentes activités d'évaluation. Les positions quant à ce choix varient entre les PIEA. Selon une première position, chaque cours comprend une évaluation finale dont la pondération est prédominante et plus élevée que chacune des évaluations antérieures. Si la pondération de l'évaluation finale n'est pas plus élevée que chacune des évaluations antérieures, l'étudiant doit obtenir la note de passage à cette évaluation finale pour réussir le cours. L'évaluation finale est l'occasion pour l'étudiant de démontrer, au terme du cours, qu'il maîtrise la compétence visée par le cours. Alors que l'étudiant est en cours d'apprentissage lors des évaluations pendant la session, il devrait en fin de session maîtriser la compétence visée et le démontrer au moyen de sa performance à l'évaluation finale. Le score atteste alors la maîtrise des compétences visées par le cours et supporte donc les implications liées au trait ou à la compétence cible. Selon une autre position, aucune activité d'évaluation ne devrait excéder, par exemple, 25 % de la note

globale d'un cours. Cette position est elle aussi justifiable par rapport à l'argument A4.2, puisqu'en situation authentique la compétence peut rarement être démontrée en une occasion unique (Perrenoud, 2004). En accordant une pondération équivalente à plusieurs évaluations différentes à plusieurs moments, dans divers contextes, l'évaluation de la compétence visée est plus authentique et plus représentative de la compétence en contexte réel. Ici encore, le score supporte les implications liées au trait ou à la compétence cible. Les deux positions sont donc justifiables par rapport à l'argumentation de la validité de l'inférence d'évaluation, mais le choix doit être explicité.

Le fait que les scores et les mentions pouvant figurer au bulletin d'études collégiales, de même que les exceptions à la notation, soient explicités dans la PIEA contribue aussi à s'assurer que les propriétés des scores supportent les implications associées au trait ou à la compétence cible. On précise ainsi les interprétations qui peuvent être faites des scores. Par exemple, le score associé à une réussite étant RE, les propriétés du score supportent les implications liées au fait que l'*étudiant satisfait aux exigences de l'épreuve uniforme de français ou de l'épreuve synthèse de programme*. Le score associé à l'équivalence étant EQ, les propriétés du score démontrent qu'il s'agit d'une équivalence et donc supportent les implications liées au fait qu'il s'agit d'une équivalence. Le score EQ en lui-même est une implication qui limite l'inférence en mentionnant que la validité de l'argument interprétatif n'a pu être vérifiée qu'en partie (inférence d'extrapolation seulement). La description des scores doit donc être complète et non équivoque.

4.5. Crédibilité de l'inférence d'évaluation (C)

Le niveau *Crédibilité* (figure 1.5) a été ajouté à la structure d'argumentation de Kane (2006) afin de représenter de façon plus exhaustive la réalité de l'évaluation des apprentissages en enseignement collégial. Ce niveau indique que l'inférence d'évaluation ne peut être jugée valide que si elle est acceptée par les acteurs impliqués. Selon Kane (2006), pour que l'argument interprétatif soit convaincant, chacune des inférences considérées individuellement doit être convaincante. Un argument a donc été ajouté au modèle initial de Kane (2006) : *A5.1 L'ensemble des arguments est respecté de façon 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 engagés pour évaluer les apprentissages. Un second argument a été ajouté en ce sens : *A5.2 Les processus d'apprentissage et d'évaluation sont connus des acteurs.*

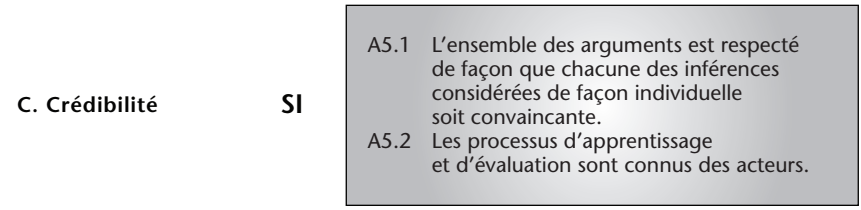


Figure 1.5.
Crédibilité de l'inférence d'évaluation

4.5.1. *L'ensemble des arguments est respecté de façon que chacune des inférences considérées de façon individuelle soit convaincante (A5.1)*

Pour que l'argument interprétatif soit convaincant, chaque inférence considérée de façon individuelle doit être convaincante. Une faille dans n'importe laquelle des inférences produites invalide l'argument interprétatif en entier, même si tous les arguments de validité des autres inférences sont respectés et que leur validité est démontrée (Crooks *et al.*, 1996 : voir Kane, 2006, p. 34).

Les moyens identifiés relativement à cet argument sont de nature métaévaluative. Ils concernent l'évaluation des pratiques d'évaluation des apprentissages en accord avec la politique institutionnelle d'évaluation des apprentissages et les arguments de validité des inférences d'évaluation. Ainsi, la commission des études vérifie que les modalités départementales d'évaluation des apprentissages sont conformes à la politique institutionnelle d'évaluation des apprentissages et le département vérifie que les pratiques d'évaluation y sont conformes. Le comité de programme prend les mesures nécessaires lorsqu'il est avisé du non-respect de la politique institutionnelle d'évaluation des apprentissages. La direction des études établit un portrait statistique de l'évaluation des apprentissages pour chaque département ou programme et de l'application de chaque argument de validité à la fin de chaque année.

4.5.2. *Les processus d'apprentissage et d'évaluation sont connus des acteurs (A5.2)*

On peut penser que l'inférence d'évaluation sera plus facilement acceptée comme valide si les processus d'apprentissage et d'évaluation sont connus des acteurs. Les moyens identifiés pour cet argument visent donc à s'assurer que les personnes et les instances impliquées au sein de l'établissement d'enseignement et toute personne extérieure

susceptible d'avoir accès à l'inférence d'évaluation connaissent les processus d'apprentissage et d'évaluation et sont ainsi mieux outillées pour évaluer elles-mêmes la crédibilité de l'inférence d'évaluation réalisée.

L'un des premiers moyens pour s'assurer que l'inférence d'évaluation est acceptée des acteurs consiste en une assurance de qualité de la part des autorités de l'établissement d'enseignement. Le conseil d'administration du collège peut attester auprès des élèves, du public, du Ministère et de la Commission d'évaluation de l'enseignement collégial la qualité, l'équité et l'équivalence des évaluations des apprentissages réalisées dans leur institution. Le fait qu'une haute instance atteste que des processus rigoureux d'évaluation ont été mis en œuvre peut en satisfaire certains, mais un tel appel à l'autorité ne permet pas de faire connaître aux acteurs les processus impliqués.

Pour s'assurer que les processus d'apprentissage et d'évaluation sont connus des acteurs, la direction des études peut indiquer dans la politique institutionnelle d'évaluation des apprentissages les droits et les responsabilités de toute personne engagée dans l'évaluation des apprentissages et veiller à la diffusion de la PIEA auprès de toute personne ou instance concernée, à l'intérieur ou à l'extérieur de l'établissement d'enseignement. La direction des études peut veiller à informer les principaux concernés, c'est-à-dire les étudiants, sur les conditions de l'obtention du diplôme et sur les exigences du marché du travail. Elle peut aussi veiller à les informer sur les processus d'évaluation en diffusant un calendrier des évaluations pour l'ensemble du collège. Enfin, elle peut informer l'étudiant personnellement sur l'évolution de ses apprentissages en lui remettant un bulletin de mi-session et de fin de session.

Le département, au moyen du plan-cadre de cours, peut informer les professeurs sur les processus d'apprentissage impliqués par chaque cours, les approches pédagogiques et les modalités d'évaluation suggérées. Le département peut chercher à s'assurer que les procédures d'évaluation sont comprises et acceptées en s'attachant à régler tous conflits en lien avec la politique institutionnelle d'évaluation des apprentissages, favorisant par la même occasion l'acceptation de l'inférence d'évaluation.

Le professeur est l'acteur le plus près de l'étudiant et le plus à même de l'aider à comprendre les processus d'apprentissage et d'évaluation impliqués dans chaque cours et dans son programme d'études. Il n'est donc pas étonnant de constater que les moyens qui lui sont associés par rapport à cet argument sont nombreux. Le professeur peut informer directement l'étudiant sur le programme, sur les activités

d'apprentissage et sur les modes d'évaluation des apprentissages grâce au plan de cours. Il peut aussi le faire à travers la planification des activités d'apprentissage et d'évaluation, ou à travers ses interventions individuelles auprès des étudiants, ses rétroactions. Ainsi, le professeur peut planifier les activités d'apprentissage et d'évaluation de façon à préparer progressivement l'étudiant à l'épreuve synthèse de programme et à l'atteinte des objectifs du cours et du programme. Il peut aussi planifier des activités d'évaluation formative qui, en plus de préparer l'étudiant à la prochaine évaluation, le renseignent sur son niveau de compétence. En plus d'aider l'étudiant à avoir une vision plus réaliste de ses performances, cela permet au professeur d'appuyer ses jugements d'évaluation quant au développement des compétences de l'étudiant au cours de la session. L'argument A5.2 renforce aussi l'importance accordée à la pratique d'informer l'étudiant sur les critères d'évaluation. Si l'étudiant connaît les critères sur lesquels sa performance est évaluée, il sera plus enclin à accepter comme valide l'inférence d'évaluation puisque celle-ci est clairement appuyée. En plus de la réponse aux critères, le professeur peut justifier son inférence d'évaluation à travers une rétroaction complète et détaillée faite à l'étudiant. Toutes ces pratiques diminuent les risques de dissonance cognitive entre la perception qu'a l'étudiant de ses propres compétences et l'inférence d'évaluation réalisée. L'étudiant qui a reçu suffisamment d'informations sur la progression de son apprentissage devrait être mieux en mesure de juger de ses performances et de sa démarche d'apprentissage. En ayant une perception réaliste de ses compétences, il est plus susceptible d'accepter comme valide l'inférence d'évaluation, qui lui apparaît crédible puisqu'en accord avec sa propre perception.

L'étudiant a lui-même la responsabilité de s'assurer de connaître ses processus d'apprentissage et les processus d'évaluation. Ainsi, il lui revient de prendre connaissance de la politique institutionnelle d'évaluation des apprentissages, des modalités et des règles d'évaluation propres au cours et au programme d'études, du plan de cours et du calendrier des évaluations. Il doit participer aux activités d'apprentissage ou veiller à faire les apprentissages attendus dans le cours. L'étudiant qui participe aux activités d'apprentissage en classe, étudie, fait les lectures demandées, etc., connaît mieux ses propres processus d'apprentissage tout en ayant une perception plus juste de ses propres compétences. S'il a une perception réaliste de ses propres compétences, l'étudiant sera mieux à même d'accepter l'inférence d'évaluation. En somme, l'étudiant doit assurer le suivi de ses propres apprentissages, chercher à connaître ses processus d'apprentissage, utiliser les informations et les ressources dont il dispose pour ajuster sa

démarche d'apprentissage et chercher l'aide appropriée au besoin. Ce sont des moyens que l'étudiant peut prendre lui-même pour s'assurer de connaître les processus d'apprentissage et d'évaluation.

4.6. Synthèse des résultats

Les moyens identifiés sont répartis selon les différents acteurs et instances de l'établissement d'enseignement. On constate que certains ont plus de responsabilités que d'autres à travers les niveaux d'inférence du modèle. Plus on monte dans les niveaux d'inférence du modèle de Kane (Il à 14 ; Kane, 2006), plus les responsabilités reviennent aux hautes instances du collège. On pouvait effectivement s'attendre à une telle répartition des responsabilités puisque les niveaux inférieurs du modèle, faisant référence à la planification du cours et des évaluations ainsi qu'à la notation, relèvent davantage des fonctions du professeur et du département, alors que les niveaux supérieurs du modèle, faisant référence à la définition du domaine et aux implications des inférences d'évaluation, relèvent davantage du département, du comité de programme, de la direction des études et du conseil d'administration du collège.

Concernant le niveau *Crédibilité* (C) qui a été ajouté au modèle initial de Kane (2006), très peu de moyens ont été identifiés pour l'argument *A5.1 L'ensemble des arguments est respecté de façon que chacune des inférences considérées de façon individuelle soit convaincante.* Il est possible que peu de moyens propres à l'argument A5.l puissent être formulés, parce que celui-ci se retrouve dans chacun des autres arguments et dans l'argument interprétatif dans son ensemble. Il s'agirait donc d'un niveau métaévaluatif, d'un argument qui doit demeurer en tête à tous les niveaux de l'inférence d'*évaluation*. À l'opposé, un très grand nombre de moyens ont été identifiés à l'argument *A5.2 Les processus d'apprentissage et d'évaluation sont connus des acteurs*. On peut qualifier cet argument de transversal, car, bien qu'il ne fasse pas précisément référence aux arguments du modèle de Kane (2006), il est fonction de leur acceptation. C'est aussi l'argument du modèle qui laisse la plus grande place à l'étudiant, puisque ce dernier a peu ou pas d'incidence dans les autres niveaux du modèle.

Les analyses effectuées à partir de la structure d'argumentation de validité de Kane (2006) ont permis de relever les moyens que peuvent mettre en place les institutions pour chercher à assurer la validité des inférences d'évaluation au regard des apprentissages des étudiants. Elles ont toutefois révélé certaines lacunes en lien avec le contenu des PIEA ou certaines incohérences relatives à la validité de l'inférence d'évaluation, qui sont discutées à la section suivante.

5. DISCUSSION

Les lacunes et les incohérences relevées dans les politiques institutionnelles d'évaluation des apprentissages, en lien avec la structure argumentative de Kane (2006), ont particulièrement trait à la mesure, à l'évaluation des compétences auxiliaires et à l'évaluation des dispenses, des substitutions et des équivalences. Ces lacunes et ces incohérences sont discutées ici de façon à alimenter la réflexion sur l'élaboration, l'application et l'évaluation des politiques institutionnelles d'évaluation des apprentissages au collégial dans une perspective de validité des inférences d'évaluation.

5.1. Les lacunes au niveau de la mesure

Les politiques institutionnelles d'évaluation des apprentissages abordent l'évaluation, mais restent muettes sur les questions relatives à la mesure. Par conséquent, elles font très peu référence au développement des procédures et des instruments de mesure. Ce faisant, elles négligent de faire état des considérations liées aux erreurs systématiques susceptibles d'être introduites dans l'inférence d'évaluation dès la construction d'un instrument ou d'une procédure de mesure.

Seules sont abordées des questions liées aux compétences linguistiques qui peuvent être des sources d'erreurs systématiques et elles sont rarement envisagées sous cet angle. Kane (2006) soulève la question des compétences auxiliaires requises par les tests, lesquelles peuvent constituer des sources d'erreurs systématiques lors de la production de l'inférence d'évaluation. Selon lui, si le niveau de compétence dans une compétence auxiliaire (par exemple la lecture) requise par un test est bas, comparativement au niveau de compétence de la population, la compétence auxiliaire en question ne devrait pas constituer une source de variance non pertinente pour la plupart des étudiants. Elle peut tout de même représenter une sérieuse source d'erreur systématique pour certains étudiants, par exemple pour ceux qui éprouvent des difficultés en lecture. En revanche, si ces compétences sont assez variables dans la population testée pour influencer substantiellement les scores, elles peuvent être considérées comme des sources de variance non pertinente. Ces compétences auxiliaires deviennent alors des sources d'erreurs systématiques qu'il faut tenter de réduire lors du développement des instruments et des procédures de mesure.

Certaines politiques institutionnelles d'évaluation des apprentissages suggèrent de procéder à la critique des outils et à la promotion des meilleurs instruments de mesure, mais elles ne mentionnent pas les critères devant fonder cette critique ni ceux qui doivent servir

à la sélection des meilleurs instruments. Les premiers tenants de l'approche par compétences ont souvent privilégié la validité de l'inférence d'extrapolation, c'est-à-dire la représentativité de l'évaluation par rapport au domaine cible. Ainsi, on a vu se multiplier dans les classes les évaluations de type « performance » visant à rapprocher la tâche d'évaluation de la performance en contexte réel. Mais ce choix comporte un inconvénient : celui de diminuer la validité de la généralisation. En effet, la tâche d'évaluation de type performance est conçue pour être plus représentative d'une tâche du domaine cible. Toutefois, comme les évaluations de performances exigent beaucoup de temps, elles incluent généralement un nombre plus restreint de tâches. Dans de telles évaluations, puisque l'inférence de généralisation est effectuée à partir d'un plus petit échantillon de tâches, la validité de l'inférence de généralisation du score observé au score univers s'en trouve diminuée. Mais l'utilisation de tests hautement standardisés présente aussi un inconvénient ; la validité de l'inférence de généralisation est certes augmentée par la représentativité de la mesure à l'univers de généralisation, mais la performance à un test standardisé s'éloigne beaucoup de la performance dans le domaine cible, en contexte réel. On se retrouve donc devant un dilemme : renforcer l'extrapolation aux dépens de la généralisation en faisant en sorte que les tâches d'évaluation soient le plus possible représentatives du domaine cible, ou renforcer la généralisation aux dépens de l'extrapolation en utilisant un grand nombre de tâches très standardisées (Kane, Cooks et Cohen, 1999 : voir Kane, 2006, p. 37). Kane (2006, p. 37) suggère de développer des tests standardisés pour contrôler l'erreur aléatoire. Mais, comme la standardisation diminue la représentativité de l'univers de généralisation par rapport au domaine cible, il suggère de développer les tests de façon à toucher aux principales habiletés comprises dans le domaine cible. Ainsi, une preuve analytique de validité peut être établie lors de la conception du test en veillant à ce que celui-ci soit le plus représentatif possible du domaine cible (Kane, 2006, p. 37).

Raîche (2008, p. 4) aspire à arriver à une correction plus objective des compétences, c'est-à-dire que les jugements d'appréciation puissent être remplacés *par des actions de vérification de la présence ou de l'absence de manifestations de la compétence*. Ainsi, il encourage l'utilisation d'échelles descriptives analytiques ou globales plutôt que d'échelles d'appréciation dans l'évaluation des productions ou des performances d'un étudiant à l'enseignement supérieur. Encore une fois, la concertation des professeurs et des départements pourrait servir à établir les principales habiletés du domaine cible qui devraient être évaluées, les critères et les standards d'évaluation.

5.2. Le problème de l'intégration de compétences auxiliaires à l'évaluation

L'évaluation des apprentissages au collégial intègre la mesure des cibles d'évaluation qui ne sont pas liées au trait, aux apprentissages ou à la compétence ciblée. Comme le mentionnent Howe et Ménard (1993), il s'agit de cibles d'évaluation telles que le respect des exigences de présentation des travaux, la motivation et l'effort à l'apprentissage, la participation active de l'étudiant en classe et la présence assidue aux cours. Les politiques institutionnelles d'évaluation des apprentissages mentionnent, par exemple, la présence aux cours ou aux activités d'évaluation, la présentation matérielle des travaux ou la qualité de la langue dont l'évaluation influence, souvent à la baisse, le résultat de l'évaluation.

Ainsi, l'évaluation des apprentissages actuellement réalisée au collégial, du moins selon les politiques institutionnelles d'évaluation des apprentissages analysées, inclut l'évaluation de certains traits ou de certaines compétences auxiliaires sans toutefois qu'elles soient impliquées dans la mesure de la compétence. Le taux d'absentéisme en est un bon exemple. L'absence aux activités d'apprentissage et d'évaluation peut entraîner le retrait du droit de l'étudiant d'être évalué et, conséquemment, un échec, et ce, peu importe le niveau de compétence de l'étudiant dans le domaine cible. La mention échec est aussi attribuée lorsque l'abandon de cours n'a pas été effectué dans les délais prescrits. Or, les propriétés de la mention échec, si elles supportent les implications associées au trait ou à la compétence mesurés, devraient signifier l'incompétence de l'étudiant par rapport au domaine cible. Cependant, ce n'est pas de la compétence de l'étudiant dans le domaine que témoigne l'échec dans ce cas précis, mais bien d'une compétence auxiliaire, soit la motivation ou tout autre trait ou compétence liés à l'absentéisme ou à l'abandon, qui seule détermine le score et biaise l'inférence d'évaluation. Cette erreur d'inférence peut nuire passablement à l'étudiant qui se voit jugé sur son incompétence dans le domaine face à un échec qui cache en réalité une tout autre interprétation. Certains établissements d'enseignement utilisent la mention EA signifiant Échec par abandon pour contourner ce problème d'interprétation du score. Au regard du modèle d'argumentation de la validité des inférences d'évaluation et des considérations éthiques, cette solution semble justifiée.

Les développements de la recherche en mesure et évaluation proposent aussi des façons d'intégrer simultanément à la mesure de la compétence une ou des mesures de la qualité des patrons de réponse de l'étudiant (Blais, Raîche et Magis, 2009 ; Raîche, Magis et Blais,

2008). Il s'agit de modélisations psychométriques qui permettent de corriger le score d'un étudiant malgré des tentatives de fraude pour augmenter son résultat, de l'inattention de sa part ou une fluctuation de son niveau de performance lors de la réalisation de la tâche d'évaluation. De cette façon, certaines sources de biais peuvent être contrôlées statistiquement et l'estimation du niveau de compétence de l'étudiant peut être plus représentative des scores univers et cibles.

Il est aussi possible de remettre en question la pertinence de l'évaluation de la qualité de la langue en tant que compétence auxiliaire d'un domaine. En incluant la qualité de la langue dans l'évaluation, on affirme qu'il s'agit d'une compétence faisant partie du domaine et que son niveau de maîtrise est une condition au jugement de la compétence ou de l'incompétence d'un individu dans le domaine cible. On peut se demander s'il est socialement acceptable de conclure à l'incompétence d'une personne dans un domaine, ou, pire, de lui refuser l'accès à la formation dans ce domaine, en raison de son échec à l'évaluation d'une compétence auxiliaire. Les compétences en langue écrite et parlée sont des compétences auxiliaires évaluées dans la quasi-totalité des domaines d'éducation au Québec. En plus d'être évaluées pendant la formation et de constituer une exigence pour l'obtention du diplôme, elles sont évaluées à l'entrée de la formation, à titre de critère d'admission. Toutefois, sur le marché du travail, le niveau de maîtrise de la langue exigé diffère selon les domaines, tandis que cette différence n'est pas toujours reflétée dans la formation et l'évaluation. L'évaluation des compétences langagières devrait être adaptée au niveau exigé dans le domaine et ces compétences, comme toute autre compétence incluse dans le domaine cible, devraient être définies par des experts du domaine. Une seule politique institutionnelle d'évaluation des apprentissages justifie l'évaluation de la langue en tant que compétence auxiliaire d'importance au sein du domaine cible. Dans certaines politiques institutionnelles d'évaluation des apprentissages, il revient aux départements de préciser les cas pour lesquels l'évaluation de la qualité de la langue est inapplicable. Il s'agit là d'une mention qui tient compte de la mise en garde de Kane (2006) quant à l'évaluation des compétences auxiliaires et qui entre dans les exceptions à considérer dans l'inférence d'évaluation.

Des solutions, des moyens spécifiques doivent être explorés pour éviter que l'évaluation des compétences auxiliaires résulte en une erreur d'inférence. La teneur de la présente recherche n'a pas permis de le faire, mais cela devrait ultimement faire l'objet d'une réflexion concertée de la part des chercheurs en mesure et évaluation des apprentissages et des acteurs des établissements d'enseignement.

5.3. Le problème de l'évaluation des dispenses, des substitutions et des équivalences

Le fait d'appliquer un modèle de validité à l'analyse des politiques institutionnelles d'évaluation des apprentissages a permis de cerner un problème récurrent en évaluation des apprentissages : l'évaluation des dispenses, des substitutions et des équivalences.

Trois des cinq politiques institutionnelles d'évaluation des apprentissages analysées suggèrent de faire appel à l'accréditation de l'établissement comme garantie pour juger de la validité des inférences de notation et de généralisation lors de l'attribution des dispenses, des substitutions et des équivalences. Dans ce cas, le service de l'établissement responsable de la reconnaissance des acquis doit supposer que les inférences de notation et de généralisation réalisées antérieurement, dans un cours, un programme ou même dans un établissement différent, sont valides pour être en mesure de faire lui-même l'inférence d'extrapolation qui l'amène à reconnaître le score attribué au cours comme valide dans son interprétation au sein du domaine cible. Pour ce faire, ce service n'a qu'une seule garantie de la validité des inférences de notation et d'extrapolation : *que les objectifs du cours aient fait l'objet d'une évaluation formelle au sein d'un établissement reconnu*. Les preuves sur lesquelles s'appuie cette garantie sont : *bulletin, attestation officielle de la réussite du ou des cours suivis, description des cours et plans de cours ou tout autre document pertinent*. Ce qui se produit alors, c'est que l'établissement introduit une valeur inconnue au sein de l'univers de généralisation qu'il suppose valide sur la seule base de l'accréditation de l'établissement d'enseignement qui fournit cette valeur et sur la similitude des objectifs et/ou des contenus planifiés. Le problème qui se pose avec cette façon de procéder est qu'il est impossible de vérifier que les arguments des inférences de notation et de généralisation sont respectés.

Pour résoudre ce problème, deux des cinq politiques institutionnelles d'évaluation des apprentissages suggèrent d'élaborer et d'administrer un examen démontrant la maîtrise par l'étudiant de la compétence et des éléments de compétence prévus au cours faisant l'objet de la demande de reconnaissance. Il s'agit là d'une solution qui permet de s'assurer de la validité de l'inférence d'évaluation, puisque l'établissement reprend alors le contrôle de chacun des niveaux d'inférence et peut ainsi vérifier que chaque argument de validité est respecté, depuis la notation (I1) jusqu'à l'implication (I4), et s'assurer de la crédibilité de l'inférence d'évaluation réalisée.

6. CONCLUSION

L'analyse du contenu des politiques institutionnelles d'évaluation des apprentissages a permis d'identifier des moyens que peuvent mettre en place les institutions pour assurer la validité des inférences d'évaluation au regard des apprentissages des étudiants. Des moyens ont donc pu être relevés pour chaque argument de la structure d'argumentation de Kane (2006) et même davantage en identifiant des façons de s'assurer de la crédibilité de l'inférence d'évaluation aux yeux des acteurs impliqués. Toutefois, les moyens évoqués dans le contenu des politiques institutionnelles d'évaluation des apprentissages sont de nature générale et correspondent plus à des lignes directrices qu'à de véritables moyens.

Pour permettre aux établissements d'enseignement collégial *d'orienter, d'encadrer et de soutenir les activités liées à l'évaluation des apprentissages* (Conseil des collèges, 1992, p. 38) et ainsi contribuer à assurer la validité des inférences d'évaluation des apprentissages, les politiques institutionnelles d'évaluation des apprentissages auraient avantage à inclure dans leur contenu des moyens spécifiques d'application et d'évaluation des arguments de validité de la structure d'argumentation de la validité des inférences d'évaluation des apprentissages.

La recherche en mesure et évaluation en éducation met à la disposition des professionnels en éducation plusieurs avancées qui constituent autant de moyens susceptibles d'être utilisés pour élaborer chacune des lignes directrices relevées dans la présente recherche. Avec la structure d'argumentation de la validité des inférences d'évaluation des apprentissages et les lignes directrices énoncées, une analyse des pratiques effectives dans les établissements d'enseignement pourrait aussi permettre d'identifier des moyens d'application et d'évaluation spécifiques.

RÉFÉRENCES

Blais, J.-G., Raîche, G. et Magis, D. (2009). La détection des patrons de réponses problématiques dans le contexte des tests informatisés. Dans J.-G. Blais (dir.), *Évaluation des apprentissages et technologie de l'information et de la communication. Enjeux, applications et modèles de mesure*. Québec, Québec : Les Presses de l'Université Laval.

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. (1971). Test validation. Dans R. L. Thorndike (dir.), *Educational measurement* (4e éd.). Washington, District de Columbia : American council on education.

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

Cureton, E. E. (1950). Validity. Dans E. F. Lindquist (dir.), *Educational measurement*. Washington, District de Columbia : American council on education.

Howe, R. et Ménard, L. (1993). *Croyances et pratiques en évaluation des apprentissages*. Rapport de recherche. Laval, Québec : Cégep Montmorency.

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

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

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

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

Lissitz, R. W. (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 Jersey : 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, 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.

Paquette-Côté, K. (2010). *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.

Paquette-Côté, K. et Raîche, G. (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*. Communication présentée au LXXVII[e] Congrès de l'Acfas, Université d'Ottawa, Ottawa, Ontario.

Paquette-Côté, K. et Raîche, G. (2011). La validité des inférences d'évaluation dans les politiques institutionnelles d'évaluation des apprentissages des établissements d'enseignement collégial québécois. Dans G. Raîche, K. Paquette-Côté et D. Magis (dir.), *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation. Tome 2 : L'évaluation*. Montréal, Québec : Presses de l'Université du Québec.

Perrenoud, P. (2004). Évaluer les compétences. *Éducateur*, numéro spécial *La note en pleine évaluation*, 8-11.

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

Raîche, G., Magis, D. et Blais, J.-G. (2008). *Multidimensional item response theory models integrating additional inattention, pseudo-guessing, and discrimination person parameters*. Communication présentée lors du congrès annuel de la Psychometric Society, Durham, New Hampshire.

Rivard, J., Banville, C., Gareau, D., Léonard, M., Mihaila, S., Paquette, G. et Rosca, I. (2006). *Logiciel Mots Plus (version 1.6). Éditeur de modèles de connaissance. Manuel de l'utilisateur.* Québec, Québec : Télé-université, LICEF.

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

Chapitre 2

Comparaison d'outils technologiques permettant l'analyse de données à l'aide des modèles de Rasch

Éric Dionne, Julie Grondin et Jean-Guy Blais

La modélisation de la mesure est une opération importante qui permet d'étayer la validité de construit en évaluation ou en recherche. Or, plusieurs logiciels permettent maintenant de modéliser les scores bruts. Dans le cadre de ce texte, nous exposerons les résultats d'une analyse comparative de deux logiciels (RUMM2020 et Winsteps) qui porte autant sur l'ergonomie de l'interface que sur les méthodes d'estimation (paramètres, adéquation modèle-données) ou sur l'information produite par ces logiciels. L'analyse illustre que les deux logiciels diffèrent quant à leur convivialité et à leur utilité pour certains types de recherche, mais que globalement chacun offre des résultats comparables tant au regard de la mesure que de la précision de cette dernière.

1. INTRODUCTION

La modélisation de la mesure est une activité à la fois théorique et pratique. De nombreux développements sont apparus au cours des dernières années dans le domaine de la modélisation théorique des scores aux tests. Chaque année, de multiples articles (Blais et Maheux, 2003 ; Loye, 2005) font état de propositions originales visant à mieux modéliser les données ou les scores obtenus à l'aide de tests, de questionnaires ou de grilles d'observation par exemple,

et ce, dans des disciplines variées comme la didactique des sciences (Boone et Scantlebury, 2006 ; Liu et Boone, 2006) ou les sciences de la santé (Hagquist, Bruce et Gustavsson, 2009 ; Massof, 2002 ; Smith, Wright, Selby et Velikova, 2007 ; Tang Wai, Wong, Chiu, Lum et Ungvari, 2005). Longtemps réservée à une élite composée d'initiés[1] maîtrisant les techniques mathématiques, la modélisation de la mesure tend à se démocratiser auprès des membres de la communauté académique et scientifique. De ce fait, nous assistons à une plus grande prise en compte par les chercheurs des aspects liés à la mesure, ce que nous ne pouvons que saluer.

Plusieurs facteurs peuvent expliquer l'intérêt grandissant pour la modélisation de la mesure. L'un de ces facteurs est assurément la plus grande disponibilité de logiciels spécialisés qui permettent maintenant de réaliser des analyses avancées sans devoir nécessairement maîtriser un langage, souvent rébarbatif, de programmation ou encore un vocabulaire permettant le codage des instructions de modélisation. À une certaine époque, seuls les spécialistes de la mesure s'intéressaient à ces outils et les maîtrisaient, mais depuis quelques années la convivialité des outils d'analyse s'est grandement améliorée. Cet état de fait permet maintenant à des « non-spécialistes » de la mesure d'utiliser à profit ces outils et ainsi d'apporter une plus-value à leurs travaux en prenant en compte les propriétés métriques des données analysées.

L'idée de produire un texte à vocation *pratique* sur les outils récents de modélisation nous occupait l'esprit depuis un bon moment. En effet, à notre connaissance, il y avait relativement peu de textes publiés qui traitaient de ce sujet (Pomplun, Omar et Custer, 2004 ; Robin, Xing et Hambleton, 1999 ; Sick, 2009 ; Skaggs, 2004) et encore moins des écrits en français. Il nous apparaissait pertinent de présenter les différents logiciels actuellement disponibles tout en soulignant les forces et les limites de chacun d'eux. C'est à ce besoin que le présent texte tentera de répondre. Toutes les personnes (étudiants diplômés, chercheurs, professionnels de recherche, professeurs, etc.) qui s'intéressent à la nature et à la qualité des données devraient trouver des pistes intéressantes dans le choix et l'utilisation des outils de modélisation que nous présentons. Évidemment, tout propos associé à des outils technologiques est irrémédiablement condamné à être un propos éphémère. En effet, les logiciels actuellement disponibles seront, tôt ou tard (mais probablement plus tôt que tard), remplacés par d'autres outils, ce qui pourrait rendre caduc l'essentiel de la réflexion contenue dans ce texte. Pour tenter d'éviter cet écueil, nous avons organisé la comparaison des logiciels en faisant reposer notre analyse

1. Le masculin est utilisé à titre épicène dans le seul but d'alléger le texte.

sur des critères qui devraient perdurer dans le temps et qui pourraient être éventuellement repris pour produire d'autres études comparatives comme celle-ci. Aussi, certaines pistes de réflexion pourraient être considérées pour d'éventuelles améliorations à apporter sur l'un ou l'autre des logiciels qui seront discutés dans ce texte. À ce sujet, nous avons choisi de restreindre l'inventaire des logiciels à ceux ayant pour vocation principale de modéliser les scores bruts au moyen d'analyses de type Rasch. Ainsi, les logiciels permettant de traiter les données au moyen des modèles à deux ou trois paramètres ne seront pas présentés.

Ce texte est divisé de la façon suivante. À la section 2, nous présentons brièvement la modélisation de type Rasch sous un angle conceptuel. La section 3 dresse un inventaire des différents logiciels actuellement disponibles et qui permettent de réaliser les analyses dont il est question dans cet écrit. Quant aux sections 4 et 5, elles présentent la méthodologie sur laquelle nous nous sommes appuyés et les résultats de notre analyse comparative. Enfin, nous concluons ce chapitre en faisant une synthèse des résultats obtenus et en indiquant quelques pistes pour d'autres publications en lien avec le sujet de ce texte.

2. LA MODÉLISATION DE TYPE RASCH

La famille des modèles de Rasch se distingue d'abord des autres propositions de modélisation de la mesure parce que ce qui est important avant tout, c'est que les données s'ajustent bien au modèle et non l'inverse, comme pour la régression multiple. Ainsi, lorsque les données ne s'ajustent pas au modèle, il faut produire de meilleures données, donc revoir les items et les remettre à l'essai, ou en produire de nouveaux plus en phase avec le construit et susceptibles de satisfaire aux exigences de la modélisation. Si les données s'ajustent adéquatement au modèle, elles produiront des estimations des paramètres sur une échelle d'intervalle avec une structure linéaire. La modélisation proposée par Rasch (1960) permet des mesures individuelles avec des unités égales, des erreurs de mesure différentes pour chaque score, le tout indépendamment des caractéristiques du test et supporté par des indicateurs de la qualité de l'ajustement pour les items et les répondants.

Le modèle unidimensionnel de Rasch pour les variables dichotomiques ($x = 0, 1$) prend la forme suivante avec θ_n comme paramètre associé à la position des répondants et b_i comme paramètre associé à la position des items :

$$P\left(X_{ni} = |\, \theta_n, b_i\right) = \frac{e^{(\Theta_n - b_i)}}{1 + e^{(\Theta_n - b_i)}}. \qquad (1)$$

La représentation visuelle du modèle est appelée la courbe caractéristique de l'item et la figure 2.1 illustre cette représentation pour la situation où la variable est dichotomique :

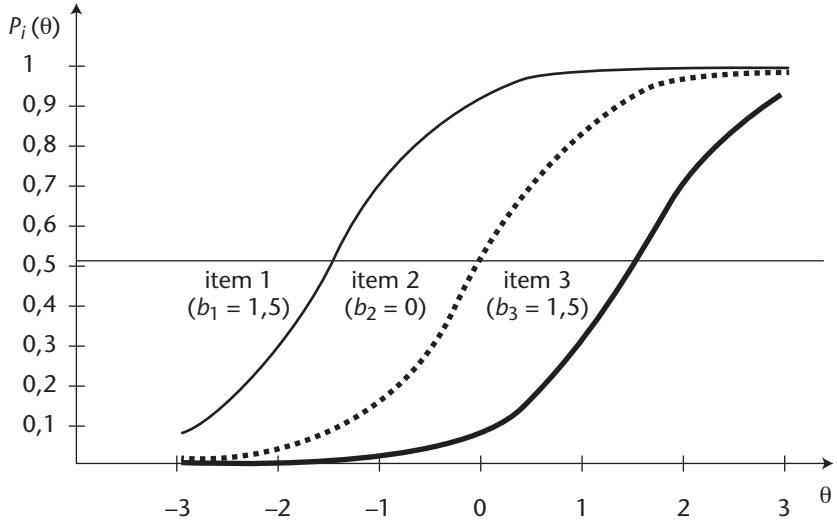


Figure 2.1.
Courbes caractéristiques de trois items pour le modèle dichotomique de Rasch

Le modèle polytomique d'Andrich (1978) est, pour sa part, l'un des modèles proposés pour des échelles de réponses ordonnées avec plus de deux catégories lorsque les échelles sont identiques pour chacun des items. Il s'agit aussi d'un modèle unidimensionnel avec un paramètre θ_n pour la position du candidat, un paramètre b_i pour la position de l'item et un paramètre F_x pour les catégories de l'échelle de réponses (x = 0, 1, 2…).

$$P\left(X_{nxi} = x \mid \theta_n, b_i, F_x\right) = \frac{e^{\left(\Theta_n - b_i - F_x\right)}}{1 + e^{\left(\Theta_n - b_i - F_x\right)}} \qquad (2)$$

Malgré leur simplicité d'application et leur versatilité, les modèles de la famille de Rasch sont aussi soumis à un certain nombre de conditions balisant les applications adéquates (Bertrand et Blais, 2004, p. 183-191). Ces conditions sont de différents ordres et nécessitent la mise en place de démonstrations empiriques qui doivent être produites

pour chacune des applications d'un modèle. La modélisation nous amène ainsi à tenir compte de quatre aspects qui forment la base de la vérification de la qualité de l'adéquation du modèle :

- la dimensionnalité de l'espace des variables latentes ;
- l'indépendance locale ;
- les ajustements statistique et résiduel aux données d'un modèle ou de plusieurs modèles concurrents ;
- le maintien de la propriété d'invariance des estimations des paramètres associés aux items et aux sujets.

Ainsi, lorsque le modèle choisi pour une application donnée est un modèle unidimensionnel, il faut produire une preuve raisonnable de cette unidimensionnalité (voir, par exemple, Tennant et Pallant, 2006, pour des détails à ce sujet). De plus, pour estimer les paramètres des modèles, nous posons comme condition qu'il y ait indépendance, pour une valeur fixée sur le continuum des candidats, entre les réponses à des items différents. En outre, comme le modèle peut très bien ne pas être le bon modèle, il faut montrer que les données s'ajustent bien au modèle. Enfin, les modèles de Rasch possèdent également une propriété théorique fondamentale, la propriété d'invariance (ou de l'objectivité spécifique). Une opération de vérification du maintien de cette propriété dans les situations de modélisation est nécessaire. Cette propriété permet d'affirmer que :

- l'estimation de l'habileté d'un individu est indépendante des items auxquels celui-ci doit répondre ;
- les estimations des caractéristiques des items sont indépendantes des caractéristiques des individus qui répondent aux items.

Plusieurs logiciels permettent maintenant aux chercheurs d'analyser leurs données à l'aide d'une modélisation de type Rasch afin d'appliquer les dimensions théoriques que nous venons de présenter. La section suivante fera l'inventaire de certains d'entre eux.

3. INVENTAIRE DES LOGICIELS PERMETTANT DE RÉALISER DES ANALYSES DE RASCH

Les outils informatiques permettant de traiter les données avec le modèle de Rasch existent déjà depuis un bon moment. Au cours des dernières années, on a remarqué une nette amélioration en ce qui concerne la convivialité des outils et les ressources disponibles afin, par exemple, d'interpréter les résultats et les statistiques disponibles dans ces outils. Il fut une époque, pas si lointaine, où les logiciels

étaient relativement peu conviviaux et souvent issus d'initiatives prises par des groupes de recherche qui voulaient développer des outils d'analyse. À ce moment-là, on sentait un désir de partager les résultats de ces développements avec la communauté des chercheurs. Les logiciels comme Conquest 1.0 et Bilog sont de bons exemples des produits offerts durant cette période. Ces logiciels exigeaient souvent la connaissance d'un langage de programmation adapté à l'outil. On trouvait peu de documentation tant sur l'utilisation des logiciels que sur la façon d'interpréter les résultats. La plupart des articles mentionnant l'utilisation de ces outils étaient en outre rédigés par des spécialistes de la mesure. L'arrivée du logiciel Winsteps a marqué, en quelque sorte, une petite révolution dans le paysage des outils d'analyse Rasch en proposant un outil convivial et bien documenté. Pour l'une des premières fois, il devenait plutôt facile de réaliser des analyses de type Rasch sans nécessairement maîtriser un langage de programmation. Dans la même lignée, RUMM2020, Conquest 2.0, T-Rasch, etc., ont suivi en proposant des outils qui se distinguent à divers points de vue (méthodes d'estimation des paramètres, types de statistiques d'ajustement, etc.). On peut remarquer que ces outils, plus conviviaux et pour lesquels une documentation plus étoffée existe, ont permis un essor de l'utilisation des modèles de Rasch auprès des non-spécialistes de la mesure. L'essor d'Internet aidant, on trouve maintenant des forums de discussion, des listes de diffusion, etc., associés à certains de ces logiciels, ce qui en facilite encore l'utilisation.

Actuellement, on peut répertorier une quarantaine de logiciels spécialisés ou de routines pouvant être exécutées sur un logiciel de traitement de données comme SAS. Une liste fréquemment mise à jour est disponible, par exemple à l'Institute for Objective Measurement (http://www.rasch.org/software.htm) ou encore sur le Rasch Measurement Special Interest Group (RaschSIG, <http://www.raschsig.org/soft.html>). On peut les regrouper selon différentes caractéristiques comme le type de données traitées. À titre d'exemple, on trouve des logiciels qui ne traitent que les données dichotomiques (T-Rasch, Bilog, RSP, Raschtest, etc.) et d'autres qui traitent des données dichotomiques et polytomiques (RUMM2020, Winsteps, Facets, Conquest, etc.). Une autre catégorisation peut s'effectuer en tenant compte des méthodes d'estimation. Certains utilisent la méthode du maximum de vraisemblance conjointe (Facets, Quest, Winsteps, etc.), alors que d'autres recourent, par exemple, à la méthode du maximum de vraisemblance conditionnelle (Winmira, LPCM-Win, etc.). Nous pouvons également souligner la présence de certains outils gratuits qui sont de plus en plus faciles d'accès, par exemple dans les librairies du logiciel R, un logiciel libre qui s'enrichit de la contribution des personnes intéressées

par la modélisation de type Rasch. Ajoutons qu'il serait aussi possible de classer les outils selon bien d'autres caractéristiques, telles que les coûts d'acquisition, les méthodes d'ajustement modèle-données ou encore les possibilités de présentation des résultats. Tout comme le menuisier doit avoir le bon outil pour accomplir une tâche donnée, le chercheur devra préciser ses besoins et choisir – parmi tous les outils disponibles – celui qui lui permettra de réaliser les analyses désirées tout en tenant compte des ressources dont il dispose.

4. MÉTHODOLOGIE

4.1. Les logiciels comparés

Dans le cadre de ce texte, nous avons choisi de comparer deux logiciels : Winsteps (Linacre, 2002) et RUMM2020 (Andrich, Sheridan et Luo, 2004). Au moment où nous terminions ce texte, une nouvelle version de RUMM2020 a été lancée, soit le RUMM2030. En raison des contraintes d'édition, nous n'avons malheureusement pas été en mesure d'utiliser cette nouvelle version pour effectuer nos analyses. Plusieurs raisons nous ont incités à choisir ces outils en particulier. D'abord, ce sont deux logiciels fréquemment utilisés et de très nombreux articles (Jackson, Draugalis, Slack et Zachry, 2002 ; Lamoureux, Pallant, Pesudovs, Hassell et Keeffe, 2006 ; Muller et Roddy, 2009 ; Pallant et Tennant, 2007) font état de l'utilisation de l'un ou l'autre de ces logiciels afin de produire des analyses de type Rasch dans des domaines variés (médecine, psychologie, éducation, etc.). Il est possible de se procurer ces outils informatiques facilement, mais la différence de prix entre les deux logiciels est considérable. Par exemple, une licence individuelle au prix éducatif pour RUMM2020 coûtait en 2010 environ 800 $ CAN, alors que la licence de Winsteps se détaillait à environ 150 $ CAN. Des licences institutionnelles ou à accès plus large (ex. : équipe de recherche de quelques personnes) sont également disponibles. Enfin, des références facilement accessibles existent pour ces outils, ce qui simplifie leur utilisation pour des personnes qui souhaitent s'initier à la modélisation Rasch sans nécessairement réaliser de développement sur les modèles de mesure en tant que tels.

4.2. Les dimensions comparées

Nous avons relevé deux grandes catégories de dimensions que nous souhaitions comparer. Dans un premier temps, nous nous attardons aux aspects plus qualitatifs liés à l'utilisation des logiciels. Cette partie traite plus particulièrement de l'ergonomie et de la convivialité pour

l'usager. Nous analysons également le niveau de connaissances requis par l'usager pour l'importation des données dans le logiciel, pour l'exploration et l'analyse des données, ainsi que les ressources disponibles pour aider l'usager dans son utilisation de ce type de logiciel. Dans un second temps, nous abordons les aspects plus quantitatifs ou techniques liés à l'analyse des données et aux résultats de la modélisation. Nous traitons donc de la question 1) des méthodes d'estimation et de la précision de la mesure et 2) de l'ajustement modèle-données. En ce qui concerne ce dernier point, nous avons procédé à une seule itération. Normalement, il faut en réaliser plusieurs (éliminer les scores de certains sujets, éliminer certains items, ajuster les catégories, etc.) afin d'obtenir la meilleure adéquation possible entre les données et le modèle. Les contraintes de rédaction nous ont empêchés de présenter ces résultats.

4.3. Les données utilisées pour les simulations

Afin de comparer les résultats des deux logiciels au regard des indicateurs quantitatifs (difficulté des items, position des sujets sur l'échelle, ajustement des données, etc.), nous avons utilisé deux ensembles de données réelles provenant d'une étude menée par Dionne (2008) sur les résultats associés à l'expérimentation d'un modèle d'évaluation des apprentissages en science. Dans le cadre de ce texte, nous avons délibérément choisi d'utiliser des données réelles, et non des données simulées, pour des raisons d'authenticité. En effet, dans la très grande majorité des cas, les utilisateurs travaillent avec des données réelles et doivent prendre des décisions dans des contextes où l'incertitude et le doute sont omniprésents. Aussi, aux fins du présent exercice, avons-nous jugé plus profitable d'utiliser des données réelles. Par conséquent, le premier ensemble contient des données dichotomiques recueillies auprès de 184 répondants et au regard de 12 items ; les données proviennent de 75 garçons et de 109 filles. Le second ensemble contient, pour sa part, des données polytomiques recueillies auprès de 189 répondants. Pour cet ensemble, il y avait 82 participants masculins et 107 participants féminins. Les scores du second ensemble sont situés sur une échelle en 5 points (0-1-2-3-4) obtenus à partir des 12 mêmes items que ceux cités précédemment pour l'ensemble de données dichotomiques. Le tableau 2.1 présente les moyennes et les écarts types des scores aux items pour chacun de ces ensembles de données.

Tableau 2.1.
Moyennes et écarts types pour les deux ensembles de données utilisés

Items	Ensemble 1 (données dichotomiques)		Ensemble 2 (données polytomiques)	
	Moyenne	Écart type	Moyenne	Écart type
1	0,63	0,49	2,60	1,31
2	0,52	0,50	2,19	1,46
3	0,49	0,50	2,23	1,27
4	0,58	0,50	1,85	1,50
5	0,30	0,46	1,89	1,25
6	0,21	0,41	1,83	1,37
7	0,11	0,32	0,68	1,18
8	0,12	0,33	1,47	1,35
9	0,25	0,43	1,91	1,43
10	0,14	0,34	1,90	1,39
11	0,53	0,50	2,09	1,41
12	0,50	0,50	2,02	1,44

5. RÉSULTATS ET DISCUSSION

5.1. Utilisation des logiciels : une comparaison qualitative

Dans cette première partie de l'analyse et de la discussion, nous comparons les deux logiciels au regard de 1) l'ergonomie et la convivialité de l'interface usager, 2) l'importation des données, 3) l'exploration et l'analyse des données et 4) l'aide offerte aux usagers.

5.1.1. *Ergonomie et convivialité pour l'usager*

Les deux logiciels étudiés fonctionnent sous le système d'exploitation Windows. Tous deux offrent donc à l'usager une interface graphique lui facilitant l'accès aux différentes options du logiciel. Dans les deux cas, l'usager a la possibilité de réaliser des analyses de données sans nécessairement devoir connaître le langage de programmation associé à l'outil. Le logiciel Winsteps offre, dès l'ouverture, la possibilité à l'usager de travailler selon un mode d'instructions données au logiciel, ou de travailler à l'aide d'une interface graphique plus simple, appelée *Setup procedure*. Cette dernière est constituée d'une fenêtre principale comportant différents menus, mais aussi et surtout, des champs à compléter, des options à choisir à l'aide de boutons radio, et des boutons qui permettent de commander des actions précises. Bien que cette interface se veuille une présentation plus conviviale

pour l'usager, certaines options ne sont peut-être pas aussi intuitives qu'on le souhaiterait. L'usager doit avoir une certaine connaissance d'analyses de type Rasch pour s'y retrouver. À cet égard, RUMM2020 se détache de Winsteps en proposant une interface graphique beaucoup plus développée. Dès l'ouverture, RUMM2020 offre à l'usager des options simples et des boîtes de dialogue lui demandant de réaliser une action précise. L'usager est ainsi guidé étape par étape par le logiciel dans la configuration de son analyse, ce qui fait de RUMM2020 un logiciel assez simple à utiliser. Toutefois, le nom de certains boutons de commande manque quelquefois un peu de clarté. De plus, pour certaines des étapes proposées, on ne navigue pas toujours aussi intuitivement qu'on le souhaiterait. En effet, en certains endroits, la navigation se fait à l'aide de la touche Entrée (Retour ou *Enter*) et, à d'autres, elle se fait à l'aide de la touche Tabulation. Cela peut être un peu déroutant et causer quelques inconvénients, dont celui d'avoir parfois le sentiment de se trouver dans une impasse et de ne plus pouvoir poser d'actions afin de poursuivre la mise en place de l'analyse. Il s'agit certainement d'un écueil dû à la « jeunesse » du logiciel, mais qui peut décourager les débutants dans la poursuite des analyses qu'ils souhaitent exécuter. Malgré tout, RUMM2020 est, selon nous, le logiciel le plus facile d'utilisation pour un débutant et c'est le logiciel qui requiert le moins de connaissances en langage de programmation.

L'analyse de l'ergonomie et de la convivialité de ces logiciels exige que nous traitions aussi du niveau de connaissances requis par l'usager pour l'importation des données dans le logiciel. En effet, de façon générale, les données d'un test, d'un questionnaire ou d'une grille d'observation sont d'abord saisies dans un logiciel comme Excel ou dans des logiciels d'analyse plus spécialisés, tels SAS ou SPSS. Pour une personne intéressée à modéliser ses données à l'aide d'une analyse de type Rasch, il importe donc de savoir s'il est possible de transférer les données de ces logiciels vers ceux qui permettent ce type d'analyse.

5.1.2. *Importation des données*

Les deux logiciels analysés ici permettent d'importer des données provenant de logiciels comme Excel, SAS ou SPSS. Dans un premier temps, les données provenant de ces logiciels doivent être sauvegardées sous un format aux tabulations délimitées (*Tab delimited file*, *****.dat). Ensuite, dans un logiciel éditeur de fichiers textes comme Notepad ou Wordpad, certaines transformations sont nécessaires. Tout d'abord, il faut s'assurer de supprimer toutes tabulations ou espaces superflues. Tous les caractères identifiant les sujets, leurs caractéristiques (sexe, âge, etc.) et leurs réponses aux items doivent être les uns à la suite

des autres, sans séparation. Il faut cependant vérifier que toutes ces variables sont alignées dans des colonnes qui leur sont propres. Par exemple, si les numéros d'identification des sujets vont de 1 à 300, il faut s'assurer que le numéro « 1 » du premier sujet se trouve dans la troisième colonne, aligné avec les unités du numéro 300. Pour ce faire, il est possible d'insérer deux espaces devant le chiffre ou de saisir « 001 » comme numéro d'identification. Cette façon de procéder (mise en colonne des données) est une condition qui s'applique à plusieurs logiciels d'analyse de données de type Rasch, puisque ces derniers doivent avoir des instructions claires sur 1) la nature des données et 2) l'organisation des données. Une fois les rangées et les colonnes bien alignées et bien définies, les données peuvent être importées dans ces logiciels d'analyse de type Rasch.

Comme nous l'avons déjà mentionné, RUMM2020 est celui qui requiert le moins de connaissances en langage de programmation. À l'ouverture du logiciel, l'usager est invité à créer un nouveau projet d'analyse. Ensuite, étape par étape, le logiciel l'amène à définir les colonnes qui permettent de regrouper les sujets, celles qui présentent leurs caractéristiques ainsi que leurs réponses aux items analysés. Si le fichier de données (.dat) a été bien préparé, le logiciel RUMM2020 est assez facile à utiliser. Une fois l'importation terminée, le projet ainsi créé et la définition de chacune des données importées sont, au moment de la sauvegarde, conservés dans un fichier de type base de données (.mdb) et sont accessibles par la suite à chaque ouverture du projet.

Bien qu'un peu moins intuitif, Winsteps est tout de même un logiciel assez simple à utiliser. Pour un débutant, l'interface *Setup procedure* permet à l'usager d'importer ses données sans avoir de connaissances en langage de programmation. À travers les champs, les boutons radio et les boutons de commande proposés, l'usager peut assez facilement repérer les colonnes de son fichier de données qui font référence aux sujets, aux caractéristiques et aux réponses à chacun des items. De plus, une partie de l'interface graphique ressemble à un tableur, ce qui permet à l'usager de voir immédiatement si les colonnes prévues pour chacune des variables sont correctes.

Winsteps permet également à un usager de niveau intermédiaire ou avancé de préparer son analyse moyennant quelques connaissances en langage de programmation. En effet, l'usager habitué de travailler avec Winsteps peut lui-même préparer les lignes de commandes dont le logiciel a besoin pour lire les données. À cet effet, il peut se servir des exemples de fichiers d'analyse proposés par Winsteps, copier les lignes de commandes dont il a besoin, les insérer dans son fichier de

données (.dat) et les ajuster en fonction de ses données. Il peut également utiliser le menu Édition (*Edit*) de Winsteps afin de travailler à partir d'un modèle générique (*template*) qui précise les différentes lignes de commandes que le logiciel permet d'utiliser. L'usager n'a ensuite qu'à compléter celles dont il a besoin. Une fois l'étape d'importation terminée, l'usager obtient, au moment de la sauvegarde, un fichier contrôle (toujours sous le format .dat), qui comprend les données ainsi que leur définition.

Après l'importation des données, notre analyse sur l'ergonomie et la convivialité des logiciels nous amène à traiter de l'analyse de données. En effet, une modélisation de type Rasch étant un processus itératif, cela implique bien souvent que l'usager doive explorer ses données et tenter plusieurs analyses. Selon le cas, certains sujets ou certains items présentant des patrons de réponses aberrants ou non conformes à la modélisation pourraient devoir être supprimés. De la même manière, pour des données polytomiques, certaines catégories de réponses pourraient nécessiter d'être regroupées. Nous pourrions ainsi passer de 5 catégories (0-1-2-3-4) à 4 catégories (0-1-2-3) en regroupant par exemple les catégories initiales (3 et 4) pour ne former qu'une catégorie (3). Enfin, l'usager pourrait souhaiter explorer différents modèles de type Rasch (*Rating Scale, Partial Credit*, ou le modèle à facettes par exemple) afin de voir lequel permet d'obtenir la meilleure adéquation modèle-données. En somme, il nous apparaissait important de montrer avec quelle facilité les logiciels étudiés ici permettent à l'usager de faire l'analyse de ses données.

5.1.3. *Exploration et analyses des données*

En ce qui a trait à l'exploration et aux analyses des données, les deux logiciels posent certaines difficultés. Tout d'abord, Winsteps requiert beaucoup de manipulations. Le fichier de contrôle (.dat) obtenu après l'importation des données dans le logiciel permet de produire un résultat unique : la modélisation de toutes les données à l'état brut. Si, après avoir analysé le résultat obtenu, l'usager décide de faire certains ajustements (supprimer des sujets ou des items, regrouper des catégories, etc.), il doit préparer un nouveau fichier d'analyse qui tient compte des nouvelles instructions et des changements proposés. En d'autres mots, il doit : 1) s'assurer d'avoir bien identifié et sauvegardé son fichier de contrôle initial (.dat) ; 2) faire une copie du fichier et lui donner un nouveau nom signifiant ; 3) ouvrir ce nouveau fichier de contrôle dans un éditeur de texte comme Notepad ou Wordpad pour apporter les ajustements souhaités ; 4) sauvegarder ces changements ; 5) procéder à une nouvelle analyse à l'aide de ce nouveau fichier de

contrôle. Si, encore une fois, l'usager souhaite apporter des ajustements afin d'en explorer les résultats, il doit reprendre les manipulations à partir de la deuxième étape. Puisqu'il s'agit d'analyses itératives, il n'est pas rare d'obtenir plusieurs dizaines de fichiers d'instruction différents, ce qui peut rendre le processus très lourd et entraîner des erreurs qui pourraient avoir des répercussions sur la validité des résultats. En effet, le nombre de fichiers d'analyse croît très rapidement et il devient alors très facile de se tromper. Pour illustrer ce propos, signalons que le moindre changement doit être consigné, par exemple le retrait d'un seul sujet, puisque tout changement, aussi mineur soit-il, peut modifier sensiblement les données obtenues. En revanche, il faut dire que cette façon de procéder offre la possibilité à l'usager de donner des noms signifiants à ces fichiers, ce qui peut l'aider à retracer ses différents essais. En outre, le fichier de contrôle de ce logiciel permet à l'utilisateur d'insérer des commentaires, c'est-à-dire de courtes phrases qui ne sont pas considérées comme des commandes par le logiciel. Ces commentaires peuvent également aider l'usager à décrire ses analyses de façon à s'y retrouver plus facilement.

Par ailleurs, RUMM2020 permet de gérer l'analyse des données sans jamais sortir du fichier-projet créé lors de l'importation des données. En effet, tout le travail effectué sur un même projet est conservé dans ce fichier. Cela facilite donc grandement la gestion des analyses par l'usager. Dans une perspective d'exploration des données, RUMM2020 se distingue nettement de son concurrent. Cependant, le défaut de ce logiciel est que le maximum de caractères alloué pour l'identification des analyses effectuées est limité à 10 ; il devient alors très difficile pour l'usager de donner un nom signifiant à chacune de ses explorations. Pour compenser, RUMM2020 offre aussi la possibilité de saisir un titre (une sorte de courte description de l'analyse), ce qui peut aider l'usager à bien identifier ses analyses. Mais, ici également, le nombre de caractères est limité. Dans certaines circonstances, cette limite de 50 caractères ne permet pas de distinguer facilement les analyses faites par l'usager.

L'analyse de données comporte un autre aspect qu'il est important de considérer : l'accès aux différentes analyses que l'usager peut effectuer. À cet égard, RUMM2020 propose encore une fois des conditions qui facilitent le traitement. En effet, toutes les options d'analyse offertes par le logiciel sont présentées sur un seul et même écran. Il suffit, à l'aide d'un bouton radio, de sélectionner l'analyse voulue, puis d'appuyer sur le bouton permettant d'exécuter la commande, et le résultat apparaît. Que l'usager souhaite obtenir les statistiques sommaires de son analyse, les données sur la qualité d'ajustement entre les données et le modèle ou la distribution des sujets (ou des items)

sur le continuum, il n'a qu'à sélectionner l'option de son choix et les résultats apparaissent instantanément. Une fois le résultat exploré, il est facile de revenir à l'écran présentant tous les choix et de demander à explorer une autre option.

Contrairement à RUMM2020 qui présente ses options sur un écran, Winsteps offre les siennes sous deux menus où une liste prédéterminée d'options apparaît. Sous le premier menu, Winsteps offre un accès rapide à environ une quarantaine de tableaux résultats (*Output Tables*) et, sous le second, il donne accès à une quinzaine de fichiers de résultats (*Output Files*). Les tableaux permettent à l'usager de cibler précisément l'analyse qui l'intéresse (ajustement pour les personnes, pour les items, graphique particulier pour l'analyse des catégories de réponses, etc.). De plus, si l'usager souhaite accéder à un tableau qui n'apparaît pas dans la liste prédéterminée (Winsteps offre environ une centaine de tableaux), une option lui permet d'en faire la requête et de l'obtenir facilement. Les fichiers de résultats permettent à l'usager d'exporter les résultats vers différents formats de fichiers (éditeur de texte, Excel, SPSS) afin de pouvoir les manipuler.

Pour terminer cette analyse sur l'ergonomie et la convivialité des logiciels, il nous reste un aspect important à examiner. En effet, pour aider l'usager à s'y retrouver plus facilement, chacun de ces logiciels fournit une aide complémentaire que l'usager peut consulter en tout temps. Nous allons donc prendre le temps de comparer les ressources offertes par chacun de ces logiciels.

5.1.4. *Ressources disponibles ou aide offerte aux usagers*

Les deux logiciels étudiés ici fournissent de l'aide en ligne à leurs utilisateurs. Étrangement, le logiciel qui jusqu'ici se démarquait de l'autre pour sa convivialité est aussi celui pour lequel l'aide est la moins abondante. En effet, l'aide offerte par le logiciel RUMM2020 se limite à une série de fichiers de documentation sous le format pdf. Bien que ces fichiers soient bien rédigés et que les saisies d'écran utilisées pour illustrer la démarche à suivre soient nombreuses, ce n'est pas toujours suffisant pour que l'usager, même expérimenté, puisse bien comprendre l'impact des options choisies pour chacune des étapes de l'analyse ni pour qu'il obtienne des réponses à ses questions.

À cet égard, Winsteps offre beaucoup plus. L'aide en ligne que les concepteurs offrent est très complète. Le logiciel contient un fichier d'aide compilé présentant un sommaire des différentes rubriques d'aide, un index des principaux termes de recherche, de même qu'un outil de recherche par mots clés. La navigation dans ces fichiers d'aide se fait facilement à l'aide de boutons permettant d'accéder à la page

précédente ou à la page suivante de la rubrique. Ce logiciel permet également d'accéder à toute une communauté d'utilisateurs par un site Web et un forum de discussion. Plusieurs cours (dont certains en ligne) sont aussi régulièrement offerts. Notons que RUMM2020 propose également des cours en ligne de façon régulière. Ces différentes options permettent donc plus facilement à l'usager de procéder à ses analyses et de trouver réponse à ses questions.

Maintenant que nous avons couvert la plupart des aspects plus qualitatifs de cette analyse, poursuivons avec les aspects plus quantitatifs. Dans les prochaines sections, nous traiterons des aspects techniques liés à l'utilisation des logiciels, comme les méthodes d'estimation et la précision de la mesure ainsi que les méthodes d'ajustement modèle-données.

5.2. Utilisation des logiciels : une comparaison quantitative

Dans cette deuxième partie de l'analyse et de la discussion, nous relevons les ressemblances et les différences de chacun des logiciels en examinant les aspects qui suivent : 1) les méthodes d'estimation et la précision de la mesure et 2) l'ajustement modèle-données.

5.2.1. *Les méthodes d'estimation et la précision de la mesure*

Afin d'estimer la mesure d'habileté des sujets, de même que le niveau de difficulté des items inclus dans l'analyse, les logiciels utilisent différentes méthodes. Winsteps, par exemple, utilise jusqu'à trois méthodes différentes (information tirée de l'aide en ligne de Winsteps (*Winsteps Help*) : la méthode du maximum de vraisemblance conjointe (*Joint maximum likelihood estimation*, JMLE), développée par Wright et Panchapakesan (1969), aussi connue sous le nom de méthode du maximum de vraisemblance inconditionnelle (*Unconditional maximum likelihood estimation*, UCON), la méthode de l'approximation normale (*Normal approximation algorithm*, PROX), développée par Cohen (1979), et la méthode du maximum de vraisemblance exclusive (*Exclusory maximum likelihood estimation*, XMLE), anciennement connue sous le nom d'algorithme extraconditionnel (*Extra-conditional* [*XCON*] *algorithm)* de Linacre (1989). Dans un premier temps, l'estimation de tous les paramètres qui ne sont pas ancrés est fixée à zéro. Ensuite, la méthode PROX est utilisée afin d'obtenir une première estimation pour chacun des paramètres. Ces dernières estimations servent alors de données initiales pour la méthode JMLE. Les estimations obtenues à l'aide de cette méthode constituent les résultats finaux fournis par le logiciel.

Dans la méthode du JMLE, les estimations de la matrice de données pour les sujets, pour les items et pour la structure de l'échelle de réponse (s'il y a lieu) sont obtenues simultanément. Un des problèmes liés à cette méthode d'estimation est qu'elle peut présenter certains biais dans l'estimation des paramètres lorsque l'échantillon de données est petit. Ce biais excède rarement la précision de la mesure, mais, lorsqu'un chercheur souhaite réaliser des inférences probabilistes exactes, il devient important de s'en soucier. Pour cette raison, Winsteps utilise également la méthode XMLE. Avant de présenter cette méthode, voyons un peu en quoi consiste la méthode PROX.

L'algorithme de la méthode PROX met l'accent sur la ressemblance entre les courbes logistiques et normales. La modélisation s'effectue donc comme si la distribution des sujets et des items était normale.

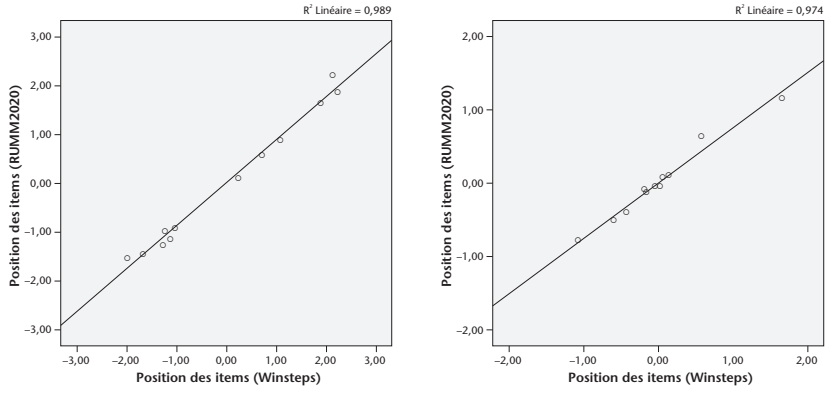
Pour sa part, la méthode XMLE est assez similaire à la méthode JMLE. La principale différence réside dans la valeur de la probabilité qui est utilisée dans les équations d'estimation. En effet, la méthode JMLE permet la présence de scores extrêmes dans les données, mais ne peut pas en estimer les paramètres. L'algorithme XMLE corrige ce problème en ne permettant pas, pour une première estimation, des vecteurs de réponses extrêmes (tout bon ou tout faux).

De son côté, RUMM2020 utilise la méthode d'estimation conditionnelle pairée (selon l'information tirée de la documentation de RUMM2020 [rmIntPoly.pdf], *Pairwise conditional estimation procedure*), une généralisation de l'équation développée par Zwinderman (1995) pour une seule paire d'items à toutes les paires d'items prises simultanément. Cette méthode d'estimation utilise une des propriétés des modèles de Rasch, la suffisance des paramètres, qui lui permet d'éliminer les paramètres des sujets lors de l'estimation des paramètres d'items. Sa principale caractéristique est le fait que l'estimation des seuils de chacune des catégories de réponse d'un item est calculée en fonction des fréquences obtenues dans toutes les catégories de réponses et non en fonction de la fréquence obtenue dans la catégorie correspondant au seuil estimé. Cela permet de renforcer la stabilité et la robustesse des estimations produites, et plus particulièrement lorsque la fréquence d'une catégorie est faible. Ainsi, la *Pairwise conditional estimation procedure* commence par estimer les paramètres d'items en les analysant deux à deux. À partir de ces estimations, les seuils entre chacune des catégories de réponse peuvent être estimés facilement. Une fois ces estimations obtenues, les paramètres pour les personnes peuvent être évalués. La méthode de la *Pairwise conditional estimation procedure* est généralement une méthode plus spécifique de l'analyse d'items dichotomique, mais elle se généralise aisément aux cas où le

nombre de catégories de réponses diffère d'un item à l'autre. De plus, cette méthode permet de prendre facilement en compte les données manquantes. En revanche, l'étendue des paramètres d'items produits par cette méthode pourrait ne pas être aussi grande que lorsqu'un algorithme basé sur une méthode inconditionnelle ou conjointe (*unconditional or joint method*) est utilisé.

Bien que les méthodes utilisées par chacun de ces logiciels soient différentes, il reste que les simulations réalisées avec nos ensembles de données produisent des résultats fort comparables. Rappelons les limites de nos comparaisons, puisque les analyses ont été faites avec seulement deux ensembles de données, ce qui réduit les généralisations possibles. Mais, quoi qu'il en soit, dans un souci d'authenticité, la comparaison des ensembles de données réelles peut donner une bonne idée des résultats qu'il est possible d'obtenir avec les deux logiciels dans des conditions habituelles de recherche. En outre, nous n'avons pas mené formellement des analyses complètes en fournissant des renseignements sur toutes les dimensions traitées (dimensionnalité, fonctionnement différentiel d'item, etc.). Le lecteur intéressé par un exemple de procédure permettant de mener à bien une analyse complète de Rasch peut se référer à l'article de Tennant et Conaghan (2007).

La figure 2.2 A présente un diagramme de dispersion qui met en relation la position des items (n = 12) pour l'ensemble de données dichotomiques des sujets (n = 184) modélisés avec Winsteps et RUMM2020. On peut constater que la position relative de l'indice de difficulté est similaire pour les modélisations faites avec les deux logiciels. Avec les données que nous avons utilisées, les différences observées sont minimes. En effet, la corrélation (r = 0,99) entre les deux ensembles de données est très forte. De plus, il est improbable que les inférences réalisées à la lumière de ces résultats permettent d'arriver à des conclusions différentes si l'on utilise l'un ou l'autre des logiciels.



A) B)

Figure 2.2.
Correspondance entre la position des items pour l'ensemble des items dichotomiques (A) et polytomiques (B) pour Winsteps et RUMM2020

Quant à la figure 2.2 B, elle présente les résultats de l'estimation de la position pour les items polytomiques en mettant en correspondance les résultats des deux logiciels. À nouveau, on ne remarque pas de différences majeures (r = 0,97) pour l'estimation de la difficulté. La différence la plus notable concerne seulement un item (item 7) qui s'éloigne de la droite d'ajustement. Winsteps situe cet item relativement plus à droite sur l'échelle (item plus difficile). Le diagramme de dispersion semble indiquer qu'il n'existe pas de différences importantes dans l'estimation de la difficulté lorsqu'on utilise les mêmes données brutes sans éliminer les scores des sujets ou des items. En ce sens, pour notre échantillon, les méthodes d'estimation respectives des logiciels donnent des résultats comparables. Bien que les résultats ne soient pas présentés ici, il est à noter que des indices de corrélation semblables s'appliquent quand on compare la position des sujets avec les deux logiciels.

5.2.2. *L'ajustement modèle-données*

Le tableau 2.2 rapporte les valeurs des statistiques d'ajustement des items pour les deux ensembles de données (dichotomiques et polytomiques). Il faut d'abord noter que Winsteps fournit plusieurs statistiques d'ajustement, dont celles identifiées par *Infit* standardisé et *Outfit* standardisé. Le logiciel offre également des statistiques *Infit* et *Outfit* présentées sous la forme d'un carré moyen (*mean square*) qui ne sont pas relevées ici. Contrairement à la statistique *Infit*, la statistique *Outfit* est sensible aux écarts importants (*outliers*) entre la valeur

prédite et la valeur réelle. La valeur standardisée provient d'un test *t* présenté sous la forme d'une distribution centrée réduite (*Z*). La valeur souhaitée devrait donc se rapprocher de la valeur 0. Toute valeur qui s'éloigne de 0 indique une difficulté d'adéquation entre les données et le modèle. En règle générale, des valeurs comprises dans l'intervalle –2 à +2 apparaissent fournir une adéquation acceptable. Des valeurs inférieures à –2 ou supérieures à +2 dénotent, quant à elles, des difficultés majeures qu'il faudrait investiguer avec attention (Linacre, 2002). En ce qui concerne RUMM2020, la statistique d'ajustement se présente sous la forme d'une valeur résiduelle qui, selon Tennant et Conaghan (2007, p. 1360), provient de la somme standardisée des différences entre les valeurs prédites et les valeurs observées, calculée sur l'ensemble des sujets (ou des items). Selon les mêmes auteurs, la statistique résiduelle serait comparable à la statistique *Outfit* standardisée proposée par Winsteps. Le logiciel offre aussi une statistique basée sur un chi carré. Une valeur résiduelle inférieure ou supérieure à 2,50 et dont la probabilité de la statistique du chi carré est inférieure à 0,05 serait un bon indicateur de problèmes d'ajustement (Covic, Pallant, Conaghan et Tennant, 2007).

Tableau 2.2
Statistiques d'ajustement pour les items de l'ensemble des données dichotomiques et polytomiques fournies par RUMM2020 et Winsteps

Items	Données dichotomiques (n = 184)			Données polytomiques (n = 189)		
	Winsteps		RUMM 2020	Winsteps		RUMM 2020
	Infit(s)	*Outfit(s)*	Résidus	*Infit(s)*	*Outfit(s)*	Résidus
1	3,48	4,48	3,10*	5,30	5,21	3,50
2	1,86	1,54	1,34	**3,51**	**2,92**	**2,71**
3	0,95	0,37	0,31	4,10	3,59	3,23
4	1,29	0,45	0,32	**4,12**	**3,36**	**2,89**
5	1,35	1,58	1,02	0,77	1,10	1,57
6	–1,44	–0,25	–0,95	**–4,03**	**–3,80**	–2,15
7	1,13	0,47	–0,25	**7,64**	**2,84**	2,26
8	**–2,03**	–1,84	–2,03	**–3,31**	**–2,41**	–1,37
9	–2,54	–2,47	–2,96	–6,87	–5,40	–4,32
10	0,23	–0,22	–0,81	**–4,86**	**–3,58**	**–2,81**
11	**–3,26**	–1,77	–1,69	**–5,18**	**–3,40**	**–3,04**
12	**–3,83**	**–2,28**	–2,15	**–3,86**	**–3,19**	–2,26

* Les valeurs résiduelles en caractères gras sont statistiquement significatives ($p \leq 0,05$).

En ce qui concerne les statistiques d'ajustement, il est difficile de comparer ce qui est offert par les deux logiciels. En effet, le calcul de ces statistiques est fort différent et donne évidemment des résultats différents. Considérant ce fait, nous avons choisi de comparer les statistiques d'ajustement en examinant les items, pour chacun des logiciels, qui montreraient des problèmes à cet égard. Le tableau 2.2 présente les différentes statistiques d'ajustement pour les deux ensembles de données. Les valeurs en gras indiquent les items qui montrent des problèmes d'ajustement. Pour l'ensemble de données dichotomiques, on remarque que les deux logiciels identifient clairement des difficultés pour les items 1 et 9. La statistique *Infit* de Winsteps indique que nous devrions également porter une attention particulière aux items 8, 11 et 12. Pour cet ensemble de données, on remarque que les indices liés aux problèmes d'ajustement donnent une information relativement similaire pour les deux logiciels. Pour ce qui est de l'ensemble de données polytomiques, on note que les statistiques *Infit* et *Outfit* de Winsteps relèvent de sérieux problèmes pour la quasi-totalité des items (11/12). Pour sa part, le logiciel RUMM2020 met en lumière sept items pour lesquels des problèmes d'ajustement sont visibles. Dans cet exemple, on voit qu'un utilisateur de l'un ou l'autre des logiciels serait alerté à peu près de la même façon concernant les problèmes d'ajustement même si RUMM2020 indique moins de problèmes.

Le tableau 2.2 présente des résultats différents entre les items dichotomiques et polytomiques. Les méthodes de calcul utilisées pour l'estimation des paramètres sont susceptibles d'être ici en cause. En effet, la principale caractéristique de la méthode utilisée par RUMM réside dans le fait que l'estimation des seuils de chacune des catégories de réponse d'un item est calculée en fonction des fréquences obtenues dans toutes les catégories de réponses et non en fonction de la fréquence obtenue dans la catégorie correspondant au seuil estimé, comme le fait Winsteps. Cela aurait pour conséquence de renforcer la stabilité et la robustesse des estimations produites, et plus particulièrement lorsque la fréquence d'une catégorie est faible. Ainsi, la différence entre les deux méthodes serait moindre dans le cas des items dichotomiques (on note d'ailleurs que ce sont sensiblement les mêmes items qui sont détectés), mais plus importante dans le cas des items polytomiques où les différences sont effectivement plus grandes et où la robustesse de RUMM se fait peut-être davantage sentir dans le fait qu'il détecte moins d'items qui ont des problèmes d'ajustement.

6. CONCLUSION

Les avancées technologiques permettent de s'intéresser de plus en plus aux résultats de la modélisation de type Rasch. En effet, sans le recours à des logiciels, il serait difficile, long et contraignant d'effectuer autant d'opérations de calcul. La disponibilité des logiciels permet également à davantage de chercheurs de s'intéresser à la prise en compte de la modélisation des scores. Ce chapitre visait à rendre compte d'une étude comparative de deux logiciels permettant de réaliser des analyses de Rasch. Comme nous avons pu le constater, chacun possède des avantages et des inconvénients, mais, ultimement, il y a peu de différences dans la nature des inférences réalisées à partir des analyses lorsqu'on utilise l'un ou l'autre de ces outils. Cependant, nous pensons que l'utilisation de ces logiciels mériterait d'être étudiée davantage, puisque les analyses et les inférences sont directement tributaires de la qualité des données issues de ces logiciels.

Les résultats présentés dans ce texte n'exposent qu'une fraction de ce qu'il faudrait réaliser comme analyse afin de bien mesurer la portée de ces outils. À titre d'exemple, il serait pertinent d'observer les résultats obtenus avec différentes distributions de scores bruts (distribution normale, asymétrique, bimodale, etc.). La question de l'ajustement modèle-données est une autre dimension qui mériterait d'être approfondie, surtout dans le contexte de faibles échantillons, fréquent en éducation.

RÉFÉRENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Andrich, D., Sheridan, B. et Luo, G. (2004). *RUMM2020 : A windows program for the Rasch unidimensional measurement model* [Logiciel]. Perth, Australie : RUMM Laboratory.

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.

Blais, J. et Maheux, P. (2003). La contribution du Rating Scale Model de la famille des modèles de Rasch à la mise au point d'une échelle de mesure de l'opinion. Dans J.-G. Blais et G. Raîche (dir.), *Regards sur la modélisation de la mesure en éducation et en sciences sociales*. Québec, Québec : Les Presses de l'Université Laval.

Boone, W. J. et Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science education*, *90*(2), 253-269.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *The British journal of mathematical and statistical psychology*, *32*, 113-120.

Covic, T., Pallant, J. F., Conaghan, P. G. et Tennant, A. (2007). A longitudinal evaluation of the Center for epidemiologic studies-depression scale (CES-D) in a rheumatoid arthritis population using Rasch analysis. *Health and quality of life outcomes*, *5*(41).

Dionne, E. (2008). *Expérimentation d'un modèle d'évaluation permettant de juger du développement d'une compétence d'investigation scientifique en laboratoire*. Thèse de doctorat inédite. Montréal, Québec : Université de Montréal.

Hagquist, C., Bruce, M. et Gustavsson, J. P. (2009). Using the Rasch model in nursing research : an introduction and illustrative example. *International journal of nursing studies*, *46*(3), 380-393.

Jackson, T. R., Draugalis, J. R., Slack, M. K. et Zachry, W. M. (2002). Validation of authentic performance assessment : a process suited for Rasch modeling. *American journal of pharmaceutical education*, *66*, 233-243.

Lamoureux, E. L., Pallant, J. F., Pesudovs, K., Hassell, J. B. et Keeffe, J. E. (2006). The impact of vision impairment questionnaire : an evaluation of its measurement properties using Rasch analysis. *Investigative ophthalmology and visual science*, *47*(11), 4732-4741.

Linacre, J. M. (1989). Extra-conditional (XCON) algorithm [now XMLE : Exclusory Maximum Likelihood Estimation]. *Rasch measurement transactions*, *3*(1), 47-48.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean ? *Rasch measurement transactions*, *16*(2), 878.

Linacre, J. M. (2010). *WINSTEPS (Version 3.69.1.7)* [Logiciel]. Chicago, Illinois : Winsteps.

Liu, X. et Boone, W. J. (2006). *Applications of Rasch measurement in science education*. Maple Grove, Minnesota : JAM Press.

Loye, N. (2005). Quelques nouveaux modèles de mesure. *Mesure et évaluation en éducation*, *28*(3), 51-68.

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

Muller, S. et Roddy, E. (2009). A Rasch analysis of the Manchester foot pain and disability index. *Journal of foot and ankle research*, *2*(1), 29.

Pallant, J. F. et Tennant, A. (2007). An introduction to the Rasch measurement model : an example using the hospital anxiety and depression scale (HADS). *British journal of clinical psychology*, *46*, 1-18.

Pomplund, M., Omar, H. et Custer, M. (2004). A comparison of Winsteps and Bilog-MG for vertical scaling with the Rasch model. *Educational and psychological measurement*, *64*(4), 600-616.

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

Robin, F., Xing, D. et Hambleton, R. K. (1999). Software review : Rasch scaling program (RSP). *Applied psychological measurement*, *23*, 90-94.

Sick, J. (2009). Rasch analysis software programs. *JALT Testing and evaluation SIG newsletter*, *13*(3).

Skaggs, G. (2004). Software use in psychometric research. *Educational measurement, issues and practice*, *23*(1), 28-33.

Smith, A. B., Wright, P., Selby, P. J. et Velikova, G. (2007). A Rasch and factor analysis of the functional assessment of cancer therapy-general (FACT-G). *Health and quality of life outcomes*, *5*, 19.

Tang Wai, K., Wong, E., Chiu, H., Lum, C. M. et Ungvari, G. S. (2005). The geriatric depression scale should be shortened : results of Rasch analysis. *International journal of geriatric psychiatry, 20,* 783-789.

Tennant, A. et Conaghan, P. G. (2007). The Rasch measurement model in rhumatology : what is it and why use it ? When should it be applied, and what should one look for in a Rasch paper ? *Arthritis and rheumatism*, *57*(8), 1358-1362.

Tennant, A. et Pallant, J. F. (2006). Unidimensionality matters ! (A tale of two Smiths ?). *Rasch measurement transactions*, *20*(1), 1048-1051.

Wright, B. et Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and psychological measurement, 29,* 23-48.

Zwinderman, A. H. (1995). Pairwise estimation in the Rasch models. *Applied psychological measurement*, *19*(4), 369-375.

Chapitre 3

Mesure de l'aptitude physique générale lors des épreuves de sélection pour les études supérieures en éducation physique et sport au Maroc et en Algérie

Jaouad Alem, Marc Cloes, Michel Guay et Nabil Kerfes

Cette recherche analyse la validité de construit de plusieurs épreuves physiques censées mesurer un facteur unique d'aptitude physique générale des étudiants se destinant à une formation supérieure en éducation physique et sport. L'échantillon est composé de 1481 candidats masculins âgés en moyenne de 20 ans dans deux formations supérieures au Maroc et en Algérie. Les analyses factorielles en composantes principales avec rotation varimax des performances aux épreuves physiques révèlent plutôt une solution en deux composantes qui se distinguent selon la durée du travail pour produire de l'énergie : la puissance musculaire glycolytique et la puissance musculaire phosphagénique. La première composante correspond à la capacité de produire du lactate en plus de 12 secondes ; elle est définie par la course de vitesse et la course de résistance. La deuxième composante correspond à la capacité de produire en moins de 7 secondes du phosphate déjà présent dans les muscles ; elle est définie par les autres épreuves physiques.

1. INTRODUCTION

Dans l'entraînement sportif de haut niveau, les intervenants sont souvent confrontés à des questions comme la détection des talents, l'identification des critères pertinents d'évaluation de l'aptitude physique générale des sportifs et la différenciation entre athlètes élites et amateurs en termes d'aptitude physique. Dès lors, on peut se demander quels sont les tests qui permettent de répondre efficacement à ces questions et comment évaluer la pertinence de ces tests.

De nombreuses publications s'intéressent à l'évaluation des caractéristiques physiques et musculaires des sportifs. En handball, par exemple, Gorostiaga et Granados (2004) ont montré que des niveaux élevés de force et de puissance musculaire, ainsi qu'une capacité aérobique élevée, constituaient des facteurs importants de la performance. Ces auteurs ont montré que l'explosivité, l'endurance et la vitesse de course ne permettent pas de différencier des handballeurs d'un niveau d'expertise mondiale et d'autres d'un niveau de deuxième division espagnole. En revanche, l'étude de Dufour et Pontier (1989) a révélé que les handballeurs élites étaient plus grands que les joueurs de niveaux inférieurs. Maso et Cazorla (2001) ont, pour leur part, établi le profil physiologique des joueurs de rugby.

Certaines qualités physiques peuvent donc constituer des critères pertinents pour différencier des joueurs de niveaux différents en termes d'aptitude physique générale. L'identification de ces qualités physiques devient alors essentielle pour la sélection des joueurs qui présentent un profil proche de celui du joueur d'élite, c'est-à-dire ceux qui seront le plus à même de répondre de manière optimale aux sollicitations imposées au cours d'une compétition.

Lors des concours annuels d'accès organisés par l'Institut national des sports du Maroc ou à l'Institut d'éducation physique et sportive de l'Université d'Alger, les candidats sont sélectionnés au moyen de tests censés les classer selon leur degré de prédisposition à s'engager dans des formations supérieures en sport. Ces formations supérieures ont pour objectif de former des éducateurs sportifs compétents pour intervenir dans le domaine de l'éducation physique et du sport. Aussi bien au Maroc qu'en Algérie, c'est sur l'aptitude physique générale que se focalisent le plus les tests de sélection. En effet, les épreuves physiques sont notées sur 40 points, alors que les trois autres épreuves (oral, écrit et spécialité) sont notées chacune sur 20 points. Les épreuves orales et écrites visent à évaluer les habiletés communicationnelles orales puis écrites des candidats, alors que le test de spécialité vise à évaluer les habiletés technico-tactiques des candidats.

La question que nous nous posons porte sur les qualités métriques de la mesure de ce construit que l'on nomme « aptitude physique générale » dans les concours marocains et algériens. Il s'agit du thème principal de cette recherche.

2. CONTEXTE THÉORIQUE

Les tests de sélection des candidats aux études en formation supérieure en sport sont censés les classer selon un profil de compétences préalablement défini par les établissements de formation.

Alem (2003), Alem, El Mezdi, Dadouchi, Kpazai et Bendefa (2008) et Alem, Hamdane, Mawfik et El Mezdi (2005) se sont demandé si ces tests de sélection tiennent compte des compétences exigées par le métier de professeur en éducation physique et sportive.

Goodlad (1990), Guyton et Farokhi (1987), Haberman (1987) ainsi que Shechtman et Godfried (1993) ont démontré que le dossier scolaire était un faible prédicteur du succès en enseignement et que, malgré le fait que les habiletés interpersonnelles, les habiletés de communication verbale et le sens du leadership constituent des facteurs plus importants pour prédire le succès en enseignement, les comités d'admission en tenaient rarement compte lorsqu'ils sélectionnaient les candidats aux études en formation à l'enseignement.

Alem (2003 ; Alem, Dadouchi *et al.*, 2009 ; Alem, Taibi et Guay, 2005) a analysé huit concours d'accès à des formations supérieures en sport au Maroc ; l'échantillon était composé de 995 candidats masculins ayant terminé leurs études préuniversitaires. Il rapporte que ces tests prédisent peu ou pas du tout le rendement dans les études, que certains des tests de sélection sont redondants et que les barèmes utilisés pour transformer les performances aux épreuves physiques doivent être réactualisés.

Thomas, Eclache et Keller (1989) notent que, comme pour les recherches en sciences de l'éducation portant sur l'existence d'un seul facteur d'intelligence générale *G*, les chercheurs en kinésiologie ont poursuivi des travaux dans le but de mettre en évidence l'existence d'une seule aptitude motrice générale. Guilford (1958) avait montré que l'idée d'un seul facteur général devrait être abandonnée. Pourtant, la théorie d'une aptitude motrice générale est réapparue dans quelques textes (Brace, 1927 ; McCloy, 1934) qui décrivent un schéma d'organisation des capacités motrices dominé par une habileté supérieure (*superability*), un genre de QI moteur ou faculté générale d'apprendre

(*motor educability*), ou encore le mythe d'une intelligence motrice générale appelée éducabilité motrice, basée sur la croyance qu'il existe une habileté unique qui détermine tous les comportements moteurs.

Drowatzky et Zuccato (1967) ont démontré que cette croyance était non fondée en examinant les corrélations entre des performances à six tests d'équilibre différents et en montrant que ces corrélations étaient faibles (entre –0,19 et 0,31). Fleishman et Parker (1962) et Lotter (1960) ont confirmé par la suite les conclusions de ces recherches.

Alors que la théorie de spécificité d'Henry (1961, 1968 : voir Aboussaïd, 1982) postule que la corrélation entre les performances à deux habiletés différentes est faible, voire nulle, d'autres chercheurs démontrent qu'il pourrait exister des habiletés communes à des tâches différentes en utilisant une technique statistique appelée l'analyse factorielle (Fleishman, 1964, 1965 ; Fleishman et Bartlett, 1969). Ainsi, Fleishman (1964) identifie deux grandes catégories d'habiletés qu'il nomme *perceptual-motor abilities* et *physical proficiency abilities*. Ces habiletés sont traduites par Schmidt (1993) par les habiletés manipulatives et les aptitudes physiques.

Les aptitudes physiques peuvent se distinguer selon la durée du travail pour produire de l'énergie : la puissance musculaire glycolytique et la puissance musculaire phosphagénique. La première composante correspond à la capacité de produire du lactate en plus de 12 secondes et la seconde, à la capacité de produire en moins de 7 secondes du phosphate déjà présent dans les muscles.

À la lecture de cette brève revue de littérature, il apparaît donc opportun de vérifier la validité de construit de la mesure de l'aptitude physique générale. Cette recherche se propose d'analyser la validité de construit des tests de sélection utilisés pour évaluer l'aptitude physique générale des candidats aux études dans les deux formations supérieures en sport que nous avons présentées précédemment. Nous allons vérifier en particulier si l'aptitude physique générale est unidimensionnelle ou bidimensionnelle.

3. MÉTHODOLOGIE

3.1. Sujets

Cette recherche s'est déroulée en prenant pour cible 1481 candidats masculins âgés en moyenne de 20 ans dans deux formations supérieures au Maroc ($n_1 = 990$) et en Algérie ($n_2 = 491$).

Au Maroc, le fichier de données est constitué de huit cohortes de candidats qui se sont présentés au concours national d'accès organisé chaque année dans un centre de formation d'entraîneurs sportifs entre 1976 et 1996.

En Algérie, le fichier de données du concours d'accès à l'Institut d'éducation physique de l'Université d'Alger en Algérie, créé en 1982, est constitué de la cohorte des candidats qui se sont présentés au concours national pour l'année universitaire 2007-2008.

3.2. Instrumentation

Au Maroc, par exemple, l'aptitude physique générale est mesurée par sept épreuves physiques : la course de vitesse (80 mètres hommes, 80 mètres femmes), la course de demi-fond (800 mètres hommes, 600 mètres femmes), le lancer du poids (5 kilogrammes hommes, 3 kilogrammes femmes), le pentabond (5 sauts consécutifs), le saut en hauteur (la détente verticale telle que mesurée par le *Sergent test*), la natation (20 mètres nage libre) et l'enchaînement de gymnastique au sol selon le programme officiel du baccalauréat. Mis à part l'épreuve de gymnastique qui est évaluée au moyen d'une table de cotation, les autres épreuves sont notées à l'aide de barèmes qui s'inspirent des tables de cotation de Letessier (1957) et de la connaissance intuitive des professeurs des niveaux extrêmes et moyens exigibles selon Aboussaïd (1982) et Nahari (1985). Selon Filliard (1995), un barème de notation correctement établi permet de transformer toutes les performances en notes standards et de réaliser un profil normalisé de l'athlète évalué.

En Algérie, l'aptitude physique générale des candidats masculins est mesurée par quatre épreuves physiques : la course de vitesse (100 mètres hommes, 80 mètres femmes), la course de demi-fond (800 mètres hommes, 600 mètres femmes), le lancer du poids (5 kilogrammes hommes, 3 kilogrammes femmes) et le saut en longueur sans élan (performance la plus élevée à trois essais).

La performance brute aux épreuves de lancer et de saut en longueur est celle qui est la meilleure à trois essais ; elle est par la suite transformée en une note sur 20 à l'aide d'un barème préétabli.

3.3. Données

Les informations recueillies dans les deux pays ont été encodées dans deux fichiers de données et compilées dans un fichier SPSS. Chaque épreuve est un item ; il y a donc sept items au Maroc et quatre items en Algérie. Les données ont été analysées avec le logiciel SPSS version 17.0.

Comme nous avions démontré auparavant que les barèmes de notation pour transformer les performances brutes en notes sur 20 points étaient caducs (Alem, Dadouchi *et al.*, 2009 ; Alem, Taibi et Guay, 2005), nous n'avons considéré que les performances brutes pour effectuer les analyses.

3.4. Les deux qualités métriques étudiées et les techniques statistiques utilisées

Nous avons analysé deux qualités métriques de la mesure de l'aptitude physique générale :

1. La consistance interne des items.

   Les mesures des performances aux épreuves physiques étant objectives, il est peu probable qu'il y ait eu des erreurs provenant des juges. Nous avons estimé la consistance interne des items. Cette qualité métrique se définit comme le degré de constance qu'offrent les réponses des individus aux items variés d'un instrument de mesure. L'intérêt de la consistance interne porte sur l'analyse de la relation entre les items et le construit. La consistance interne a été analysée par le coefficient alpha de Cronbach. Bien qu'il existe plusieurs autres façons de calculer un indice de fidélité, cet indice est le plus populaire selon Crocker et Algina (1986).

2. La validité de construit ou la validité hypothéticodéductive.

   La validité de construit consiste à démontrer que l'instrument est capable de reconnaître des différences (la validité de différenciation) et des ressemblances (validité de convergence). Un construit est un concept délibérément inventé par un chercheur dans un but précis. Pour être en mesure de mesurer un construit, il faut utiliser des définitions opérationnelles ou encore des définitions qui spécifient les indicateurs témoignant de ce construit. Pour démontrer la validité de construit de la mesure de l'aptitude physique générale, nous avons utilisé la technique des analyses factorielles confirmatoires avec rotation varimax. Il s'agit d'une méthode qui regroupe empiriquement les items d'un instrument en une ou plusieurs dimensions indépendantes aussi appelées facteurs. Cette méthode statistique permet de confirmer un regroupement préalable d'items comme les différentes épreuves physiques qui sont à l'étude.

Ainsi, on devrait s'attendre à ce que la structure factorielle observée soit consistante avec la structure théorique. Puisque la structure théorique est un facteur unique d'aptitude physique générale, la structure factorielle devrait révéler un facteur ou encore une dimension unique. C'est précisément ce que nous tentons de vérifier.

4. RÉSULTATS

4.1. Analyse de fidélité : la consistance interne des items

Le tableau 3.1 présente le degré d'interrelation entre les notes aux épreuves physiques selon les deux bases de données. Nous constatons que, dans le cas des données algériennes, les épreuves physiques sont plus interreliées (α de Cronbach = 0,73 contre α = 0,69). Comme le soulignent Alem, Taibi et Guay (2005) ainsi qu'Alem, Dadouchi, Kerfes et Cloes (2009), cela est dû en partie au fait que les épreuves physiques et les barèmes utilisés au Maroc pour transformer les performances en notes sur 20 points doivent être actualisés.

Tableau 3.1.
Le degré de consistance interne entre les épreuves physiques programmées au Maroc et en Algérie

	α de Cronbach si l'item est supprimé	
	Maroc (n_1 = 990) 0,69 (7 items)	Algérie (n_2 = 491) 0,73 (4 items)
Gymnastique	0,68*	
Natation	0,69	
Course de vitesse	0,60	0,62
Pentabond	0,63	
Détente verticale	0,64	0,67
Course de résistance	0,68	0,71
Lancer du poids	0,65	0,68

* Les performances brutes aux items ont été transformées en des notes sur 20 points par des barèmes préétablis.

4.2. Analyse de validité : la validité de construit

Les graphiques des éboulis des analyses factorielles exploratoires avec rotation varimax des épreuves physiques révèlent des solutions factorielles en deux composantes distinctes et non pas un facteur unique qui mesure l'aptitude physique générale.

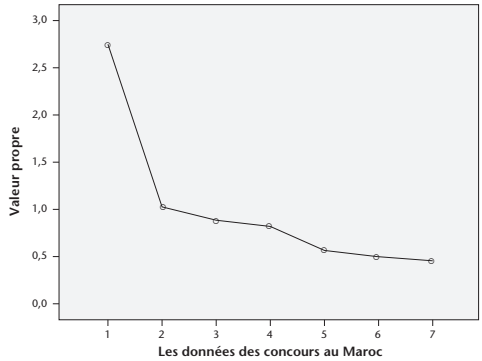


Figure 3.1.
Les graphiques des éboulis des deux bases de données

Presque 54,00 % de la variance totale est expliquée par ces deux composantes pour les données du Maroc et 76,40 % pour les données de l'Algérie. Le tableau 3.2 présente les deux composantes distinctes qui émergent des analyses factorielles confirmatoires.

Tableau 3.2.
Les deux composantes distinctes révélées par les analyses factorielles exploratoires puis confirmatoires avec rotation varimax

Les épreuves physiques	Composantes Maroc		Composantes Algérie	
	1	2	1	2
Lancer du poids	0,69		0,86	
Saut vertical sans élan	0,49	0,54	0,82	
Pentabond (ou saut horizontal sans élan)	0,37	0,67		
Natation (25 m nage libre)	0,70			
Enchaînement imposé en gymnastique	0,55			
Course de résistance		0,80		0,93
Course de vitesse		0,80	0,48	0,69

Les analyses factorielles en composantes principales des performances aux épreuves physiques ne révèlent pas une solution factorielle en une composante, mais plutôt une solution en deux composantes. Après analyse des composantes émergentes, en particulier des saturations les plus élevées dans le tableau 3.2, il apparaît que les deux

composantes se distinguent selon la durée du travail pour produire de l'énergie : la puissance musculaire glycolytique et la puissance musculaire phosphagénique.

La première composante correspond à la capacité de produire du lactate en plus de 12 secondes ; elle est définie par la course de vitesse et la course de résistance. La deuxième composante correspond à la capacité de produire en moins de 7 secondes du phosphate déjà présent dans les muscles ; elle est définie par les autres épreuves physiques.

La base de données algérienne comportait une variable *expérience sportive* qui catégorisait les candidats algériens en trois classes : les sujets qui ont une expérience sportive limitée, intermédiaire ou élevée. Nous avons donc effectué des analyses de variance (ANOVA) pour vérifier si le concours algérien classait bien les candidats en trois niveaux distincts pour chacune des quatre épreuves physiques et pour la moyenne aux quatre épreuves physiques (notée sur 20 points). Cette analyse permet d'explorer la validité de différenciation de la mesure de l'aptitude physique générale.

Le tableau 3.3 présente les moyennes obtenues aux quatre épreuves physiques selon les trois niveaux d'expérience sportive pour les données algériennes. Remarquons que les différences sont systématiquement significatives au seuil de 1 % aussi bien pour chacune des épreuves physiques que pour la moyenne de celles-ci (F Brown-Forsythe $(2 ; 488) = 47{,}51$, $p < 0{,}05$).

Le concours algérien est donc bien en mesure de reconnaître des groupes de niveaux différents et la validité de différenciation est donc démontrée.

5. DISCUSSION DES RÉSULTATS

Deux constats ressortent des analyses :

1. Les coefficients α de Cronbach obtenus sont plutôt moyens pour les items du Maroc, mais ils sont acceptables pour les données algériennes.
2. Par ailleurs, les analyses factorielles avec rotation varimax ne révèlent pas un facteur unique d'aptitude physique générale, mais plutôt deux facteurs indépendants, et ce, aussi bien pour la base de données du Maroc que pour celle d'Algérie. Après analyse des items qui saturent sur chacune des composantes, il apparaît que la première composante factorielle correspond à la puissance musculaire glycolytique ou encore à la capacité de produire du lactate en plus de 12 secondes. Elle est définie

Tableau 3.3.
Les moyennes à chacune des quatre épreuves physiques et à l'ensemble des épreuves physiques selon les trois niveaux d'expérience sportive des candidats masculins algériens

		n	Moyenne	Écart type	Valeur du test *F* pour comparer les moyennes
Performance en course de demi-fond (800 m garçons, 600 m filles)	Expérience sportive limitée	243	2,59	0,32	*F* Brown-Forsythe[1] (2, 488) = 13,53*
	Expérience sportive intermédiaire	200	2,54	0,29	
	Haut niveau d'expérience sportive	47	2,35	0,19	
	Total	490	2,55	0,31	
Performance en course de vitesse (100 m garçons, 80 m filles)	Expérience sportive limitée	244	13,83	0,89	*F* Brown-Forsythe (2, 488) = 28,77*
	Expérience sportive intermédiaire	200	13,64	0,82	
	Haut niveau d'expérience sportive	47	12,82	0,60	
	Total	491	13,65	0,88	
Performance en saut en longueur sans élan	Expérience sportive limitée	244	2,15	0,19	*F* Anova (2, 488) = 23,57*
	Expérience sportive intermédiaire	200	2,17	0,19	
	Haut niveau d'expérience sportive	47	2,40	0,17	
	Total	491	2,18	0,20	
Performance en lancer du poids	Expérience sportive limitée	244	8,065	1,13	*F* Anova (2, 488) = 17,65*
	Expérience sportive intermédiaire	200	8,25	1,14	
	Haut niveau d'expérience sportive	47	9,15	1,27	
	Total	491	8,24	1,19	
Moyenne sur 20 aux 4 tests physiques	Expérience sportive limitée	244	8,70	2,22	*F* Brown-Forsythe (2, 488) = 47,51*
	Expérience sportive intermédiaire	200	9,23	2,25	
	Haut niveau d'expérience sportive	47	12,11	1,84	
	Total	491	9,24	2,40	

* $p < 0{,}001$
1. Nous avons comparé les moyennes avec le test alternatif de Brown-Forsythe (1974) lorsque les variances des groupes comparés étaient présumées inégales.

par la course de vitesse et la course de résistance. La deuxième composante factorielle correspond à la puissance musculaire phosphagénique ou encore à la capacité de produire en moins de 7 secondes du phosphate déjà présent dans les muscles ; elle est définie par les autres épreuves physiques.

6. CONCLUSION

En conclusion, l'étude des déterminants physiologiques et musculaires de la performance aux épreuves physiques des candidats qui se présentent aux concours d'entrée pour les études d'éducation physique et sport au Maroc et en Algérie invite les formateurs à repenser le construit de l'aptitude physique générale.

Selon nous, le concept *aptitude physique générale* est caduc. Cela confirme bien le mythe non fondé de l'existence d'une intelligence motrice générale (Brace, 1927 ; McCloy, 1934). En effet, nos analyses démontrent clairement que les différents items qui mesurent l'aptitude physique générale mesurent deux construits distincts plutôt qu'un facteur global et unique.

Des recherches additionnelles sur la mise au point de tests plus valides pour mesurer les différentes dimensions de l'aptitude physique seraient souhaitables. Si les institutions de formation des professionnels de l'éducation physique et sportive estiment qu'il est important de sélectionner leurs candidats sur la base de leur aptitude physique, il conviendrait de trouver des tests et des critères de sélection plus valides pour mesurer ce construit. Le construit de l'aptitude physique générale est au moins bidimensionnel selon cette recherche. Le protocole de sélection devrait être capable de différencier clairement les candidats sur au moins les deux dimensions distinctes suivantes : la puissance musculaire glycolytique et la puissance musculaire phosphagénique.

RÉFÉRENCES

Aboussaid, A. (1982). *Les épreuves d'athlétisme du concours d'accès à l'Institut national des sports Moulay Rachid*. Mémoire de maîtrise inédit. Rabat, Maroc : Institut royal de la formation des cadres.

Alem, J. (2003). *La valeur de l'appréciation par simulation (APS) pour prédire le succès initial en enseignement des candidats aux études en éducation*. Thèse de doctorat inédite. Québec, Québec : Université Laval.

Alem, J., Dadouchi, F., Kerfes, N. et Cloes, M. (2009). *La validité de construit de sept épreuves physiques qui mesurent l'aptitude physique générale sportive des candidats pour une formation supérieure en sport*. Actes du XII^{e} Congrès

international de psychologie du sport, Marrakech, Maroc : International Society of Sport Psychology (ISSP), <http://www.issp2009.com/abstracts_submission/index.php>, consulté le 8 février 2010.

Alem, J., El Mezdi, F., Dadouchi, F., Kpazai, G. et Bendefa, O. (2008). *La sélection des candidats aux études dans les formations à l'enseignement : quels critères de sélection ? Quels indicateurs de succès en enseignement ?* Actes du Ier Colloque scientifique de l'instance nationale d'évaluation du système éducation formation (INESEF-CSE). Souissi-Rabat, Maroc : Faculté de médecine, Université Mohamed V.

Alem, J., Hamdane, K., Mawfik, N. et El Mezdi, F. (2005). *La valeur des tests de sélection dans les programmes d'études universitaires scientifiques pour prédire l'engagement aux études et l'engagement professionnel : recension des écrits*. Actes de la IIIe Rencontre nationale AIPU-Maroc, Colloque de l'Association internationale de la pédagogie universitaire. Kénitra, Maroc : AIPU.

Alem, J., Taibi, M. et Guay, M. (2005). *Étude de la validité prédictive de 4 tests de sélection qui composent le concours d'accès à un institut de formation d'entraîneurs de sport*. Actes du XVIIIe Colloque international de l'Association pour le développement des méthodologies d'évaluation en éducation. Reims, France : Université de Reims.

Brace, D. K. (1927). *Measuring motor ability*. New York, New Jersey : Barnes.

Brown, M. B. et Forsythe, A. B. (1974). The small sample behavior of some statistics with test of equality of several means. *Tecnometrics*, *16*, 129-132.

Crocker, L. et Algina, J. (1986). *Introduction to classical and modern test theory*. New York, New Jersey : Harcourt Brace Jovanovich.

Drowatzky, J. N. et Zuccato, F. C. (1967). Interrelationships between selected measures of static and dynamic balance. *Research quarterly*, *38*, 509-510.

Dufour, A. B. et Pontier J. (1989). Morphologie des handballeurs français selon les niveaux et les postes de jeu : un exemple d'application de la méthode Longi. *Cahiers d'anthropologie et biométrie humaine*, *7*(1-2), 69-80.

Filliard, J. R. (1995). *Tables de cotation de la valeur physique. Garçon et filles 10-22 ans* (2e éd.). Paris, France : INSEP.

Fleishman, E. A. (1964). *The structure and measurement of physical fitness*. Englewood Cliffs, New Jersey : Prentice Hall.

Fleishman, E. A. (1965). The description and prediction of perceptual motor skill learning. Dans R. Glaser (dir.), *Training research and education*. New York, New Jersey : Wiley.

Fleishman, E. A. et Bartlett, C. J. (1969). Human abilities. *Annual review of psychology*, *20*, 349-380.

Fleishman, E. A. et Parker, J. F. (1962). Factors in the retention and relearning of perceptual motor skill. *Journal of experimental psychology*, *64*, 215-226.

Goodlad, J. I. (1990). *Teachers for our nation's schools*. San Francisco, Californie : Jossey-Bass.

Gorostiaga, E. M. et Granados, C. (2004). Differences in physical fitness and throwing velocity among elite and amateur male handball players. *International journal of sports medicine*, *25*, 1-8.

Guilford, J. P. (1958). A system of the psychomotor abilities. *American Journal of psychology*, *71*, 164-174.

Guyton, E. et Farokhi, E. (1987). Relationships among academic performance, basic skills, subject matter knowledge and teaching skills of teacher education graduates. *Journal of teacher education*, *38*, 37-42.

Haberman, M. (1987). *Recruiting and selecting teachers for urban schools*. Reston, Virginie : Association of teacher educators.

Henry, F. M. (1961). Reaction time-movement time correlations. *Perceptual and motor skills*, *12*, 63-66.

Letessier, J. (1957). *Table de cotation des performances sportives*. Paris, France : ESP.

Lotter, W. S. (1960). Interrelationships among reaction times and speeds of movement in different limbs. *Research quarterly*, *31*, 147-155.

Maso, F. et Cazorla, G. (2001). Exigences physiologiques nécessaires à la pratique de rugby de haut niveau. *Science et sport*, *17*, 297-301.

McCloy, C. H. (1934). The measurement of general motor capacity and general motor ability. *Research quarterly*, *5*(supplément 5), 45-61.

Nahari, M. (1985). *Essai d'une évaluation de la formation supérieure en sport. Concours d'entrée et polyvalence*. Mémoire de maîtrise inédit. Rabat, Maroc : Institut royal de la formation des cadres.

Schmidt, R. A. (1993). *Apprentissage moteur et performance*. Paris, France : Vigot.

Shechtman, Z. et Godfried, L. (1993). Assessing the performance and traits of teacher education students by a group assessment procedure : a study of concurrent and construct validity. *Journal of teacher education*, *44*(2), 130-138.

Thomas, R., Eclache, J.-P. et Keller, J. (1989). *Les aptitudes motrices. Structure et évaluation*. Paris, France : Vigot.

Chapitre 4

La simulation comme technique d'enseignement et d'évaluation en sciences infirmières

Un état de la question

Marie-Ève Latreille, Éric Dionne et Diana Koszycki

La simulation est une technique d'enseignement et d'évaluation de plus en plus courante dans le domaine médical. Elle est un bon exemple d'une évaluation basée sur la performance, que ce soit en contexte formatif ou certificatif. Dans ce texte, nous exposerons le fruit d'une recension des écrits sur les différents modèles de simulation en évoquant les avantages et les limites de chacun d'eux, particulièrement dans le contexte des sciences infirmières. Notre synthèse tend à démontrer que les simulations ajoutent une plus grande valeur à l'enseignement et permettent d'obtenir des données originales sur le degré de maîtrise des compétences.

1. INTRODUCTION

Notre système de soins de santé à l'échelle mondiale s'affaiblit en raison, entre autres, des contraintes financières, du vieillissement de la population, de la pénurie de professionnels de la santé, des coûts élevés liés à la technologie moderne et à l'avancement médical. De réels efforts sont indispensables pour assurer la viabilité et l'efficience des systèmes de santé. Au Canada, plus de 172 milliards de dollars ont été dépensés en soins de santé pour l'année 2008, soit 60 % de plus qu'il y a dix ans (Institut canadien d'information sur la santé, 2009). On constate alors que nos

milieux hospitaliers sont sous tension et en perpétuel changement afin de s'adapter. Les heures de travail sont longues. À titre d'exemple, les infirmières ont réalisé, en 2005, l'équivalent de 10 000 postes temps plein en heures supplémentaires (Association des infirmières et infirmiers du Canada, 2006). En plus d'une augmentation des heures de travail et de surtemps, le champ de pratique des infirmières a été élargi en raison des avancées médicales. Les infirmières font à présent face à une augmentation de la complexité des patients hospitalisés, ce qui signifie aussi une charge de travail plus lourde. Parmi ces avancées médicales, citons les instruments chirurgicaux spécialisés, les transplantations complexes d'organes et les appareils visant à maintenir les fonctions vitales d'un patient, dont la respiration et la filtration du sang. Par conséquent, il est nécessaire de se doter de professionnels de la santé bien formés et hautement performants. Ces derniers éprouvent de la difficulté à réaliser correctement leur travail et perçoivent que la formation n'est plus aussi appropriée.

Les établissements d'enseignement, plus particulièrement les facultés de médecine et des soins infirmiers, subissent directement les contrecoups de ce phénomène. On exige instamment que la formation de praticiens et d'infirmières soit à la hauteur des milieux complexes de santé. De plus, la pénurie de professionnels de la santé oblige les facultés à trouver des moyens pour augmenter le nombre d'étudiants diplômés par année afin de remplacer les départs à la retraite. Les facultés sont également assujetties à des contraintes financières et l'espace limité des salles de classe ne facilite pas la réalisation des requêtes. Or, les facultés des sciences de la santé ont dû combiner leurs efforts pour redéfinir la façon même d'enseigner aux étudiants afin de promouvoir l'apprentissage et améliorer la formation. Une solution innovatrice pour contrer ces problèmes a récemment été envisagée par les facultés des sciences de la santé afin d'introduire une nouvelle approche d'enseignement et de mieux préparer leurs étudiants à la réalité des milieux hospitaliers. Cette technique d'enseignement nouvellement adoptée s'intitule la simulation de haute fidélité. Elle permet aux étudiants de développer les connaissances et les compétences nécessaires tout en minimisant les contraintes évoquées précédemment. Cette nouvelle technique particulièrement alléchante pour les facultés leur a enfin permis de fournir un enseignement plus réaliste, authentique et représentatif des milieux actuels de soins de santé. En fait, la technique de simulation tente de recréer des situations réelles qui se présentent sur une base quotidienne ou quasi quotidienne. Bien que la réussite scolaire soit fondamentale, la simulation permet à l'étudiant de perfectionner la prise en charge d'un patient par le biais d'un jugement clinique et apporte une meilleure vision quant

à la collecte et à l'analyse des données. Par conséquent, la simulation est sans contredit un complément intéressant au développement des compétences professionnelles en sciences infirmières.

Sur le plan de l'enseignement, la simulation de haute fidélité élimine plusieurs contraintes et offre de multiples bénéfices aux étudiants. Comme l'ont relevé Medley et Claydell (2005), la simulation de haute fidélité effectuée en laboratoire est un outil éducationnel qui permet aux étudiants de développer et de pratiquer leurs compétences dans un environnement sécuritaire en leur donnant l'occasion de prendre des décisions, d'exercer leur jugement critique et de travailler en équipe. Les facultés des sciences de la santé se sont donc rapidement procuré l'équipement nécessaire et elles ont intégré la simulation de haute fidélité dans leurs programmes d'études. Toutefois, les rares recherches disponibles ne nous permettent pas de savoir si cette méthode est réellement efficace ni dans quelle proportion. Day-Black et Watties-Daniels (2006) notent que, malgré la hausse en popularité des simulations de haute fidélité, la recherche sur ce type d'enseignement n'est pas suffisante pour guider l'utilisation de cette méthode d'enseignement et d'évaluation en sciences infirmières. Le contexte associé à la mise en place de ces simulations est une condition importante pour assurer la validité de construit de cette méthode. Gaba (1992) soutient également qu'aucune industrie n'attendra d'obtenir la preuve des bénéfices de la simulation avant son implantation lorsque des vies humaines dépendent des compétences de ses opérateurs. Il est à noter que les simulations ont fait leurs preuves dans d'autres secteurs, notamment dans le secteur de l'aérospatiale et de l'aviation. À ce jour, nous n'avons encore aucune certitude en ce qui concerne le transfert de connaissances et de compétences acquises en simulation vers les milieux cliniques réels. Feingold, Calaluce et Kallen (2004) ont découvert que seulement la moitié des étudiants qui avaient participé à des simulations de haute fidélité croyaient que les compétences apprises pourraient être transférées dans une situation clinique réelle. Les auteurs soulignent ensuite qu'il pourrait être difficile de répondre à cette question du transfert puisque cela demanderait de réaliser une recherche comparative, ce qui poserait trop de risques pour la sécurité des patients. Wellard, Woolf et Gleeson (2007), pour leur part, évoquent la difficulté d'établir des liens solides entre la capacité des étudiants à faire des liens entre la théorie et la pratique quand on les place en contexte de laboratoire visant le développement de compétences cliniques. Autrement dit, on ne sait que peu de choses sur la validité prédictive de ces méthodes d'enseignement.

Dans son article, Harder (2010) conclut qu'il existe un manque de preuve et d'évaluation pour déterminer si la simulation est réellement une méthode d'enseignement efficace. Ce dernier point retient particulièrement notre attention car, dans notre recension d'écrits, nous avons constaté, tout comme Harder, que les études effectuées au cours des dix dernières années sur la simulation ont principalement porté sur l'application de la simulation comme méthode d'acquisition de connaissances et de compétences. Très peu d'études ont vérifié l'efficacité de la simulation sur l'évaluation et l'amélioration des résultats des étudiants. Cette efficacité semble pour l'instant être tenue pour acquise.

Dans la prochaine section, nous nous pencherons sur la définition de la simulation de haute fidélité pour ensuite nous attarder sur les différentes techniques d'évaluation d'une simulation en évoquant les avantages et les limites de chacune d'elles, particulièrement dans le contexte des sciences infirmières.

2. CADRE THÉORIQUE

2.1. La simulation de haute fidélité

La simulation de haute fidélité est la reproduction, en laboratoire, d'une situation clinique réelle à l'aide d'un simulateur électronique de patients, appelé « mannequin », ou d'un acteur appelé « patient standardisé ». Le mannequin et le patient standardisé reproduisent les conditions médicales envisageables d'un patient et répondent physiologiquement et physiquement aux actions des étudiants. Pour qu'elle soit définie comme étant de haute fidélité, la simulation doit répondre de façon réaliste aux interventions cliniques effectuées par les étudiantes (Bearnson et Wiker, 2005). Les mannequins de haute fidélité et les patients standardisés répondent à ce critère en réagissant positivement ou négativement aux actions, aux interventions cliniques et aux décisions prises par les étudiants, telles que l'administration de médicaments ou d'oxygène.

La simulation ne concerne pas seulement l'utilisation de mannequins ou d'acteurs ; elle vise le développement d'une situation qui transmet à l'étudiant la sensation de prendre soin d'un véritable patient. Elle nécessite la combinaison de la théorie, des évaluations physiques, de la pharmacologie, de la pathophysiologie, du développement des compétences psychomotrices, etc., afin que l'étudiant développe son jugement clinique (Gore, Hunt et Raines, 2008). C'est ainsi que la simulation permet à l'étudiant d'exercer sa pensée critique et d'utiliser une démarche de soins infirmiers afin de surveiller le patient simulé,

de résoudre différents problèmes et d'améliorer son état. À l'aide de la simulation, il est donc à présent possible pour un étudiant d'expérimenter une situation de soins complexes sans compromettre la sécurité du patient. La popularité de la simulation et l'intérêt qu'elle suscite découlent de plusieurs facteurs, entre autres de sa capacité à fournir aux étudiants un apprentissage expérientiel qui consiste en la transformation de leur expérience vécue en savoirs, comparativement à l'assimilation d'informations verbales ou écrites (Chevrier et Charbonneau, 2000).

En 2005, Jeffries mit au point le cadre de référence d'une simulation dans l'éducation des sciences infirmières, basé sur les recherches théoriques et empiriques de la simulation en sciences infirmières, en médecine ainsi que dans d'autres domaines médicaux et non médicaux (Jeffries, 2007). Ce cadre est formé de cinq composantes ayant différentes caractéristiques. Ce modèle comprend 1) le professeur, 2) l'étudiant, 3) les pratiques éducationnelles à incorporer dans l'enseignement, 4) les caractéristiques du design de la simulation et 5) les objectifs et résultats prévus de la simulation (voir la figure 4.1).

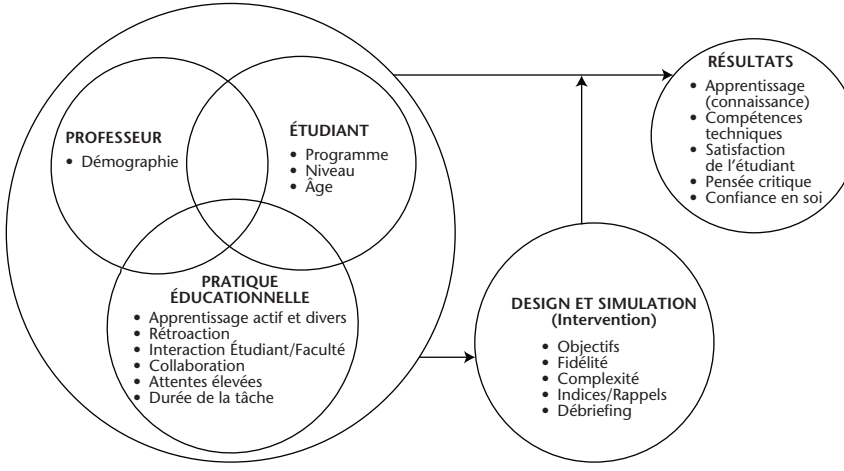


Figure 4.1.
Le cadre de référence de la simulation en éducation des sciences infirmières

En plus de fournir un apprentissage par l'expérience, la simulation possède une composante de réflexion, soit le débriefing. Le débriefing est la dernière et la plus importante des étapes d'une simulation ; elle permet à l'étudiant d'effectuer un transfert et une réflexion profonde sur l'expérience vécue. Cette dernière étape permet aussi au professeur

de revoir les compétences particulières des étudiants et leurs prises de décision en relation avec les objectifs du scénario. Ces objectifs consistent à offrir des soins sécuritaires, à bien communiquer avec le patient, à utiliser la pensée critique pour résoudre le problème de santé et à utiliser les connaissances et compétences techniques pour prodiguer les soins infirmiers appropriés en vue de traiter le patient. Durant le débriefing, le professeur offre aussi une rétroaction en énumérant les réussites des étudiants avant d'aborder ce qu'ils ont appris de leurs erreurs ou de leurs omissions (Henneman, Cunningham, Roche et Cumin, 2007). Dans leurs recherches, Issenberg et Scalese (2007) ont découvert que fournir de la rétroaction aux étudiants à propos de leurs performances était l'élément le plus important d'une simulation en vue d'un apprentissage efficace et réussi. Cette période de réflexion offre aux étudiants l'occasion d'évaluer leurs actions, leurs décisions, leurs communications et leurs habiletés à cheminer durant la simulation lors de situations imprévues (Henneman et Cunningham, 2005).

À l'heure actuelle, la simulation est surtout utilisée pour évaluer les étudiants dans un contexte d'évaluation formative ou sommative ; une évaluation formative permet de faire le point sur la progression de l'étudiant au cours de l'apprentissage. Dans un tel contexte d'évaluation, les étudiants pourraient être placés dans des situations cliniques complexes, avoir par exemple à gérer un patient en acidocétose diabétique. Cette situation offrirait aux étudiants la possibilité de transférer les connaissances et compétences acquises en théorie sur le diabète et de les mettre en pratique durant une séance de simulation en laboratoire. Pour sa part, l'évaluation sommative, habituellement effectuée en contexte certificatif, permet d'évaluer les acquisitions de l'étudiant afin de lui attribuer une note. Lors de ce type d'évaluation, une échelle non standardisée, mais établie à partir des compétences et critères essentiels à la profession infirmière, permet de déterminer le score de l'étudiant. Un exemple d'évaluation sommative serait l'injection d'un médicament par voie intramusculaire dans le muscle deltoïde d'un mannequin.

2.2. Les différentes techniques d'évaluation de simulations

Plusieurs recherches ont tenté de déterminer les avantages de l'enseignement et de l'apprentissage acquis par l'entremise de la simulation. Il a été particulièrement difficile de trouver la façon de procéder afin d'évaluer l'efficacité de la simulation. Voici une recension des méthodes de recherche quantitatives et qualitatives utilisées : 1) l'utilisation de pré- et post-tests, 2) l'utilisation de l'ECOS – Examen clinique objectif structuré et 3) l'évaluation de quatre critères d'aptitudes, soit

l'acquisition de connaissances, les compétences techniques, la satisfaction de l'étudiant et la confiance en soi. Une ouverture sur une méthode d'évaluation, nommée le « test de concordance de script », sera discutée à la fin de cette section. Le TCS (test de concordance de script) pourrait devenir un instrument de mesure intéressant à explorer pour mesurer l'efficacité de la simulation dans le futur.

2.2.1. *Test de connaissance dans un devis pré- et post-test*

Bien que les pré- et post-tests ne soient pas une méthode d'évaluation de simulation, mais bien un devis expérimental, ils sont fréquemment utilisés pour évaluer l'efficacité de la simulation. Voici quelques exemples d'études où la méthode de pré- et post-test a été utilisée pour déterminer l'acquisition de connaissances et de compétences des étudiants à la suite de l'application de la simulation de haute fidélité. Hoffman, O'Donnell et Kim (2007) ont eu recours à cette méthode pour évaluer les connaissances de base des étudiants de sciences infirmières au baccalauréat. À l'aide de l'instrument de mesure des connaissances de base (*Basic knowledge assessment tool* : BKAT-6), ils ont testé les connaissances de base avant et après l'exposition de stages cliniques ainsi que de simulation en laboratoire de 29 participants répartis en différents groupes. Le BKAT-6 est un test papier-crayon standardisé comprenant une centaine de questions qui mesure le rappel de l'information de base et son application dans des situations de pratique. Trois mois après l'exposition à la simulation de haute fidélité, les chercheurs ont pu démontrer que la combinaison de simulation et de stages cliniques améliore les connaissances de base des étudiants. Une représentation statistique pour appuyer les conclusions de ces chercheurs aurait été nécessaire, mais ces données ne figuraient pas dans l'article de recherche.

Une étude similaire a été effectuée par Morgan, Cleave-Hoff, Desousa et Lam-McCulloch (2006). Ces chercheurs ont voulu vérifier si la simulation de haute fidélité améliorait la performance de 299 étudiants en anesthésiologie sur les examens de simulation et les examens écrits. Comme prétest, les étudiants ont répondu à un questionnaire de pharmacologie comprenant 10 questions à choix multiples visant à établir leurs connaissances sur la gestion des arythmies cardiaques. Après avoir été soumis à quatre scénarios de simulation, les étudiants étaient évalués à l'aide d'une liste de contrôle et d'une échelle d'évaluation globale. Finalement, les étudiants ont repris le quiz de pharmacologie à choix multiples comme post-test. Les résultats au post-test ont révélé que la simulation avait significativement amélioré les résultats des étudiants. En revanche, les chercheurs ont souligné

que seuls les apprentissages à court terme étaient évalués et qu'ils ne pouvaient déterminer si l'application des connaissances des étudiants acquises à la suite de la simulation se prolongerait dans le temps pour devenir des apprentissages à long terme.

Les pré- et post-tests nous permettent donc de vérifier la différence de performance entre deux groupes, témoin et expérimental, à la suite de la simulation. En revanche, les inconvénients sont nombreux. Par exemple, cette méthode ne permet pas d'observer un transfert de connaissances dans les milieux cliniques ni de vérifier les compétences d'un professionnel de la santé, telles que la communication interdisciplinaire et le raisonnement clinique (Hoffman, O'Donnel et Kim, 2007).

2.2.2. *Examen clinique à l'aide d'objectifs structurés (ECOS)*

Introduit à la fin des années 1970 par Harden et Gleeson (1979), l'examen clinique à l'aide d'objectifs structurés (*objective structured clinical examination*, OSCE) se compose de stations qui permettent l'évaluation de la performance d'un étudiant à différents niveaux. Les stations peuvent être utilisées pour tester les connaissances théoriques, les compétences techniques, l'habileté de communication, la pensée critique, la résolution de problèmes et plus encore. Les examens cliniques à l'aide d'objectifs structurés permettent donc d'évaluer les compétences d'un étudiant sur plusieurs facettes et de donner une meilleure image du rendement de celui-ci.

Habituellement, les examens cliniques à l'aide d'objectifs structurés sont composés de 15 à 20 stations. Un étudiant aura quelques minutes à consacrer à chacune des stations pour effectuer l'exercice demandé. Les exercices à effectuer peuvent être, par exemple, la démonstration d'une technique infirmière spécifique ou des questions de type papier-crayon. Le pointage accordé à l'étudiant est effectué à partir d'une grille d'évaluation remplie par un examinateur qui surveille les stations. L'évaluation de l'examen vient alors de sources objectives, ce qui contribue à sa validité (Alinier, 2003). L'objectivité de l'examen tient à son format standardisé, qui fait en sorte que les étudiants sont tous évalués de la même façon et sur les mêmes sujets par l'entremise des stations. Peden, Cairncross, Harden et Crooks (1985) relèvent également l'objectivité de cette méthode d'évaluation en soulignant que la variabilité entre les examinateurs est réduite à l'aide des grilles d'évaluation. Le désavantage de cette méthode d'évaluation réside dans le fait qu'elle nécessite beaucoup de ressources financières ainsi que du personnel de soutien. De plus, elle peut être complexe à élaborer et à mettre en application.

Alinier, Hunt et Gordon (2004) ont réparti 67 étudiants de deuxième année du programme des sciences infirmières en un groupe expérimental (29 étudiants) et en un groupe témoin (38 étudiants). Ils ont ensuite constaté lors de la première session de l'examen clinique à l'aide d'objectifs structurés que les deux groupes avaient obtenu des scores similaires, soit 49,59 % pour le groupe témoin et 50,19 % pour le groupe expérimental. Ils ont ainsi pu démontrer que les deux groupes possédaient des connaissances et des compétences similaires en début d'étude. Les étudiants participèrent ensuite à deux sessions de simulation avant d'être réévalués lors d'un second examen clinique à l'aide d'objectifs structurés. La moyenne des étudiants après le second examen se situait à présent à 56,35 % pour le groupe témoin et à 63,62 % pour le groupe expérimental. À partir d'une analyse statistique, les auteurs ont découvert que les deux groupes d'étudiants avaient amélioré leurs scores de 6,76 % et 13,43 %, ce qui donne une différence entre les deux groupes de 6,67 % en faveur du groupe expérimental. Les chercheurs ont ensuite conclu que la simulation avait permis aux étudiants du groupe expérimental d'améliorer leurs connaissances et compétences de base en sciences infirmières, comparativement au groupe témoin. De plus, un test t ($p < 0,05$) des scores individuels à l'examen clinique à l'aide d'objectifs structurés a révélé que le groupe expérimental avait amélioré ses scores de façon significative, comparativement au groupe témoin.

Alinier, Hunt, Gordon et Harwood (2006) ont reproduit la même étude pour évaluer l'influence des scénarios utilisés en simulation sur la formation des connaissances et compétences des étudiants en sciences infirmières à l'aide de 99 participants répartis aléatoirement dans des groupes expérimental et témoin. Les chercheurs ont fait participer le groupe expérimental à deux sessions de simulation en soins intensifs en plus de leur faire poursuivre les cours en sciences infirmières et les stages cliniques selon un apprentissage traditionnel. Les chercheurs se sont assurés que les simulations ne permettaient pas de mieux préparer les étudiants de ce groupe à l'examen clinique à l'aide d'objectifs structurés. Pour sa part, le groupe témoin avait poursuivi selon un enseignement traditionnel les cours en sciences infirmières ainsi que les stages cliniques sur une période de six mois, sans session de simulation. À la suite de l'évaluation du premier examen clinique à l'aide d'objectifs structurés, composée de 15 stations, le groupe témoin avait obtenu une moyenne de 48,82 % et le groupe expérimental, une moyenne de 47,54 %. Lors de la réévaluation des connaissances et compétences du deuxième examen clinique à l'aide d'objectifs structurés, le groupe témoin avait une moyenne de 56,00 % et le groupe expérimental, une moyenne de 61,71 %. La moyenne des

scores avait ainsi augmenté de 7,18 % et 14,18 % respectivement, une différence de 7 % en faveur du groupe expérimental. La différence entre les moyennes, établie à partir d'un test t ($p < 0{,}001$), était statistiquement significative. Malheureusement, cet article n'offrait que des données statistiques relativement à l'amélioration des deux groupes et ne précisait pas sur quelles dimensions les étudiants avaient réussi à s'améliorer.

Les examens cliniques à l'aide d'objectifs structurés semblent efficaces comme méthode d'évaluation, mais ils n'ont malheureusement pas été introduits dans le curriculum des sciences infirmières, ce qui demande une participation volontaire aux recherches. La préparation et la conception des examens cliniques à l'aide d'objectifs structurés constituent un travail méticuleux exigeant des concepteurs qu'ils indiquent clairement les aptitudes à évaluer. Bien que les participants trouvent qu'il est stressant de participer à des examens cliniques à l'aide d'objectifs structurés, ces examens sont généralement bien appréciés et les étudiants jugent cette méthode d'apprentissage efficace (Alinier, 2003).

2.2.3. *Recherches sur l'acquisition de connaissances, les compétences techniques, la satisfaction de l'étudiant et la confiance en soi*

Plusieurs recherches ont tenté, à l'aide de sondages menés auprès des étudiants, de dégager la perception qu'ils avaient de la simulation de haute fidélité. Les thèmes les plus utilisés étaient l'acquisition de connaissances et de compétences techniques, la satisfaction des étudiants et leur confiance en soi.

Bambini, Washburn et Perkins (2009) ont effectué une étude sur la méthode d'enseignement et d'apprentissage de la simulation de haute fidélité afin d'augmenter l'efficacité individuelle des étudiants en sciences infirmières. Les résultats relevés à partir des données qualitatives recueillies indiquent que les étudiants sont d'avis que la simulation de haute fdélité est une expérience d'apprentissage valable. La simulation leur permettait d'augmenter leur confiance en eux, de concevoir les attentes à leur égard ainsi que les conduites qu'ils devaient idéalement adopter dans les milieux de soins cliniques.

Cioffi (2001) a souligné que l'apprentissage actif accroît la motivation et l'intérêt lors de l'apprentissage d'un étudiant. Ce concept est ce qui rend la simulation de haute fidélité populaire et il possède un haut taux de satisfaction chez les étudiants. Feingold, Calaluce, et Kallen (2004) le confirment : la majorité des sujets de leur étude ont estimé que les simulations pouvaient recréer des expériences cliniques réelles.

Les étudiants de cette étude ont aussi relevé que la simulation de haute fidélité permettait de tester leurs compétences techniques et leurs jugements cliniques et, en retour, qu'elle a su renforcer leur apprentissage. De plus, les expériences de simulation ont été très estimées autant par les professeurs que par les étudiants.

Il semble y avoir un plus grand nombre de recherches qualitatives sur l'efficacité de la simulation de haute fidélité en raison des difficultés engendrées pour bien l'évaluer à l'aide de méthodes quantitatives. L'information recueillie provient alors d'instruments qualitatifs. La collecte des données, effectuée auprès des participants après la simulation, rend compte des sentiments et des expériences personnelles des étudiants et des membres de la faculté. Un manque de données quantitatives concrètes, fidèles et valides, se fait donc sentir. Kneebone (2003) note ainsi un manque de recherche quantitative sur les résultats globaux des étudiants par rapport aux compétences cliniques, aux habiletés de communication et à la confiance en soi des étudiants après la formation par simulation.

Par conséquent, il est difficile de cerner l'efficacité de la simulation de façon quantitative lorsque les étudiants sont aussi exposés à plusieurs expériences en stages cliniques, ceux-ci ayant eux aussi un effet important sur la performance des étudiants. De plus, il serait éthiquement inacceptable de ne pas fournir la méthode d'enseignement basée sur la simulation à un groupe témoin lorsque celle-ci est intégrée dans les curricula des sciences infirmières. Il devient ainsi très difficile de séparer les étudiants en groupes témoin et expérimental pour poursuivre un devis de recherche expérimental.

2.2.4. *Test de concordance de scripts (TCS)*

Le test de concordance évalue les scripts des participants. Lambert, Gagnon et Charlin (2005) définissent les scripts comme des réseaux de connaissances permettant d'agir en situation clinique. Ils précisent que le script pertinent à une maladie contient un ensemble de signes et de symptômes ainsi que les éléments qui les relient entre eux. D'après eux, lors d'un entretien entre l'expert et son patient, le clinicien envisage diverses hypothèses. Les scripts permettent alors à un professionnel de la santé d'établir et de formuler un diagnostic, de décider d'une intervention ou d'un traitement.

Lors de l'évaluation par test de concordance de scripts, l'étudiant fait la lecture d'un cas, choisit ensuite une hypothèse parmi plusieurs à l'aide d'une échelle de type Likert et détermine l'intervention ou l'évaluation appropriée pour ce patient. Les situations couramment

rencontrées dans la pratique sont habituellement complexes, incomplètes et imparfaites, et c'est à partir de ce genre de situations que l'on construit le test de concordance de scripts. Avant l'administration du test aux étudiants, un panel composé d'experts du milieu effectuent le test et les résultats obtenus servent à déterminer le score de chaque question. Le test de concordance de scripts est donc un instrument qui mesure le niveau de concordance entre les scripts des participants et ceux des experts (Charlin et St-Jean, 2002).

Le test de concordance de scripts permet d'évaluer le raisonnement et la compétence clinique, mais cette technique comporte certaines limites (Gibot et Bollaert, 2008). Le choix des experts est une étape délicate et l'on y procède avec attention afin d'obtenir le champ précis de compétence à évaluer (Sibert, Charlin, Corcos, Gagnon, Grise et Van der Vleuten, 2002). Le test de concordance de scripts étant un examen écrit, il ne permet pas d'évaluer certaines compétences cliniques telles que les habiletés de collecte des données durant une entrevue et lors de l'examen physique ou encore les habiletés relationnelles (Gibot et Bollaert, 2008).

Waldner et Olson (2007) ont mis l'accent dans leur étude sur la difficulté à mesurer la performance clinique des étudiants, comme les connaissances, le jugement critique ou l'auto-efficacité, après qu'ils ont été exposés à une simulation, à partir d'un test écrit quantifiable. Le test de concordance de scripts peut donc être utilisé avec succès pour évaluer les connaissances avant et après une activité d'apprentissage telle que la simulation.

Le test de concordance de scripts est un outil pour évaluer le jugement des professionnels de la santé et il peut être adapté à diverses disciplines. Des études en radiologie, en gynécologie, en médecine générale, en pharmacie et pour devenir sages-femmes ont été développées à l'aide de cet instrument de mesure.

3. AVANTAGES/LIMITES DES DIFFÉRENTS TYPES D'ÉVALUATION DE SIMULATION

Lors de notre recension d'écrits, deux articles de recherche ont particulièrement attiré notre attention. Les auteurs de ces deux articles ont effectué une méta-analyse et recensé des centaines de recherches afin d'établir l'efficacité de la simulation comme méthode d'enseignement. Ces mêmes auteurs révèlent, en début d'articles, que de plus en plus de recherches sont effectuées sur la simulation comme méthode d'enseignement, mais que peu de chercheurs se concentrent sur son

efficacité pour en encourager l'utilisation dans nos curricula. Lors de leurs études, des critères d'admissibilité très stricts ont été appliqués et seulement une vingtaine d'études ont été retenues et analysées.

Le premier article de Laschinger, Medves, Pulling, McGraw, Waytuck, Harrison et Gambeta (2008) relate une recension exhaustive de plus de 1 118 recherches, dont seulement 23 ont été retenues en raison des critères d'admissibilité stricts. Les recherches retenues comprenaient 15 études en médecine et 6 études en sciences infirmières. Les auteurs ont déterminé par la suite que les études ne pouvaient clairement établir les bénéfices et les avantages de la simulation ni son efficacité. Plusieurs recherches se contredisaient, d'autant plus qu'elles n'évaluaient pas la même population ni les mêmes niveaux d'étude.

Au cours de sa recherche, Harder (2010) trouva un total de 23 recherches quantitatives portant sur l'évaluation de la performance des étudiants lors de l'utilisation de la simulation dans les milieux de la médecine, des sciences infirmières et interdisciplinaires. Au total, 10 études ont utilisé des pré- et post-tests et sept études ont utilisé des examens cliniques à l'aide d'objectifs structurés pour évaluer la performance de leurs étudiants. Harder conclut que ces résultats appuient son hypothèse que la simulation est une méthode d'enseignement qui influence positivement l'apprentissage des étudiants. L'auteur souligne aussi qu'il y avait un manque d'outils d'évaluation structurés pour évaluer les simulations. Conséquemment, des outils basés sur l'évaluation des aptitudes en milieu clinique ou des examens cliniques à l'aide d'objectifs structurés ont été modifiés et adaptés afin d'évaluer les étudiants lors de simulations (Alinier, Hunt et Gordon, 2004). Les efforts doivent être ainsi tournés vers les mesures d'évaluation d'une simulation.

4. CONCLUSION

Lorsque des enjeux critiques tels que l'augmentation de complexité des patients en milieu hospitalier se font ressentir et que la performance de nos professionnels de la santé a une influence majeure sur ces vies humaines, il est essentiel de réagir rapidement afin de trouver une solution. Il va sans dire que la simulation de haute fidélité est une méthode d'enseignement comprenant de nombreux avantages, tant pour les facultés des sciences de la santé que pour leurs étudiants ou pour le public en général. Issenberg et Scalese (2007) reconnaissent que la simulation de haute fidélité est une méthode d'enseignement

intéressante permettant l'enseignement d'une variété de compétences, qu'elles soient psychomotrices ou cognitives, telles que le développement du jugement clinique.

Étant une nouvelle méthode d'enseignement, en élaboration dans nos curricula, la simulation de haute fidélité est un sujet d'actualité. Les recherches et les tests d'évaluation en sont à leurs débuts et se concentrent surtout sur l'application de la simulation comme technique d'acquisition de connaissances et de compétences. Malheureusement, et comme nous l'avons déjà mentionné, très peu de recherches se sont penchées sur l'efficacité de la simulation afin d'évaluer l'amélioration des connaissances et des compétences des étudiants, ce qui nuit quelque peu à sa promotion dans nos curricula. De plus, certaines inquiétudes subsistent à l'égard du transfert des connaissances lors de la simulation et de son application dans les milieux cliniques. Il est aussi particulièrement difficile de mesurer quantitativement l'efficacité de la simulation ou de procéder à son évaluation. Il existe manifestement un manque de validité à ce niveau.

Enfin, de la nouveauté dans les outils d'évaluation pourrait apporter plus de réponses à la question d'efficacité. Des développements récents liés au jugement clinique ont été démontrés à l'aide des tests de concordance de scripts. Le test de concordance de scripts nous permettrait donc d'acquérir des informations quantitatives intéressantes sur le jugement clinique des étudiants. Nous aurions ainsi la possibilité d'évaluer et de déceler l'effet d'interventions éducatives, telles que des stages cliniques ou de la simulation de haute fidélité, pour mettre en évidence la progression de nos étudiants.

Cet article renferme néanmoins quelques limites. Faute d'espace, nous n'avons pu examiner certaines composantes importantes de la simulation, dont la perception du stress et le niveau de confiance en soi de l'étudiant qui a reçu un enseignement à l'aide de cette méthode. Nous anticipons aussi certains obstacles avec le test de concordance de scripts lorsqu'il s'agira de vérifier si l'amélioration du jugement clinique des étudiants est liée à l'enseignement théorique reçu en classe ou bien aux expériences vécues en stages cliniques.

RÉFÉRENCES

Alinier, G. (2003). *Nursing students' and lecturers' perspectives of OSCE, incorporating simulation.* Department of nursing and paramedic sciences, University of Hertfordshire, Royaume-Uni.

Alinier, G., Hunt, B. et Gordon, R. (2004). Determining the value of simulation in nurse education : study design and initial results. *Nurse education in practice*, *4*(3), 200-207.

Alinier, G., Hunt, B., Gordon, R. et Harwood, C. (2006). Effectiveness of intermediate-fidelity simulation training technology in undergraduate nursing education. *Journal of advanced nursing*, *54*(3), 359-369.

Association des infirmières et infirmiers du Canada (2006). *Toward 2020 : visions for nursing*. Ottawa, Ontario : Association des infirmières et infirmiers du Canada.

Bambini, D., Washburn, J. et Perkins, R. (2009). Outcomes of clinical simulation for novice nursing students : communication, confidence, clinical judgment. *Nursing education perspectives*, *30*(2), 79-82.

Bearnson, C. S. et Wiker, M. (2005). Human patient simulators : a new face in baccalaureate nursing education at Brigham Young University. *Journal of nursing education*, *44*(9), 421-425.

Charlin, B. et St-Jean, M. (2002). Test de concordance de script : un outil pour évaluer le jugement en médecine. *Bulletin du centre d'études et de formation en enseignement supérieur de l'Université de Montréal*, *6*, 4-5.

Chevrier, J. et Charbonneau, B. (2000). Le savoir-apprendre expérientiel dans le contexte du modèle de David Kolb. *Revue des sciences de l'éducation*, *26*(2), 287-323.

Cioffi, J. (2001). Clinical simulations : development and validation. *Nurse education today*, *21*, 477-486.

Day-Black, C. et Watties-Daniels, A. D. (2006). Cutting edge technology to enhance nursing classroom instruction at Coppin state university. *The ABNF journal*, *17*(3), 103-106.

Feingold, E. C., Calaluce, M. et Kallen, A. M. (2004). Computerized patient model and simulated clinical experiences : evaluation with baccalaureate nursing students. *Journal of nursing education*, *43*(4), 156-163.

Gaba, D. M. (1992). Improving anaesthesiologist's performance by simulating reality. *Anaesthesiology*, *74*, 491-494.

Gibot, S. et Bollaert, P.-E. (2008). Le test de concordance de script comme outil d'évaluation formative en réanimation médicale. *Pédagogie médicale*, *9*(1), 7-18.

Gore, T., Hunt, W. C. et Raines, H. K. (2008). Mock hospital unit simulation : a teaching strategy to promote sage patient care. *Clinical simulation in nursing*, *4*, e57-e64.

Harden, R. M. et Gleeson, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination. *Medical education*, *13*, 41-54.

Harder, N. (2010). Use of simulation in teaching and learning in health sciences : a systematic review. *Journal of nursing education*, *49*(1), 23-28.

Henneman, E. A. et Cunningham, H. (2005). Using clinical simulation to teach patient safety in an acute/critical care nursing course. *Nurse educator*, *30*, 172-177.

Henneman, E. A., Cunningham, H., Roche, J. P. et Cumin, E. M. (2007). Human patient simulation : teaching students to provide safe care. *Nurse educator*, *32*(5), 212-217.

Hoffman, L. R., O'Donnell, M. J. et Kim, Y. (2007). The effects of human patient simulators on basic knowledge in critical care nursing with undergraduate senior baccalaureate nursing students. *Simulation in healthcare*, *2*(2), 110-114.

Institut canadien d'information sur la santé (2009). *Les soins de santé au canada 2009 : revue de la dernière décennie*. Ottawa, Ontario : Institut canadien d'information sur la santé.

Issenberg, S. B. et Scalese, J. R. (2007). Best evidence on high-fidelity simulation : what clinical teachers need to know. *The clinical teacher*, *4*, 73-77.

Jeffries, R. P. (2007). *Simulation in nursing education : from conceptualization to evaluation*. New York, New Jersey : National league for nursing.

Kneebone, R. (2003). Simulation in surgical training : education issues and practical implications. *Medical education*, *37*, 267-277.

Lambert, C., Gagnon, R. et Charlin, B. (2005). *Le test de concordance de scripts : un outil pour évaluer le raisonnement clinique des résidents en radio-oncologie*. Mémoire inédit de maîtrise en sciences de l'éducation, Université de Montréal.

Larew, C., Lessans, S., Spunt, D., Foster, D. et Covington, B. G. (2006). Innovations in clinical simulation : application of Benner`s theory in an interactive patient care simulation. *Nurse educator perspectives*, *27*, 16-21.

Laschinger, S., Medves, J., Pulling, C., McGraw, R., Waytuck, B., Harrison, B. M. et Gambeta, K. (2008). Effectiveness of simulation on health profession students' knowledge, skills, confidence and satisfaction. *International journal of evidence-based healthcare*, *6*(3), 278-302.

Medley, F. C. et Claydell, H. (2005). Using simulation technology for undergraduate nursing education. *Journal of nursing education*, *44*(1), 31-34.

Morgan, J. P., Cleave-Hogg, D., Desousa, S. et Lam-McCulloch, J. (2006). Applying theory to practice in undergraduate education using high fidelity simulation. *Medical teacher, 28*(1), e10-e15.

Peden, R. N., Cairncross, G. R., Harden, M. R. et Crooks, J. (1985). Assessment of clinical competence in therapeutics : the use of the OSCE. *Medical teacher, 7*, 217–223.

Sibert, L., Charlin, B., Corcos, J., Gagnon, R., Grise, P. et Van der Vleuten, C. (2002). Le test de concordance de script comme outil d'évaluation formative en réanimation médicale. *Pédagogie médicale*, *9*, 7-18.

Waldner, M. et Olson, J. (2007). Taking the patient to the classroom : applying theoretical frameworks to simulation in nursing education. *International journal of nursing education scholarship*, *4*(1), 1-14.

Wellard, J. S., Woolf, R. et Gleeson, L. (2007). Exploring the use of clinical laboratories in undergraduate nursing programs in regional Australia. *International journal of nursing education scholarship*, *4*(1), 1-11.

Chapitre 5

Portrait d'un contexte pour évaluer les apprentissages en arts plastiques et en danse contemporaine au collégial[1]

Diane Leduc, Jean-Guy Blais et Gilles Raîche

Avant la Révolution tranquille, l'évaluation des apprentissages dans les classes d'arts portait davantage sur le produit fini et, dans les classes de danse, sur l'observation des capacités physiques. Au cours de la décennie 1960, dans un souffle d'émancipation, le système scolaire québécois a été revu en profondeur. Issus de cette reconstruction, les cégeps apportent un renouveau et permettent aux jeunes de faire des études techniques ou préuniversitaires. L'enseignement des arts et avec lui les pratiques d'évaluation des apprentissages n'échappent pas à ces nouvelles structures et hériteront d'impératifs auxquels les professeurs devront nécessairement s'adapter. La réforme de 1993 provoquera de nouveaux ajustements dans l'enseignement collégial, obligeant cette fois les professeurs à prendre le tournant des approches par compétences.

1. Ce texte est issu de nos travaux soutenus financièrement par une bourse postdoctorale et une subvention du programme sur la persévérance et la réussite scolaires du Fonds québécois de recherche sur la société et la culture (FQRSC) et du ministère de l'Éducation, du Loisir et du Sport (MELS) ainsi que par une subvention pour le soutien des équipes de recherche du FQRSC.

1. INTRODUCTION

Au Québec, les années 1960 sont synonymes de Révolution tranquille. La Belle Province se tourne alors vers l'avenir en instituant la Commission royale d'enquête sur l'enseignement, nommée commission Parent. Celle-ci est en réalité l'aboutissement d'un vaste mouvement de société, amorcé au début des années 1950, visant à réformer en profondeur tout le système éducatif québécois. Dans son programme électoral de 1960, Jean Lesage s'engage à créer une commission d'enquête sur l'éducation, ce qu'il fait dès le début de son mandat de premier ministre de la province de Québec en 1961 (Corbo, 2006). Le travail des commissaires est à la mesure des lacunes du système en place : énorme et multiple. En effet, plusieurs éléments sont à prendre en considération pour penser cette reconstruction du système éducatif québécois. Il y a la sous-scolarisation de la population, un système injuste pour les jeunes des milieux modestes, pour la population des régions et pour les filles et l'accès restreint, surtout pour les francophones, aux études supérieures. À cela s'ajoute le fouillis des structures scolaires avec en tête de liste des programmes de formation qui ne permettent pas de répondre aux besoins en main-d'œuvre spécialisée, des filières de formation étanches tel le cours classique, offert uniquement dans les collèges privés, unique voie vers l'université (Corriveau, 1991).

Pour répondre à ces constats, la commission Parent propose la création d'un système scolaire unifié, intégré et public depuis la maternelle jusqu'à l'université, placé sous l'autorité d'un ministère de l'Éducation soutenu par un organisme consultatif, le Conseil supérieur de l'éducation (le ministère de l'Éducation du Québec a porté de 2004 à 2012 le nom de ministère de l'Éducation, du Loisir et du Sport). Il s'agit d'assurer le droit à l'éducation pour tous jusqu'aux portes de l'université (Corbo, 2006). Ainsi, de cette réforme sont nés le remplacement des différents types d'écoles secondaires publiques par une seule école dite polyvalente, la mise en place d'une maternelle publique et gratuite, l'augmentation des fonds destinés aux universités pour leur permettre de soutenir et de promouvoir l'évolution des connaissances dans tous les domaines et, ce qui nous intéresse particulièrement, l'ajout d'un ordre d'enseignement entre le secondaire et l'université. Constituant l'une des pièces maîtresses de cette réforme scolaire québécoise, les cégeps (acronyme de collèges d'enseignement général et professionnel) transforment le paysage éducatif postsecondaire. Depuis leur création en 1967, ils ont pour mission de parfaire et de confirmer la formation générale de l'étudiant et d'amorcer ou de compléter sa formation professionnelle (Gouvernement du Québec, 1967).

Sitôt le rapport Parent déposé, les acteurs du milieu artistique exigent une enquête sur l'enseignement des arts, ce qu'ils obtiendront avec la commission Rioux. Dans la foulée de cette seconde commission, l'enseignement des arts au Québec s'est transformé en une nouvelle plate-forme pédagogique basée sur une vision progressiste des arts dans la société. En effet, le rôle des arts dans l'appareil éducatif est entièrement revu : l'art est intégré à l'éducation, un encadrement administratif est mis en place, la formation professionnelle est restructurée et intégrée aux collèges et aux universités, etc. De ces changements organisationnels profonds, l'évaluation des arts en salle de classe héritera d'impératifs auxquels les professeurs devront nécessairement se plier, ce que provoquera également la réforme de 1993 dans l'enseignement collégial, obligeant cette fois les cégeps à prendre le tournant des approches par compétences.

Pour bien comprendre les enjeux de cette nouvelle réforme, il importe de situer les raisons fondamentales qui ont poussé le gouvernement du Québec vers la création des cégeps. L'un des éléments les plus significatifs est probablement le fait que dès l'après-guerre, jusqu'à l'aube des années 1970, un vent d'émancipation souffle sur tout le territoire québécois. Au moment où le gouvernement s'interroge sur la voie à prendre, le Québec est passé, en quelques années, d'un monde où la religion occupait toutes les sphères à une société presque totalement laïcisée (Laurin, 2007). Ce tournant dans l'histoire de l'éducation au Québec est l'inspiration initiale de notre démarche consistant à poser un bref regard sur le contexte dans lequel les professeurs en arts du collège évaluent les apprentissages. En guise d'introduction à un court portrait sur l'évaluation des apprentissages en arts plastiques et en danse contemporaine au cégep, nous vous proposons, dans le texte qui suit, un aperçu historique sur l'enseignement de ces disciplines.

2. APERÇU DE L'ENSEIGNEMENT ET DE L'ÉVALUATION EN ARTS PLASTIQUES ET EN DANSE AVANT LES CÉGEPS

2.1. Les arts plastiques

Avant même qu'ils ne soient définis dans le système scolaire du Québec, le dessin et les métiers d'art faisaient partie intégrante de la vie quotidienne en répondant à des besoins matériels, économiques et culturels. La colonie de la Nouvelle-France a besoin d'ouvriers pour les industries manufacturières, d'ébénistes, d'orfèvres, d'architectes, etc. Il est possible de retracer une première volonté d'encadrer l'enseignement des arts dans la loi de 1876 qui avait pour but de mettre en

place un système uniforme pour l'enseignement du dessin au primaire. Les trois leçons de dessin par semaine proposées par le Conseil de l'instruction publique d'alors se révéleront toutefois inefficaces (Couture, Joyal, Landry, Lemerise, Lussier et Wallot, 1980). Conséquemment, s'inspirant des méthodes en vogue dans les écoles du Vieux Continent, le gouvernement du Québec nommera, en 1892, monsieur Lefèvre pour organiser le dessin dans la province. Ainsi, jusqu'au milieu du XX^{e} siècle, les programmes officiels de formation contiennent un cours de dessin d'environ une heure par semaine, obligatoire dès la 7^{e} année (soit 12-13 ans ; aujourd'hui, la septième année a disparu de l'enseignement primaire). Il faut aussi noter qu'à cette époque au Québec l'école publique francophone est assujettie au clergé catholique et ce n'est qu'en 1942 qu'elle sera obligatoire jusqu'à 14 ans (Graveline, 2007).

L'évaluation se fait alors en fonction des éléments suivants : l'habileté ou la dextérité, le réalisme d'exécution et la propreté. Les formes géométriques sont à l'honneur et les exercices ne font appel qu'à l'observation d'objets usuels. Il faut noter que c'est le titulaire de classe qui donne les cours de dessin et que ceux-ci sont souvent synonymes de récompense, selon l'intérêt que le professeur y porte. Ces pratiques sollicitent peu l'intérêt et la motivation des élèves (Couture *et al*., 1980).

Au niveau des programmes complémentaire et supérieur (équivalent du secondaire actuel), on intègre graduellement des leçons sur l'histoire de l'art, données par des membres du clergé ou par des professeurs des classes ordinaires qui ont un *talent naturel*. Nous pouvons facilement déduire que leur formation académique spécialisée était à peu près inexistante (Couture *et al*., 1980). C'est d'ailleurs sur ce point que se jouera la suite : le manque de compétence des professeurs en la matière forcera la venue, dès 1928, dans le système scolaire de quelques artistes issus de l'École des beaux-arts de Montréal (ÉBAM). Pilier de la formation publique en arts au Québec, cette école a été fondée en 1923 sur le modèle de l'École des beaux-arts de Nantes, par le gouvernement Taschereau désireux de développer une culture française à l'image du Québec moderne. C'est à Louis Athanase David que revient la tâche de mettre en place les programmes, orientés surtout vers les arts appliqués, qui répondent à des objectifs économiques et de modernisation du système scolaire. Sous sa gouverne, les budgets augmenteront considérablement et des bourses d'études ainsi que des prix artistiques seront créés.

Toutefois, la fondation de cette école ne s'est pas faite sans heurt. Le clergé défendra bec et ongles la préservation des valeurs canadiennes-françaises et sera imperméable aux transformations que vit la société

québécoise (Graveline, 2007). L'École des beaux-arts de Montréal échappe au contrôle ecclésiastique en proposant une formation axée sur les arts décoratifs et l'architecture. Jusqu'en 1946, elle a une visée utilitariste : former des professeurs de dessin, des artisans et des ouvriers d'art pour répondre aux besoins nouveaux d'une société souhaitant se moderniser et développer son potentiel industriel, commercial et agricole. Graduellement, et non sans complications, l'École des beaux-arts de Montréal orientera ses formations vers l'histoire de l'art et la peinture, laissant à l'École du meuble le développement des formations techniques. Entre ces deux écoles se jouera la distinction entre les beaux-arts et les arts appliqués.

Après la Deuxième Guerre mondiale, la société québécoise amorce une lente évolution. Le monolithisme idéologique qui prévaut est ébranlé par de nombreux mouvements d'opposition (syndicaux, radicalistes, étudiants) et le pouvoir du gouvernement Duplessis, qui fait tout en son pouvoir pour conserver le contrôle sur le système scolaire, est contesté. L'École des beaux-arts de Montréal vit aussi des transformations, notamment avec l'engagement d'Alfred Pellan comme professeur, qui prône une liberté quant au choix des thèmes (non pieux) par l'artiste (Flibotte, 1987). Sa pensée s'oppose à celle de Borduas, professeur à l'École du meuble. Ces deux artistes et ces deux écoles seront au cœur des bouleversements artistiques provoqués par deux manifestes publiés en 1948, *Prismes d'yeux* et *Refus global*, ce dernier remettant en question les valeurs traditionnelles (la foi catholique et l'attachement aux valeurs ancestrales) et proposant le *refus* d'un repliement sur soi, exprimant un profond besoin de libération.

Peu à peu, l'École des beaux-arts de Montréal abandonne les arts domestiques au profit des disciplines strictement artistiques et fraie avec le champ intellectuel. Les nouveaux professeurs connaissent les idées naissantes sur l'enseignement des arts. C'est aussi une période où Piaget et les méthodes actives exercent une grande influence sur l'éducation. Ainsi, vers 1960, les objectifs pédagogiques sont orientés vers la maîtrise de moyens d'expression et non vers des moyens de reproduction. La pédagogie artistique fait son entrée grâce à Irène Senécal, pionnière de la didactique des arts au Québec dont les influences se font sentir partout dans la province (Couture *et al.*, 1980). Sa conception de l'enseignement des arts entraînera plusieurs changements, dont la transformation du cours de dessin en un cours d'arts plastiques qui développe davantage les facultés créatrices des enfants. Senécal saura intégrer avec doigté l'art aux matières scolaires et stimuler l'imagination créatrice.

Concernant l'évaluation des apprentissages entre 1923 et 1964, les cours de dessin se donnent principalement dans les écoles des beaux-arts, en atelier, et l'on évalue surtout le produit fini. Les objectifs pédagogiques visent la maîtrise des moyens d'expression liés à l'épanouissement de la personnalité. Lorsque ces écoles sont intégrées au système d'enseignement, il ne s'agit plus de former des artistes, mais de développer leur créativité, ce qui a un impact sur les manières d'évaluer. Par ailleurs, l'enseignement des arts devient une discipline de pointe : beaucoup d'emplois pour les professeurs d'arts sont créés en peu de temps, le cours de pédagogie artistique reçoit une reconnaissance officielle, par une résolution du rapport Parent, établissant le *brevet spécialisé dans l'enseignement des arts plastiques* (Couture *et al.*, 1980) et les méthodes d'enseignement dites actives puisées dans la pratique artistique sont à l'ordre du jour des classes d'arts plastiques. Tandis que les objectifs des cours deviennent plus spécifiques, l'évaluation des apprentissages s'oriente vers l'expérience créatrice et l'intention artistique.

2.2. La danse

L'historienne de la danse québécoise, Iro Valaskakis-Tembeck, a trouvé des traces d'un studio de ballet, vraisemblablement le plus ancien, autour de 1737 : Louis Renault y aurait enseigné à Montréal la danse classique jusqu'en 1749 (Valaskakis-Tembeck, 1991). Quelques écoles, que nous pouvons compter sur les doigts de la main, offrent des cours entre 1787 et 1863. Il faut dire qu'au Québec la danse est réprouvée par le clergé qui la limite à trois types : la danse traditionnelle folklorique, les bals et réceptions dansées ainsi que les revues musicales (Séguin, 1986). Les cours sont principalement de ballet, de maintien et de folklore. Ne subissant pas la pression ecclésiastique, les anglophones dansent plus que les francophones. De plus, les anglophones du Québec sont sensibilisés à la danse et à son enseignement principalement grâce aux universités américaines qui offrent des diplômes en danse à partir de 1926. L'Université McGill donne même des cours de danse créative et d'interprétation accrédités dès 1929 dans un programme d'éducation physique. Autour de 1930, les cours se développent autour de la danse sociale et à claquettes, et la venue d'immigrants, dont Loie Fuller, Frank Norman, Ruth St-Denis, Ezzak Ruvenoff et George Scheffler, favorise l'essor de la danse. Certains d'entre eux ne sont que de passage dans les salles de spectacles, alors que d'autres ouvrent des écoles et demeurent à Montréal. Toutefois, aucune troupe de danse n'existe et c'est le vaudeville qui est le plus populaire.

Vers 1940, Maurice Lacasse-Morenoff et Gérald Crevier ouvrent des écoles et forment notamment Fernand Nault et Françoise Sullivan, qui deviendront des figures marquantes pour le développement de la danse au Québec : le premier, au sein des Grands Ballets canadiens et la seconde, avec une pratique personnelle ouvrant la porte à la danse moderne. Ils comptent également sur l'appui d'une nouvelle vague d'immigrantes, Ruth Sorel, Elizabeth Leese et Séda Zaré, qui apportent dans leurs valises de nouvelles pratiques chorégraphiques. En somme, jusqu'à la moitié du XX^e^ siècle, l'enseignement de la danse se fait surtout dans des écoles privées qui forment de futurs danseurs ou qui offrent des cours pour les loisirs de la population. La formation des maîtres est à peu près inexistante, mis à part celle dispensée par Elizabeth Leese à des professeurs en danse classique et moderne dans son école entre 1945 et 1958 (Valaskakis-Tembeck, 1991).

La vague de contestation sociale qui monte à l'aube de 1960 au Québec entraînera dans son sillage l'enseignement de la danse. La danse de cette décennie est à l'image de la Révolution tranquille qui a cours au Québec : en pleine ébullition, mais également en expérimentation et en assimilation. En effet, les véritables pionnières de la danse au Québec, dont font partie Ludmilla Chiriaeff, Françoise Sullivan et Jeanne Renaud, sont peu nombreuses, mais elles rapportent de leurs voyages tout un bagage chorégraphique à explorer et à reconstruire. Elles défricheront la terre de la danse québécoise en rompant définitivement avec la pression du clergé, en s'inspirant des émotions, en véhiculant des messages sociaux. La première créera les Grands Ballets canadiens et aura toujours en tête d'intégrer la danse dans le cursus scolaire. Madame Chiriaeff sera notamment l'instigatrice d'un programme en concentration danse classique à l'école secondaire Pierre-Laporte (Forget, 2006). Les secondes traceront le chemin de la danse contemporaine, en passant par le moderne, en formant plusieurs danseurs et chorégraphes qui, à leur tour, transformeront le paysage de la danse au Québec. Entre 1970 et 1980, deux écoles importantes associées à deux compagnies de danse ouvriront leurs portes, le Groupe de la Place Royale et le Groupe Nouvelle Aire. Ces écoles constituent certes des terrains d'exploration fertile, mais elles sont axées sur la formation de danseurs et de chorégraphes professionnels plutôt que sur la formation des professeurs. De plus, elles ne sont pas incluses dans le système scolaire. Ce n'est qu'en 1981 que la danse apparaît dans les écoles primaires et secondaires.

En ce qui concerne l'évaluation en danse, jusqu'au milieu du siècle les écoles privées ne décernent pas de diplôme, à moins que le candidat n'ait reçu une formation basée sur une méthode reconnue (Cechetti, Dalcroze, Delsarte, Graham, etc.). Comme la relation maître-élève est

primordiale et que la danse s'apprend par imitation, l'évaluation se fait surtout par l'observation des capacités physiques, de l'expressivité, de la capacité à mémoriser corporellement et du sens de l'effort. Le processus d'audition pour faire partie d'une troupe ou pour entrer dans une école est déjà bien en place dans la culture de la danse. Il s'agit essentiellement de comparer les candidats entre eux en considérant les éléments précédemment nommés. L'évaluation se fait alors principalement par l'observation. Par exemple, pour sélectionner des candidats pour faire partie de la troupe des Grands Ballets canadiens, Ludmilla Chiriaeff évalue l'adhésion physique au style, la qualité de l'interprétation des œuvres de répertoire, les lignes corporelles et la performance globale (Forget, 2006). D'ailleurs, en 1972, elle est parmi les premières à proposer des qualifications pour les professeurs en danse classique au ministère de l'Éducation afin qu'il leur délivre un permis d'enseignement.

2. LA CRÉATION DES CÉGEPS

Entre 1965 et 1966, à l'initiative des étudiants de l'École des beaux-arts de Montréal soutenus par la suite par des professeurs et après moult péripéties, le gouvernement Lesage crée officiellement la commission Rioux, chargée d'étudier toutes les questions relatives à l'*enseignement des arts* (Corbo, 2006, p. 27), y compris les structures administratives et l'organisation matérielle. Cette commission se veut une réponse aux limites du rapport Parent au regard de l'enseignement des arts. Cet enseignement est éclaté, désorganisé entre le privé et le public et entre les écoles et les diffuseurs. Les acteurs de la commission Rioux reprendront le chantier abandonné par le rapport Parent en ce qui concerne les arts. Pour les revendicateurs, la place des arts dans un système moderne d'éducation n'est pas suffisante. Afin de réaliser l'homme polyvalent souhaité dans le rapport Parent, il faut accorder une plus grande place aux arts dans toutes les écoles et à tous les ordres d'enseignement et développer le passage d'une formation de la personnalité à une formation professionnelle. De plus, il est recommandé d'enraciner davantage l'enseignement des arts dans la réalité de la société québécoise, en pleine mouvance, et de développer une meilleure compréhension du rôle de l'art dans la société. La mission s'annonce donc colossale (Corbo, 2006).

Dès le début des travaux de la commission Rioux, plusieurs problèmes sont répertoriés : l'absence d'orientations générales et de pensée pédagogique, le manque de coordination entre les programmes et le manque de reconnaissance des diplômes, l'isolement des écoles d'art par rapport au réseau scolaire public, etc. Ces constatations du

milieu de l'enseignement des arts obligent les commissaires à réfléchir profondément à la place de l'art dans la société québécoise avant de formuler leurs recommandations au gouvernement. Et cette réflexion ne peut être ancrée que dans un remaniement complet du milieu de l'éducation artistique qui favorisera l'émergence d'une vision nouvelle sur la nature de l'art et sa fonction dans la société.

Au terme de la commission Rioux, en pleine tourmente de 1968, l'enseignement des arts sera sous la responsabilité du ministère de l'Éducation et il sera intégré à tous les ordres d'enseignement. La nouvelle Université du Québec à Montréal accueillera en son sein les programmes de l'École des beaux-arts de Montréal et l'Université Laval ceux de l'École des beaux-arts de Québec. De nouveaux programmes en psychopédagogie habiliteront les étudiants à enseigner les arts plastiques et les arts seront de plus en plus intégrés à l'environnement visuel de la collectivité (par exemple avec des œuvres majeures dans le métro de Montréal ou dans les jardins et les tours à bureaux).

Le rapport Rioux n'a certes pas fait l'unanimité, d'autant plus que sa mise en œuvre difficile a coïncidé avec plusieurs événements importants : décès du premier ministre Johnson, qui force des élections, mouvement de contestation étudiante, élaboration de projets de loi sur l'éducation, etc. Toutefois, bien que les recommandations du rapport Rioux n'aient pas toutes été mises en place, il ne faut pas en sous-estimer l'influence philosophique et conceptuelle *souterraine* sur tout l'enseignement des arts (Couture et Lemerise, 1990). Les conséquences de la commission Rioux sont variées et nombreuses. Parmi les aspects positifs, nous pouvons compter la valorisation accrue des arts dans la société ainsi que l'établissement de programmes de formation des maîtres en arts et de programmes d'éducation artistique dès le primaire, et ce, jusqu'au cégep. Parmi les aspects moins réjouissants, nous relevons un manque de moyens, une perte du contact intime entre le maître et l'élève, des conflits entre professeurs-artistes et pédagogues en arts qui perdurent et un changement de discours de l'expérience artistique vers la didactique.

3.1. Les particularités du réseau collégial québécois

Répondant à plusieurs constats sur les lacunes du système éducatif québécois d'alors, les cégeps participent à la démocratisation de l'éducation. Les commissaires de la commission Parent proposent de mettre en place une structure éducationnelle offrant une formation technique et une solide formation générale afin que chaque jeune Québécois puisse poursuivre les études les plus longues possible (Corriveau, 1991).

Il s'agit en somme de rendre l'enseignement public équivalent à l'enseignement privé et à unifier tout le système éducatif. Responsables de cet enseignement, les cégeps et, par le fait même, les professeurs qui y travaillent contribuent au développement de leur région. Entre 1967 et 1972, 46 cégeps étaient créés dans la foulée de ce virage important pour le réseau scolaire du Québec.

Malgré un démarrage difficile en raison des contestations étudiantes de 1968 et de la syndicalisation massive des professeurs durant la décennie 1970, aujourd'hui, pour chaque région du Québec, les cégeps constituent un pôle non seulement éducatif, mais aussi culturel, social, économique, sportif, scientifique et technologique. La mise en place du réseau des cégeps a permis à toutes les régions et à plusieurs communautés de bénéficier de ces apports. En 2010, 48 cégeps, dont 5 anglophones, forment le niveau collégial public au Québec. À ces cégeps publics s'ajoutent 11 autres établissements publics relevant d'autres ministères (tels que les conservatoires de musique et l'Institut technologique agroalimentaire) et 25 établissements privés subventionnés. Ils assurent la transition entre, d'une part, le secondaire et l'université et, d'autre part, le monde scolaire et le marché du travail.

Les programmes de formation technique (trois ans d'études menant au marché du travail) ou préuniversitaire (deux ans d'études menant à l'université) sont sous l'égide de départements qui regroupent généralement les professeurs d'une même discipline. Depuis 1993, les cégeps ont adopté l'approche par compétences afin de renouveler la formation, d'ajuster les programmes en fonction du marché du travail et de mettre en œuvre une véritable stratégie de réussite des études (Barbeau, 1995). C'est une décentralisation d'une partie du pouvoir du ministère de l'Éducation vers les collèges qui s'opère puisque les cégeps deviennent habilités à élaborer leurs propres programmes d'études par compétences. Ils doivent rendre des comptes à la Commission d'évaluation de l'enseignement collégial (CÉEC), organisme gouvernemental autonome dont la mission couvre la plupart des dimensions de l'enseignement collégial, avec un accent particulier mis sur les apprentissages et les programmes d'études.

3.2. La formation collégiale en arts plastiques

Le programme préuniversitaire en arts plastiques se donne dans 25 cégeps. Toutefois, plus de 41 cégeps offrent des formations, techniques ou préuniversitaires, dans des disciplines aussi variées que les arts visuels et médiatiques, le cinéma, les arts du cirque, la photographie,

les métiers d'arts, la musique et l'art dramatique. C'est donc dire qu'une majorité d'établissements collégiaux abordent les arts et que chacun propose une vision personnalisée des disciplines artistiques. Si nous considérons uniquement les arts plastiques, la formation permet à l'étudiant de s'initier en atelier à l'utilisation de différents matériaux, d'expérimenter la création à travers les formes, les dimensions et les couleurs et de réfléchir au sens de sa démarche pour arriver, éventuellement, à se définir en tant qu'artiste. L'intention est de développer la créativité de l'étudiant qui se caractérise par la curiosité, l'esprit de recherche et le sens de l'esthétique et de la critique. L'étudiant est appelé, tout au long de son parcours, à produire des œuvres d'art et un portfolio et à inscrire son travail en regard des grandes tendances de l'art contemporain. Les compétences s'échelonnent de la reproduction de procédés techniques à la réalisation et la diffusion d'une œuvre personnelle, en passant par l'interprétation du monde sensible et la reconnaissance des œuvres à différentes époques.

3.3. La formation collégiale en danse

Actuellement au Québec, quatre cégeps offrent une formation préuniversitaire en danse (Drummondville, Montmorency, Sherbrooke, Saint-Laurent) et deux proposent une formation technique en interprétation (Sainte-Foy et Vieux-Montréal). À l'exception de ces deux derniers cégeps, qui visent la maîtrise technique de la danse contemporaine, les formations offertes permettent d'atteindre un niveau de maîtrise en danse suffisant pour être admis dans les programmes universitaires en danse et dans les programmes professionnels menant à une carrière d'interprète, de chorégraphe ou de professeurs. Il s'agit d'apprendre les techniques de danse, l'interprétation et la création chorégraphique. Les cours sont donc 1) techniques (mieux connus sous l'expression *classe de danse*), 2) théoriques – l'histoire, la somatique, l'anatomie, la gestion de carrière – et 3) artistiques, où la composition, l'interprétation et l'appréciation sont enseignées. Chaque année, la production d'un spectacle de fin d'études vient couronner le cheminement des étudiants. C'est l'occasion idéale de réaliser la synthèse des apprentissages acquis et d'en faire la démonstration en concevant, produisant et diffusant un spectacle complet. Le ministère de l'Éducation, du Loisir et du Sport du Québec impose six compétences aux programmes de danse qui vont de la maîtrise des bases du mouvement dansé jusqu'à la création, l'interprétation et l'appréciation d'une chorégraphie.

4. REGARD SUR L'ÉVALUATION DES APPRENTISSAGES EN ARTS PLASTIQUES ET EN DANSE DEPUIS 1970

Au début des années 1980, un premier virage s'impose pour les collèges : la technologie fait son entrée avec, en tête de file, l'ordinateur. Ce renouveau technologique, précédant le renouveau pédagogique, apportera avec lui non seulement une évolution, mais aussi des retards notamment sur les plans du perfectionnement des professeurs, du renouvellement du corps professoral et de l'aménagement des classes pour s'adapter aux nouvelles réalités (Lamy, 1984). Les impacts de ce tournant obligé sur l'évaluation des apprentissages ne semblent pas avoir eu le temps de se matérialiser puisque, dès 1993, s'amorce une réflexion pour un second virage, soit celui vers les approches par compétences. Ce renouveau pédagogique a été mis en place en 1993, amenant avec lui la révision progressive des programmes d'études (s'échelonnant sur une douzaine d'années) et forçant les différents acteurs à revoir un certain nombre de pratiques, notamment au regard de l'évaluation des apprentissages. Notons que quelques recherches se sont intéressées aux changements des pratiques évaluatives issus de cette réforme (Chbat, 2004 ; Corriveau, 2005 ; Isabel, 2000 ; Leroux, 2010 ; Pineault, 2001).

Dans ce nouveau contexte, les professeurs du collégial ont maintenant la responsabilité d'évaluer l'atteinte de compétences complexes par leurs étudiants plutôt que l'atteinte de simples objectifs d'apprentissage (ministère de l'Éducation, 2003). Dans bien des cas, des changements importants doivent être apportés aux pratiques évaluatives, celles-ci étant considérées, entre autres dans les politiques, comme des savoir-faire méthodologiques qui reposent sur une collecte d'informations, une interprétation et un jugement attestant le degré d'atteinte de la compétence par l'étudiant (Leroux, 2010). De plus, les récentes politiques d'évaluation des apprentissages au primaire et au secondaire du ministère de l'Éducation du Québec (Gouvernement du Québec, 2003) soulignent l'importance du passage du paradigme de l'enseignement à celui de l'apprentissage et de la régulation de la démarche du professeur avec celle de l'étudiant. Selon ces politiques, l'évaluation devrait se faire en vue d'une contribution à la réussite éducative des étudiants (Maltais, Ross et Lafleur, 2010). Ce changement de vision suppose que l'évaluation des apprentissages ne soit plus considérée comme un moment distinct de l'enseignement, mais plutôt comme une interaction dynamique entre les actions du professeur et les apprentissages de l'étudiant. Au collégial s'ajoute l'obligation, à la suite de la réforme de 1993, d'administrer des épreuves synthèses à la fin de chacun des programmes d'études préuniversitaires et techniques. Ces épreuves synthèses de programme (ÉSP) doivent avoir une visée intégratrice de toutes les compétences développées dans le

programme d'études, y compris celles associées à la formation générale (langue première, langue seconde, philosophie et éducation physique). Cette visée d'intégration des compétences se retrouve également dans l'obligation d'évaluer les apprentissages par une épreuve finale dans chacun des cours dont la pondération doit être suffisante pour refléter cette intégration.

En somme, de nos jours, plusieurs documents de référence encadrent les droits, les devoirs et les obligations des intervenants concernés par l'évaluation des apprentissages. Toutefois, comme l'acte d'évaluer se situe avant tout dans la salle de classe, il suppose des prises d'informations régulières et parfois nombreuses pour rendre compte de la progression des apprentissages. Les professeurs doivent dorénavant utiliser des outils plus formels, en lien avec les politiques de leur institution, et structurer davantage leur démarche évaluative. En arts, l'évaluation des apprentissages porte autant sur la démarche que sur la production. Les professeurs évaluent aussi l'appréciation des œuvres d'art et utilisent de plus en plus l'évaluation formative, l'autoévaluation et l'évaluation par les pairs (Leroux, 2010).

Selon Chaîné et Bruneau (1998), l'évaluation des apprentissages, en arts plastiques comme en danse, peut porter sur deux éléments principaux : le faire et la perception esthétique. Évaluer en arts exige inévitablement de considérer ces deux éléments qu'Ardouin (1997) nomme la conduite esthétique. Dès lors, ce sont les attitudes qui deviennent le lieu d'apprentissage : comment l'étudiant se comporte-t-il face à l'œuvre artistique, dansée ou picturale, en tant qu'artiste et en tant qu'observateur ? Dans cette perspective, le faire renvoie à la dimension pratique et technique, par exemple exécuter une séquence dansée en respectant les consignes données, alors que la perception esthétique concerne la compétence à interpréter, à apprécier, à observer et à questionner ses pairs. Ainsi, la perception esthétique résulte du sens que l'on donne à ce que l'on voit et perçoit dans une œuvre d'art. Pour arriver à produire ce sens, l'étudiant doit faire appel à sa mémoire affective et sensorielle ; il doit l'affiner et développer ses habiletés de concentration et d'expression en plus d'expliquer et de rationaliser ce qu'il a fait ou ce qu'il voit. Comme nous le constatons, le professeur en arts au collégial doit actuellement évaluer des apprentissages beaucoup plus complexes que ceux que l'on pouvait faire dans les salles de classe des années 1940. À cela s'ajoute le fait que les domaines artistiques posent souvent des problèmes particuliers pour l'évaluateur, tels que l'obtention de traces matérielles et les biais associés à la subjectivité (Hadji, 1997). Ceux-ci doivent être considérés dans les interventions pédagogiques, puisqu'ils participent à la complexité d'évaluer et donc à celle d'enseigner pour que la réussite des évaluations reflète les apprentissages.

Une documentation sur l'évaluation des apprentissages en arts plastiques au primaire et au secondaire existe, sans être abondante, principalement en raison des exigences ministérielles. Toutefois, bien que des recherches soient menées pour ces niveaux (notamment Gagnon-Bourget, 2000 ; Gosselin, 2006 ; Gosselin, Potvin, Gingras et Murphy, 1998 ; Lavender, 1996 ; Richard, 2005 ; Smith-Autard, 1994), leurs fruits sont plus difficilement accessibles aux professeurs des collèges. L'une des raisons est que les objectifs de formation diffèrent. Au primaire et au secondaire, il s'agit surtout d'encourager chez les élèves des aptitudes à créer et à développer, ce que Gosselin (2006) appelle un *équilibre émotivorationnel* (p. 18) ; il ne s'agit pas d'en faire des artistes. Au collégial, la formation vise plutôt à ce que l'étudiant développe une capacité à créer, aiguise son sens critique et acquière des habiletés techniques suffisantes pour participer à des activités artistiques telles que des expositions. Évidemment, la formation au collégial est plus spécifique et prépare les jeunes adultes à s'engager dans une voie artistique. Plusieurs professeurs se servent du portfolio professionnel, d'apprentissage ou de présentation en format papier et, de plus en plus, en format électronique (Leroux, 2010). Malheureusement, l'évaluation des apprentissages en arts plastiques au collégial est un terrain de recherche trop peu documenté et il est nécessaire de s'inspirer des travaux menés sur les enseignements primaire et secondaire pour tenter de développer ce champ.

En danse, la littérature sur l'évaluation s'articule surtout autour de recherches réalisées un peu partout dans le monde en classe de danse au primaire et au secondaire (dont Lavender, 1996 ; Smith-Autard, 1994). À l'échelle du Québec, une poignée de chercheurs s'intéressent à l'évaluation des apprentissages dans ce domaine, généralement par le biais de l'éducation esthétique (Bruneau, 1993 ; Chaîné et Bruneau, 1998 ; Émond et Raymond, 2006 ; Lord, 1998). Tout comme en arts plastiques, bien que ces auteurs abordent l'évaluation en salle de classe au primaire et au secondaire, certains de leurs travaux peuvent nous être utiles pour parler des pratiques évaluatives en danse au collégial.

Par exemple, Bruneau (1993) expose le malaise de certains professeurs de danse au secondaire à évaluer leurs élèves, malgré leur brevet en enseignement de la danse. L'expertise du professeur repose en grande partie sur sa formation pratique et sur sa connaissance du milieu de la danse : *ses compétences pratiques présument de ses compétences pédagogiques* (Bruneau, 1993, p. 695). De plus, la danse est encore souvent enseignée comme un langage technique et physique qui s'appuie essentiellement sur la tradition orale et sur la transmission d'un corps à un autre. Dans une perspective d'évaluation des apprentissages, le professeur en danse se voit dans l'obligation de rendre compte de

l'atteinte d'objectifs, de témoigner de la progression des apprentissages, de reconnaître les principes d'appréciation, d'encadrer les élèves dans la réalisation de compositions dansées, etc. Bref, l'évaluation des apprentissages en classe de danse est complexe et fort différente de la réalité d'une pratique artistique professionnelle. Il en va de même pour les professeurs de danse au cégep, souvent moins formés que ceux du primaire et du secondaire en évaluation des apprentissages.

Face aux multiples batailles que les professeurs doivent mener simplement pour conserver leur fragile programme de danse, optionnel au secondaire, Émond et Raymond (2006) constatent qu'ils cèdent à la tentation de ne pas changer leurs pratiques évaluatives fortement ancrées dans le paradigme enseignement. La réforme scolaire au Québec des années 2000, imposée au primaire et au secondaire, a provoqué un changement de vision important pour les professeurs : ils ne vérifient plus seulement les talents techniques des élèves, mais aussi leurs réponses aux attentes, leurs démarches dans la résolution de problèmes, leur engagement dans la tâche d'évaluation. Autrement dit, le passage vers le paradigme apprentissage se fait avec certaines résistances chez les professeurs en danse au secondaire. Ceux-ci, comme ceux en arts plastiques, doivent considérer l'évaluation différemment et voir en elle une condition nécessaire à l'apprentissage (Sarrasin, 2006). Chez les professeurs du collégial, une étude (Leduc, Blais et Raîche, 2011) révèle au contraire une ouverture et une volonté de leur part d'adapter leurs pratiques évaluatives aux nouvelles réalités et exigences institutionnelles tout en tenant compte de multiples facteurs, comme l'hétérogénéité des classes (notamment sur le plan du degré de maîtrise du mouvement) et les connaissances inégales des étudiants sur la danse. Les difficultés en matière d'évaluation ressenties par les professeurs en arts au collégial sont également soulignées par d'autres auteurs et mettent généralement en cause un sentiment de manque de compétence, d'instrumentation et de soutien ainsi qu'une non-valorisation publique de l'éducation artistique (Corbo, 2006 ; Flibotte, 1987 ; Grégoire, 1987). Ces préoccupations au sujet de l'évaluation sont légitimes et compréhensibles, surtout si l'on considère le manque de formation en évaluation des apprentissages chez ces professeurs.

5. CONCLUSION

Au cours de la décennie 1960, on assiste au Québec à l'aboutissement d'un vaste mouvement de société visant à réformer en profondeur le système éducatif de la province. Les commissions Parent et Rioux font un état des lieux et déterminent les actions à réaliser pour concrétiser cette vision d'un Québec moderne. Avec ces changements viennent la

création des cégeps et la restructuration des écoles d'arts. Le paysage global actuel du système éducatif québécois est à peu près le même que celui instauré à la suite des recommandations du rapport Parent. Cependant, quelques soubresauts en teintent les structures. En font foi le remplacement des programmes en 1980 par les nouvelles orientations ministérielles et l'ajout des approches par compétences au milieu des années 1990 pour les cégeps et au tournant des années 2000 pour le primaire et le secondaire.

L'instauration des approches par compétences dans les cégeps a obligé les professeurs du collégial à voir l'évaluation des apprentissages sous un nouvel angle et à la considérer comme une action sensible à la progression des étudiants et comme une opération complexe et permanente, voire continue pour le professeur. Par l'intermédiaire des politiques institutionnelles d'évaluation des apprentissages (PIEA), également mises en place autour de 1995, les orientations de l'établissement en matière d'évaluation sont précisées, un processus commun est défini et l'encadrement des pratiques évaluatives est soutenu. L'évaluation a alors comme fonction l'aide à l'apprentissage.

Pour les professeurs du collégial en arts plastiques et en danse contemporaine, l'approche par compétences implique notamment la création d'outils pédagogiques leur permettant d'obtenir de la part des étudiants plus de traces et de données observables pour évaluer la précision des critères sur lesquels ils feront porter leur jugement et la mise en place de divers moyens de communication avec les étudiants. Tout en tenant compte des politiques et des nouvelles manières de faire, ils doivent varier leurs modalités d'évaluation pour s'assurer que les compétences à créer, à interpréter et à apprécier sont bien atteintes. Ils doivent enfin dépasser les difficultés qui se présentent dans la pratique évaluative en arts pour intégrer davantage la fonction formative de l'évaluation dans leur enseignement.

Malgré certaines critiques encore exprimées aujourd'hui sur la légitimité de cet ordre d'enseignement, la formation collégiale, marquée par une certaine fragilité par sa position intermédiaire entre le secondaire et l'université, est dorénavant bien ancrée dans le paysage de l'éducation au Québec. Nous l'avons mentionné, à la fin du XX^{e} siècle le ministère de l'Éducation du Québec a imposé l'approche par compétences au primaire et au secondaire. Onze années se sont écoulées depuis. Les élèves qui ont vécu tout leur parcours scolaire avec ces approches centrées sur les compétences entrent maintenant au cégep. Les professeurs les attendent bien préparés, mais le défi est grand et les inquiétudes sont nombreuses.

RÉFÉRENCES

Ardouin, I. (1997). *L'éducation artistique à l'école*. Paris, France : ESF éditeur.

Barbeau, D. (1995). *Analyse de déterminants et d'indicateurs de la motivation scolaire d'élèves du collégial*. Actes du XV[e] Colloque de l'AQPC. Rivière-du-Loup, Québec : Association québécoise de pédagogie collégiale.

Bruneau, M. (1993). L'évaluation des apprentissages en danse : une utopie ? *Revue des sciences de l'éducation*, *19*(4), 695-713.

Chaîné, F. et Bruneau, M. (1998). Introduction : de la pratique artistique à la formation d'enseignants en art. *Revue des sciences de l'éducation*, *24*(3), 475-486.

Chbat, J. (2004). *Les attitudes et les pratiques pédagogiques du collégial*. Montréal, Québec : Collège André-Grasset, Direction pédagogique, Service de recherche (PAREA).

Corbo, C. (2006). *Le rapport Rioux et l'enseignement des arts au Québec, 1966-1968*. Québec, Québec : Septentrion.

Corriveau, G. (2005). *L'équilibre des compromis ou les attentes négociées au cœur de l'évaluation en enseignement des sciences au collégial*. Mémoire de maîtrise inédit. Trois-Rivières, Québec : Université du Québec à Trois-Rivières.

Corriveau, L. (1991). *Les cégeps : question d'avenir*. Québec, Québec : Institut québécois de recherche sur la culture.

Couture, F., Joyal, B., Landry, L., Lemerise, S., Lussier, C. et Wallot, A. (1980). *L'enseignement des arts au Québec*. Montréal, Québec : Université du Québec à Montréal.

Couture, F. et Lemerise, S. (1990). A social history of art and public art education in Quebec : the 1960's. *Studies in art education*, *31*(4), 226-233.

Émond, L. et Raymond, C. (2006). L'évaluation des apprentissages en arts : dénouer ou trancher le nœud gordien ? *Vie pédagogique*, *141*, 40-44.

Forget, N. (2006). *Chiriaeff : danser pour ne pas mourir*. Montréal, Québec : Québec Amérique.

Flibotte, C. (1987). *L'enseignement des arts plastiques au Québec de 1930 à 1987*. Mémoire de maîtrise inédit. Montréal, Québec : Université du Québec à Montréal.

Gagnon-Bourget, F. (2000). *Engager le dialogue matériel en classe d'arts plastiques*. Actes du colloque du CRED. Sherbrooke, Québec : Collectif de recherche en éducation artistique.

Gosselin, P. (2006). Des aptitudes sollicitées et développées à différents moments de la démarche de création. *Vie pédagogique*, *141*, 17-20.

Gosselin, P., Potvin, G., Gingras, J.-M. et Murphy, S. (1998). Une représentation de la dynamique de création pour le renouvellement des pratiques en éducation artistique. *Revue des sciences de l'éducation*, *24*(3), 647-666.

Gouvernement du Québec (1967). *L'enseignement collégial et les collèges d'enseignement général et professionnel : documents d'éducation, 3*. Québec, Québec : ministère de l'Éducation.

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

Graveline, P. (2007). *Une histoire de l'éducation au Québec*. Montréal, Québec : Bibliothèque québécoise.

Grégoire, R. (1987). *L'enseignement des arts tel qu'il se pratique, selon 30 enseignants et enseignantes, dans les écoles publiques francophones du Québec*. Québec, Québec : Conseil supérieur de l'éducation.

Hadji, C. (1997). *L'évaluation démystifiée : mettre l'évaluation scolaire au service des apprentissages*. Paris, France : ESF éditeur.

Isabel, B. (2000). *Les changements de pratiques d'évaluation des apprentissages chez les enseignants de philosophie et de français dans le contexte du renouveau de l'enseignement collégial : une étude de 11 cas dans un collège*. Thèse de doctorat inédite. Montréal, Québec : Université du Québec à Montréal et Université du Québec à Rimouski.

Lamy, F. (1984). *Problématique en arts appliqués. Rapport de recherche*. Québec, Québec : ministère de l'éducation, Direction générale de l'enseignement collégial.

Laurin, M. (2007). *Anthologie de la littérature québécoise*. Anjou, Québec : Éditions CEC.

Lavender, L. (1996). *Dancers talking dance : critical evaluation in the choreography class*. Albuquerque, Mexique : University of Mexico Press.

Leduc, D., Blais, J.-G. et Raîche, G. (2011). *L'évaluation en arts à l'enseignement supérieur au Québec : vers une intégration des pratiques*. Actes du IV[e] congrès Questions de pédagogie dans l'enseignement supérieur (QPES). Angers, France.

Leroux, J.-L. (2010). *L'évaluation des compétences au collégial : un regard sur des pratiques évaluatives*. Saint-Hyacinthe, Québec : Cégep de Saint-Hyacinthe.

Lord, M. (1998). Enseigner la danse à l'école secondaire : lier la théorie et la pratique d'éducation esthétique. *Revue des sciences de l'éducation*, *24*(3), 585-603.

Maltais, J.-F., Ross, P. et Lafleur, M. (2010). *Fiches d'information du Cégep Limoilou sur le renouveau pédagogique*. Montréal, Québec : Fédération des cégeps, Carrefour de la réussite.

Pineault, M.-C. (2001). *Pratiques pédagogiques et approche par compétences dans le programme Sciences humaines du cégep de Rimouski*. Mémoire de maîtrise inédit. Rimouski, Québec : Université du Québec à Rimouski.

Richard, M. (2005). *Culture populaire et enseignement des arts : jeux et reflets d'identité*. Québec, Québec : Presses de l'Université du Québec.

Sarrasin, L. (2006). L'évaluation : au cœur de l'enseignement et de l'apprentissage. Entrevue avec Christian Rousseau. *Vie pédagogique*, *141,* 47-49.

Séguin, R.-L. (1986). *La danse traditionnelle au Québec*. Québec, Québec : Presses de l'Université du Québec.

Smith-Autard, J. M. (1994). *The art of dance in education*. London, Ontario : A & C Black publishers.

Valaskakis-Tembeck, I. (1991). *Danser à Montréal : germination d'une histoire chorégraphique*. Québec, Québec : Presses de l'Université du Québec.

Chapitre 6

Authenticité des tâches d'évaluation en milieu scolaire

État des lieux

Pascal Ndinga

Afin de faire le point sur le concept de l'authenticité des tâches d'évaluation en éducation, dix ans après son introduction dans la foulée de l'implantation du renouveau pédagogique au Québec, nous avons réalisé une revue de littérature. Les écrits consultés et l'analyse qui en a découlé nous ont permis de constater l'état de quasi-friche de cette question, tant sur le plan de la pratique que sur celui de la recherche. Ces écrits se situent encore au stade de la caractérisation. De plus, la quasi-totalité de ceux-ci affiche une filiation avérée avec les assertions énoncées par Wiggins (1989), signe du piètre niveau de développement scientifique dans ce domaine. La recherche doit transcender la caractérisation ambiante, élaborer et valider des outils d'évaluation afin d'entreprendre des études corrélationnelles et expérimentales. Ce défi incombe particulièrement aux chercheurs du domaine de l'évaluation en éducation.

1. INTRODUCTION

L'authenticité des tâches d'évaluation (signifiance des tâches, au primaire et au secondaire) est l'un des concepts phares de la réforme, introduits dans les écoles du Québec dans la foulée de

l'implantation du renouveau pédagogique et des approches par compétences. Une décennie après l'entrée en vigueur de ce courant pédagogique, quel portrait peut-on en dresser ? Il nous semble pertinent de procéder à une recension des écrits afin d'esquisser un état des lieux sur ce sujet. Cette synthèse des connaissances sur le concept de l'authenticité des tâches d'évaluation vise aussi à jeter un nouvel éclairage sur les réalisations s'y rapportant dans le domaine de la recherche en éducation. Ainsi, on cernera mieux les défis à relever à cet égard. Dans les lignes qui suivent, nous traiterons d'abord de la définition du concept de l'authenticité des tâches d'évaluation et des réalisations relevées dans la littérature. Suivront les défis d'ordre pratique, observés dans l'application des tâches authentiques en milieu scolaire. L'état des lieux de la recherche viendra compléter le tableau de cette synthèse des connaissances sur le concept de l'authenticité des tâches d'évaluation. Enfin, une conclusion résumera le tout.

Mais avant d'examiner la définition du concept de l'authenticité des tâches, telle que donnée par les différents auteurs dont nous avons consulté les écrits, exposons en quelques lignes la méthode de recherche et d'analyse qui a présidé à cet exercice.

2. MÉTHODE DE RECHERCHE ET D'ANALYSE DOCUMENTAIRE

Précisons d'entrée de jeu que nous avons limité notre travail au domaine de la recherche en éducation, c'est-à-dire au niveau des activités scolaires. Ainsi, les mots clés suivants furent utilisés tantôt en anglais, tantôt en français. Il s'agit de *authentic task; authentic assessment; authentic evaluation, significance task* (en anglais) ou en français : authenticité des tâches, évaluation authentique, tâches authentiques, tâches signifiantes. Ces termes ont aussi fait l'objet de combinaison avec des mots comme éducation, apprentissages, etc., afin de répertorier le plus grand nombre de documents ou d'articles possible. Les moteurs de recherche utilisés vont du répertoire en ligne aux bases de données. Ainsi, nous avons consulté ERIC, Google Scholar, etc. Des documents répertoriés, nous n'avons retenu que ceux qui étaient pertinents, c'est-à-dire ceux qui abordent le concept de l'authenticité des tâches sous un des aspects suivants : la conception et la validation d'outils, l'évaluation proprement dite, l'analyse des impacts de l'utilisation des tâches authentiques, etc. C'est sur cette base que fut réalisée cette recension des écrits. À présent, voyons ce qu'il en est de la définition.

3. DÉFINITION DU CONCEPT ET RÉALISATIONS

La littérature disponible en évaluation consacre très peu d'espace au concept de *tâche authentique* proprement dit. En revanche, il est courant de trouver des écrits qui traitent du concept d'évaluation authentique. Ainsi, c'est essentiellement à travers les écrits portant sur l'évaluation authentique que nous tenterons de cerner la nature des tâches authentiques appliquées à l'école. Relevons, d'entrée de jeu, qu'une tâche évaluative ou une situation d'évaluation constitue une *tâche spécifique assignée à un élève dans le but d'apprécier sa performance et de porter un jugement* (Legendre, 2005, p. 1318).

Dans leur présentation des cinq dimensions de la structure de l'évaluation authentique, Gulikers, Bastiaens et Kirschner (2004, p. 71) ont aussi défini comme suit la tâche authentique : « *An authentic task is a problem task that confronts students with activities that are also carried out in professional practice* […] *as a task that resembles the criterion task with respect to the integration of knowledge, skills, and attitudes, its complexity, and its ownership.* »

Cette définition met en exergue les critères fondamentaux suivant lesquels une tâche complexe peut être considérée comme authentique. Il est à relever aussi le caractère défiant de la tâche puisqu'à travers elle l'étudiant doit être confronté à la nécessité de démontrer son savoir, son savoir-être ainsi que son savoir-faire en mobilisant de façon autonome toutes les ressources pertinentes (internes ou externes) à la résolution efficace du problème. Dans cette perspective, la tâche ainsi proposée doit nécessairement être intégratrice. C'est ce qu'a semblé soutenir Jonassen (1992, p. 140) dans ce qui suit : « *Authentic tasks are those that have real word relevance and utility, that integrate those tasks accross the curriculum, that provide levels of difficulty or involvement.* »

Herrington, Oliver et Reeves (2006) ont abondé dans le même sens en proposant 10 caractéristiques d'activités authentiques. Celles-ci furent conçues pour les environnements d'apprentissage en ligne. Selon ces chercheurs, les tâches authentiques comportent des activités complexes devant être accomplies par les apprenants durant une période appréciable (une journée, une semaine) et non seulement en quelques heures, comme le seraient les tâches ponctuelles. Or, le temps alloué aux tâches complexes constitue l'une des limites importantes, selon Montgomery (2002). Les professeurs se plaignent du peu de temps dont ils disposent pour déployer et faire réaliser les apprentissages d'un contenu curriculaire de plus en plus chargé. À cela s'ajoute la volonté politique et sociale d'intégrer les élèves en difficulté dans les classes ordinaires, d'une part, et le principe du respect du rythme d'apprentissage de chaque élève, indépendamment de ses caractéristiques

particulières, d'autre part. Le défi à relever pour opérationnaliser les tâches authentiques s'annonce plutôt difficile. Dans les circonstances, il est pertinent de se demander s'il est réaliste d'envisager l'application des taches authentiques dans tous les milieux scolaires.

Pour Paris et Ayres (2000, p. 193), l'évaluation authentique (*authentic assessment*) signifie : « façons multiples d'évaluer l'apprentissage, les résultats, la motivation et les attitudes des élèves qui sont pertinentes par rapport aux objectifs fixés, aux programmes et aux méthodes adoptés en classe ». On notera ici la portée contextuelle et globale de l'évaluation authentique. En d'autres termes, l'évaluation authentique ne se limite pas au produit et au processus, mais elle englobe tous les facteurs et toutes les variables qui concourent à la réalisation du produit. Cela confère à l'évaluation authentique l'attribut de *terme générique* fréquemment employé sur le plan international pour décrire une série de nouvelles approches d'évaluation (Torrance, 1995, p. 1). Par ailleurs, ce dernier auteur rallie notamment Montgomery (2002) en soutenant que les tâches d'évaluation destinées aux élèves en contexte d'évaluation authentique devraient être plus pratiques, réalistes et défiantes que celles des tests traditionnels de type papier-crayon.

On le voit, à mesure que se précise la définition du terme *tâches authentiques*, l'analogie à établir avec *le concept de tâches complexes* est de plus en plus évidente. Ainsi, traitant de l'évaluation en situation authentique, la Société de gestion du réseau informatisé des commissions scolaires (2000) a défini une douzaine de caractéristiques relatives aux situations d'apprentissage ou d'évaluation authentique. Ces caractéristiques se présentent comme suit :

1. La situation tient compte des intérêts des élèves.
2. La situation tient compte des connaissances antérieures des élèves.
3. Les élèves doivent résoudre des problèmes réels ou simulés susceptibles d'être rencontrés à l'école ou dans la vie à l'extérieur de l'école.
4. L'élève doit faire une ou plusieurs tâches qui permettront d'observer sa démarche et lui demanderont de réaliser une ou des productions.
5. La ou les tâches sollicitent plusieurs compétences.
6. Pour réaliser la ou les tâches, l'élève doit mobiliser plusieurs ressources : notions, stratégies, attitudes, etc.
7. Les élèves font appel à leur créativité et produisent des réponses originales.

8. La situation incite les élèves à travailler en équipe ou à collaborer entre eux.
9. Les élèves ont accès à diverses ressources : livres, personnes, logiciels, etc.
10. Les productions sont destinées à un public (élèves de la classe, élèves des autres classes, parents, etc.).
11. Les élèves ont le temps nécessaire pour réaliser leur tâche. La durée est variable : quelques périodes, jours, semaines, mois, etc.
12. Le professeur utilise plusieurs critères pour juger de l'efficacité de la démarche et de la qualité de la production. Les critères d'évaluation sont connus des élèves.

À travers ces caractéristiques et en référence à la définition précitée de Legendre (2005), seules deux d'entre elles (5 et 6) traitent spécifiquement et explicitement des tâches d'évaluation. Une lecture attentive de celles-ci laisse voir nettement la nature propre à une tâche complexe. En effet, rappelons-le, la tâche complexe requiert la mobilisation de plusieurs ressources et implique plusieurs compétences. Incidemment, cette description des situations d'évaluation par la Société de gestion du réseau informatisé des commissions scolaires fut produite au moment de l'entrée en vigueur du renouveau pédagogique au Québec. Cet organisme demeure actif dans la mise en place et l'animation d'ateliers de formation en milieu scolaire sur l'évaluation des compétences et son corollaire, les tâches complexes.

Signalons en outre que l'ensemble des caractéristiques de la situation d'évaluation authentique, déclinées par la Société de gestion du réseau informatisé des commissions scolaires (2000), semble avoir été directement inspiré des conditions relatives à l'évaluation authentique, telles que décrites dix ans plutôt par Wiggins (1989). Pour s'en rendre compte, exposons d'abord les principales facettes de cette évaluation authentique, selon une traduction faite par Perrenoud (2001, p. 26) :

1. L'évaluation n'inclut que les tâches contextuelles.
2. L'évaluation porte sur des problèmes complexes.
3. L'évaluation doit contribuer à ce que les étudiants développent davantage leurs compétences.
4. L'évaluation exige l'utilisation fonctionnelle de connaissances disciplinaires.
5. Il n'y a aucune limite de temps fixée arbitrairement lors de l'évaluation des compétences.
6. La tâche et ses exigences sont connues avant la situation d'évaluation.

7. L'évaluation exige une certaine forme de collaboration avec des pairs.
8. La correction prend en considération les stratégies cognitives et métacognitives utilisées par les étudiants.
9. La correction ne tient compte que des erreurs importantes dans l'optique de la construction des compétences.
10. Les critères de correction sont déterminés en faisant référence aux exigences cognitives des compétences visées.
11. L'autoévaluation fait partie de l'évaluation.
12. Les critères de correction sont multiples et donnent lieu à plusieurs informations sur les compétences évaluées.

En comparant ces deux séries de caractéristiques, il est aisé de voir leur affinité. Par exemple, la notion du temps de réalisation de la tâche, abordée au numéro 11 par la Société de gestion du réseau informatisé des commissions scolaires (2000), fut traitée au numéro 5 dans la série de caractéristiques proposée par Wiggins (1989). De même, les critères d'évaluation des tâches qu'on retrouve au numéro 12 à la Société de gestion du réseau informatisé des commissions scolaires (2000) furent présentés aux numéros 6 et 10 de la proposition de Wiggins (1989). On pourrait poursuivre cet exercice de comparaison des propositions et l'appariement des caractéristiques qui les composent se ferait aisément.

Retenons dans ce qui précède que la définition de l'évaluation authentique se fait en déclinant ses différentes facettes ou caractéristiques. À travers celles-ci, on voit se profiler les critères des tâches authentiques. Ainsi, tant dans la série des caractéristiques de l'évaluation authentique proposées par la Société de gestion du réseau informatisé des commissions scolaires (2000, p. 5, 6, et 11) que dans celle élaborée par Wiggins (1989, p. 3, 6, 10 et 12), les critères des tâches authentiques sont évoqués en filigrane. À travers ces caractéristiques de l'évaluation authentique, on relève trois conditions majeures : 1) les élèves sont placés dans une relative autonomie d'exécution de la tâche ; 2) le produit fini ou leurs réalisations varient d'un élève à l'autre ; 3) l'usage de critères préétablis est nécessaire pour assurer une évaluation équitable de ces réalisations. Telles sont les conditions *sine qua non* pour reconnaître qu'une tâche est complexe (Tarackdjian, 2003).

Pour Perrenoud (2001), cette conception de l'évaluation authentique, bien que pertinente pour les connaissances, peut être un moyen efficace de réconcilier observation formative et évaluation certificative. Selon lui, il ne s'agit plus de simuler des situations de travail (fictives, mais réalistes) en s'appuyant notamment sur le progrès technologique, mais plutôt d'accorder au formateur la place qui lui revient

afin qu'il procède à la fois à l'observation formative et à l'évaluation certificative. Ainsi, grâce à leur rôle déterminant auprès des apprenants ou des étudiants, les formateurs rigoureux et qualifiés pour observer et évaluer, qui pratiquent une observation continue, sont mieux placés que quiconque pour réaliser une évaluation authentique.

C'est en tant que formateur qu'il est possible de conjuguer adéquatement l'observation et l'évaluation certificative, car celles-ci s'opèrent sur la base des mêmes données. Cette assertion situe l'évaluation authentique au niveau même de la collecte des données brutes, reflet direct et immédiat de la réalité vécue (Legendre, 2005). Perrenoud soutient que ce qui est conforme à la réalité, vraisemblable, est authentique. Ainsi, l'authenticité est contraire à toute fabrication ou idéalisation théorique. Elle peut donc découler d'une situation simulée, pourvu que celle-ci soit foncièrement réaliste. Plus une simulation tend à refléter la réalité (vécu professionnel ou parascolaire quelconque), plus elle tend vers l'authenticité.

Ce réalisme dans les tâches d'évaluation constitue l'une des conditions *sine qua non* pour en reconnaître le caractère authentique. Des auteurs l'ont évoqué (explicitement ou implicitement selon les cas) en tentant de définir les tâches authentiques. En effet, l'authenticité des tâches d'évaluation est indispensable pour atteindre le niveau d'expertise requis pour la résolution de problème, car cela permet de s'occuper tant des compétences disciplinaires que du processus réflexif, qui caractérise une performance éprouvée, selon Gielen, Dochy et Dierick (2003 : voir Gulikers, Bastiaens, et Kirschner, 2006). C'est précisément dans ce créneau que se fonde l'évaluation des compétences, à savoir le fait de prendre en considération la qualité du produit réalisé, mais aussi le processus ayant mené à la réalisation de ce produit. Cette considération comporte des avantages salutaires. De fait, selon Janesick (2006, p. 5-6), une tâche d'évaluation authentique est conçue de manière à fournir une approche plus riche, plus forte et plus complexe que la tâche d'un test traditionnel pour comprendre la progression de l'étudiant. Cette auteure a défini comme suit les caractéristiques de la tâche authentique : *1) Requires quality performance or product ; 2) Is connected to the student's world ; 3) Is complex and multilayered ; 4) Is continuing with multiple tasks ; 5) Provides complex feedback continually recurring ; as the student self-adjusts, performance is improved ; 6) Looks for higher-order skills with a demonstration of knowledge.*

Ici aussi, cette description traduit précisément les conditions essentielles de la tâche authentique, déclinées plus tôt par Wiggins (1998), que Janesick (2006, p. 4) a cité : *1) It is realistic. The assessment task should follow closely the ways in which a person's abilities are "tested"*

in the real world; 2) It requires judgment and innovation; 3) It asks the student to do the subject; 4) It replicates or simulates actual "test" in the workplace, personal life, and civic life; 5) It assesses the student's ability and skills to effectively and efficiently use a repertoire of many skills to complete a problem or task; 6) It allows many opportunities to practice, rehearse, consult, get feedback, and refine actual performances and productions.

À ce réalisme dominant des tâches authentiques, Montgomery (2001, p. 4) ajoute la visée au potentiel élevé de défis, nécessaire dans la résolution des problèmes de haut niveau de complexité. Sa définition de la tâche authentique se lit de la façon qui suit. *Authentic task: this is a real-life activity, performance, or challenge that mirrors those faced by experts in the particular field; it is complex and multidimensional and requires higher levels of cognitive thinking such as problem solving and critical thinking.* Ainsi, l'expression « tâche authentique », en plus de renvoyer au concept de *tâche complexe*, concernerait davantage les tâches de niveau de complexité élevé. En outre, tout en considérant le produit et le processus, l'évaluation des tâches authentiques tient également compte du contexte global dans lequel la performance est réalisée. Dans ce même ordre d'idées, St-Germain (2000, p. 5) a tracé un portrait prescriptif des tâches authentiques qui comporteraient huit caractéristiques précises :

1. Les tâches authentiques s'attardent à la qualité du rendement, mais aussi à sa justification. L'évaluation ne se limite pas à l'exactitude des réponses, mais aussi à l'explication, à la justification et à la démonstration que peut en faire l'élève.
2. Elles ne doivent pas constituer des pièges, mais plutôt être connues à l'avance de l'élève. Celui-ci en connaît les critères d'évaluation.
3. Elles doivent être rattachées à des situations réelles et ancrées dans le vécu de l'élève.
4. Elles doivent être de véritables défis en intégrant plusieurs éléments de savoir et de jugement. Elles comportent de multiples facettes et permettent une ouverture à la créativité.
5. En intégrant plusieurs critères et exigences complexes, elles sont évidemment plus difficiles à corriger. Elles ne doivent toutefois pas écarter la validité au profit de la fiabilité.
6. Comme elles mesurent les apprentissages sur une longue période de temps, elles sont répétées souvent afin de réellement évaluer si l'élève a acquis les connaissances de façon durable. Elles donnent aussi à l'élève le droit de faire des erreurs et de se corriger ultérieurement.

7. Elles ont une valeur évidente aux yeux de l'élève et lui paraissent valables à première vue, puisqu'elles ont été validées en tenant compte de défis fondés sur la discipline. Elles suscitent donc l'intérêt et la persévérance.
8. Elles ne sont pas limitées à la seule vérification du rendement, mais servent surtout à l'élève et à son apprentissage. L'élève peut confirmer les résultats et s'ajuster au besoin grâce à cette rétroaction.

Selon Clauw, Dufays, Thyrion, Vercruysse, Carlier et Paquay (2006, p. 118), un apprentissage authentique implique des situations qui ont un sens pour l'apprenant. Ils ont ainsi relevé les indicateurs favorables à un tel apprentissage : « Elles doivent interpeller l'élève, lui poser un défi, lui paraître utiles, lui permettre de contextualiser les savoirs, l'amener à une réflexion épistémologique et mettre en évidence l'apport des autres disciplines dans la résolution de la tâche. » Fondamentalement, ces indicateurs s'apparentent à ceux contenus dans la grille d'analyse de la tâche authentique proposée par Sauvé (2000), lesquels découlent des caractéristiques des tâches authentiques décrites par plusieurs autres auteurs (Lusignan 2002 ; Perrenoud, 2001 ; Société de gestion du réseau informatisé des commissions scolaires, 2000 ; St-Germain, 2000).

Précisons qu'une constante subsiste à travers la kyrielle de définitions des tâches authentiques ainsi répertoriées. Il s'agit de la parenté flagrante entre celles-ci et les assertions avancées par Wiggins (1989). La figure suivante résume les caractéristiques des tâches authentiques.

En somme, la recension des écrits des dix dernières années, que nous avons pu consulter, sur l'authenticité des évaluations révèle une parenté flagrante entre eux. La quasi-totalité de ceux-ci gravite autour du scénario descriptif de Wiggins (1989). En effet, la littérature semble unanime pour accorder à Wiggins (1989) la paternité des fondements de la forme actuelle du concept d'évaluation authentique (Le Mauff, Bail, Gargot, Garnier, Guyot, Honnorat et Huez, 2005 ; Lusignan 2002 ; Lussier et Allaire, 2004 ; Perrenoud, 2001 ; Raîche, 2006 ; Scallon, 2004). Ces chercheurs soutiennent que les apprentissages réalisés par les apprenants demeurent intimement liés aux contextes où ils ont été acquis ; selon leur prémisse, les pratiques scolaires, lorsqu'elles ne sont pas étroitement liées aux situations de la vie courante ou du milieu réel de travail, sont insignifiantes pour les apprenants et ne favorisent guère l'acquisition des connaissances qu'ils auront forcément besoin de mobiliser dans des situations sociales courantes. Ainsi, ils préconisent la promotion en classe des situations reproduisant les conditions

rencontrées dans la vie courante, extrascolaire (Clauw *et al.*, 2006). Cette assertion suggère deux situations distinctes, notamment celle de la formation générale et celle de la formation professionnelle. On imagine bien que, par sa nature, la tâche authentique est plus aisée à réaliser en formation professionnelle qu'elle ne l'est en formation générale.

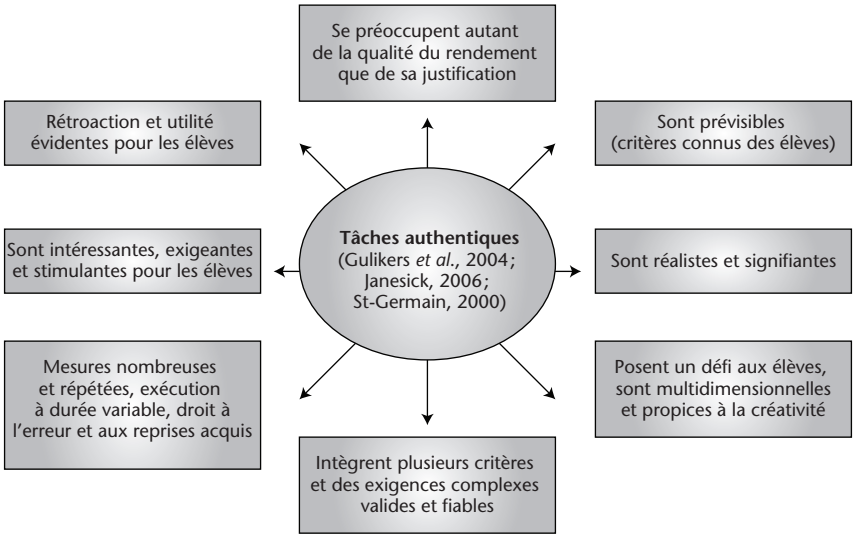


Figure 6.1.
Caractéristiques des tâches authentiques

En effet, l'authenticité par la vraisemblance de la tâche, c'est-à-dire le fait que la tâche soit conforme à la réalité du milieu du travail, semble pertinente dans les domaines de formation professionnelle. Par exemple, il paraît relativement aisé et approprié de simuler un incendie pour les élèves pompiers, une prise d'otage pour les élèves policiers, un pilotage d'avion avec un simulateur de vol pour les pilotes de ligne en aviation ou une salle d'attente débordée pour un résident en médecine, etc. À cet égard, Le Mauff *et al.* (2005, p. 66) considèrent le contexte professionnel comme une condition *sine qua non* pour l'authenticité de l'évaluation des compétences des internes en médecine générale. Pour ces chercheurs, une telle évaluation ne peut être univoque. Elle peut recourir à un dispositif intégré dans un système pédagogique, car il peut être utile d'évaluer spécifiquement les connaissances, les stratégies ou la performance ou le résultat de l'action de l'interne. Ce faisant, les professeurs doivent recourir à trois modalités, trois positionnements d'évaluation complémentaires, trois types d'outils, trois compétences

pédagogiques distinctes et complémentaires (la plus pertinente étant celle qui intègre les trois). En d'autres mots, il s'agit d'une évaluation complexe dont les différentes dimensions se complètent dans chaque situation, et ce, de façon continue. Dans ce contexte, quoi de mieux que l'immersion professionnelle ou les stages en milieu de travail pour assurer l'acquisition et le développement des compétences ? C'est là aussi que les tâches reflètent davantage la réalité du milieu de travail de l'apprenant, gage de l'authenticité de celles-ci.

Cela semble moins évident cependant dans l'enseignement régulier. En effet, simuler une situation problème en mathématique ou en français, par exemple, appelle à considérer le milieu de vie des élèves. On parle alors de signifiance des tâches. La question qui se pose est de savoir si cette signifiance de la tâche est partagée par l'ensemble des élèves. On peut en douter, en particulier dans les milieux urbains de plus en plus multiculturels et diversifiés. Les élèves viennent de milieux de vie variés et leurs expériences de vie varient également. Or, pour que les tâches d'évaluation soient signifiantes pour les élèves, il est impératif qu'elles soient le plus possible associées à leur milieu de vie (Scallon, 2004). En somme, les défis de la pratique des tâches authentiques sont multiples et leur manifestation a été rapportée dans la littérature. La section suivante en donne un aperçu.

4. LES DÉFIS D'ORDRE PRATIQUE INHÉRENTS AUX TÂCHES AUTHENTIQUES EN MILIEU SCOLAIRE

Quelques difficultés d'ordre pratique sont rapportées dans la littérature relativement à l'évaluation authentique. Par exemple, une épreuve authentique exige beaucoup de temps pour son déploiement, mais ne permet d'évaluer qu'une infime partie du contenu (Messick, 1995). Cela pose le problème de la représentativité du contenu disciplinaire dans une épreuve d'évaluation authentique. De plus, la variabilité de temps d'évaluation authentique (Société de gestion du réseau informatisé des commissions scolaires, 2000) peut se révéler une source de biais. Celle-ci rend incomparables les résultats, et ce, tant entre les élèves au cours d'une séance qu'entre les performances successives d'un même individu dans une série d'observations.

Le respect du rythme de travail de chaque apprenant (prescrit par le ministère de l'Éducation, du Loisir et du Sport, MELS) combiné à l'exigence temporelle dans la réalisation des tâches authentiques semble complexifier la visée réaliste, indispensable à l'authenticité des tâches en milieu scolaire. Comment concilier alors les variations de rythmes d'apprentissage chez les élèves, d'une part, avec l'impératif

de l'équité et de l'égalité des critères d'évaluation, d'autre part ? On est en droit de se demander si la récente annonce du ministère de l'Éducation, du Loisir et du Sport du Québec visant à prioriser l'évaluation des connaissances par rapport à celle des compétences n'est pas une manifestation de ces difficultés pratiques. En effet, en réponse à l'entente intervenue entre l'Alliance des professeurs et des professeures de Montréal (syndicat) et la Commission scolaire de Montréal (CSDM, 2011) visant à prioriser l'évaluation des connaissances par rapport à l'évaluation des compétences, la ministre de l'Éducation a confirmé l'intention du Ministère de renforcer cette démarche. À cet égard, la déclaration suivante de la présidente de la Commission scolaire de Montréal illustre bien le malaise inhérent à l'évaluation des compétences transversales : « Aussi pertinente soit-elle, l'évaluation des compétences transversales est trop compliquée pour le moment. » Pour sa part, l'Alliance faisait écho au piètre résultat observé par ses membres quant à la pratique de l'évaluation des compétences, en général, et des compétences transversales, en particulier.

Par ailleurs, la mise en place d'une évaluation authentique peut être onéreuse sur le plan logistique. En effet, pour qu'elle soit réaliste (authentique), une simulation peut nécessiter un investissement matériel important. On peut penser par exemple au coût d'un simulateur de vol pour la formation de pilotes de ligne ou de contrôleurs aériens. Il arrive aussi qu'on ait recours à des spécialistes pour le montage ou le contrôle durant le déroulement et le démontage sécuritaire et efficace de l'épreuve. Ainsi, en dépit de l'importance du caractère réaliste d'une tâche d'évaluation authentique, il peut être difficile, voire impossible, de satisfaire totalement à cette condition. On recourt alors à des approximations qui laissent libre cours à l'interprétation. En un mot, l'authenticité est aussi une question de degré d'appréciation.

Il est à noter que l'authenticité est aussi tributaire de la perception. Autrement dit, il n'est pas sûr que les tâches dites authentiques ou signifiantes, selon le concepteur, soient perçues comme telles par les élèves visés. À travers les écrits que nous avons consultés, la perception des élèves concernés par l'évaluation ne semble pas encore avoir été abordée. Tout se passe comme si le concepteur des tâches authentiques, à travers la connaissance présumée des élèves visés, se porte aussi garant de leurs perceptions à cet égard. En milieu scolaire, on s'attend donc à ce que le professeur fasse preuve de suffisamment d'empathie pour cerner précisément la perception potentielle de l'élève et la concilier avec les tâches à proposer afin que ce dernier les considère, à son tour, comme raisonnablement réalistes ou signifiantes.

Or, un élève ne perçoit pas l'importance des situations de la même façon que son professeur (Gulikers *et al*., 2004, 2006). Chacun d'eux a une réalité différente sur laquelle se fonde cette perception. De plus, l'authenticité se révèle à travers des situations d'évaluation réalistes, c'est-à-dire qui se rapprochent du vécu des examinés ou des situations auxquelles ils seront confrontés dans l'exercice d'une profession (Scallon, 2004, p. 137). On mesure alors la tâche qui incombe au professeur-concepteur de tâches authentiques ou signifiantes, notamment en ce qui a trait aux procédés d'observation ou de mesure déployés dans un tel contexte afin qu'ils acquièrent un caractère d'authenticité. Cette authenticité repose sur le niveau de signification (réalisme) de la situation proposée, la qualité qui la rend comparable à une situation de la vie quotidienne. Mais de quelle réalité s'agit-il ? Celle de l'élève ou celle du milieu dans lequel il évolue ? On conviendra que cerner la réalité d'un élève ou de quelques-uns (un petit groupe d'une dizaine d'individus) est dans l'ordre du possible. Or, de nos jours, il est utopique de penser à un groupe composé de seulement une dizaine d'individus. Bref, dans un cas comme dans l'autre, chaque individu ayant sa perception et par extension sa réalité, il est fort probable qu'il soit difficile de trouver un consensus eu égard à l'authenticité perçue. Devant ces défis d'ordre pratique, il est pertinent de se demander où en est la recherche dans ce domaine. La section suivante tente de faire le point à ce sujet.

5. ÉTAT DES LIEUX DE LA RECHERCHE

D'après notre revue de la littérature, aucune étude ne rapporte une expérience concrète de l'application des tâches authentiques en milieu scolaire. Tout au plus avons-nous constaté une intention annoncée par Clauw *et al*. (2006). Ces chercheurs ont projeté d'étudier l'effet des caractéristiques des pratiques des professeurs dans un contexte d'apprentissage authentique sur la motivation des élèves. Le fruit de ces travaux n'est pas encore accessible. On ne connaît pas non plus l'étendue réelle de la pratique des tâches authentiques dans les évaluations en milieu scolaire. Autrement dit, la proportion des professeurs utilisateurs des tâches authentiques dans leurs évaluations est une donnée inconnue. Des projets de recherche en cours comme celui conduit par le Centre des applications des modèles de réponses aux items (CAMRI), visant à étudier l'intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques dans un contexte d'approches par compétences, sont encore dans leur phase préliminaire (élaboration et validation des outils de collecte des données).

Quant aux outils de mesure de l'authenticité, la récolte n'est guère abondante. En fait, aucun outil ne permet actuellement de mesurer le niveau d'authenticité des tâches d'évaluation. Signalons néanmoins que Sauvé (2000) a esquissé les paramètres d'une grille d'analyse de la tâche authentique, constituée de six caractéristiques globales de la tâche authentique : 1) liée à la vie courante ; 2) significative et stimulante, 3) souple et adaptable ; 4) réaliste ; 5) cohérente et 6) rigoureuse. Cette grille d'analyse reprend l'ensemble des caractéristiques de l'évaluation authentiques définies par la Société de gestion du réseau informatisé des commissions scolaires (2000), qui découlent elles-mêmes de celles décrites par Wiggins (1989). Aussi, la grille d'analyse apparaît non seulement globale, mais aussi complexe (non opérationnelle) quant à son application. En effet, chaque composante de cette grille constitue un univers complexe en soi, qui doit être cerné pour ne pas se détourner de l'objectif premier de la tâche : son authenticité. Par exemple, concilier la souplesse et l'adaptabilité de la tâche tout en assurant sa signifiance et son caractère stimulant relève de l'exploit, par définition difficilement accessible à tout le monde, car cela nécessite minimalement un certain apprentissage et une expérience pratique de longue date. De plus, si l'on y ajoute l'exigence temporelle de planification et d'organisation de ce type d'activité, il est à craindre que l'authenticité des tâches ne se réalise que partiellement, même dans les conditions les plus optimistes.

Enfin, à travers cette brève revue de la littérature, il appert qu'aucune étude ne s'est attardée à la question de l'impact réel de l'authenticité des tâches sur le rendement scolaire ; il en est de même pour les variables liées au rendement scolaire, notamment la motivation. En d'autres mots, ce champ de recherche demeure, minimalement, à défricher.

6. CONCLUSION

Le but de cet exercice consistait à faire une revue de littérature portant sur l'authenticité des tâches, dix ans après l'implantation de la réforme au Québec, qui introduisait ce concept dans le contexte d'une évaluation authentique. Les écrits consultés à cet effet nous indiquent que, tant sur le plan de la pratique que de la recherche, les avancées demeurent fort marginales. En effet, les écrits relatifs à l'authenticité des tâches en évaluation se situent encore au stade de la caractérisation ; ils se limitent à décrire, définir ou circonscrire ce qu'est une tâche authentique. De plus, la quasi-totalité de ces écrits semble afficher une parenté notoire avec les assertions énoncées par Wiggins (1989),

suggérant de fait qu'ils en découlent. Cela tend à confirmer la stabilité temporelle du concept de l'authenticité des tâches d'évaluation en éducation.

Dans le domaine de la pratique, il se trouve que les nombreuses caractéristiques des tâches authentiques contribuent aussi à en complexifier l'application. À l'instar de la diversité de ces caractéristiques, les difficultés pratiques relatives à l'authenticité des tâches sont de sources variées. Par exemple, le réalisme exigé dans ces tâches reste difficilement opérationnel dans un milieu cosmopolite en raison, entre autres, du principe du respect du rythme d'apprentissage de chaque élève. Ce principe rend incomparables les résultats des évalués. De plus, le temps exigé pour l'exécution d'une tâche authentique pose de sérieux problèmes de validité dans la mesure où elle se limite à ne considérer qu'une infime partie du contenu dont on doit observer la maîtrise chez l'apprenant. En somme, les contraintes liées à chaque dimension de la tâche authentique, d'une part, et, d'autre part, celles relatives au temps imparti dans le monde de l'enseignement représentent un défi colossal à la pratique des tâches authentiques.

Quant au domaine de la recherche, aucune étude ne rapporte d'expérience concrète et complète de l'application des tâches authentiques en milieu scolaire. On ne connaît pas non plus l'étendue réelle de la pratique des tâches authentiques chez les professeurs, ni leur effet sur certaines variables importantes du domaine de l'éducation, notamment la motivation, le rendement scolaire, etc. La littérature ne rapporte pas non plus l'existence d'instruments de mesure de l'authenticité des tâches. Seule une grille d'analyse (Sauvé, 2000), qui ressemble étonnamment à une liste de vérification, permet de s'assurer du respect des critères relatifs à l'authenticité des tâches d'évaluation édictés par Wiggins (1989). La recherche en matière de l'authenticité des tâches d'évaluation est encore au stade embryonnaire. Tout reste donc à faire. Elle doit aller au-delà de la caractérisation actuelle, élaborer et valider des outils de mesure pour entreprendre ensuite des études corrélationnelles et expérimentales. À cet égard, le terrain demeure en friche. Tel est le portrait des défis posés à la recherche, en général, et à la recherche en éducation et en évaluation en particulier.

La figure 6.2 ci-après dresse un portrait (synthèse) de l'état des lieux de la pratique et de la recherche sur les tâches authentiques en évaluation des apprentissages, dix ans après l'implantation du renouveau pédagogique au Québec.

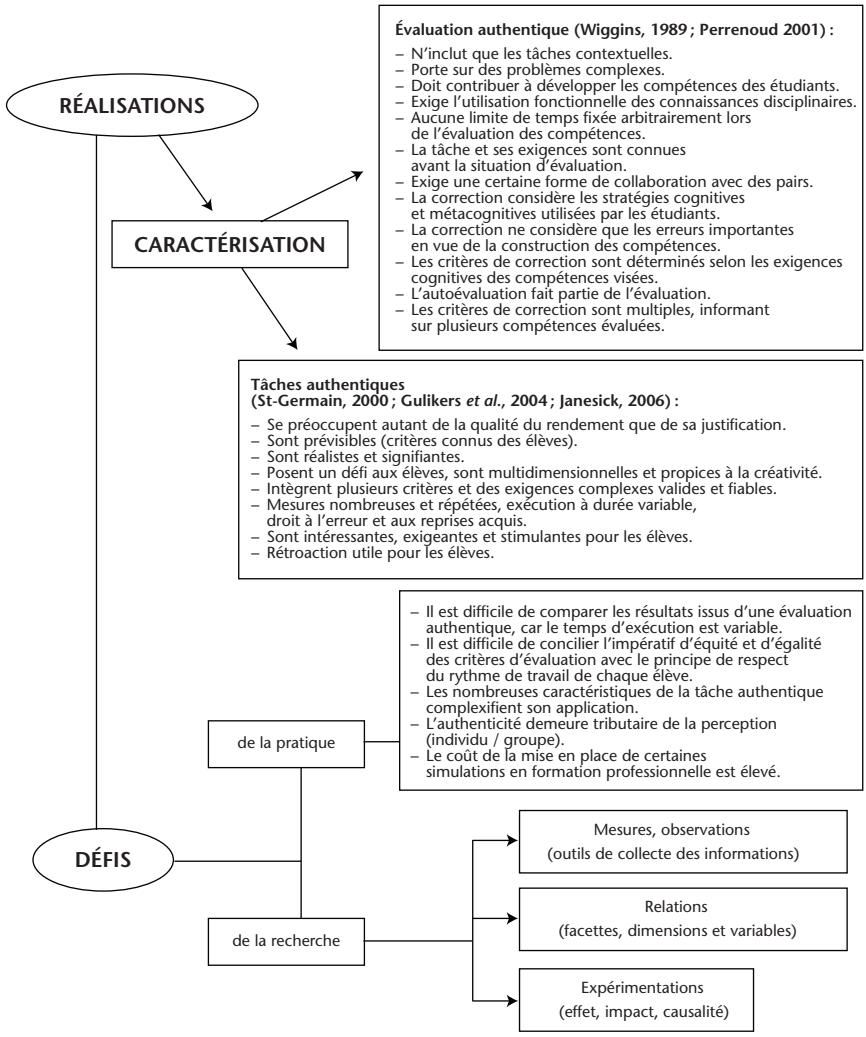


Figure 6.2.
Dix ans d'évaluation des tâches d'authenticité

RÉFÉRENCES

Alliance des professeurs et des professeures de Montréal (syndicat) et Commission scolaire de Montréal (2011). *Fin de l'évaluation des compétences transversales*. Montréal, Québec : Commission scolaire de Montréal.

Clauw, C., Dufays, J.-L., Thyrion, F., Vercruysse, B., Carlier, G. et Paquay, L. (2006). *Comment les enseignants du secondaire supérieur favorisent-ils un apprentissage contextualisé authentique ? Revue de littérature et recherche exploratoire dans des classes de français et d'éducation physique*. UCL, en collaboration avec L. Mottier Loppez. Document inédit. Genève, Suisse : Groupe de recherche interdisciplinaire en formation des enseignants et en didactique, Université de Genève.

Gulikers, J. T. M., Bastiaens, P. A. et Kirschner, T. J. (2004). A five-dimensional framework for authentic assessment. *Educational technology research and development*, *52*(3), 67-85.

Gulikers, J. T. M., Bastiaens, P. A. et Kirschner, T. J. (2006). Authentic assessment, student and teacher perceptions : the practical value of the five-dimensional framework. *Journal of vocational education and training*, *58*(3), 337-357.

Herrington, J., Oliver, R. et Reeves, T. C. (2006). Authentic tasks online : a synergy among learner, task, and technology. *Distance education*, *27*(2), 233-247.

Janesick, V. J. (2006). *Authentic assessment*. New York, New York : Peter Lang.

Jonassen, D. H. (1992). Evaluating constructivistic learning. Dans D. H. Jonassen (dir.), *Constructivism and the technology of instruction : a conversation*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

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

Le Mauff, P., Bail, P., Gargot, F., Garnier, F., Guyot, H., Honnorat, C. et Huez, J.-F. (2005). L'évaluation des compétences des internes de médecine générale : aspects théoriques, réflexions pratiques. *Exercer*, *73*, 63-69.

Lusignan, G. (2002). Planifier des situations complexes d'apprentissage pour aider les élèves à développer des compétences. *Vie pédagogique*, *123*, 21-23.

Lussier, O. et Allaire, H. (2004). L'évaluation « authentique » (*authentic task assessment*). *Pédagogie collégiale, 17*(3).

Messick, S. (1995). Standard of validity and the validity of standards in performance assessment. *Educational measurement : issues and practice*, *14*(4), 5-8.

Montgomery, K. (2001). *Authentic assessment : a guide for elementary teachers*. New York, New Jersey : Longman.

Montgomery, K. (2002). Authentic task and rubrics : going beyond traditional assessments in college teaching. *College teaching*, *50*(1), 34-39.

Paris, R. A. et Ayres, L. R. (2000). *Réfléchir et devenir : apprendre en autonomie, des outils pour l'enseignant et l'apprenant*. Traduction de la première édition américaine par M. Aussanaire-Gracia. Bruxelles, Belgique : De Boeck.

Perrenoud, P. (2001). Évaluation formative et évaluation certificative : postures contradictoires ou complémentaires ? *Formation professionnelle suisse*, *4*, 25-28.

Raîche, G. (2006). L'intégration des pratiques d'évaluation des apprentissages aux pratiques pédagogiques dans le contexte des approches par compétences. *Vivre le primaire*, *19*(2), 43-45.

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

Sauvé, P. (2000). Les tâches et l'évaluation authentique : le journal de Valérie… la suite. *Virage express*, *2*(8), 1-2.

Société de gestion du réseau informatisé des commissions scolaires (2000). *Service de consultation en développement pédagogique*. Document inédit. Montréal, Québec : Société de gestion du réseau informatisé des commissions scolaires.

St-Germain, C. (2000). La logique des compétences et de l'évaluation. *Virage express*, *2*(8), 5.

Tarackdjian, É. (2003). *Évaluation des apprentissages*. Notes de cours, document inédit. Montréal, Québec : Université de Montréal.

Torrance, H. (1995). *Evaluating authentic assessment*. Buckingham, Angleterre : Open University Press.

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

Chapitre 7

Impact de la méthode d'estimation du niveau d'habileté et du choix des premiers items sur l'efficacité de l'administration adaptative du TCALS II

David Magis et Gilles Raîche[1]

Le TCALS II (test de classement en anglais, langue seconde, au collégial, 2e version) est un questionnaire constitué d'un nombre fixe de 85 items administré aux étudiants de la province de Québec qui entament des études au niveau collégial. Une version adaptative informatisée de ce test est envisagée pour la première fois dans cette étude. Deux problématiques sont regardées de plus près : le choix d'une méthode d'estimation du niveau d'habileté optimale et la sélection des premiers items du test. Ces deux problématiques sont étudiées simultanément par le biais de simulation Monte-Carlo à partir de plusieurs règles d'arrêt liées à la longueur du test. On en conclut que le choix des premiers items affecte peu l'estimation du niveau d'habileté, tandis que des différences notables apparaissent entre les quatre méthodes d'estimation comparées. Certaines conclusions et recommandations sont mentionnées pour la poursuite ultérieure de ces travaux.

1. Cette recherche a été soutenue par une bourse postdoctorale du Fonds national de la recherche scientifique (FNRS), en Belgique, par le Conseil national de recherches du Canada et par le ministère de l'Éducation, du Loisir et du Sport du Québec.

1. INTRODUCTION

Depuis les premiers travaux réalisés dans les années 1980 (Green, 1983a, 1983b ; Weiss, 1983), le testing adaptatif informatisé (TAI) s'est imposé comme une méthode efficace d'administration de tests psychométriques ou d'épreuves destinées à l'évaluation des apprentissages. Le testing adaptatif informatisé peut être décrit rapidement comme suit : l'apprenant est placé devant un ordinateur et reçoit une série d'items qui lui sont présentés l'un après l'autre. Le choix de l'item administré dépend des réponses de l'apprenant aux items précédemment administrés. En fait, à chaque réponse enregistrée, le niveau d'habileté de l'apprenant est estimé au moyen de méthodes issues de la théorie classique des tests (TCT) ou, plus fréquemment, de la théorie de la réponse à l'item (TRI). Le test adaptatif se termine lorsque l'étudiant a répondu à un nombre déterminé d'items ou lorsque la précision sur l'estimation de son niveau d'habileté est jugée suffisante.

Cette approche adaptative a plusieurs avantages sur l'administration de questionnaires (de type papier-crayon) où les items administrés sont fixés à l'avance. Tout d'abord, un test adaptatif réduit le risque de fraude, puisque chaque test est individualisé en fonction des réponses fournies par l'apprenant. Ensuite, il a été maintes fois démontré (Wainer, 2000) qu'un test adaptatif fournit la même précision d'estimation du niveau d'habileté avec un nombre moindre d'items qu'un test fixe similaire. Cela s'explique par le fait qu'un test adaptatif permet de sélectionner l'item le plus informatif pour le niveau d'habileté estimé provisoire. Le testing adaptatif informatisé permet aussi l'utilisation de différentes règles d'arrêt et de sélection du prochain item. Enfin, il fournit une estimation immédiate du niveau d'habileté.

Toutefois, le testing adaptatif informatisé présente également des inconvénients. Le premier est d'ordre technique et financier : l'administration d'un testing adaptatif implique la disponibilité d'un grand nombre d'ordinateurs performants ainsi que la constitution d'une banque d'items suffisamment large pour permettre une sélection adéquate des items les plus informatifs. La construction, la calibration et la confidentialité d'une telle banque d'items impliquent un coût non négligeable. De plus, l'utilisation pratique du testing adaptatif nécessite une plateforme informatique et un programme adapté, convivial et facile d'utilisation. Malheureusement, il existe peu de logiciels de testing adaptatif informatisé disponibles actuellement. Enfin se pose le problème de la transformation d'un test fixe en test adaptatif : les chercheurs et spécialistes de l'éducation ayant généralement investi beaucoup de temps et d'efforts à développer de tels tests fixes, il serait utile de savoir si ces tests peuvent être aisément transformés en version adaptative afin de profiter des avantages inhérents à cette technique.

Le sujet d'étude de ce texte est axé principalement sur cette dernière problématique. En particulier, nous nous sommes intéressés à la possibilité d'appliquer de façon adaptative le *test de classement en anglais, langue seconde, au collégial* (TCALS II ; Laurier, Froio, Pearo, et Fournier, 1998). Il s'agit d'un test utilisé depuis la fin des années 1990 dans le réseau collégial au Québec pour déterminer le niveau de connaissance en anglais des étudiants francophones entrant au collège. L'étudiant est ensuite dirigé vers un cours en anglais langue seconde adapté à son niveau d'habileté. Le test est plutôt long (85 items) et une version adaptative informatisée pourrait être envisagée. Deux problématiques sont plus particulièrement étudiées lors de ce passage de la version fixe à la version adaptative. Premièrement, nous étudions si le choix initial du ou des items a un impact sur la précision de l'estimation du niveau d'habileté à l'issue du test adaptatif. Deuxièmement, nous examinons différentes méthodes d'estimation du niveau d'habileté et nous les comparons en termes de précision et de biais d'estimation. Des conclusions et recommandations sont alors émises au sujet de ces deux problématiques dans l'optique d'une utilisation future du TCALS II sous forme adaptative.

2. CONTEXTE THÉORIQUE

Dans cette section, nous présentons brièvement les grands principes du testing adaptatif informatisé, puis nous décrivons les caractéristiques du TCALS II (version fixe). Les problématiques d'intérêt y sont également détaillées.

2.1. Le testing adaptatif informatisé

Un test adaptatif est composé de quatre phases :

a) la *phase initiale*, au cours de laquelle le ou les premiers items sont sélectionnés et administrés au répondant ;

b) la *phase de test*, qui consiste à sélectionner itérativement le prochain item à administrer, à obtenir la réponse et à mettre à jour le patron de réponses ainsi que l'estimateur du niveau d'habileté du répondant ;

c) la *phase d'arrêt*, qui interrompt le processus adaptatif lorsque le critère d'arrêt choisi est satisfait ;

d) la *phase finale*, qui fournit l'estimateur final du niveau d'habileté sur la base des items administrés.

L'élément central de ces quatre phases est la constitution de la banque d'items. Une banque d'items est une grande collection d'items disponibles, parmi lesquels sont sélectionnés ceux à administrer. En testing adaptatif informatisé, ces items sont préalablement calibrés en estimant leurs paramètres d'items selon un modèle de réponse à l'item sous-jacent et sélectionné à l'avance. Il existe plusieurs types de modèles de réponses à l'item et nous nous concentrerons ici sur le modèle logistique à trois paramètres (*three-parameter logistic* ou 3PL). Ce modèle est décrit comme suit (Birnbaum, 1968) :

$$Pr\left(X_{ij}=1\,|\,\theta_i\right)=c_j+\left(1-c_j\right)\frac{\exp\left[a_j\left(\theta_i-b_j\right)\right]}{1+\exp\left[a_j\left(\theta_i-b_j\right)\right]}. \qquad (1)$$

Dans l'équation 1, X_{ij} est la réponse de l'apprenant i à l'item j, codée 1 pour une bonne réponse et 0 pour une réponse incorrecte ; q_i est le niveau d'habileté de l'apprenant ; et a_j, b_j et c_j sont les paramètres d'items : respectivement la discrimination, la difficulté et la pseudo-chance (ou asymptote inférieure). Par la suite, l'indice de sujet i sera omis et la probabilité décrite par l'équation 1 sera simplement notée $P_j(\theta)$. Le calibrage des items consiste à estimer les paramètres a_j, b_j et c_j et à les considérer, au moment de l'administration du test adaptatif, comme des valeurs fixes connues. L'estimation de q_i, au contraire, se fait lors de l'administration du test adaptatif.

Lors de la phase initiale, afin d'initialiser le test, au moins un item doit être sélectionné dans la banque d'items. En général, un seul item est sélectionné en fixant d'abord une valeur initiale q_0 du niveau d'habileté (souvent 0), puis en sélectionnant l'item *a*) dont le niveau de difficulté est le plus proche du niveau d'habileté initial q_0, ou *b*), qui est le plus informatif au niveau d'habileté initial q_0. Par informatif, nous entendons que l'item sélectionné est celui dont la fonction d'information (au sens de Fisher) est maximale (parmi les items de la banque d'items) à la valeur initiale q_0 du niveau d'habileté. La fonction d'information de Fisher $I_j(\theta)$ pour l'item j est donnée par Baker (1992) :

$$I_j(\theta)=\frac{P_j'(\theta)^2}{P_j(\theta)Q_j(\theta)}, \qquad (2)$$

où $P_j'(\theta)$ est la dérivée première de $P_j(\theta)$ par rapport à θ et où $Q_j(\theta) = 1 - P_j(\theta)$. Ainsi, pour déterminer l'item le plus informatif au niveau d'habileté q_0, il faut calculer l'information (2) à la valeur q_0 pour tous les items de la banque et sélectionner celui correspondant à la valeur de l'information la plus élevée. Notons qu'une approche plus générale

consisterait à sélectionner plusieurs items initiaux, par exemple en fixant une séquence de valeurs initiales du niveau d'habileté permettant de balayer une étendue plus large de niveaux d'habileté initiaux (par exemple de –2 à 2 par pas de 1, soit 5 items initiaux). Cette approche constitue justement l'une des problématiques de recherche décrites ci-après.

La phase de test est la partie itérative et adaptative du testing adaptatif informatisé. Une fois l'item initial (ou les items initiaux) administré, et la ou les réponses de l'apprenant enregistrées, le niveau d'habileté est estimé de façon provisoire. Il existe plusieurs méthodes d'estimation du niveau d'habileté : par le maximum de vraisemblance (*maximum likelihood*, ML), le maximum *a posteriori* (MAP), la moyenne *a posteriori* (*expected a posteriori*, EAP) ou le maximum de vraisemblance pondéré (*weighted likelihood*, WL). Ces méthodes diffèrent quant à leur approche conceptuelle (maximisation d'une fonction ou calcul de la moyenne d'une distribution) et quant à leur paradigme (fréquentiste ou bayésien). Nous n'entrerons pas dans les détails dans ce texte, mais nous suggérons au lecteur intéressé de consulter Magis et Raîche (2012), par exemple. Notons toutefois que, pour un grand nombre d'items administrés, les méthodes ci-dessus fournissent des estimateurs du niveau d'habileté proches les uns des autres. Par contre, de grandes disparités entre les estimateurs peuvent apparaître lorsque peu d'items sont administrés, ce qui est certainement le cas au début d'un test adaptatif. Il sera dès lors intéressant de comparer ces méthodes d'estimation du niveau d'habileté dans le contexte du testing adaptatif informatisé, ce qui constitue une seconde problématique de recherche décrite ci-après. Enfin, bien qu'il soit possible de changer de méthode d'estimation du niveau d'habileté au cours de la phase de test, en général la même méthode d'estimation est utilisée tout au long du test adaptatif.

Avec le patron de réponses, la liste d'items déjà administrés et l'estimation provisoire de l'habileté, l'item suivant à administrer peut être sélectionné parmi les items de la banque encore disponibles. Il existe un grand nombre de méthodes de sélection de l'item suivant à administrer. Les plus connues sont la méthode de maximisation de l'information de Fisher (*Maximum Fisher information*, MFI) et la règle d'Urry (*Urry's rule*). Dans le premier cas, l'item ayant la plus grande information au sens de Fisher (2) au niveau d'habileté provisoire est sélectionné (Baker, 1992 ; Wainer, 1990). Dans le second cas, l'item sélectionné est celui dont le niveau de difficulté est le plus proche du niveau d'habileté provisoire (Urry, 1970). Il existe d'autres méthodes plus techniques : minimisation de l'espérance mathématique de la variance *a posteriori* (*minimum expected posterior variance*, MEPV) ; maximisation

de l'information pondérée par la vraisemblance (*maximum likelihood weighted information*, MLWI ; Veerkamp et Berger, 1997) ou pondérée par la distribution *a posteriori* de l'habileté (*maximum posterior weighted information*, MPWI ; van der Linden, 1998) ; maximisation de l'espérance mathématique de l'information (*maximum expected information*, MEI ; van der Linden, 1998). Bien qu'il n'y ait aucune étude formelle comparant ces méthodes de sélection de l'item suivant, il semble que celles-ci ne diffèrent pas beaucoup quant aux résultats finaux aux tests adaptatifs. De plus, la même méthode de sélection est habituellement utilisée tout au long de la phase de test. Le lecteur intéressé trouvera plus de détails sur ces procédures dans Choi et Swartz (2009).

Le processus itératif de la phase de test (sélection de l'item suivant, administration de celui-ci, mise à jour du patron de réponses et estimation du niveau d'habileté) s'arrête lorsqu'un certain critère d'arrêt est réalisé. Il s'agit de la troisième phase, soit la phase d'arrêt. Trois critères sont couramment utilisés. Le premier est simplement un critère de longueur, qui fixe l'arrêt du test adaptatif lorsqu'un nombre prédéfini d'items ont été administrés. Le deuxième critère en est un de précision : le test adaptatif s'achève lorsque la précision de l'estimateur provisoire du niveau d'habileté est suffisamment grande ou, en d'autres termes, lorsque l'erreur type de l'estimateur du niveau d'habileté est suffisamment petite. À cette fin, on fixe une valeur minimale de l'erreur type et le test adaptatif s'arrête lorsque l'erreur type de l'estimateur provisoire du niveau d'habileté devient inférieure à cette valeur limite. Enfin, le troisième est un critère de classification : un seuil de classification est fixé à l'avance, et le test adaptatif s'arrête lorsque l'estimateur provisoire du niveau d'habileté est supérieur ou inférieur à ce seuil avec un certain niveau de confiance. Par *niveau de confiance*, il faut entendre que l'on construit, à chaque étape du test adaptatif, un intervalle de confiance pour le niveau d'habileté (en utilisant les estimateurs provisoires du niveau d'habileté et de leur erreur type). Le critère de classification s'applique donc lorsque les deux bornes de cet intervalle de confiance sont inférieures ou supérieures au seuil de classification. Cette approche est surtout utilisée pour distinguer les répondants en fonction d'un seuil de classification préétabli. Notons que certains de ces critères peuvent être combinés. Par exemple, on peut exiger d'atteindre une précision de l'estimateur du niveau d'habileté fixé à l'avance sans dépasser un nombre maximal d'items à administrer.

Enfin, la phase finale du test consiste à estimer, une dernière fois, le niveau d'habileté de l'apprenant en considérant le patron de réponses complet comme patron final. À cette étape, la méthode d'estimation peut être différente de celle utilisée lors de la phase de test, bien

qu'en général les deux méthodes soient identiques. De plus, l'erreur type de l'estimateur final du niveau d'habileté est fournie pour obtenir une indication de la précision de celui-ci.

Pour être tout à fait complet, on ne saurait passer sous silence deux problématiques d'intérêt dans le contexte du testing adaptatif informatisé : le contrôle de l'exposition des items (*item exposure*) et celui de l'équilibre du contenu (*content balancing*).

Un problème pratique en testing adaptatif informatisé est la surexposition des items aux répondants. En effet, il arrive souvent qu'un item très informatif soit plus souvent administré qu'un item moins informatif. Cela peut poser des problèmes de confidentialité et de validité du test, dans la mesure où ces items surexposés deviennent connus des répondants. Afin de limiter cette surexposition, plusieurs approches ont été proposées. La plus simple et la plus facile à mettre en œuvre est la méthode *randomesque* (Kingsbury et Zara, 1989). Celle-ci consiste, lors de la phase de test, à sélectionner non pas un seul item mais un ensemble d'items (dont le nombre est prédéfini) et de choisir au hasard l'item à administrer parmi ceux-ci. Cette approche garantit que l'item suivant est choisi parmi les items les plus informatifs, mais l'introduction d'une sélection au hasard diminue le risque d'une surexposition des items les plus informatifs.

Un autre problème surgit lorsque la banque d'items possède une certaine structure sous-jacente, par exemple lorsque les items appartiennent à des sous-groupes bien définis (selon leur contenu ou leur mode d'administration, par exemple). Il semble alors naturel d'imposer une certaine répartition des items administrés en fonction du nombre d'items de ces sous-groupes afin d'éviter une sous-représentation des sous-groupes d'items dont le nombre est plus important ou, à l'inverse, une surreprésentation des sous-groupes d'items dont le nombre est le moins élevé. Kingsbury et Zara (1989) ont proposé une méthode simple de contrôle de l'équilibre de contenu : en fixant à l'avance les pourcentages d'items devant provenir de chaque sous-groupe de la banque d'items, l'item suivant est sélectionné parmi le sous-groupe le moins représenté par rapport aux pourcentages prédéfinis d'items par sous-groupe. Cette méthode est simple et efficace. De plus, elle ne modifie en rien le processus général du testing adaptatif informatisé.

2.2. Le TCALS II

Nous présentons dans cette section le test de classement en anglais, langue seconde, au collégial (TCALS II). Il s'agit d'un questionnaire de 85 items à choix multiples administré aux étudiants canadiens

francophones au moment de leur entrée aux études collégiales dans la province de Québec (Laurier *et al*., 1998). L'objectif de ce questionnaire est d'évaluer le niveau d'habileté des élèves en anglais, langue seconde, afin de les répartir en groupes de niveaux plus ou moins semblables (Raîche, 2002). Les 85 items du TCALS II sont répartis en huit sous-groupes et ils sont identiques pour tous les collèges du Québec. Ce test fut administré pour la première fois en 1998 et il est toujours en application actuellement sous forme fixe (papier-crayon).

Laurier *et al*. (1998) ont établi que le TCALS II possède un haut niveau de fidélité avec un coefficient α de Cronbach de l'ordre de 0,96. L'unidimensionnalité du test a également été vérifiée (voir aussi Raîche [2002], pour une analyse différente de la dimensionnalité du TCALS II, mais menant à des conclusions identiques). Finalement, le TCALS II a été calibré sur la base des résultats de la première administration au Collège de l'Outaouais en 1998 (Raîche, 2002) à l'aide d'un modèle logistique à trois paramètres. Les paramètres sont disponibles dans la librairie *catR* du logiciel R (Magis et Raîche, 2011, 2012b). Notons que le TCALS II est majoritairement composé d'items faciles : les niveaux de difficulté varient de –3,457 à 0,840 avec une moyenne de –1,104. De plus, les niveaux de discrimination varient de 0,407 à 3,983 avec une moyenne de 1,983, ce qui indique que les items du test sont en majorité suffisamment discriminants. Enfin, les niveaux de pseudo-chance estimés varient de 0,063 à 0,384 avec une moyenne de 0,20.

Dans la suite de cette étude, les 85 items seront regroupés en cinq sous-groupes afin de mieux contrôler l'équilibre du contenu lorsque, dans le futur, une version adaptative du TCALS II sera envisagée. Le regroupement choisi est le suivant : *audio1* (items 1 à 12, compréhension à l'audition de phrases), *audio2* (items 13 à 33, compréhension à l'audition de dialogues et de courts textes), *ecrit1* (items 34 à 46, exercices écrits de vocabulaire), *ecrit2* (items 47 à 63, exercices écrits de grammaire) et *ecrit3* (items 64 à 85, exercices écrits d'autres types).

2.3. Problèmes d'intérêt

Comme il a été mentionné plus haut, le problème d'intérêt principal est d'étudier l'éventualité et la faisabilité d'une administration adaptative informatisée du TCALS II. Plusieurs problématiques peuvent être envisagées, mais dans cette étude nous nous limiterons aux deux suivantes.

Premièrement, le choix initial d'un ou de plusieurs items influence-t-il les résultats du test adaptatif sur base du TCALS II ? Plusieurs stratégies de démarrage du test seront envisagées par le biais de la sélection d'un nombre d'items différent. Deuxièmement, peut-on mettre en évidence une différence significative dans les estimateurs finaux du niveau d'habileté en fonction des différentes méthodes d'estimation du niveau d'habileté disponibles ? À cette fin, nous comparerons plusieurs méthodes d'estimation du niveau d'habileté lors de l'administration du TCALS II adaptatif et nous étudierons des différences potentielles entre les résultats fournis par ces estimateurs.

Au-delà de l'aspect immédiatement utile en ce qui concerne le TCALS II, étant donné l'absence quasi totale de références sur l'administration des premiers items en TAI, cette étude sera également une bonne approche empirique pour tirer des conclusions plus larges. De plus, l'objectif est également de formuler des recommandations pratiques quant à l'administration future du TCALS II sous forme adaptative.

3. MÉTHODOLOGIE

L'étude a été réalisée par le biais de simulations informatisées de type Monte-Carlo. La librairie *catR* (Magis et Raîche, 2011) du logiciel R a été exploitée en ce sens. Le code complet est disponible dans l'annexe. Nous décrivons ci-dessous les diverses options retenues pour ces simulations.

3.1. Phase initiale

Quatre types de phases initiales ont été sélectionnés. Toutes ces phases se basent sur la sélection des items initiaux les plus informatifs aux niveaux d'habileté préalablement fixés. Elles diffèrent selon les niveaux initiaux fixés du niveau d'habileté. Dans le premier cas, un seul niveau d'habileté est sélectionné : il est fixé à zéro. Dans le deuxième cas, les niveaux d'habileté sont fixés à –1 et 1. Dans le troisième cas, trois items initiaux sont sélectionnés par rapport aux niveaux d'habileté fixés à –1, 0 et 1. Finalement, dans le quatrième cas, quatre items initiaux sont sélectionnés comme étant les plus informatifs par rapport aux niveaux d'habileté fixés à –1,5, 0,5, 0,5 et 1,5. Remarquons que dans les quatre cas les niveaux d'habileté initiaux sont symétriques autour de la valeur zéro.

3.2. Phase de test

Seule la méthode de sélection selon la maximisation de l'information de Fisher (MFI), prenant l'item le plus informatif à la valeur de l'estimateur du niveau d'habileté, a été considérée dans cette étude. Ce critère est en effet le plus couramment utilisé en plus d'être rapide d'exécution. De plus, l'objet de l'étude ne repose pas sur le choix optimal de ce critère en particulier.

Quatre méthodes d'estimation du niveau d'habileté sont considérées : ML, MAP avec une loi normale centrée réduite comme distribution *a priori* du niveau d'habileté, MAP avec une fonction *a priori* non informative de Jeffreys ainsi que WL (Magis et Raîche, 2012a). Notons que, lors de l'exécution d'un test adaptatif avec la banque d'items du TCALS II, la même méthode d'estimation du niveau d'habileté a été utilisée tout au long du test.

3.3. Phase d'arrêt

Le critère d'arrêt qui a été considéré est celui de la longueur du test, et trois longueurs maximales ont été sélectionnées : 10 items, 15 items et 20 items. Cette approche garantit l'arrêt du test adaptatif dans n'importe quelle configuration (la banque d'items comprenant 85 items). La variation de la longueur du test devrait mettre en évidence les différences de précision en fonction des méthodes d'estimation du niveau d'habileté.

3.4. Phase finale, exposition des items, équilibre du contenu

Pour la phase finale du test adaptatif, les mêmes méthodes d'estimation du niveau d'habileté que celles utilisées lors de la phase de test ont été considérées. Ainsi, nous ne permettons pas de modification à la méthode d'estimation du niveau d'habileté tout au long du test.

L'exposition des items a été contrôlée grâce à l'approche *randomesque* (Kingsbury et Zara, 1989) en fixant le nombre d'items optimaux à trois. De plus, le contrôle de l'équilibre du contenu a été réalisé en imposant les proportions d'items par sous-groupes de la façon suivante : 14 % pour les items *audio1*, 25 % pour les items *audio2*, 15 % pour les items *ecrit1*, 20 % pour les items *ecrit2* et 26 % pour les items *ecrit3*. Ces proportions correspondent approximativement aux proportions d'items du TCALS II repris dans chaque sous-groupe décrit plus haut.

Ainsi, à chaque étape du test, trois items au plus sont présélectionnés parmi les plus informatifs dans le sous-groupe ciblé par l'équilibre du contenu : l'item suivant à administrer est choisi au hasard parmi ces items. Bien entendu, s'il reste moins que trois items disponibles dans le sous-groupe ciblé, seuls les items disponibles sont présélectionnés.

3.5. Simulations et extraction de résultats

Le schéma de simulations décrit ci-dessus conduit à 48 situations distinctes : quatre règles initiales, quatre règles de sélection de l'item suivant (une par méthode d'estimation) et trois critères d'arrêt. Pour chaque combinaison de ces trois critères, 1000 tests adaptatifs ont été simulés pour chacune des valeurs réelles du niveau d'habileté variant entre –3 et 3 par pas de 1 (soit 7000 tests adaptatifs pour chacune des 48 situations de l'étude). De chaque test adaptatif généré, seul l'estimateur final du niveau d'habileté a été retenu et analysé.

L'efficacité de chaque méthode d'estimation du niveau d'habileté et de chaque sélection initiale des items a été comparée comme suit. Pour chaque situation de test et chaque valeur réelle de l'habileté, soit θ, les 1000 estimations de l'habileté ($\hat{\theta}_1$, …, $\hat{\theta}_{1000}$) sont résumées en calculant le biais :

$$Biais_{\theta} = \frac{1}{1000} \sum_{i=1}^{1000} \left(\hat{\theta}_j - \theta \right). \tag{3}$$

Ensuite, la racine du carré moyen résiduel (*root mean squared error*, RMSE) est obtenue :

$$RMSE_{\theta} = \sqrt{\frac{1}{1000} \sum_{i=1}^{1000} \left(\hat{\theta}_j - \theta \right)^2}. \tag{4}$$

Le biais est une mesure de déviation de l'estimateur moyen du niveau d'habileté par rapport à la vraie valeur de ce niveau d'habileté. Un biais positif indique une surestimation du niveau d'habileté, tandis qu'une valeur négative indique une sous-estimation. Le carré moyen résiduel, quant à lui, est une mesure synthétique d'exactitude (le biais) et de précision (l'erreur type) des méthodes d'estimation du niveau d'habileté. Plus ce carré moyen est petit, mieux se comporte la méthode d'estimation du niveau d'habileté dans la situation étudiée.

4. RÉSULTATS

Les résultats sont résumés en six figures, trois pour les valeurs du biais de l'estimateur du niveau d'habileté et trois pour les RMSE correspondantes. Selon la statistique étudiée (biais ou RMSE), les trois figures correspondent à chacun des critères d'arrêt (après 10 items, 15 items ou 20 items). Chaque figure est divisée en quatre panneaux, correspondant à chacune des règles initiales. La séquence de vraies valeurs du niveau d'habileté (de –3 à 3 par pas de 1) est représentée en abscisses et les courbes de biais (ou de RMSE) en ordonnées pour chaque méthode d'estimation. La même échelle de valeurs du biais (et de la RMSE) est conservée pour permettre une comparaison directe des courbes de biais et de RMSE en fonction des trois critères d'arrêt.

Les courbes de biais selon les quatre méthodes d'estimation du niveau d'habileté et selon les quatre règles initiales sont représentées aux figures 7.1, 7.2 et 7.3, respectivement, lorsque le test s'arrête après 10 items, 15 items et 20 items. Lorsque la longueur du test augmente, on constate une diminution générale du biais dans tous les cas de figure, ce qui est un résultat attendu. En effet, l'augmentation du nombre d'items administrés implique une augmentation de la quantité d'information disponible : on obtient donc une meilleure précision dans l'estimation du niveau d'habileté, indépendamment des règles initiales et des méthodes d'estimation.

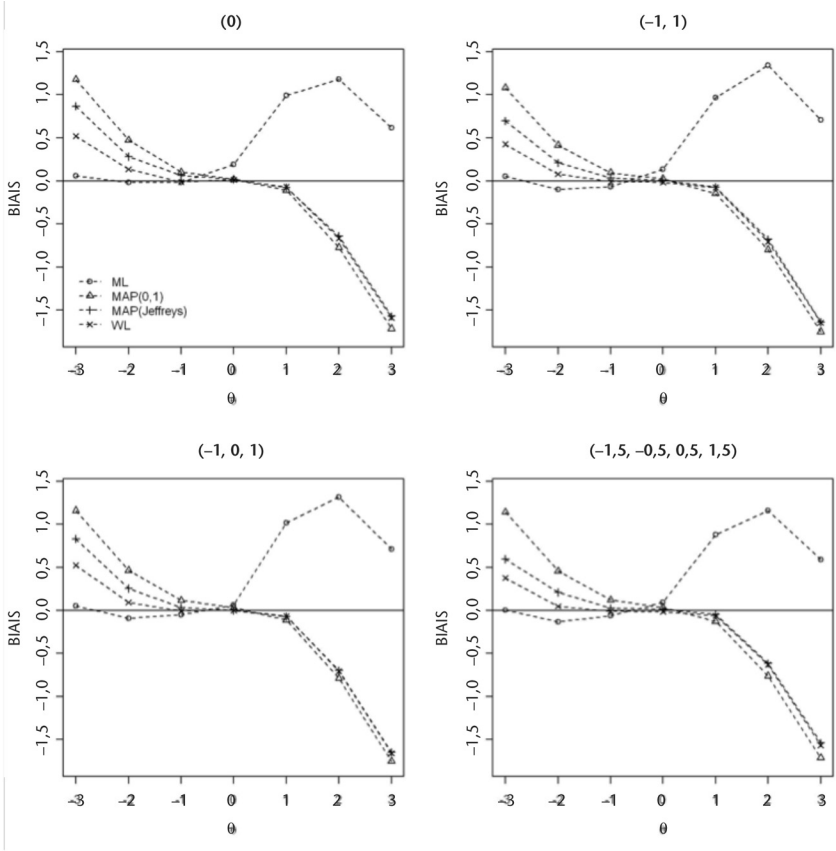


Figure 7.1.
Courbes de biais pour les quatre méthodes d'estimation finales du niveau d'habileté et les quatre règles initiales du test avec le critère d'arrêt de 10 items. Les règles initiales sont représentées par les valeurs initiales du niveau d'habileté (entre parenthèses) choisies pour la sélection des premiers items du test.

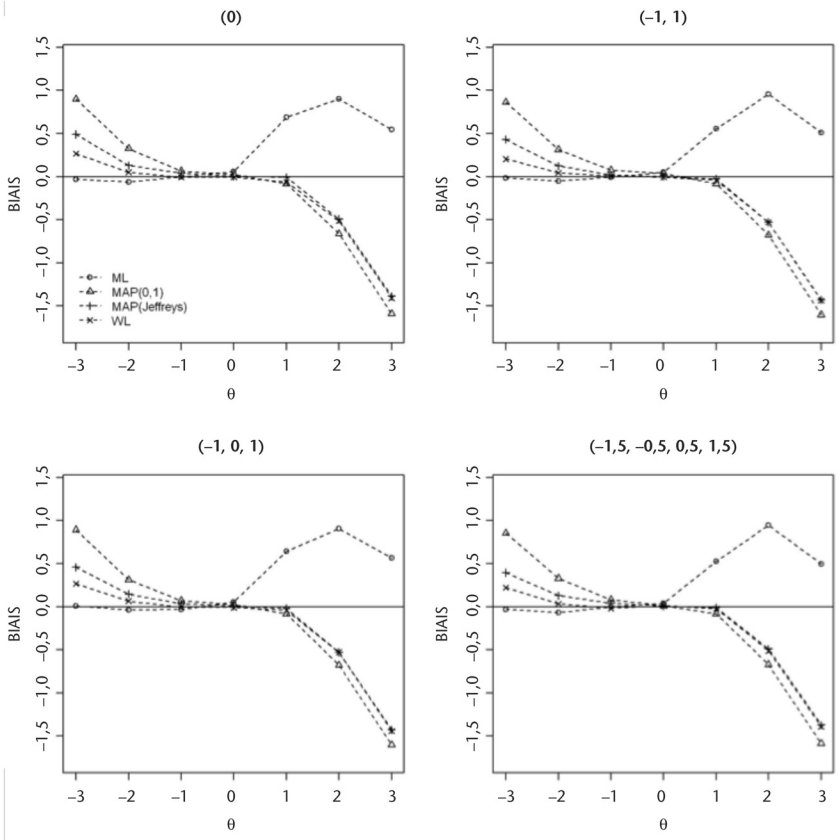


Figure 7.2.
Courbes de biais pour les quatre méthodes d'estimation finales du niveau d'habileté et les quatre règles initiales du test avec le critère d'arrêt de 15 items. Les règles initiales sont représentées par les valeurs initiales du niveau d'habileté (entre parenthèses) choisies pour la sélection des premiers items du test.

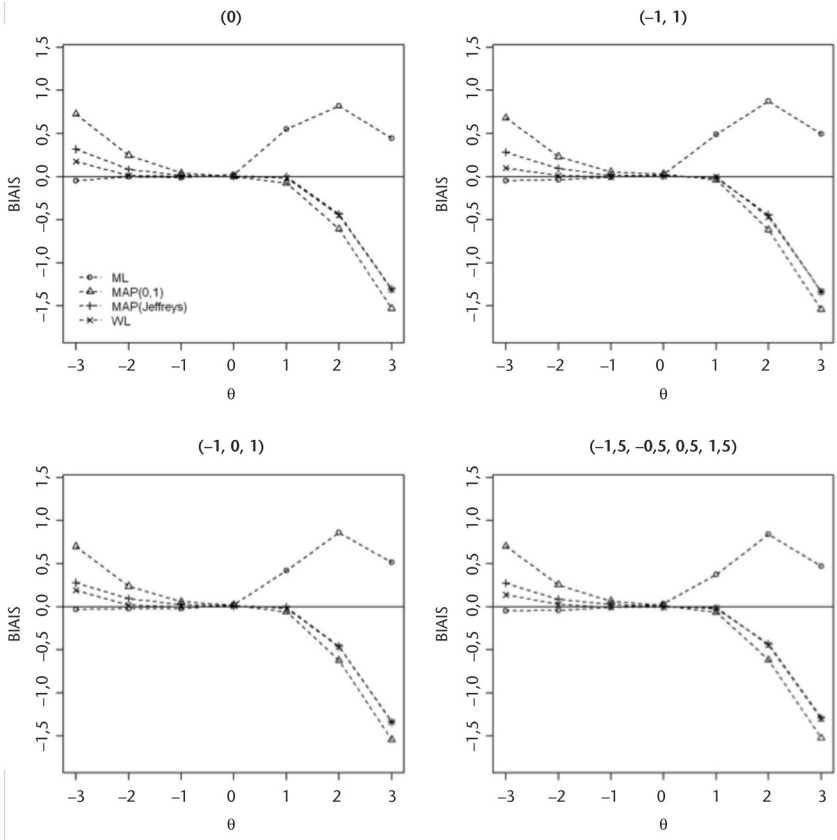


Figure 7.3.
Courbes de biais pour les quatre méthodes d'estimation finales du niveau d'habileté et les quatre règles initiales du test avec le critère d'arrêt de 20 items. Les règles initiales sont représentées par les valeurs initiales du niveau d'habileté (entre parenthèses) choisies pour la sélection des premiers items du test.

La méthode d'estimation selon ML affiche, en termes de biais, un comportement différent des trois autres méthodes d'estimation. En effet, son biais est quasi nul pour les valeurs négatives du niveau d'habileté, alors que les trois autres méthodes d'estimation donnent des valeurs de biais positives. Par contre, pour les valeurs positives du niveau d'habileté, le biais selon ML est largement positif, alors que celui selon les trois autres méthodes d'estimation est largement négatif. Cet effet n'est pas surprenant, dans le sens où Lord (1983) avait déjà remarqué les tendances opposées des courbes de biais selon ML et MAP avec une distribution *a priori* normale. De plus, la méthode

d'estimation MAP avec *a priori* de Jeffreys donne des valeurs similaires à celles obtenues avec MAP avec une distribution *a priori* normale, ce qui n'est pas surprenant, car seule la distribution *a priori* diffère selon la méthode. Enfin, le comportement des courbes de biais associées à WL et MAP (avec *a priori* de Jeffreys) est conforme aux résultats de Magis et Raîche (2012) : ces méthodes d'estimation donnent des valeurs similaires pour les valeurs positives du niveau d'habileté, et WL donne des estimateurs du niveau d'habileté inférieurs à ceux obtenus selon MAP pour des valeurs négatives du niveau d'habileté (bien que fort proches, ce qui explique une différence de biais très réduite).

Finalement, il est important de mentionner que le choix des items initiaux des tests adaptatifs ne semble pas influencer beaucoup les valeurs du biais selon les méthodes d'estimation du niveau d'habileté, et ce, quelle que soit la longueur des tests. Cela tend à indiquer que même le choix d'un seul item conduit à des courbes de biais similaires à celles obtenues en choisissant plusieurs items initiaux.

Les figures 7.4 à 7.6 présentent les courbes des RMSE selon les quatre méthodes d'estimation du niveau d'habileté et, pour chaque règle initiale, pour des longueurs de tests respectives de 10 items, 15 items et 20 items. Les valeurs du RMSE diminuent lorsque la longueur du test augmente, ce qui était attendu compte tenu de l'augmentation de la quantité d'information disponible pour l'estimation des niveaux d'habileté. De plus, on retrouve la forme typique des courbes de RMSE, c'est-à-dire des courbes quadratiques en fonction des valeurs du niveau d'habileté : le minimum se situant approximativement autour de la valeur moyenne (soit zéro) du niveau d'habileté. Notons toutefois que les courbes associées à la méthode ML affichent une forme légèrement différente avec un sommet pour les valeurs du niveau d'habileté variant entre 1 et 2. Cela est dû à un accroissement important du biais de l'estimateur du niveau d'habileté obtenu à partir de la méthode d'estimation selon ML.

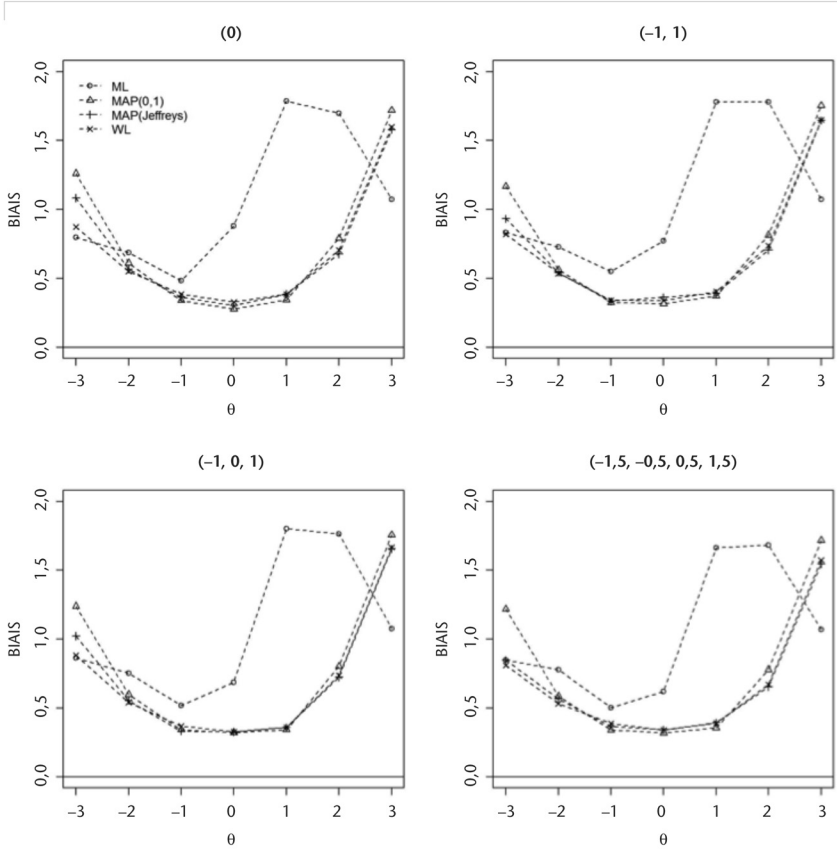


Figure 7.4.
Courbes des racines carrées des carrés moyens résiduels (RMSE) pour les quatre méthodes d'estimation finales du niveau d'habileté et les quatre règles initiales du test, avec le critère d'arrêt de 10 items. Les règles initiales sont représentées par les valeurs initiales du niveau d'habileté (entre parenthèses) choisies pour la sélection des premiers items du test.

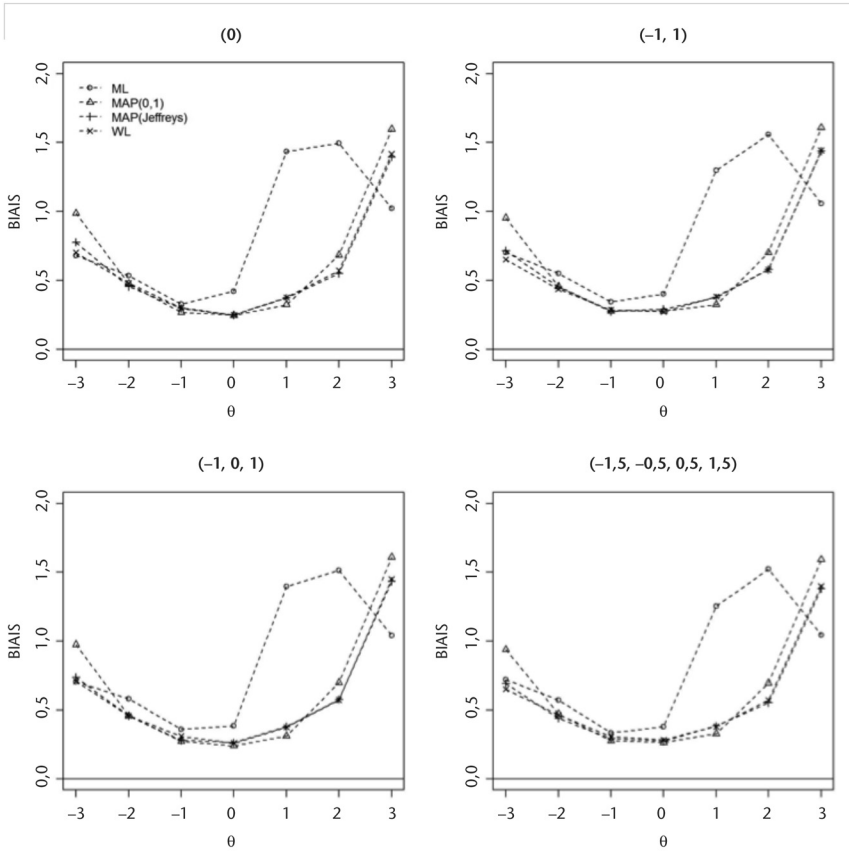


Figure 7.5.
Courbes des racines carrées des carrés moyens résiduels (RMSE) pour les quatre méthodes d'estimation finales du niveau d'habileté et les quatre règles initiales du test, avec le critère d'arrêt de 15 items. Les règles initiales sont représentées par les valeurs initiales du niveau d'habileté (entre parenthèses) choisies pour la sélection des premiers items du test.

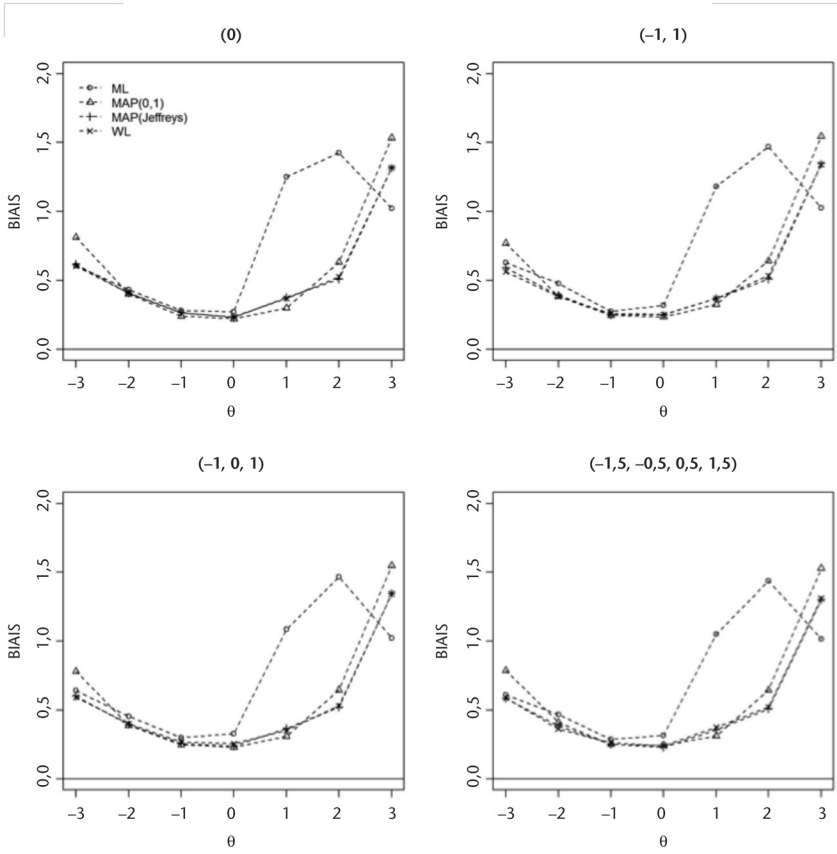


Figure 7.6.
Courbes des racines carrées des carrés moyens résiduels (RMSE) pour les quatre méthodes d'estimation finales du niveau d'habileté et les quatre règles initiales du test, avec le critère d'arrêt de 20 items. Les règles initiales sont représentées par les valeurs initiales du niveau d'habileté (entre parenthèses) choisies pour la sélection des premiers items du test.

De plus, les trois autres méthodes d'estimation (MAP avec distribution *a priori* normale, MAP avec *a priori* non informatif de Jeffreys et WL) présentent des courbes de RMSE très proches les unes des autres, et cela, à l'exception des valeurs extrêmes de l'habileté (–3 et 3) où l'écart est plus important, quoique limité. La méthode d'estimation selon MAP avec distribution *a priori* normale semble présenter des valeurs de RMSE légèrement supérieures à celles des deux autres méthodes d'estimation, mais cela n'a rien d'étonnant puisque les courbes de biais indiquent également un biais légèrement plus important pour cette

méthode. L'écart entre les valeurs des RMSE s'explique donc par un écart similaire entre les valeurs de biais. Cela révèle aussi que MAP avec distribution *a priori* normale n'a pas une variabilité (une erreur type) plus importante que les deux autres méthodes d'estimation.

Notons aussi que les valeurs de RMSE sont plus importantes pour les grandes valeurs du niveau d'habileté (de l'ordre de 2 à 3) que pour les valeurs négatives correspondantes (de 2 à –3). Cette dissymétrie des courbes de RMSE est là aussi imputable à un biais (négatif) plus important pour les valeurs positives du niveau d'habileté que le biais (positif) pour les petites valeurs du niveau d'habileté. Une fois encore, la dissymétrie s'explique par les courbes de biais.

Enfin, on constate à nouveau que le choix des premiers items du test n'influence pas les courbes de RMSE. Cette constatation était déjà valable pour les courbes de biais : il est logique de la retrouver à nouveau pour les valeurs de RMSE. En somme, le choix d'un ou de plusieurs items pour initialiser le test adaptatif n'a pas (ou a peu) d'influence sur la qualité de l'estimateur final du niveau d'habileté, indépendamment de la méthode d'estimation du niveau d'habileté utilisée et de la longueur du test, aspects étudiés à l'intérieur de cette recherche.

5. DISCUSSION

Cette étude avait pour but d'évaluer deux aspects importants du testing adaptatif informatisé sur base de la banque d'items du TCALS II : le choix d'une méthode d'estimation du niveau d'habileté approprié et le choix optimal des premiers items du test. Dans le premier cas, il est apparu que trois des quatre méthodes d'estimation du niveau d'habileté se comportent de façon similaire pour le calcul de l'estimateur finale du niveau d'habileté : les deux méthodes d'estimation selon MAP considérées et la méthode selon WL. Par contre, la méthode d'estimation selon ML semble plus biaisée que les trois autres, surtout pour les valeurs positives du niveau d'habileté. Par conséquent, sa RMSE est aussi plus importante. Étant donné que, de plus, cette méthode d'estimation du niveau d'habileté ne surpasse pas les trois autres dans le cas de valeurs négatives du niveau d'habileté (sauf dans le cas du biais avec peu d'items administrés), il est recommandé de ne pas utiliser cette méthode d'estimation du niveau d'habileté pour des analyses futures du TCALS II sous forme adaptative. Toutefois, le choix d'une méthode d'estimation initiale ne peut être clairement exprimé ici, puisque, sur la base des résultats obtenus aux simulations, les trois autres méthodes d'estimation constituent de bonnes compétitrices avec des performances semblables.

Il est évident que le fait de permettre la sélection de plusieurs items initiaux selon une séquence prédéfinie et éventuellement variable selon les sujets est une bonne stratégie pour limiter la surexposition des items et décourager les tentatives de fraude. Néanmoins, cette étude a montré les limites de telles stratégies en termes d'amélioration de la qualité des estimateurs finaux du niveau d'habileté. Il est en effet apparu très clairement que le choix d'un ou de plusieurs items au début du test n'a presque pas d'incidence sur la qualité de l'estimation du niveau d'habileté des sujets. Cela ne doit toutefois pas remettre en cause l'utilisation d'une telle approche, mais le gain réel d'une stratégie de sélection de plusieurs items initiaux doit être évalué en fonction d'un autre critère objectif (que celui de l'estimation finale du niveau d'habileté).

Il est également intéressant de constater que, indépendamment de la méthode d'estimation du niveau d'habileté ou du choix des items initiaux du test, les valeurs du biais (en valeur absolue) et de la RMSE sont plus importantes pour les valeurs positives du niveau d'habileté. On peut alors en déduire que le TCALS II adaptatif est plus adapté aux faibles niveaux d'habileté. Ce qui n'est pas un fait nouveau, car Raîche (2002) avait déjà remarqué la dissymétrie des niveaux de difficulté et d'information des items du TCALS II, en faveur des faibles niveaux d'habileté. Autrement dit, le TCALS II est un test plutôt facile et ainsi peu informatif pour les niveaux d'habileté moyens à élevés. Il n'est dès lors pas surprenant de retrouver la même tendance dans sa version adaptative, le manque d'items informatifs pour les niveaux d'habileté élevés se traduisant dans ce cas par un biais plus élevé (et une RMSE aussi plus élevée). Le choix optimal d'une méthode d'estimation du niveau d'habileté ou des items initiaux ne peut malheureusement pas pallier cette lacune du TCALS II.

Une extension intéressante de cette étude consisterait à utiliser un autre critère d'arrêt que celui de la longueur du test, par exemple en fixant une erreur type minimale à l'estimateur du niveau d'habileté (critère de précision). Cette extension pourrait relever des différences plus marquées entre les méthodes d'estimation du niveau d'habileté et établir la longueur du test adaptatif requise pour atteindre ce niveau de précision. De plus, cela fournirait une base pratique importante pour les utilisations futures du TCALS II sous format adaptatif. Il serait en effet possible de prévoir le nombre d'items à administrer en fonction du niveau d'habileté pour atteindre un degré de précision souhaité lors de l'estimation finale du niveau d'habileté.

Pour terminer, il faut souligner que les résultats obtenus ne s'appliquent qu'au TCALS II. Un test dont les items présenteraient des valeurs des paramètres d'items différentes se comporterait différemment. Il convient également de souligner que le nombre d'items constituant la banque d'items est très limité : il est évident que la mise en œuvre d'une version adaptative du TCALS II ou d'un autre test adaptatif informatisé exigerait un plus grand nombre d'items disponibles.

Annexe

Code R (annoté) pour les simulations informatisées

```
#  Chargement de la librairie catR et lecture de la banque d'items du
#  TCALS II
require(catR) ; data(tcals)
## ……
#  Fonction permettant de générer K patrons de réponses pour des
#  répondants de niveaux d'habileté TH, selon les quatre critères
#  (start, test, stop, final) choisis parmi les critères définis
#  ci-dessus, et qui retourne les K estimations finales du
#  niveau d'habileté
Gen <- function(K, TH, start, test, stop, final){
 res <- NULL
 for (i in 1:K){
  pr  <- randomCAT(trueTheta = TH, itemBank = BANK,
                cbControl = cbList, start = start,
                test = test, stop = stop, final = final)
  res <- c(res, pr$thFinal)
  }
 return(res)
 }
# ……
preparation <- function(theta) {
 theta        <- theta
 RES         <- NULL
 for (i in 1:4) {
  for (j in 1:4) {
   for (k in 1:3) {
    for (l in 1:length(theta)) {
     RES <- rbind(RES,c(i,j,k,theta[l],NA,NA))
     }
    }
   }
  }
 colnames(RES) <- c("METH", "START", "STOP", "TH", "BIAS", "RMSE")
 return(RES)
}
# ……

# ……
#  Fonction d'initialisation avec des variables globales
#  ……
initialisation <- function(theta) {
 # Création de la banque d›items au format adéquat et avec contrôle
 # de l'équilibre de contenu
 BANK <<- createItemBank(tcals, cb=TRUE)

# Création des quatre phases initiales des tests (selon le nombre
# d'items initiaux)
start1 <<- list(nrItems=1, theta=0,
           startSelect="MFI")
```

```
start2 <<- list(nrItems=2, theta=0, halfRange=1,
           startSelect="MFI")
start3 <<- list(nrItems=3, theta=0, halfRange=1,
           startSelect="MFI")
start4 <<- list(nrItems=4, theta=0, halfRange=-1.5,
           startSelect="MFI")

test1 <<- list(method="ML", itemSelect="MFI", randomesque=3)
test2 <<- list(method="BM", priorDist="norm", priorPar=c(0,1),
          itemSelect="MFI", randomesque=3)
test3 <<- list(method="BM", priorDist="Jeffreys",
          itemSelect="MFI", randomesque=3)
test4 <<- list(method="WL", itemSelect="MFI", randomesque=3)
cbList <<- list(names=c("Audio1", "Audio2", "Written1", "Written2",
                  "Written3"),
           props=c(0.14,0.25,0.15,0.2,0.26))
stop1 <<- list(rule="length", thr=10)
stop2 <<- list(rule="length", thr=15)
stop3 <<- list(rule="length", thr=20)

final1 <<- list(method="ML")
final2 <<- list(method="BM", priorDist="norm", priorPar=c(0,1))
final3 <<- list(method="BM", priorDist="Jeffreys")
final4 <<- list(method="WL")
RES <<- preparation(theta)
}
## .....
## .....
#  Génération des figures 1 à 4 (au format pdf ou à l'écran)
#  .....
plotFigures <- function(RES, stop=1, type="BIAS", pdf=FALSE) {
if (!(type%in% c("BIAS", "RMSE"))) {
stop(«type doit prendre la valeur 'BIAS' ou 'RMSE'»)
}
if (!(stop%in% 1:4)) {
stop("stop doit prendre la valeur 1, 2, 3 ou 4")
}
if (dev.cur()!= 2) dev.new()
nItems <- switch(stop,
              "1" = 10,
              "2" = 15,
              "3" = 20)
fileName <- paste("STOP1_", type, "_", nItems, ".pdf", sep="")
if (pdf) pdf(file=fileName, width=10, height=10)
mainInfo <- c("(0)","(-1, 1)","(-1, 0, 1)",
            "(-1.5, -0.5, 0, 0.5, 1.5)")
yLimBIAS <- round(range(RES$BIAS), 1)
yLimRMSE <- c(0, round(max(RES$RMSE), 1))
if (type == "BIAS") yLim <- yLimBIAS else yLim <- yLimRMSE
minTheta <- round(min(RES$TH),1)
```

```
if (pdf) cexLegend <- 1 else cexLegend <- 0.7
par(mfrow=c(2,2))
for (start in 1:4) {
plot (RES[RES$METH==1 & RES$START==start & RES$STOP==stop,4],
    RES[RES$METH==1 & RES$START==start & RES$STOP==stop,type],
    type="o", lty=2, ylim=yLim, xlab=expression(theta),
    ylab=type,
    main=mainInfo[start], cex.main=cexLegend)
lines(RES[RES$METH==2 & RES$START==start & RES$STOP==stop,4],
    RES[RES$METH==2 & RES$START==start & RES$STOP==stop,type],
    type="o", lty=2, pch=2)
lines(RES[RES$METH==3 & RES$START==start & RES$STOP==stop,4],
    RES[RES$METH==3 & RES$START==start & RES$STOP==stop,type],
    type="o", lty=2, pch=3)
lines(RES[RES$METH==4 & RES$START==start & RES$STOP==stop,4],
    RES[RES$METH==4 & RES$START==start & RES$STOP==stop,type],
    type="o", lty=2, pch=4)
abline(h=0)
if (start == 1 & type == "BIAS") {
legend(minTheta, yLimBIAS[1]+.7,
     c("ML","MAP(0,1)","MAP(Jeffreys)","WL"), lty=rep(2,4),
     pch=1:4, bty="n", cex=cexLegend)
  }
 if (start == 1 & type == "RMSE") {
  legend(minTheta, yLimRMSE[2],
       c("ML","MAP(0,1)","MAP(Jeffreys)","WL"), lty=rep(2,4),
       pch=1:4, bty="n", cex=cexLegend)
  }
 }
if (pdf) dev.off()
}
## ..........

## ..........
# Fonction de simulation des patrons de réponses, calculs des
# valeurs de biais et de RMSE et encodage des résultats dans la
# matrice RES créée ci-avant ainsi que la production des figures
# ..........
simulation <- function(nSubjects, theta=-3:3, save=TRUE,
                  append=FALSE, figures=TRUE, pdf=FALSE) {
cat(«Départ : «, as.character(Sys.time()), «\n», sep=»»)
initialisation(theta)
nValues <- 4*4*3*length(theta)
for (i in 1:nValues) {
TEST <- test1
FINAL <- final1
if (RES[i,1]==2) {
TEST <- test2
FINAL <- final2
}
if (RES[i,1]==3) {
TEST <- test3
FINAL <- final3
```

```
}
if (RES[i,1]==4) {
TEST <- test4
FINAL <- final4
}
START <- start1
if (RES[i,2]==2) START <- start2
if (RES[i,2]==3) START <- start3
if (RES[i,2]==4) START <- start4
STOP <- stop1
if (RES[i,3]==2) STOP <- stop2
if (RES[i,3]==3) STOP <- stop3
TH <- RES[i,4]
pr <- gen(nSubjects, TH=TH, start=START, test=TEST, stop=STOP,
       final=FINAL)
RES[i,5] <- mean(pr-TH)
RES[i,6] <- sqrt(mean((pr-TH)^2))
cat("(",i,") ", as.character(Sys.time()), "\n", sep="")
}
fileName <- paste("RES _ PUQ _ ", nSubjects, ".txt", sep="")
if (save & append & file.exists(fileName)) {
write.table(RES, fileName, sep=" ", dec=".", quote=FALSE,
        append=append, col.names=FALSE)
} else
if (save) {
write.table(RES, fileName, sep=" ", dec=".", quote=FALSE,
append=append)
}
RES <- data.frame(RES)
if (figures) {
for (stop in 1:3) {
plotFigures(RES, stop=stop, type="BIAS", pdf=pdf)
plotFigures(RES, stop=stop, type="RMSE", pdf=pdf)
}
}

invisible(RES)
}
## .....

## .....
# Résultats et production des figures
# .....
N <- 1000
RES <- simulation(nSubjects=N, theta=-3:3, append=FALSE,
              figures=TRUE)
## .....
# Pour obtenir les résultats préalablement produits
# 1000 correspond à la valeur de nSubjects
# .....
RES <- read.table(paste(«RES _ PUQ _», N, «.txt»,
                  sep=»»))
```

RÉFÉRENCES

Baker, F. B. (1992). *Item response theory : parameter estimation techniques*. New York, New Jersey : Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models. Dans F. M. Lord et M. R. Novick (dir.), *Statistical theories of mental test scores*. Reading, Massachusetts : Addison-Wesley.

Choi, S. W. et Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied psychological measurement*, *32*, 419-440.

Green, B. F. (1983a). Adaptive testing by computer. Dans R. B. Ekstrom (dir.), *Measurement technology, and individuality in education : new directions for testing and measurement*. San Francisco, Californie : Jossey-Bass.

Green, B. F. (1983b). The promise of tailored tests. Dans H. Wainer et S. Messick (dir.), *Principals of modern psychological measurement*. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Kingsbury, G. G. et Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied measurement in education*, *2*, 359-375.

Laurier, M., Froio, L., Pearo, C. et Fournier, M. (1998). *L'élaboration d'un test provincial pour le classement des étudiants en anglais langue seconde au collégial*. Québec, Québec : Direction générale de l'enseignement collégial, ministère de l'Éducation du Québec.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*, 233-245.

Magis, D. et Raîche, G. (2011). *catR : an R package to generate IRT adaptive tests*. R package version 2.0.

Magis, D. et Raîche, G. (2012a). On the relationships between Jeffreys modal and weighted likelihood estimation of ability under logistic IRT models. *Psychometrika*, *77*, 163-169.

Magis, D. et Raîche, G. (2012b). Random generation of response patterns under computerized adaptive testing with the R package *catR*. *Journal of statistical software*, *48*(8), 1-31.

Raîche, G. (2002). *Le dépistage du sous-classement aux tests de classement en anglais, langue seconde, au collégial*. Gatineau, Québec : Collège de l'Outaouais.

Urry, V. W. (1970). *A Monte Carlo investigation of logistic test models*. Thèse de doctorat inédite. West Lafayette, Indiana : Purdue University.

Van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216.

Veerkamp, W. J. J. et Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of educational and behavioral statistics*, *22*, 203-226.

Wainer, H. (1990). *Computerized adaptive testing : a primer*. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Wainer, H. (2000). *Computerized adaptive testing : a primer* (2e éd.). Mahwah, New Jersey : Lawrence Erlbaum Associates.

Weiss, D. (1983). *New horizons in testing : latent trait theory and computerized adaptive testing*. New York, New Jersey : Academic Press.

LISTE DES CONTRIBUTEURS

Jaouad ALEM
Université Laurentienne, Sudbury, Ontario, Canada
jalem@laurentienne.ca

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

Marc CLOES
Université de Liège, Liège, Belgique
marc.cloes@ulg.ac.be

Éric DIONNE
Université d'Ottawa, Ottawa, Ontario, Canada
eric.dionne@uottawa.ca

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

Michel GUAY
Université Laurentienne, Sudbury, Ontario, Canada
mguay@laurentienne.ca

Nabil KERFES
Université d'Alger, Alger, Algérie
Kerfes23@yahoo.fr

Diana KOSZYCKI
Université d'Ottawa, Ottawa, Ontario, Canada
dkoszyck@uottawa.ca

Marie-Ève LATREILLE
Université d'Ottawa, Ottawa, Ontario, Canada
mlatr085@uottawa.ca

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

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

Hélène MEUNIER
Université du Québec à Montréal, Montréal, Québec, Canada
meunier.h@uqam.ca

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

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

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

RÉSUMÉS EN ANGLAIS

CHAPTER 1

How college institutional assessment policies warrant inference validity of learning assessment

Moyens relevés dans les politiques institutionnelles d'évaluation des apprentissages (PIEA) pour assurer la validité des inférences en évaluation des apprentissages au collégial

Karine Paquette-Côté et Gilles Raîche

This exploratory research is a first attempt to validate the Kane's interpretive argument (2006) by the application of this structure to the analysis of institutional student evaluation policies (ISEP) of the college system in Quebec. One of its goals was to identify ways that can be established by the institutions to help ensure the validity of evaluation inferences in the light of student learning. A content analysis of institutional student evaluation policies (ISEP) and a schematic model were carried out starting from the Kane's interpretive argument structure (2006), leading to the development of guidelines to seek to ensure the argumentation of the validity of evaluation inferences in the context of the evaluation of learning at the college level.

CHAPTER 2
Comparison of technological tools for Rash models data analysis
Comparaison d'outils technologiques permettant l'analyse de données à l'aide des modèles de Rasch
Éric Dionne, Julie Grondin et Jean-Guy Blais

The modeling of measurement is an important operation that can support the construct validity in evaluation or research. Many types of software can now model the raw scores. As a part of this presentation, we present the results of a comparative analysis of two programs (RUMM2020 and Winsteps), which is as much about the ergonomics of the interface as on the estimation methods (parameters, data-models fit) or on the information produced by such softwares. The analysis shows that the two softwares differ in their user-friendliness and their utility for certain types of research, but that overall each one offers comparable results both at the measurement level and at the accuracy of the latter.

CHAPTER 3
General physical ability measurement for higher education in physical education and sport in Morocco and Algeria
Mesure de l'aptitude physique générale lors des épreuves de sélection pour les études supérieures en éducation physique et sport au Maroc et en Algérie
Jaouad Alem, Marc Cloes, Michel Guay et Nabil Kerfes

The aim of this research is to analyze the construct validity of several physical tests designed to measure the physical qualities of students interested to follow a higher education programme in physical education and sport. We have considered a sample of 1481 male students being in average 20 years old from two different countries (Morocco and Algeria). Principal component factorial analyses with varimax rotation of performance physical tests reveal rather a solution of two components that differ according to the duration of the work to produce energy, as follows: the muscular glycolitic power output and the muscular phosphagenic power output. The first component corresponds to the capacity to produce lactate within more than twelve seconds; it is defined by the speed race and the resistance race. The second component is the ability to produce phosphate which is already in muscles within less than seven seconds; it is defined by the other physical tests.

CHAPTER 4
Simulation as a teaching and assessment technic in nursing: a review
La simulation comme technique d'enseignement et d'évaluation en sciences infirmières: un état de la question
Marie-Ève Latreille, Éric Dionne et Diana Koszycki

In health care, simulation is a teaching strategy and an evaluation instrument that is frequently used. It is a good example of a performance-based

evaluation in a formative or summative context. In this article, we are exposing the results of several writings in regards to a variety of simulation models while demonstrating the different advantages and limits relating to nursing studies. Our findings tend to demonstrate that simulations do provide greater value in teaching techniques and allow us to obtain new data in regards to the level of understanding of the student's knowledge.

CHAPTER 5
CEGEP's assessment context in arts and contemporary dance education
Portrait d'un contexte pour évaluer les apprentissages en arts plastiques et en danse contemporaine au collégial
Diane Leduc, Jean-Guy Blais et Gilles Raîche

Prior to the Quiet Revolution, learning assessments in art classes mostly dealt with the finished product and, in dance classes, on the observation of physical abilities. During the 1960s, in a process of emancipation, the Quebec school system was extensively reviewed. Stemming from this reconstruction, the CEGEPs brought a breath of fresh air and allowed young people to pursue technical and undergraduate studies. Arts education and learning assessment practices did not escape these new structures and inherited imperatives to which teachers had to adapt. The 1993 reform led to more adjustments in college education, thereby forcing teachers to switch to skills-based approaches.

CHAPTER 6
School context and assessment task authenticity: a review
Authenticité des tâches d'évaluation en milieu scolaire : état des lieux
Pascal Ndinga

To update on the assessment "task authenticity" concept in the education field, we conducted a literature review 10 years after the concept's introduction following the implementation of the new school reform in Quebec. The documents consulted and the resulting analysis allowed us to notice the quasi-fallow state of this issue, both on the practical and research fronts. These documents are still at the characterization stage. Furthermore, almost all of the documents display a proven link to the statements made by Wiggins (1989), proof of the lack of scientific development in this field. Research must transcend ambient characterization and develop and validate assessment tools to conduct correlational and experimental studies. This challenge is particularly incumbent on education assessment researchers.

CHAPTER 7

Impact of ability estimation method and choice of first items on adaptive administration efficacy of the TCALS II

Impact de la méthode d'estimation du niveau d'habileté et du choix des premiers items sur l'efficacité de l'administration adaptative du TCALS II

David Magis et Gilles Raîche

The English as a Second Language Placement Test at the college level, 2nd version (TCALS II), is a questionnaire composed of a fixed number of 85 items administered to Quebec students beginning college-level studies. A computer-adaptive version of this test is being considered for the first time in this study. Two problems are studied more closely : the choice of a method used to assess the optimal skills level and the selection of the first items of the test. Both problems are studied simultaneously through a Monte Carlo simulation and using various stopping rules related to the length of the test. We concluded that the choice of the first items have little impact on the assessment of the skills level, while there are significant differences between the four assessment methods compared. Certain conclusions and recommendations are drawn for the pursuit of this work.

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 la suite d'un colloque organisé en mai 2010 à Montréal au Canada lors du 78e congrès annuel de l'Association francophone pour le savoir (ACFAS). Ce colloque visait à faire le point sur les avancées accomplies ainsi que les défis qui se posent dans le domaine de la mesure et de l'évaluation en éducation afin de formuler des propositions et des stratégies susceptibles de répondre aux préoccupations actuelles et de rendre les jugements d'évaluation plus valides. Quatre grands axes composaient ce colloque : la mesure, l'évaluation, la technologie et les aspects pratiques de la mesure et de l'évaluation. C'est ce quatrième axe qui fait l'objet de ce troisième volume. Au regard de la mesure, il aborde l'évaluation assistée par ordinateur et les tests adaptatifs, l'utilisation des logiciels permettant l'estimation des paramètres d'items et de personnes dans le contexte des modèles de Rasch, de même que l'application de l'analyse factorielle exploratoire à la validation conceptuelle. En ce qui a trait à l'évaluation, il s'attarde aux politiques institutionnelles d'évaluation des apprentissages, aux modèles de simulation pour soutenir l'évaluation des apprentissages, à la notion d'authenticité des tâches d'évaluation et, enfin, à l'évolution des pratiques d'évaluation des apprentissages en arts plastiques et en danse.

GILLES RAÎCHE est professeur au Département d'éducation et pédagogie à l'Université du Québec à Montréal. Rédacteur en chef de la Revue des sciences de l'éducation, *il est aussi directeur du Collectif pour le développement et les applications en mesure et évaluation (Cdame).*

PASCAL NDINGA est professeur au Département d'éducation et pédagogie à l'Université du Québec à Montréal. Il est directeur du programme de baccalauréat en enseignement au secondaire et président de l'ADMEE-Canada. Il est spécialiste en évaluation des apprentissages.

HÉLÈNE MEUNIER est doctorante en éducation à l'Université du Québec à Montréal. Spécialisée en évaluation des apprentissages, elle s'intéresse aux principes et aux applications du portfolio comme instrument d'évaluation des apprentissages.

ONT COLLABORÉ À CET OUVRAGE

Jaouad Alem
Jean-Guy Blais
Marc Cloes
Éric Dionne
Julie Grondin
Michel Guay
Nabil Kerfes
Diana Koszycki
Marie-Ève Latreille
Diane Leduc
David Magis
Hélène Meunier
Pascal Ndinga
Karine Paquette-Côté
Gilles Raîche

www.puq.ca

9 782760 535930
ISBN 978-2-7605-3593-0