MESURE ET ÉVALUATION

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

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

VOLUME 1 – La mesure

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 1 – La mesure

DANS LA MÊME COLLECTION

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, 196 pages

Membre de L'ASSOCIATION NATIONALE DES ÉDITEURS DE LIVRES

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

Diffusion/Distribution :

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

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

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

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

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

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

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

VOLUME 1 – La mesure

Presses de l'Université du Québec

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

Vedette principale au titre :

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

Comprend des réf. bibliogr.

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

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

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

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

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

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

Mise en pages : Info 1000 mots

Conception de la couverture : Richard Hodgson

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

TABLE DES MATIÈRES

Chapitre 1

Une solution numérique au test de Cattell pour déterminer le nombre de composantes principales à retenir

Gilles Raîche, David Magis, Theodore A. Walls,
Martin Riopel et Jean-Guy Blais

La détermination du nombre de composantes à retenir a toujours été une préoccupation pour les utilisateurs de l'analyse en composantes principales. Cattell suggère une approche graphique et subjective. Ne serait-il toutefois pas possible d'établir une solution numérique au test de Cattell ? À cette fin, un indice est développé. Celui-ci consiste en un facteur d'accélération tributaire de la dérivée seconde calculée à chacune des composantes principales. On donnera deux exemples d'application de cet indice.

La détermination du nombre de composantes ou de facteurs à retenir a été une préoccupation constante pour les utilisateurs de l'analyse en composantes principales (ACP) et de l'analyse factorielle exploratoire (AFE). Plusieurs stratégies ont été proposées. Ajar (1978, p. 5-16 ; 1982, p. 46-49) classe celles-ci en deux catégories : psychométrique et statistique. La première de ces catégories, soit celle des stratégies psychométriques, ne tient pas compte de l'erreur échantillonnale (Ajar, 1978, p. 6) et a plutôt recours à des stratégies empiriques pour régler le problème du nombre de composantes à retenir. La catégorie des stratégies statistiques, pour sa part, tient compte de l'erreur échantillonnale (Ajar, 1978, p. 10) et repose sur des tests d'hypothèse. Pour cette raison, elle est dite statistique. Dans ce chapitre, on ne prend en considération que la catégorie des stratégies psychométriques.

Parmi les stratégies psychométriques, il y en a trois qui sont adoptées plus fréquemment par les chercheurs et les praticiens. La première et la troisième sont d'ailleurs intégrées dans les logiciels les plus utilisés pour effectuer des analyses en composantes principales ou des analyses factorielles exploratoires : SAS, SPSS et Systat, par exemple. La première de celles-ci repose sur la décision de retenir le nombre de composantes principales d'après l'importance des valeurs propres associées à chacune de celles-ci. Cette stratégie a été employée par Kaiser (1960), se basant sur les travaux de Guttman (1954), et elle consiste, en particulier, à déterminer le nombre de composantes principales en fonction du nombre de valeurs propres supérieures à l'unité.

Dans une deuxième stratégie, Horn (1965), ainsi que Montanelli et Humphreys (1976 ; Franklin, Gibson, Robertson, Pohlmann et Fralish, 1995), suggèrent, pour leur part, une variante de la stratégie de Kaiser où le nombre de composantes principales est déterminé en fonction du nombre de valeurs propres supérieures aux valeurs propres obtenues au hasard dans un échantillon de matrices de corrélations. Il s'agit d'une méthode qui requiert de nombreuses opérations mathématiques, car elle exige d'effectuer un grand nombre d'observations aléatoires afin d'obtenir plusieurs matrices de corrélations. La puissance actuelle des ordinateurs facilite toutefois de plus en plus son application et elle est utilisée de plus en plus souvent. La méthode de Horn, elle, n'est pas disponible, à notre connaissance, par défaut, dans la plupart des logiciels utilisés par les chercheurs ou les praticiens ; il est généralement nécessaire d'en effectuer la programmation.

Enfin, Cattell (1966) suggère une troisième approche, soit de déterminer le nombre de composantes principales en fonction du point de rupture de la courbe des valeurs propres. Il s'agit d'une approche graphique de la détermination du nombre de composantes principales à retenir. Le test de Cattell présuppose l'appréciation par des juges, appréciation qui n'est malheureusement pas exacte (Ajar, 1978, p. 8-9 ; Tabachnick et Fidell, 2001, p. 621 ; Zwick et Velicer, 1986, p. 434). Ne serait-il toutefois pas possible d'éviter que cette dernière méthode repose uniquement sur l'appréciation de juges et, ainsi, qu'elle fournisse plutôt une solution numérique ? C'est ce que nous nous proposons d'établir dans ce chapitre.

Apportons quelques précisions au sujet du test de Cattell. Ensuite, nous en proposerons une solution numérique au moyen d'un indice. Enfin, nous donnerons deux exemples d'utilisation de cet indice.

1. TEST DE CATTELL

Cattell, comme il le souligne (1966, p. 249), après une expérience étalée sur 30 ans où il a effectué plus d'une centaine d'analyses en composantes principales et d'analyses factorielles exploratoires, en vient à suggérer une stratégie de détermination du nombre de composantes principales à retenir qu'il nomme le test de l'éboulis (*scree test*) (1966, p. 249). Cette appellation provient de l'allure de la représentation graphique des valeurs propres de chacune des composantes principales en fonction du rang de celles-ci dans une analyse en composantes principales. Une telle représentation affiche, au départ, une pente abrupte pour tendre vers une ligne droite autour de laquelle les valeurs propres varient de façon irrégulière. Une telle courbe fait penser à l'accumulation de débris qui tombent rapidement pour rebondir ensuite au pied d'une montagne de déchets.

Cattell suggère de retenir uniquement les valeurs propres qui surplombent le bas de la courbe, soit celles supérieures à la valeur propre où apparaît un point de rupture avec la courbe abrupte initiale. À noter que ces valeurs propres doivent tout de même être égales ou supérieures à l'unité.

La figure 1.1 illustre une représentation graphique typique de la courbe des valeurs propres en fonction du rang des composantes principales. Il s'agit d'un exemple tiré des travaux de Banville, Richard et Raîche (2004). Les valeurs propres ont été obtenues à partir de la matrice des corrélations polychoriques entre les 11 styles d'enseignement en éducation physique décrits par Mosston et Ashworth (1990, 2002).

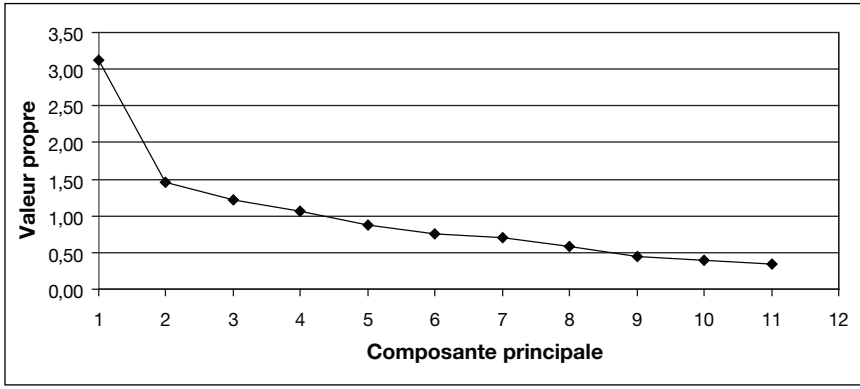


Figure 1.1
Valeurs propres associées à la matrice des corrélations polychoriques issue des travaux de Banville, Richard et Raîche (2004), *N* = 370

Selon la stratégie proposée par Cattell, à la figure 1.1, le point de rupture pourrait se situer à la deuxième valeur propre. Une seule valeur propre est supérieure à la deuxième valeur propre : une seule composante principale est alors retenue. Ce choix est toutefois quelque peu arbitraire, car, dans ce cas particulier, un autre observateur pourrait prétendre qu'il est exagéré de considérer que le point de rupture se situe à la deuxième valeur propre.

Ce type de situation a mené les chercheurs et les praticiens à déterminer la localisation du point de rupture selon deux approches ; celles-ci n'étant d'ailleurs pas toujours bien précisées dans leurs écrits. Selon la première approche, le point de rupture serait situé à l'endroit où on observe le changement le plus important dans la pente de la courbe, soit la plus importante accélération. Ces chercheurs et praticiens ne tiennent compte alors uniquement que d'un premier aspect de la description du test par Cattell : la pente abrupte. Selon cette interprétation, à la figure 1.1, on pourrait retenir une seule composante principale.

D'autres optent plutôt pour la recherche de la valeur propre qui permettrait d'observer une ligne droite où les valeurs propres varient aléatoirement autour de celle-ci. Ils s'intéressent alors au second aspect de la description de Cattell et ainsi aux valeurs propres associées au bas de la courbe. Selon cette seconde interprétation, toujours à la figure 1.1, il semble difficile de déterminer le nombre de composantes principales à retenir : une ou deux éventuellement. Il faut souligner que, selon nous, cette interprétation semble être celle que Cattell a appliquée. Hoyle et Duvall en présentent une excellente illustration (2004, p. 304-305).

On voit bien que ces deux interprétations du test de Cattell sont difficiles à appliquer. Une simple observation de la représentation graphique de la courbe des valeurs propres en fonction du rang de la composante principale peut ainsi mener à une appréciation arbitraire. On comprend alors pourquoi l'application de ce test a été considérée par plusieurs comme étant trop subjective et que la fidélité des appréciations des juges ait été estimée insuffisante par certains (Ajar, 1978, p. 8-9 ; Tabachnick et Fidell, 2001, p. 621 ; Zwick et Velicer, 1986, p. 434). Hoyle et Duvall (2004, p. 305) indiquent que des coefficients de fidélité des appréciations des juges variant de 0,60 à 0,90 ont été observés chez des juges qui ont reçu une formation : la valeur moyenne de ces coefficients est égale à 0,80. Toutefois, dans la plupart des cas, ceux et celles qui appliquent cette interprétation du test de Cattell ne

tiennent pas compte de l'expérience des juges. Il est alors, bien sûr, impossible de vérifier l'exactitude du coefficient de fidélité qui risque d'ailleurs d'être assez faible.

Compte tenu de ce qui précède, il serait pertinent de proposer une solution mathématique qui permettrait de déterminer le point de rupture et pourrait être implantée à l'intérieur des programmes informatiques utilisés pour effectuer les analyses en composantes principales. Cet indice remplacerait alors l'appréciation subjective de la représentation graphique des valeurs propres. C'est ce que nous visons spécifiquement dans ce chapitre. Nous proposons maintenant une solution numérique à la première des deux interprétations du test de Cattell, soit un indice qui détermine la localisation de la fin de la pente abrupte, indice que nous désignons sous le nom de facteur d'accélération (*FA*).

2. APPROCHE THÉORIQUE

Pour réaliser l'objectif spécifique de cette recherche, deux étapes sont proposées : 1) un développement numérique ; 2) deux exemples d'application.

2.1. Développement numérique

Tout d'abord, il s'agit de proposer une stratégie numérique qui permet de calculer la vitesse de changement de la pente de la courbe des valeurs propres. Cette valeur servira alors d'indice pour localiser la fin de la pente abrupte dans le test de l'éboulis, soit le facteur d'accélération.

2.2. Exemples d'application et critères des exemples

Par la suite, deux exemples d'application du facteur d'accélération sont présentés. Le premier est tiré d'une étude réalisée par Banville, Richard et Raîche (2004) sur l'utilisation par Mosston des styles d'enseignement en éducation physique. Le second correspond à une adaptation de l'interprétation d'une analyse factorielle par Laforge (1981). Dans le premier exemple, le nombre de composantes principales à retenir n'est pas connu. Dans le second exemple, ce nombre est connu et il est égal à trois. À l'intérieur de tableaux, les valeurs propres, le pourcentage de variance expliquée et le facteur d'accélération sont calculés pour chacune des composantes principales. Les calculs sont effectués à partir de la librairie *nFactors* du logiciel R (Raîche, Riopel et Blais, 2006 ; Song, Walls et Raîche, 2008).

3. DÉVELOPPEMENT THÉORIQUE

3.1. Développement numérique du facteur d'accélération – FA ($f'(\lambda)$)

Selon la première approche de localisation du point de rupture des valeurs propres, celui-ci serait situé à l'endroit où on observe le changement le plus important dans la pente de la courbe des valeurs propres λ, soit la plus importante accélération. L'accélération d'une courbe est définie par la dérivée seconde de cette courbe, $f'(\lambda)$.

La courbe des valeurs propres n'étant pas caractérisée par une équation précise, il est toutefois nécessaire de recourir à une approximation numérique. Il est possible d'obtenir la dérivée seconde de cette fonction par l'utilisation des polynômes de Taylor de divers degrés. Les polynômes de Taylor permettent d'obtenir une approximation suffisante pour la plupart des courbes. La dérivée seconde de la fonction étudiée ici ne nécessite pas une précision supérieure à celle donnée par un polynôme de Taylor de cinquième degré.

Selon Yakovitz et Szidarovszky (1986, p. 82), assumant qu'une fonction peut être dérivée au moins quatre fois, on a :

$$f(x_0+h)=f(x_0)+f'(x_0)h+\frac{f''(x_0)}{2!}h^2+\frac{f'''(x_0)}{3!}h^3+\frac{f''''(\varsigma_1)}{4!}h^4 \qquad (1)$$

et

$$f(x_0-h)=f(x_0)-f'(x_0)h+\frac{f''(x_0)}{2!}h^2-\frac{f'''(x_0)}{3!}h^3+\frac{f''''(\varsigma_2)}{4!}h^4 \qquad (2)$$

où $f'(x_0)$, $f''(x_0)$ et $f'''(x_0)$ sont respectivement les dérivées première, seconde et troisième au point x_0, tandis que $f'''(\varsigma_1)$ et $f'''(\varsigma_2)$ correspondent à l'erreur de l'estimation définie par la dérivée quatrième aux points ς_1 et ς_2, avec $\varsigma_1 \in [x_0, x_0+h]$ et $\varsigma_2 \in [x_0, x_0-h]$. La variable h permet de définir un point approché dans le voisinage de x_0.

En réarrangeant les termes des équations 1 et 2, la dérivée seconde est égale à :

$$f''(x_0)=\frac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2}-\frac{h^2}{4!}\left[f''''(\varsigma_1)+f''''(\varsigma_2)\right] \qquad (3)$$

Le premier terme de la fonction, $\frac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2}$, définit une approximation de la dérivée seconde, tandis que le second terme, $\frac{h^2}{4!}\left[f''''(\varsigma_1)+f''''(\varsigma_2)\right]$, correspond à l'erreur d'approximation de cette dérivée seconde. Pour obtenir une approximation de la dérivée seconde, on ne conservera donc que le second terme et ainsi :

$$f''(x_0)=\frac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2} \tag{4}$$

Puisque la dérivée seconde représente l'accélération de la fonction, il s'agira donc du facteur d'accélération (*FA*) auquel on fera correspondre plus spécifiquement les termes de la façon suivante : $x_0 = i$, soit la composante principale *i*, et $f(x_0) = \lambda_i$, la valeur propre associée à la composante principale. Aussi, on peut simplifier l'équation, car ici *h* est toujours égal à 1,00 (donc plus ou moins une valeur propre). On obtient alors le facteur d'accélération :

$$FA = f''(\lambda_i) = \lambda_{i+1} - 2\lambda_i + \lambda_{i-1} \tag{5}$$

Pour indiquer le nombre de composantes principales à retenir, on prendra la valeur maximale de ce facteur, avec la contrainte $\lambda_{i-1} \geq 1{,}00$. Les *i*-1 composantes correspondront au nombre de composantes à retenir.

3.2. Exemples d'application

Le premier exemple provient de l'étude réalisée par Banville, Richard et Raîche (2004) sur l'utilisation des styles d'enseignement par Mosston. C'est le même exemple qui a été présenté, plus haut, à la figure 1.1. Le tableau 1.1, outre les valeurs propres et le pourcentage de variance expliquée, fournit la valeur du facteur d'accélération à chacune des valeurs propres.

Le facteur d'accélération *FA*, avec la contrainte $\lambda_{i-1} \geq 1{,}00$, mène à ne retenir qu'une seule composante principale puisque celui-ci est maximal à la deuxième composante principale (1,42). On notera que cette décision ne permet cependant d'expliquer que 28,36 % de la variance totale.

Tableau 1.1
Valeurs propres associées à la matrice des corrélations polychoriques issue des travaux de Banville, Richard et Raîche (2004), N = 370[1]

Composante principale	λ	Variance cumulative expliquée (%)	FA2 $f''(\lambda_i)$
1	**3,12**	28,36	**na**
2	1,46	41,64	1,42
3	1,22	52,73	0,08
4.	1,06	62,36	–0,02
5	0,88	70,36	0,06
6	0,76	77,27	0,06
7	0,70	83,64	–0,05
8	0,59	89,00	–0,03
9	0,45	93,09	0,09
10	0,40	96,73	0,00
11	0,35	100,00	na

1. La valeur en gras est celle qui permet d'indiquer le nombre de composantes principales à retenir.
2. Il n'est pas possible de calculer la valeur de cet indice aux première et dernière composantes principales.

Le second exemple est tiré de l'adaptation d'une interprétation d'une analyse factorielle proposée par Laforge (1981, p. 185). Il s'agit tout simplement d'une matrice des corrélations de Pearson obtenue à partir de la production aléatoire d'observations dont le contenu et la dimensionnalité, contrairement à l'exemple précédent, sont connus à l'avance. Diverses opérations mathématiques ont été effectuées sur trois variables produites aléatoirement selon une distribution N(0,1) : largeur (L), hauteur (H) et profondeur (P). Les opérations mathématiques sont présentées au tableau 1.2 : 10 variables ont pu ainsi être créées et 1 000 observations ont été effectuées.

Tableau 1.2
Opérations mathématiques effectuées sur les variables utilisées dans le deuxième exemple (adapté de Laforge, 1981, p. 185)

1	2	3	4	5	6	7	8	9	10
L	H	P	2(L+P)	LH	HP	LP	$\sqrt{L^2 + P^2}$	LHP	2(L+H)

La figure 1.2 et le tableau 1.3 présentent les résultats du calcul des valeurs propres de la matrice des coefficients de corrélation de Pearson obtenue. Dans ce cas-ci, le facteur d'accélération (0,50) suggère encore de ne retenir qu'une seule composante principale. Celle-ci explique 25,10 % de la variance totale. On pourrait alors interpréter cette composante principale comme une composante représentant la taille globale des observations.

La même situation se retrouve fréquemment lorsque l'on effectue une analyse en composantes principales sur une matrice de corrélations issue d'un test d'aptitude. On se retrouve alors, dans plusieurs cas, avec une seule composante principale qui peut être interprétée comme une composante de difficulté des items (McDonald et Ahlawat, 1974).

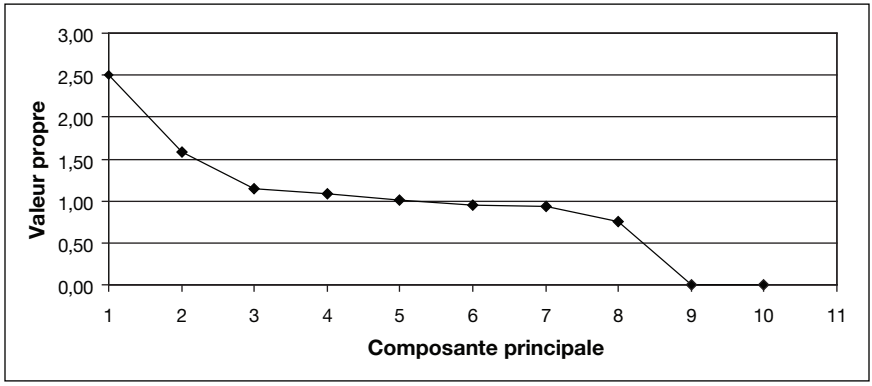


Figure 1.2
Valeurs propres associées au deuxième exemple (adapté de Laforge, 1981, p. 185), N = 1 000

Tableau 1.3
Valeurs propres associées au deuxième exemple (adapté de Laforge, 1981, p. 185), N = 1 000

Composante principale	λ	Variance cumulative expliquée (%)	FA $f''(\lambda_i)$
1	**2,51**	25,10	**na**
2	1,58	40,90	0,50
3	1,15	52,40	0,37
4	1,09	63,30	–0,02
5	1,01	73,40	0,02
6	0,95	82,90	0,05
7	0,94	92,30	–0,17
8	0,76	99,90	–0,58
9	0,00	100,00	0,76
10	0,00	100,00	na

CONCLUSION

Une solution numérique à une des deux interprétations usuelles du test de Cattell a été proposée. Selon cette interprétation, le point de rupture serait situé à l'endroit où on observe le changement le plus important dans la pente de la courbe des valeurs propres, soit la plus importante

accélération. À cette fin, un indice a été développé et deux exemples d'application ont été présentés. Il s'agit d'un facteur d'accélération (*FA*), facteur qui est fonction de la dérivée seconde calculée à chacune des composantes principales.

Cet indice permet une prise de décision à partir d'une solution numérique au test de Cattell : le problème de la variabilité de l'appréciation par plusieurs juges peut ainsi être évité. Il est alors possible d'automatiser le calcul de cet indice à l'intérieur des logiciels d'analyse en composantes principales ou d'analyse factorielle exploratoire. Il faut tout de même souligner que l'utilisation de plus d'une stratégie de détermination du nombre de composantes principales à retenir est toujours de mise : il ne s'agit donc pas de se limiter uniquement à ce facteur d'accélération lors de la prise de décision.

On ne peut toutefois pas porter un jugement sur l'adéquation des solutions obtenues à partir de l'indice développé ni le comparer à d'autres stratégies psychométriques ou statistiques. Ce n'était d'ailleurs pas notre intention : le travail reste donc à faire. À cette fin, il serait alors opportun de comparer l'efficacité de diverses stratégies de détermination du nombre de composantes à retenir, dont celle du facteur d'accélération, avec diverses structures factorielles déterminées à l'avance. Le nombre de variables et la taille des échantillons devraient aussi être pris en compte.

Il faudrait aussi, ultérieurement, penser à faire l'étude de l'erreur type de cet indice et, éventuellement, lui associer un ou des tests d'hypothèse. De cette façon, la taille de l'échantillon pourrait être prise en considération lors de son utilisation. Par exemple, on pourrait déterminer la taille de l'échantillon nécessaire pour assurer la validité des interprétations obtenues à partir du facteur d'accélération.

RÉFÉRENCES

Ajar, D. (1978). *L'invariance factorielle et le problème de l'échantillonnage des sujets*. Thèse de doctorat inédite. Montréal, Québec, Université de Montréal.

Ajar, D. (1982). Le problème de la détermination du nombre de facteurs en analyse factorielle. *Revue des sciences de l'éducation, 8*(1), 45-62.

Banville, D., Richard, J.-F. et Raîche, G. (2004). Utilisation des 11 styles d'enseignement de Mosston chez des éducateurs physiques francophones du Canada en fonction de caractéristiques démographiques. *Avente, 10*(2), 32-44.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research, 1*(2), 245-276.

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T. et Fralish, J. S. (1995). Parallel analysis : a method for determining significant principal components. *Journal of vegetation science, 6*(1), 99-106.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*(2), 149-161.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.

Hoyle, R. H. et Duvall, J. L. (2004). Determining the number of factors in exploratory and confirmatory factor analysis. Dans D. Kaplan (dir.), *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, Californie : Sage.

Kaiser, H. F. (1960). The application of electronic computer to factor analysis. *Educational and psychological measurement, 20*, 141-151.

Laforge, H. (1981). *Analyse multivariée : pour les sciences sociales et biologiques avec applications des logiciels BMD, BMDP, SPSS, SAS*. Saint-Laurent, Québec : Éditions Études vivantes.

McDonald, R. P. et Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of mathematical and statistical psychology, 27*, 82-99.

Montanelli, R. G. et Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal : a Monte Carlo study. *Psychometrika, 41*(3), 341-348.

Mosston, M. et Ashworth, S. (1990). *The spectrum of teaching styles : from command to discovery*. White Plains, New York : Longman.

Mosston, M. et Ashworth, S. (2002). *Teaching physical education*. New York, New York : Benjamin Cummings.

Raîche, G., Riopel, M. et Blais, J.-G. (2006). *Non graphical solutions for the Cattell's scree test*. Communication présentée au congrès annuel de la Psychometric Society, Montréal.

Song, J., Walls, T. et Raîche, G. (2008). *Application of non-graphical solutions for Cattell's scree test to determine the number of factors to be retained in the adolescent smoking consequences questionnaire (ASCQ)*. Communication présentée au congrès annuel de la Psychometric society, Durham.

Tabachnick, B. G. et Fidell, L. S. (2001). *Using multivariate statistics*. Boston, Massachusetts : Allyn and Bacon.

Yakovitz, S. et Szidarovszky, F. (1986). *An introduction to numerical computation*. New York, New York : Macmillan.

Zwick, W. R. et Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological bulletin, 99*(3), 432-442.

Chapitre 2

Validité de la précision de la mesure d'une stratégie de testing adaptatif informatisé (TAI) sous trois conditions de représentativité du domaine

Patrick Charles, Réjean Auger, Jean-Guy Blais et Serge P. Séguin

Cette recherche porte sur la validité de la mesure en éducation. Plus particulièrement, elle étudie la précision de la mesure d'une stratégie de testing adaptatif informatisé (TAI) sous trois conditions de représentativité du domaine. Les données proviennent des résultats d'une stratégie de testing conventionnel de forme papier-crayon de 156 étudiants, parmi lesquels 110 ont été choisis de manière aléatoire aux fins de simulation. La précision des scores a été appréciée à l'aide de l'erreur standard d'estimation (RMSE) et l'erreur systématique (BIAIS). Les résultats indiquent que la précision de la mesure augmente sensiblement et le biais diminue avec l'augmentation du nombre d'items. Cependant, pour assurer une meilleure validité dans l'interprétation des résultats des tests, on s'interroge sur la pertinence d'augmenter invariablement le nombre d'items lorsque la représentativité du domaine du test a été assurée.

Avant l'entrée en vigueur de la réforme de l'éducation au Québec en l'an 2000, le ministère de l'Éducation, du Loisir et du Sport (2004), alors le ministère de l'Éducation du Québec (MEQ), administrait ses épreuves uniques en histoire, en sciences et en mathématique à l'ordre d'enseignement secondaire selon une stratégie de testing où l'on présente les mêmes items à tous les élèves. Cette stratégie, dite de testing conventionnel, se base sur la modélisation

de la théorie classique des tests (TCT). Elle offre la possibilité d'assurer *a priori* la représentativité du domaine et conséquemment, la validité du contenu de l'instrument de mesure. Cependant, en conformité avec les préoccupations de la réforme de l'éducation (Ministère de l'Éducation du Québec, 2004) et du renouveau pédagogique (Ministère de l'Éducation, du Loisir et du Sport, 2005) qui préconisent, d'une part, une approche par compétences, et d'autre part, un apprentissage différencié et individualisé, la mesure de la validité, de façon plus globale, des instruments d'évaluation employés avec des grands groupes d'étudiants devient une des priorités du ministère de l'Éducation, du Loisir et du Sport (2008 ; ministère de l'Éducation du Québec, 2003). À cet égard, les modèles de la théorie des réponses aux items (TRI) fournissent une structure théorique pour le calibrage et la validation des items utilisés dans la construction des tests standardisés. Ainsi, le testing adaptatif informatisé (TAI) découlant de la théorie des réponses aux items pourrait être envisagé pour construire des tests pour chaque individu et s'intègrerait alors dans les démarches de la mesure et de l'évaluation des apprentissages scolaires.

La validité des instruments d'évaluation préoccupe également la plupart des juridictions au Canada responsables de la formation et de la certification des professionnels, par exemple, le Conseil canadien des directeurs de l'apprentissage (2010), la Commission de la construction du Québec (2008) et l'*Industry training authority* en Colombie-Britannique (2008) élaborent des normes pour l'interprétation des résultats de leur épreuve. Aux États-Unis, Kane (2008) explique l'obligation de valider les instruments d'évaluation par des exigences légales et sociales qui imposent que les décisions basées sur les scores des tests soient scientifiquement fondées. Dans cette perspective, plusieurs organismes, dont *The code of fair testing practices in education* (2004), l'*American educational research association,* l'*American psychological association* et *The national council on measurement in education* (1999) élaborent des standards docimologiques en vue de renforcer la validité des outils d'évaluation.

Le changement de paradigme qui accompagne la réforme de l'éducation au Québec (Ministère de l'Éducation, du Loisir et du Sport, 2005), les nouvelles exigences de qualification et de certification des travailleurs spécialisés ainsi que la mobilité de la main-d'œuvre entre les pays de contextes de formation différents amènent les chercheurs à repenser le concept de la validité des résultats de l'évaluation en éducation (Kane, 2008 ; Lissitz et Samuelsen, 2007). Selon Ramseier (2008), la validité des tests doit s'insérer dans les procédures de la mesure, dans les processus et dans l'élaboration des protocoles de recherche. De son côté, Mislevy (2007) explique comment le concept de validité

émerge dans les activités de planification de nouvelles formes de tests et dans le développement des procédures d'évaluation. D'après lui, cette approche qui s'élabore à partir des devis de recherche est une nouvelle théorie pour la planification de l'évaluation, c'est-à-dire qu'il faut unifier ses différentes composantes : un plan de travail, une terminologie, des représentations, des structures de données et des procédures, à travers lesquels nous essayons de mieux comprendre la pratique du testing et développer de nouveaux modèles.

Pour sa part, Messick (1988) définissait la validité d'un instrument de mesure en tant que jugement général fondé sur des évidences empiriques, d'un cadre conceptuel ainsi que sur l'adéquation entre la pertinence des inférences et des actions basées sur les scores des tests. Dans la logique de la définition de Messick (1988), *The National council of measurement in education* (1999) propose de définir la validité d'un test comme étant le degré selon lequel les évidences empiriques et la théorie supportent l'interprétation des scores en regard de l'utilisation que l'on envisage d'en faire. Cependant, cette conception de la validité popularisée par Messick (1988) n'a pas toujours fait l'unanimité dans la communauté scientifique. Plusieurs auteurs, dont Borsboom, Mellenbergh et van Heerden (2004), Lissitz et Samuelsen (2007) la contestent : la définition de la validité qui combine à la fois la théorie et la pratique reste imprécise ; elle ne facilite pas l'étude ni la compréhension de la théorie qui sous-tend le concept. En outre, Lissitz et Samuelsen (2007) soutiennent que la validité est une caractéristique intrinsèque du test ; elle n'est pas une propriété de l'interprétation ni de l'utilisation des scores. Ils recommandent donc la séparation des propriétés internes et externes de l'instrument de mesure pour faciliter l'étude de sa validité.

En revanche, dans une réplique à Lissitz et Samuelsen (2007), Kane (2008) explique pourquoi la validité d'un test doit inclure une évaluation de la proposition de l'interprétation et de l'utilisation des scores. De plus, Kane (1996) avait assimilé l'étude de la validité de l'interprétation des scores à la problématique de la précision de la mesure. D'après lui, une façon adéquate d'évaluer la précision de la mesure consiste à comparer l'erreur de mesure avec la tolérance sur cette erreur. Les deux questions importantes que l'on doit se poser pour apprécier la précision d'un score sont : *quelle* est la marge d'erreur acceptable sur un score ? Et, cette marge d'erreur est-elle assez grande pour constituer un sérieux problème de validité ? Kane concluait alors que la précision de la mesure fait partie intégrante de la validité de l'interprétation des scores.

En tenant compte des propositions de Kane (1996) qui relient la précision de la mesure d'un test à la validité de l'interprétation de ses scores, on étudie dans ce chapitre la validité d'une stratégie d'un testing adaptatif informatisé en comparant la précision des scores de trois tests de longueurs différentes, lorsque la représentativité du domaine est assurée.

1. CADRE THÉORIQUE

Dans cette section sur le contexte théorique, nous proposons de définir d'abord les concepts (le niveau d'habileté q, la théorie des réponses aux items, la représentativité du domaine et le testing adaptatif informatisé). D'autre part, nous ferons une recension des écrits sur le sujet et présenterons les objectifs de la recherche.

Les dictionnaires définissent une habileté comme étant la capacité d'un individu à raisonner sur une situation complexe et à résoudre des problèmes. Cette définition de l'habileté cadre bien avec les préoccupations de la présente recherche. D'autre part, selon Legendre (2005), la représentativité du domaine d'un test indique l'ensemble des items de l'instrument de mesure en regard de ce que l'on veut mesurer, une définition qui se rapproche de celle de la validité de contenu, un type de validité interne. Elle indique aussi la correspondance entre l'échantillon des questions qui composent le test et l'univers des situations possibles d'où est tiré cet échantillon. La représentativité du domaine se manifeste par une distribution équilibrée des items de tous les éléments de compétences du programme d'études que l'on veut mesurer.

Plusieurs problèmes sont reliés à la stratégie du testing conventionnel standardisé de forme papier-crayon. Les résultats de cette stratégie sont analysés le plus souvent selon la théorie classique des tests (TCT). Selon cette modélisation, les indices de difficulté et de discrimination d'un item ainsi que le coefficient de fidélité sont fortement dépendants de l'échantillon d'individus utilisé pour les calculer. Weiss et Betz (1973) énumèrent différents problèmes reliés à la stratégie du testing de forme papier-crayon : les erreurs sont dues à l'administration du test, à l'ordre des items, au temps de passation limité imposé aux candidats et aux divers intervenants. Pour toutes ces raisons, il convient d'étudier d'autres modélisations qui pourraient permettre de surmonter ces inconvénients dans l'évaluation des apprentissages afin d'assurer la validité des scores.

Ainsi, la modélisation de la théorie des réponses aux items constitue un système de testing qui s'intéresse à la relation entre chaque sujet et chaque item, ceux-ci constituant des entités autonomes par rapport aux grands ensembles dont ils sont membres (Blais, 1994). Dans cette modélisation, l'accent est mis sur le rapport individu-item plutôt que sur le rapport population-test. Plusieurs études théoriques et empiriques (Auger et Séguin, 1996 ; Blais, 1994) ont montré que la modélisation de la théorie des réponses aux items permet d'individualiser le testing, notamment par des stratégies dites adaptatives. Le testing adaptatif informatisé découlant de la théorie des réponses aux items semble approprié pour obtenir la même précision de l'estimation du niveau d'habileté q de tous les examinés. D'ailleurs, Wright et Stone (1979) recommandent d'utiliser la théorie des réponses aux items pour estimer les caractéristiques métriques des items et préciser la mesure pour chaque individu, même dans le contexte d'une stratégie de testing conventionnel. Pour sa part, Lord (1980) explique comment la modélisation de la théorie des réponses aux items permet de tenir compte des diverses caractéristiques des items (difficulté b_i, discrimination a_i, pseudo-chance c_i) pour l'estimation du niveau d'habileté des sujets. La probabilité pour un sujet de niveau d'habileté q de réussir l'item i est donnée par l'équation 1 :

$$P_i\left(u=1\right)=c_i+(1-c_i)\left[\frac{1}{1+e^{-Da_i(\theta-b_i)}}\right] \qquad (1)$$

où u prend la valeur 1 lors d'une bonne réponse à l'item et 0 lors d'une mauvaise réponse tandis que D correspond à une constante égale à 1,702 pour permettre d'approximer la probabilité associée à une distribution normale centrée réduite.

Cependant, lorsque l'on cherche à définir l'item i en fonction du seul paramètre de difficulté b_i, pendant que le paramètre de discrimination a_i reste constant et que la pseudo-chance c_i est égale à 0, on obtient alors un modèle logistique à un paramètre, appelé le modèle de Rasch (Bertrand et Blais, 2004). Ce dernier est représenté par l'équation 2 :

$$P_i\left(u=1\right)=\left[\frac{1}{1+e^{-D(\theta-b_i)}}\right] \qquad (2)$$

Lorsque l'on applique le modèle de Rasch, il est courant d'utiliser un test d'ajustement statistique pour examiner dans quelle mesure l'ensemble des données répond aux exigences spécifiques de ce modèle. Gustafsson (1980) a montré que lorsque le nombre d'observations est

suffisamment grand, le test d'ajustement statistique va toujours indiquer une incompatibilité entre n'importe quel ensemble de données et n'importe quel modèle de mesure. Alors, l'important est de savoir de quelle façon le modèle en question permet de construire une représentation utile de la structure des données. Par contre, d'autres chercheurs ont soulevé des interrogations importantes quant à l'utilisation du modèle de Rasch. Par exemple, Wainer (1992) pense que la possibilité de choisir une réponse au hasard doit être considérée comme une facette fondamentale du testing adaptatif informatisé. Tout modèle de la théorie des tests qui n'en tient pas compte pourrait conduire à des résultats douteux.

Cependant, le choix du modèle de Rasch pour l'estimation et l'interprétation des paramètres d'items de cette recherche a été fait pour des raisons théoriques et pratiques. D'abord, d'un point de vue théorique, les diverses stratégies du testing adaptatif minimisent le nombre d'items trop difficiles ou trop faciles pour chaque élève et, par conséquent, limitent dans une certaine mesure l'effet du hasard dans les réponses. Ensuite, on pourrait d'emblée utiliser le modèle de Rasch et assimiler le paramètre de pseudo-chance à l'erreur de mesure. De plus, d'un point de vue pratique, durant la passation d'un test standardisé à choix multiples, l'élève qui ne maîtrise pas la matière peut faire le bon choix de réponse en fonction de ses connaissances parcellaires ou antérieures, ce qui ne relève pas nécessairement du hasard. Tous ces motifs nous amènent donc à choisir le modèle de Rasch.

Plusieurs recherches sur les stratégies du testing adaptatif informatisé étudient l'impact du nombre d'items d'un test sur l'information obtenue à partir des scores des personnes étudiées. Kingsbury et Zara (1991) ont montré que l'utilisation d'une stratégie de testing adaptatif contraignante, c'est-à-dire un test qui impose un certain contenu, augmente de 5 % à 11 % la longueur du test en comparaison du testing adaptatif traditionnel pour un même niveau de précision et de 43 % à 104 % pour les tests adaptatifs qui imposent un nombre minimum d'items pour chaque élément de compétences d'un programme d'études. Le test adaptatif traditionnel, c'est-à-dire axé sur l'aspect psychométrique seulement, peut être de 30 % à 51 % plus court que le test adaptatif contraignant pour obtenir la même qualité d'information.

De plus, Leung, Chang et Hau (2003), dans une étude de simulation utilisant les stratégies de testing adaptatif informatisé, comparent trois méthodes de sélection des items pour équilibrer les contenus des tests : un testing adaptatif informatisé contraignant, un testing adaptatif informatisé contraignant modifié et un modèle polyto-

mique multinomial modifié. Ils ont étudié plusieurs conditions de représentativité du domaine, c'est-à-dire des tests de longueurs différentes ainsi que l'exposition ou la fréquence d'apparition des items dans plusieurs versions des tests. Les résultats indiquent que les trois méthodes produisent des estimés du niveau d'habileté qui sont fortement corrélés avec les véritables valeurs du niveau d'habileté q des sujets. Ces corrélations augmentent avec l'augmentation du nombre d'items du test. Les auteurs n'ont pas observé d'effets systématiques des méthodes d'équilibre de contenu des tests, d'une part, avec la précision de la mesure, et d'autre part, avec la répétition des items dans plusieurs versions des tests. Ils ont observé cependant que le modèle multinomial modifié semble être plus performant que les deux autres et utilise moins souvent certains items de la banque.

Plusieurs études sur le testing adaptatif informatisé, dont celle de Hambleton et Fennessy (1994), ont montré qu'il est possible d'obtenir une précision optimale de la mesure individuelle par l'imposition d'une contrainte quant au contenu spécifique du test. Linn (1990), dans ses recherches sur les applications de la théorie des réponses aux items, a fait remarquer que les considérations quant au contenu spécifique des tests de performance n'ont pas retenu suffisamment l'attention des chercheurs. Il pense que ces considérations méritent une plus grande attention que celle qui leur a été accordée jusqu'à ce jour. Gershon (2005), pour sa part, a montré qu'avec le développement de l'informatique, les images graphiques et les présentations multimédias permettent de mesurer des concepts qui ne sont pas mesurables avec la stratégie du testing conventionnel de forme papier-crayon. Il a observé, en outre, que la stratégie du testing adaptatif informatisé possède l'avantage de faciliter la passation de l'examen pour les candidats par rapport au testing conventionnel, même aux fins de certification.

En résumé, l'histoire montre que les psychométriciens, depuis toujours, se préoccupent de l'estimation des paramètres d'items, de la précision des scores et de l'erreur de mesure. Ce sont les chercheurs et plus particulièrement les praticiens qui, en plus de la précision des scores, s'intéressent à la représentativité du domaine et à la validité de l'interprétation des scores des tests.

Ainsi, l'objectif de la présente recherche étant d'étudier la précision de la mesure d'une stratégie de testing adaptatif informatisé sous trois conditions de représentativité du domaine, elle s'articule d'emblée sous la modélisation de la théorie des réponses aux items et selon une stratégie de testing adaptatif informatisé choisie pour sa capacité à montrer la représentativité du domaine. En adoptant l'approche de la validité développée par Messick (1988), plus spécifiquement sous

l'angle étudié par Kane (1996), cette recherche vise à répondre à la question suivante : dans l'estimation du niveau d'habileté d'un sujet à l'aide d'une stratégie de testing adaptatif informatisé, quel est l'ordre de grandeur de la précision de la mesure et du biais en fonction du nombre d'items, une fois assurée la représentativité du domaine ?

2. MÉTHODE

Dans cette section on examine les éléments méthodologiques de la recherche : les sujets, la stratégie du testing adaptatif informatisé par strate de contenu, les procédures de simulation, les trois conditions de représentativité du domaine et les indices statistiques, l'erreur standard d'estimation (RMSE, *root mean squared error*) et l'erreur systématique (BIAIS).

2.1. Les sujets

Les sujets proviennent des cohortes de l'automne 1998, 1999 et 2000 de la faculté des sciences de l'éducation de l'Université du Québec à Montréal (UQAM). Cette faculté compte 2 000 étudiants à temps plein et à temps partiel dans ses différents programmes de baccalauréat en enseignement. Les données qui sont analysées ont été obtenues à partir des résultats d'un examen sommatif administré, selon une stratégie de testing conventionnel de forme papier-crayon, à tous les étudiants du programme de baccalauréat en enseignement secondaire de cette faculté. Compte tenu de l'effectif de sujets inscrits à ce programme, un échantillon de 156 étudiants a été constitué, parmi lesquels 110 ont été choisis de manière aléatoire pour la simulation.

2.2. L'instrumentation

Les 29 items de la stratégie du testing adaptatif informatisé représentent 100 % de représentativité du domaine du test. Ils proviennent d'une banque initiale de 34 items d'un examen standardisé sommatif à choix multiples de fin de session d'un cours de mesure et évaluation des apprentissages scolaires du programme de baccalauréat en enseignement secondaire. Cinq items du testing conventionnel de forme papier-crayon ont été rejetés parce qu'ils ne répondent pas aux exigences du modèle de Rasch. Le tableau 2.1 présente la distribution des 29 items retenus, le pourcentage de chaque regroupement, les numéros et le nombre d'items par dimension ainsi que leur coefficient de difficulté moyen.

Tableau 2.1
Distribution des 29 items

Contenu notionnel – habiletés	Cadre conceptuel $C_1 = 24\%$	Objets d'évaluation $C_2 = 10\%$	Instrumentation $C_3 = 42\%$	Interprétation $C_4 = 24\%$	
Définir (h_1) = 31 %	8, 13, 14 $b_{11} = -1{,}220$	3 $b_{12} = -2{,}5600$	21, 23 $b_{13} = -2{,}0750$	9, 10,12 $b_{14} = -3{,}8570$	$b_{h1} = -2{,}44$
Distinguer (h_2) = 41 %	4, 5, 6, 7 $b_{21} = -3{,}145$	1, 2 $b_{22} = -1{,}1150$	15, 20, 24, 25 $b_{23} = 0{,}7825$	11, 27 $b_{24} = -1{,}7350$	$b_{h2} = -1{,}26$
Interpréter (h_3) = 10 %			16, 17, 26 $b_{33} = 1{,}150$		$b_{h3} = 1{,}15$
Appliquer (h_4) = 14 %			18, 19, 29 $b_{43} = -1{,}630$	28 $b_{44} = -2{,}950$	$b_{h4} = -1{,}96$
Apprécier (h_5) = 4 %				30 $b_{54} = 3{,}40$	$b_{h5} = 3{,}40$
	$b_{c1} = -2{,}320$	$b_{c2} = -1{,}596$	$b_{c3} = -0{,}205$	$b_{c4} = -2{,}084$	

2.3. Considérations éthiques

En l'absence d'études empiriques précises sur le testing adaptatif informatisé comportant des sujets humains, les études de simulation apparaissent comme une source importante de données dont on peut se servir pour évaluer et comparer les procédures du testing adaptatif informatisé. Cette étude utilise des données secondaires qui n'ont pas été recueillies aux fins de cette recherche. Ces données ont été fournies de manière à ne pas nous permettre d'identifier les sujets.

2.4. Déroulement

La stratification retenue pour le testing adaptatif informatisé est celle des quatre regroupements du contenu (C_1, cadre conceptuel ; C_2, objets d'évaluation ; C_3, instrumentation ; C_4, interprétation), vu le petit nombre de quatre strates possibles. Le logiciel de simulation *POSTSIM* a été retenu pour sa capacité à simuler un testing adaptatif informatisé sur les données réelles. L'estimation du niveau d'habileté des sujets est initialement basée sur le modèle de Bayes. Cette simulation continue jusqu'à ce que le sujet obtienne une bonne et une mauvaise réponse. À ce moment, le programme de simulation *POSTSIM* passe de la méthode bayesienne à la méthode maximale de vraisemblance et continue à estimer le niveau d'habileté des sujets dans le but d'éviter des problèmes potentiels de biais qui seraient dus à l'effet bayesien.

La même stratégie du testing adaptatif informatisé est simulée sur les quatre regroupements séparément. Elle utilise la méthode de vraisemblance maximale pour l'estimation des valeurs du niveau d'habileté pour chaque individu. La simulation d'un testing adaptatif informatisé sur cette banque d'items consiste pour chaque individu à passer un testing adaptatif informatisé sur le premier regroupement de contenu C_1 selon la condition de représentativité retenue ; puis à faire un autre testing adaptatif informatisé pour un deuxième regroupement de contenu C_2 en utilisant l'estimation du niveau d'habileté résultant du premier regroupement comme point d'entrée du deuxième regroupement et ainsi de suite. L'estimation finale du niveau d'habileté de chaque sujet pour le test combine les estimations des simulations des quatre regroupements (C_1, C_2, C_3, C_4) de contenu. Leung, Chang et Hau (2003) utilisent quatre regroupements importants d'un programme de mathématique (les nombres, la mesure, les statistiques et la géométrie) pour étudier l'effet des stratégies d'un testing adaptatif informatisé sur les contenus. Pour leur part, Finney, Smith et Wise (1999) ont montré que le testing adaptatif informatisé par strate basé sur les indices de difficulté empirique augmente l'efficience du test et la précision de l'estimation du niveau d'habileté de l'individu comparé à une approche traditionnelle de testing. En outre, ils ont montré que la performance du TAI par strate augmente à mesure que le nombre de strates diminue. De plus des études réalisées par Bouvette (1997) et Coutu (1996) dans un contexte de testing conventionnel en sciences humaines au Québec avec un échantillon d'élèves de quatrième secondaire, en utilisant les épreuves uniques du ministère de l'Éducation du Québec, ont montré que quelle que soit la théorie de mesure utilisée, la théorie des réponses aux items ou la théorie classique des tests, la difficulté de l'item ne varie pas en fonction des habiletés mesurées. Ce sont davantage la typologie des items et leur contenu qui influencent leur difficulté. D'où le choix des tests standardisés à choix multiples et la stratification par contenu.

Aux fins de cette étude, nous retenons trois niveaux de représentativité pour chaque regroupement. Le tableau 2.2 indique le nombre d'items par regroupement pour chacune des 110 simulations selon les représentations de 50 %, 75 % ou 100 %. Par exemple, le testing adaptatif informatisé à 50 % des items de la banque donne 13 items : trois items proviennent du regroupement C_1, un du C_2, six du C_3 et trois du C_4. Alors qu'un testing adaptatif informatisé à 75 % contient 21 items et à 100 %, il contient les 29 items.

Tableau 2.2
Nombre d'items simulés par strate selon une représentation de 50 %, 75 % et 100 %

C_1	C_2	C_3	C_4
TAI* CONV	TAI CONV	TAI CONV	TAI CONV
50 % 75 % 100 %	50 % 75 % 100 %	50 % 75 % 100 %	50 % 75 % 100 %
3 5 7	1 2 3	6 9 12	3 5 7

* TAI = test adaptatif informatisé ; CONV = test papier-crayon conventionnel.

Les critères d'évaluation utilisés pour l'interprétation et l'utilisation des scores s'en trouvent donc orientés vers des mesures d'estimation qui prennent en compte la variation des estimés du niveau d'habileté selon les conditions imposées (pourcentage d'items à l'intérieur de chaque strate). La variation des estimés du niveau d'habileté doit également être appréciée par rapport aux estimés du niveau d'habileté obtenus en contexte de testing conventionnel où la représentativité du domaine correspond à une situation théorique optimale.

2.5. Les méthodes d'analyse

La performance de la stratégie du testing adaptatif informatisé pour la représentativité du domaine sera appréciée en terme de biais, de l'erreur standard d'estimation (RMSE) et du nombre total d'items dans le test. Dans cette étude, le continuum du niveau d'habileté (–4,0004 à + 3,009) sur lequel les valeurs du niveau d'habileté des sujets sont distribuées est une métrique commune à partir de laquelle on a constitué six groupes, chacun devant contenir un nombre suffisant de sujets. Sur une base empirique, chaque groupe est théoriquement formé de 0,71 écart-type des valeurs du niveau d'habileté observées au testing conventionnel à 100 % d'items. La validité de la précision des scores est évaluée à l'aide de l'erreur standard d'estimation (RMSE) et du biais. À l'instar d'autres études (Chen, Ankenmann et Chang, 2000 ; De Ayala et Sava-Bolesta, 1999 ; Guemin, 1999 ; Leung, Chang et Hau, 2003 ; Wang et Vispoel, 1998) sur le testing adaptatif informatisé, le nombre d'items d'un test ainsi que les indices statistiques, la racine carrée du carré de l'erreur d'estimation et le biais, ont été choisis pour apprécier la précision de la mesure.

L'erreur standard d'estimation se compose d'une partie associée au biais et d'une partie associée à l'erreur type (SE, *standard error*). L'équation suivante décrit la relation entre ses composantes :

$$RMSE^2 = BIAIS^2 + SE^2 \qquad (3)$$

La racine carrée du carré de l'erreur standard d'estimation est définie par l'équation suivante :

$$RMSE\left(n,\theta_0\right)=\sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\hat{\theta}_{n,k}-\theta_0\right)^2} \tag{4}$$

où θ_0 est la valeur théorique du niveau d'habileté qui est comparée à la valeur du niveau d'habileté estimé $\hat{\theta}_{n,k}$ à la k^e simulation tandis que n est le nombre total de simulations.

Le biais indique le signe de l'erreur. Une valeur positive indique une surestimation moyenne du niveau d'habileté. Une valeur négative indique une sous-estimation moyenne. Le biais se définit par l'équation suivante :

$$BIAIS(n,\theta_0)=\frac{1}{n}\sum_{k=1}^{n}(\hat{\theta}_{n,k}-\theta_0) \tag{5}$$

et l'erreur type (SE) par :

$$SE(n,\theta_0)=\sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\hat{\theta}_{n,k}-\bar{\hat{\theta}}_n\right)^2} \tag{6}$$

où, $\bar{\hat{\theta}}_n$ indique la moyenne du niveau d'habileté pour le nombre total de simulations :

$$\bar{\hat{\theta}}_n=\frac{1}{n}\sum_{k=1}^{n}\hat{\theta}_{n,k} \tag{7}$$

L'erreur type (SE), quant à elle, n'apporte aucune information nouvelle pour l'appréciation du testing adaptatif informatisé. D'ailleurs, les études de Wang et Vispoel (1998) ont montré que l'erreur échantillonnale indique une erreur aléatoire sur la mesure plus importante lorsqu'on utilise la méthode de vraisemblance maximale pour l'estimation du niveau d'habileté du sujet que la méthode bayesienne d'estimation. L'erreur échantillonnale obtenue de manière empirique par la méthode de Bayes est généralement très proche et souvent plus petite que le niveau théorique de l'erreur type (SE). C'est d'ailleurs ces deux méthodes d'estimation du niveau d'habileté, d'abord bayesienne et ensuite vraisemblance maximale, qu'utilise le logiciel de simulation *POSTSIM*, d'où le rejet de l'erreur type (SE) comme mesure de précision dans cette recherche.

3. RÉSULTATS

Le testing traditionnel de forme papier-crayon avec 29 items représente la stratégie du testing adaptatif informatisé à 100 % d'items, c'est-à-dire la condition optimale de représentativité du domaine. Les conditions de 50 % et 75 % de représentativité seront comparées entre elles et à la condition optimale en termes de précision et de biais. La figure 2.1 présente les courbes de l'erreur standard d'estimation pour les deux conditions de représentativité, soit 50 % et 75 % du testing adaptatif informatisé en fonction des six sous-groupes du niveau d'habileté des sujets qui ont été constitués pour les quatre regroupements.

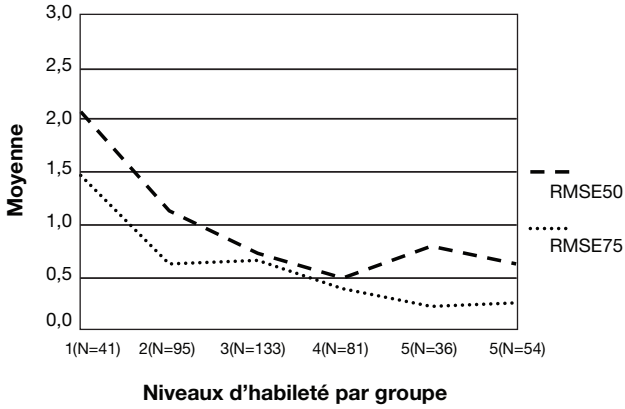


Figure 2.1
Représentation graphique des valeurs du RMSE pour l'ensemble des quatre regroupements

L'erreur standard d'estimation dans la condition à 75 % de représentativité est toujours inférieure à celle dans la condition à 50 % quelle que soit la valeur du niveau d'habileté considérée. En d'autres termes, la précision des estimations augmente avec l'augmentation du nombre d'items. Cette précision dans les estimations du niveau d'habileté est généralement moins grande pour des valeurs inférieures du niveau d'habileté. De plus, l'écart entre les erreurs pour les conditions de représentativité à 50 % et à 75 % chez les sujets qui sont proches du niveau d'habileté moyen est plus faible, c'est-à-dire les sous-groupes 3 et 4 constitués respectivement de 33 et 81 sujets. Ce résultat est important puisque le point de césure autour de la moyenne du continuum du niveau d'habileté sur laquelle on se base pour prendre des décisions quant à l'échec et au succès des candidats présente la plus grande précision.

La figure 2.2 présente les courbes du biais pour les deux conditions de représentativité du domaine des estimations des valeurs du niveau d'habileté d'une stratégie de testing adaptatif informatisé. Les valeurs du biais pour les sujets dont le niveau d'habileté est inférieur à la moyenne se situent entre 0,00 et 1,50, alors qu'elles atteignent un maximum de 0,70 pour les sujets dont le niveau d'habileté est élevé. On observe un biais positif, c'est-à-dire une surestimation de l'estimation des valeurs du niveau d'habileté pour les sujets les plus faibles et une valeur de biais négative, c'est-à-dire une sous-estimation des valeurs du niveau d'habileté, pour les sujets les plus forts. Les surestimations et sous-estimations des valeurs du niveau d'habileté sont plus importantes pour les tests plus courts (50 % du nombre total d'items) que pour les tests plus longs (75 % du nombre total d'items).

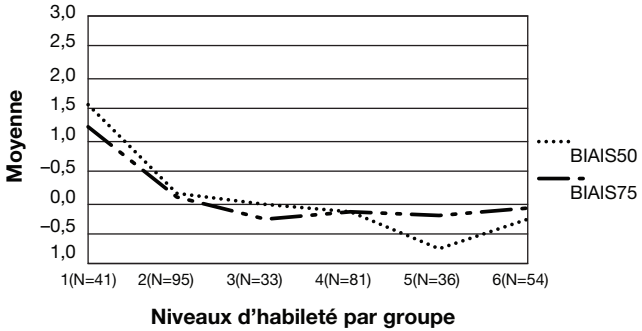


Figure 2.2
Représentation graphique des valeurs du biais pour l'ensemble des quatre regroupements

En résumé, la simulation d'un testing adaptatif informatisé contraignant par strate de contenu sur des données réelles permet d'une part, de représenter le domaine d'un programme d'études avec un plus petit nombre d'items, et d'autre part, d'obtenir une meilleure précision dans l'estimation du niveau d'habileté que ne le permet le testing conventionnel papier-crayon. Cette précision des estimés du niveau d'habileté augmente avec l'augmentation du nombre d'items, d'où une amélioration de la validité de leur interprétation.

4. DISCUSSION

Les résultats de cette recherche confirment ceux de Finney, Smith et Wise (1999), de Leung, Chang et Hau (2003), ainsi que de Kingsbury et Zara (1991) sur le testing adaptatif informatisé : l'augmentation du nombre d'items d'un testing adaptatif informatisé contraignant

par strate de contenu permet d'obtenir une meilleure précision de la mesure individuelle, tout en assurant une représentativité du domaine. D'autre part, la grandeur de l'estimation bayesienne du biais provenant des données simulées paraît conforme aux résultats des études de Tate et King (1994). Le biais subit une augmentation linéaire pour les sujets dont les habiletés sont situées de plus en plus loin de la valeur moyenne de la distribution des thêtas. L'effet bayesien sur le biais est dû à la formule du biais qu'utilise le logiciel de simulation *POSTSIM*. Pour une estimation de thêta donnée, la valeur du biais augmente avec la diminution du nombre d'items d'un test.

Les limites de cette étude par simulation sur les données réelles sont imposées par l'échantillon de sujets peu représentatif, 110 sujets choisis aléatoirement parmi un échantillon de 156 qui ont passé l'examen, ainsi que par une banque d'items réduite, 29 items sur les 34 disponibles du testing conventionnel. D'autres auteurs, dans de meilleures conditions, dont Leung, Chang et Hau (2003), ont utilisé une banque de 700 items calibrés. Par ailleurs, il est important de souligner que la procédure de sélection des items par le logiciel *POSTSIM* contribue sensiblement à l'augmentation de l'erreur de mesure de l'estimé final du niveau d'habileté à cause, d'une part, des deux procédures d'estimation utilisée, et d'autre part, de la simulation de quatre testings adaptatifs informatisés, un par regroupement de contenu, au lieu d'un seul testing.

CONCLUSION

Notre objectif était d'étudier l'effet de trois conditions de représentativité du domaine d'une stratégie de testing adaptatif informatisé sur la validité de l'interprétation de la précision de la mesure en éducation. L'analyse de l'erreur standard d'estimation et du biais pour les conditions de représentativité du domaine à 50 % et 75 % du nombre d'items des quatre regroupements permet de dégager un certain nombre de conclusions. Il apparaît qu'un minimum de 13 items soit suffisant pour bien estimer les valeurs du niveau d'habileté des individus lors d'un testing adaptatif informatisé stratifié par contenu. On a constaté qu'en augmentant le nombre d'items de 13 à 21, on obtient un gain minimal en précision et que pour certains sous-groupes de sujets, ce gain est même négligeable, surtout pour ceux qui sont situés autour des valeurs moyennes du continuum du niveau d'habileté, où l'on doit décider de la réussite ou de l'échec. Cependant, en augmentant le nombre d'items, on obtient une meilleure représentativité du domaine et par conséquent une plus grande validité dans l'interprétation des valeurs du niveau d'habileté. Nous avons aussi constaté qu'il existe

une sous-estimation du niveau d'habileté variant de 0,3 à 0,5 pour les individus qui ont un niveau d'habileté supérieur à la moyenne et une surestimation pour ceux qui ont un niveau d'habileté inférieur à la moyenne. Il nous semble que l'adéquation entre l'estimation du niveau d'habileté et les conditions optimales quant au niveau de difficulté des items selon la modélisation de la théorie de réponses des items et le seuil de passage au testing adaptatif informatisé ainsi qu'au testing conventionnel soit aussi une piste à explorer.

RÉFÉRENCES

American educational research association, American psychological association and The national council of measurement in education (1999). *Standards for educational and psychological testing*. Washington, Columbia : American educational research association.

Auger, R. et Séguin, S. P. (1996). Validité globale d'une stratégie de testing adaptatif de maîtrise pour fins de certification scolaire au Québec. *Revue canadienne de l'éducation, 21*(2), 143-154.

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

Blais, J.-G. (1994). Compte rendu de Fundamentals of item response theory, de Ronald K. Hambleton, H. Swaminathan et H. Jane Rogers (Newbury Park, Californie : Sage publications Inc., 1991). *Mesure et évaluation en éducation, ADMEE, 17*(1).

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

Bouvette, M. (1997). *La typologie et les paramètres d'item dans le contexte de la TRI*. Mémoire de maîtrise inédit. Université du Québec à Montréal, Montréal.

Chen, S.-Y., Ankenmann, R. D. et Chang, H.-H. (2000). A comparaison of item selection rules at the early stages of computerized adaptive testing. *Applied psychological measurement*, *24*(3), 241-255.

Code of fair testing practices in education (2004). Washington, Columbia : Joint committee on testing practices.

Commission de la construction du Québec (CCQ) (2008). *L'analyse des résultats de l'examen de qualification provinciale du métier de carreleur*. Montréal, Québec : Direction de la formation professionnelle.

Conseil canadien des directeurs de l'apprentissage (2010). Renforcer le programme du Sceau rouge. Ottawa, Ontario : Ressources humaines et développement social du Canada.

Coutu, G. (1996). *La Typologie des items comme facteurs pouvant en influencer les paramètres : une étude portant sur les épreuves uniques du MEQ en histoire 414 entre 1988 et 1993*. Mémoire de maîtrise inédit. Université du Québec à Montréal, Montréal.

De Ayala R. J. et Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied psychological measurement, 23*(1), 3-19.

Finney, S. J., Smith, R. W. et Wise, S. L. (1999). *The effects of judgment-based stratum classifications on the efficiency of stratum scored CATs*. Communication présentée en 1999 à l'Annual meeting of the national council on measurement in education, Montréal, Québec.

Gershon, R. C. (2005). Computerized adaptive testing. *Journal of applied measurement, 6*(1),109-127.

Guemin L. (1999). *Conditional standard errors of measurement for tests composed of testlets*. Communication présentée en 1999 à l'Annual meeting of the national council on measurement in education, Montréal, Québec.

Gustafsson, J.-E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of mathematical and statistical psychology, 33*, 205-233.

Hambleton, R. K. et Fennessy, L. (1994). Progrès techniques dans le développement d'examens d'accréditation. Traduit de l'anglais par Réjean Auger. *Mesure et évaluation en éducation, 17*(2), 83-105.

ITA (2008). *ITA Multiple assessment pathways (MAP) project : a project to develop alternative methodologies for assessing challengers seeking trade qualification*. Colombie-Britannique, Canada : Government of British Colombia.

Kane, M. T. (1996). The precision of measurement. *Applied measurement in education*, *9*(4), 355-379.

Kane, M. (2008). Terminology, emphasis, and utility in validation. *Educational researcher, 37*(2), 76-82.

Kingsbury, G. G. et Zara, A. R. (1991). A comparaison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied measurement in education*, *4*(3), 241-261

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

Leung, C.-K., Chang, H.-H. et Hau, K.-T. (2003). Computerized adaptive testing : a comparison of three content balancing methods. *The Journal of technology, learning and assessment, 2*(5), 3-15.

Linn, R. L. (1990). Has item response theory increased the validity of achievement test scores ? *Applied measurement in education, 3*(2), 115-141.

Lissitz, R. W. et Samuelsen. K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational researcher, 36*(8), 437-448.

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

Messick, S. (1988). The once and future issue of validity : assessing the meaning and consequences of measurement. Dans H. Wainer et H.I. Braum (dir.), *Test validity*. Hillsdale, New Jersey : Lawrence Erlbaum.

Ministère de l'Éducation du Québec (2003). *Grille d'évaluation de la compétence à écrire un texte argumentatif : épreuve unique d'écriture, français langue d'enseignement, 5e année du secondaire (129-510). 2002-2003*. Direction générale de la formation des jeunes. Direction de l'évaluation. Québec, Québec : Gouvernement du Québec.

Ministère de l'Éducation du Québec (2004). *Programme de formation de l'école québécoise pour l'enseignement secondaire*. Québec, Québec : Gouvernement du Québec.

Ministère de l'Éducation, du Loisir et du Sport du Québec (2008). *Mieux soutenir le développement de la compétence à écrire : rapport du Comité d'experts sur l'apprentissage de l'écriture. Janvier 2008*. Québec, Québec : Gouvernement du Québec.

Ministère de l'Éducation, du Loisir et du Sport (2005). *Le renouveau pédagogique : ce qui définit le changement préscolaire-primaire-secondaire*. Québec, Québec : Gouvernement du Québec.

Mislevy, R. J. (2007). Validity by design. Educational researcher. *American educational research association, 36*(8), 463-469.

Ramseier, E. (2008). Validation of competence models for developing education standards : Methodological choices and their consequences. *Mesure et évaluation en éducation*, *31*(1).

Tate, R. L. et King. F. J. (1994). Factors which influence precision of school-level IRT ability estimate. *Journal of educational measurement, 31*(1), 1-15.

Wainer, H. (1992). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational measurement : issues and practice*, *12*(1),15-20.

Wang, T. et Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of educational measurement*, *35*(2), 109-135.

Weiss, D. J. et Betz, N. E. (1973). *Ability measurement : conventional or adaptive ?* Research report 73-1. Minneapolis, Minnesota : Psychometric methods Program, Department of Psychology, University of Minnesota.

Wright, B. D. et Stone, M. H. (1979). *Best test design : Rasch measurement*. Chicago, Illinois : MESA.

Chapitre 3

Comparaison empirique des méthodes classiques de détection du fonctionnement différentiel d'items en psychométrie[1]

David Magis, Paul De Boeck et Gilles Raîche

Un défi majeur de la psychométrie moderne est la détection, dans une procédure de test, des items dont le degré de difficulté varie selon les sous-groupes de sujets. Afin de détecter ces items fonctionnant différemment, plusieurs méthodes ont été proposées. Cependant aucune étude comparative de l'efficacité de ces techniques n'a été réalisée à ce jour. Nous proposons une étude par simulations pour combler cette lacune. Il ressort de l'étude qu'aucune de ces méthodes n'est plus efficace que les autres, bien que des différences méthodologiques significatives soient observées.

Un objectif de la théorie de la réponse à l'item est de pouvoir tirer des conclusions valides sur la caractéristique que l'on souhaite étudier à partir d'un test formé d'un ensemble d'items. Il arrive cependant que des sujets de deux groupes différents et ayant le même niveau d'habileté ont néanmoins une probabilité différente de répondre correctement à un item donné. On parle alors d'un item fonctionnant différemment. Cette problématique est identifiée sous l'appellation fonctionnement différentiel d'items. Un tel item ne mesure donc pas uniquement la caractéristique étudiée,

1. Les recherches relatives à cette étude ont été menées grâce au support financier du Fonds de recherche GOA/2005/04 de l'Université catholique de Louvain (Belgique) et du Conseil national de recherches en sciences humaines du Canada (CRSH).

mais également un ou plusieurs paramètres de nuisance additionnels (Ackerman, 1992). La détection des items fonctionnant différemment dans un test est un objectif important en vue d'inférer des résultats valides et fiables pour les sujets considérés. Généralement, les items clairement identifiés comme fonctionnant différemment sont retirés du test ou modifiés pour corriger ce phénomène indésirable.

Le cadre général considéré tout au long de ce chapitre repose sur trois conventions fondamentales. Premièrement, les réponses aux items sont de type dichotomique, de sorte que les modèles usuels de réponse à l'item (comme le modèle de Rasch) sont justifiés. Il existe une littérature abondante sur la détection du fonctionnement différentiel d'items dans le cadre de réponses polytomiques (entre autres, Chang, Mazzeo et Roussos, 1996 ; Potenza et Dorans, 1995) ; cette généralisation se situe cependant en dehors du champ de notre étude. Deuxièmement, l'échantillon de sujets étudiés se subdivise en deux sous-groupes, le groupe de référence et le groupe focal, et le fonctionnement différentiel d'un item particulier n'est caractérisé que par une différence de fonctionnement entre ces deux sous-groupes uniquement. Là aussi, il existe des généralisations de l'application des méthodes classiques lorsque plus de deux groupes de sujets sont en présence (Kim, Cohen et Park, 1995 ; Penfield, 2001), mais nous n'étudierons pas non plus cette généralisation dans ce chapitre. Enfin, la troisième convention porte sur la distinction entre le fonctionnement différentiel uniforme et non uniforme. On parle de fonctionnement différentiel uniforme lorsque la différence entre les probabilités de succès pour des sujets de même habileté ne varie pas selon le niveau d'habileté ; l'avantage conféré par l'item à un des deux groupes est constant quel que soit le niveau d'habileté considéré. Si par contre l'écart entre les probabilités de succès varie selon le niveau d'habileté, alors le fonctionnement différentiel de l'item étudié est non uniforme. La plupart des méthodes développées traitent du fonctionnement différentiel uniforme, bien que certaines techniques aient été développées spécifiquement pour la détection du fonctionnement différentiel non uniforme (Finch et French, 2007 ; Mazor, Clauser et Hambleton, 1994 ; Narayanan et Swaminathan, 1996 ; Penfield, 2003). Dans ce chapitre cependant, seul l'aspect uniforme du fonctionnement différentiel sera pris en compte.

Dans ce contexte bien précis, de nombreuses méthodes ont été suggérées pour permettre l'identification des items fonctionnant différemment et leur retrait du test considéré ; Clauser et Mazor (1998) ainsi que Millsap et Everson (1993) proposent d'excellents aperçus et analyses de la plupart de ces concepts. L'objectif principal de ce chapitre est de confronter six méthodes classiques de détection du fonctionnement différentiel d'items au moyen de simulations de données. Ces

six méthodes sont la méthode de Mantel-Haenszel, la standardisation, la méthode SIBTEST, le test du rapport de vraisemblance, la méthode de Raju et la régression logistique. Si les trois premières méthodes sont basées sur l'analyse de tables de contingence, le test du rapport de vraisemblance et la méthode de Raju utilisent quant à elles les modèles de réponse à l'item tandis que la régression logistique peut être vue comme un lien entre les deux grands types de méthodes (Camilli et Shepard, 1994). Le choix particulier de ces techniques est dû au fait que ce sont les méthodes classiques les plus simples et les plus couramment utilisées en pratique. En outre, elles ont été étudiées intensivement et servent très souvent de méthodes de référence pour étudier l'efficacité de nouvelles techniques. Cependant, une étude globale comparant ces six méthodes entre elles n'a encore jamais été réalisée.

Le reste de ce chapitre comprend 4 sections. Dans la section 1, nous décrivons brièvement chacune des six méthodes et nous énumérons les principales études comparatives qui ont été réalisées dans le passé. Ensuite, dans la section 2, nous présentons notre démarche de comparaison des techniques par l'entremise de simulations de données ; les divers paramètres de production de données ainsi que les outils de comparaison de l'efficacité des méthodes y sont présentés. La section 3 est consacrée à la présentation et l'analyse des résultats des simulations. Les conclusions que l'on en tire sont analysées plus en détail dans la section 4. Finalement, dans la conclusion, on signale plusieurs méthodes récentes de détection du fonctionnement différentiel d'items.

1. CONTEXTE THÉORIQUE

Nous commençons par présenter succinctement les six méthodes de détection du fonctionnement différentiel d'item, citées auparavant. Une revue de la littérature reprenant les plus importantes études est également proposée. Les objectifs de la recherche sont présentés à la fin de la section.

1.1. Méthodes de détection du fonctionnement différentiel

L'ordre de présentation des six méthodes est à peu de chose près l'ordre chronologique des dates de la première publication de celles-ci. Dans chaque cas, nous recherchons les items fonctionnant différemment en les testant un à un. Les méthodes sont donc utilisées de façon itérative. Afin de ne pas alourdir le texte, nous ne présentons que les caractéristiques principales de chacune des méthodes.

1.1.1. *Méthode de Mantel-Haenszel*

Cette méthode a été proposée par Holland et Thayer (1988) et consiste à tester l'absence de relation entre le type de réponse à l'item (correcte ou incorrecte) et l'appartenance au groupe de sujets (focal ou de référence), en fonction du score total du test. Soit K le nombre total d'items du test, pour chaque valeur du score total au test (entre 0 et K), l'information est regroupée dans une table de contingence à double entrée, la première entrée étant l'appartenance des sujets (au groupe focal ou au groupe de référence) et la seconde étant leur réponse à l'item concerné (correcte ou incorrecte). Le tableau 3.1 décrit la situation.

Tableau 3.1
Table de contingence des réponses à l'item étudié pour le j-ème score total

Groupe	Réponse		
	Correcte	Incorrecte	Total
Référence	A_j	B_j	n_{Rj}
Focal	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

Dans ce tableau, T_j désigne le nombre total de sujets ayant le j-ème score total au test considéré ; n_{Fj} et n_{Rj} désignent les nombres respectifs de sujets du groupe focal et du groupe de référence ; m_{1j} et m_{0j} sont les nombres totaux de réponses respectivement correctes et incorrectes à l'item étudié et A_j, B_j, C_j et D_j sont les comptages respectifs de sujets selon leur groupe (focal ou de référence) et leur réponse à l'item. La statistique de Mantel-Haenszel (1959) est alors calculée ; elle s'écrit avec les notations du tableau 3.1 :

$$MH = \frac{\left(\left| \sum_j A_j - \sum_j E(A_j) \right| - 0.5 \right)^2}{\sum_j V(A_j)} \qquad (1)$$

où les sommes portent sur tous les scores totaux j observés et où $E(A_j)$ et $V(A_j)$ sont donnés par :

$$E(A_j) = \frac{n_{Rj}\, m_{1j}}{T_j} \text{ et } V(A_j) = \frac{n_{Rj}\, n_{Fj}\, m_{1j}\, m_{0j}}{T_j^2\, (T_j - 1)} \qquad (2)$$

La statistique (1) permet de vérifier si l'association entre l'appartenance au groupe de sujets et la réponse à l'item, conditionnellement au score total, est significative. Un test de signification statistique est alors établi, sur la base d'une loi du χ^2 à un degré de liberté. Une valeur trop importante de la statistique mène à la classification de l'item comme fonctionnant différemment (Holland et Thayer, 1988).

1.1.2. *Standardisation*

L'approche par standardisation, proposée par Dorans et Kullick (1986), présente de grandes similitudes avec le test de Mantel-Haenszel (Dorans, 1989). Le principe de la standardisation consiste à comparer les proportions de réponses correctes à l'item dans chaque groupe de sujets, pour toutes les valeurs du score total obtenu. Nous sommes ainsi amenés à calculer la *p*-différence standardisée *P-DIF* :

$$P - DIF = \frac{\sum_j \omega_j \, (P_{Fj} - P_{Rj})}{\sum_j \omega_j} \qquad (3)$$

où $P_{Fj} = C_j / n_{Fj}$ et $P_{Rj} = C_j / n_{Rj}$ sont les pourcentages de réussite respectivement dans les groupes focal et de référence. La *p*-différence standardisée est en fait une moyenne pondérée de ces différences de pourcentages de réussite à l'item. Les poids ω_j sont habituellement définis comme étant les pourcentages de sujets du groupe focal ayant un score total *j* (entre 0 et *K*), bien que d'autres systèmes de poids aient été suggérés (Dorans et Kullick, 1986).

Plus la statistique *P-DIF* s'éloigne de zéro et plus l'item considéré indique un fonctionnement différentiel, car dans ce cas, les pourcentages de réussite dans les groupes de sujets deviennent très différents. Il n'existe aucun critère théorique permettant de déterminer un seuil précis de décision : en général, un item est considéré comme fonctionnant différemment lorsque $|P\text{-}DIF| > 0{,}05$ ou 0,10. Une autre règle de décision, basée sur deux seuils distincts, est parfois utilisée (Dorans et Kullick, 1986), bien que globalement moins précise.

1.1.3. *Test du rapport de vraisemblance*

L'idée de faire appel aux modèles de réponse à l'item et au test du rapport de vraisemblance pour la détection du fonctionnement différentiel d'items est due à Thissen, Steinberg et Wainer (1988). Le principe consiste à estimer et comparer deux modèles de réponse à l'item (dont la structure est fixée à l'avance, par exemple un modèle de

Rasch). Le premier modèle, appelé modèle compact, impose des paramètres d'items identiques dans chaque groupe de sujets ; tandis que le second modèle, appelé modèle étendu, inclut une interaction supplémentaire entre l'item étudié et l'appartenance aux groupes de sujets. Ensuite, les deux modèles sont comparés à l'aide du test de rapport de vraisemblance classique (le modèle compact étant un cas particulier du modèle étendu). Cela revient donc à vérifier la signification statistique du terme d'interaction item-groupe (et donc à vérifier l'existence d'un fonctionnement différentiel de l'item).

La statistique *RV* du rapport de vraisemblance est, à une transformation près, le rapport des vraisemblances maximisées des deux modèles :

$$RV = -2 \log \frac{\text{vraisemblance maximale du modèle compact}}{\text{vraisemblance maximale du modèle étendu}} \qquad (4)$$

Cette statistique est confrontée au quantile d'une loi χ^2 à un degré de liberté pour vérifier l'hypothèse que l'item ne fonctionne pas différemment. De trop grandes valeurs de *RV* mènent au rejet de cette hypothèse.

1.1.4. *Régression logistique*

Swaminathan et Rogers (1990) ont introduit l'utilisation de la régression logistique pour la détection du fonctionnement différentiel d'items. Le principe de cette approche consiste à estimer les paramètres du modèle logistique suivant :

$$\text{logit } (\pi_i) = \beta_0 + \beta_1 S_i + \beta_2 G_i + \beta_3 (SG)_i$$

où π_i désigne la probabilité que le sujet i réponde correctement à l'item étudié, S_i le score total du test, et G_i le groupe (de référence ou focal) auquel le sujet i appartient. Le terme $(SG)_i$ désigne une interaction entre le groupe de sujets et le score total. Les paramètres $\{\beta_0, \beta_1, \beta_2, \beta_3\}$ sont estimés par maximum de vraisemblance ; l'item ne fonctionne pas différemment si et seulement si les paramètres β_2 et β_3 sont tous deux égaux à 0 (autrement dit, si la probabilité de répondre correctement à l'item ne dépend pas du groupe de sujets). Il existe plusieurs tests statistiques (test de Wald, test du rapport de vraisemblance, etc.) permettant de vérifier si l'hypothèse $H_0 : \beta_2 = \beta_3 = 0$ est acceptable.

Notons que la présence du terme G_i dans le modèle indique un fonctionnement différentiel uniforme, tandis que l'ajout de l'interaction $(SG)_i$ permet de modéliser un fonctionnement différentiel non uniforme, ce qui est un des avantages de cette méthode.

1.1.5. *Méthode de Raju*

La méthode suivante est due à Raju (1988, 1990) et ses recherches théoriques sur l'aire séparant deux courbes caractéristiques d'un item. Le principe de cette technique consiste à sélectionner un modèle de réponse à l'item et à estimer les paramètres d'items séparément dans chaque groupe de sujets. Ensuite, après une transformation des paramètres pour les ramener à une même échelle de mesure (Cook et Eignor, 1991), l'aire entre les deux courbes caractéristiques de l'item (une courbe par groupe de sujets) est calculée. Une valeur trop élevée indique un fonctionnement différentiel de l'item.

Dans le cas du modèle de Rasch, les paramètres d'items se réduisent au seul paramètre de difficulté et l'aire entre les deux courbes caractéristiques est alors égale à la différence entre les deux difficultés des items dans chaque groupe de sujets (Raju, 1988). Il est alors possible d'obtenir la statistique Z_j pour l'item j, basée sur cette différence de paramètres :

$$Z_j = \frac{\hat{\beta}_{jF} - \hat{\beta}_{jR}}{\sqrt{s_{jR}^{\ 2} + s_{jF}^{\ 2}}} \tag{5}$$

Dans la formule (5), $\hat{\beta}_{jF}$ et $\hat{\beta}_{jR}$ représentent les difficultés (estimées) du j-ième item dans le groupe focal et dans le groupe de référence, respectivement, tandis que s_{jF} et s_{jR} sont les erreurs types (estimées) de ces paramètres de difficulté. La statistique Z_j suit une loi normale standard sous l'hypothèse d'un item ne fonctionnant pas différemment (Raju, 1990). Ainsi, un item sera classifié comme fonctionnant différemment si la valeur de $|Z|$ est supérieure au quantile d'une loi normale standard (typiquement 1,96).

Notons que la formule (5) de la statistique Z peut être aisément adaptée à l'utilisation d'autres modèles comme le modèle logistique à deux ou trois paramètres, bien que cela sorte du cadre de ce chapitre. De plus, dans les limites de l'utilisation du modèle de Rasch, la méthode de Raju est équivalente à l'approche de Lord (1980) qui consiste à tester l'hypothèse d'égalité des paramètres d'items des deux groupes de sujets (test χ^2 de Lord).

1.1.6. *Méthode SIBTEST*

La dernière méthode que l'on présente brièvement est désignée par l'acronyme SIBTEST pour *Simultaneous Item Bias TEST* (Shealy et Stout, 1993). Elle peut être vue comme une généralisation de l'approche par standardisation avec deux améliorations significatives : il est possible

de tester la présence du fonctionnement différentiel dans un groupe d'items et plus seulement dans un seul item. Cette méthode propose une statistique ayant une distribution normale standard sous l'hypothèse que le groupe d'items ne fonctionne pas différemment. Nous présentons la méthode SIBTEST sous ce nouvel angle. Cependant, afin de rester cohérentes avec les méthodes précédentes, les analyses ultérieures seront effectuées en testant chaque item un à un. Une différence importante avec les méthodes de Mantel-Haenszel et de standardisation est également à signaler : le score total est calculé sur la base de tous les items, excepté les items testés.

La statistique SIBTEST est de la forme

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)} \tag{6}$$

Dans l'équation (6), le numérateur $\hat{\beta}_U$ est donné par

$$\hat{\beta}_U = \sum_j F_j \, (P_{Rj} - P_{Fj}) \tag{7}$$

(avec les mêmes notations que précédemment, et avec F_j étant la proportion de sujets ayant le j-ième score total provenant du groupe focal) et le dénominateur $\hat{\sigma}(\hat{\beta}_U)$ est l'écart type estimé de $\hat{\beta}_U$ (pour ce dernier, la formule complète se trouve dans Shealy et Stout (1993). La formule (7) est similaire à la statistique *P-DIF* donnée par la formule (3). Cependant la sommation dans $\hat{\beta}_U$ ne s'effectue que pour les scores totaux obtenus sur la base des items non testés, comme on l'a signalé auparavant.

La statistique B suit une loi normale standard sous l'hypothèse que les items vérifiés ne fonctionnent pas différemment (Shealy et Stout, 1993), ce qui permet d'obtenir une règle de décision immédiate (similaire à l'approche de Raju).

1.2. Études comparatives

Différentes études de ces méthodes (ainsi que d'autres, non citées dans ce texte) ont été réalisées. Les plus anciennes sont celles de Ironson et Subkoviak (1979), Rudner, Getson et Knight (1980) ainsi que celle de Shepard, Camilli et Averill (1981). Après l'introduction des techniques devenues classiques, certains auteurs ont étudié leur efficacité sans les comparer aux autres méthodes (Ankenmann, Witt et Dunbar, 1999 ; Cohen, Kim et Wollack, 1996 ; Jiang et Stout, 1998 ; Jodoin et Gierl, 2001 ; Raju, Bode et Larsen, 1989 ; Uttaro et Millsap, 1994). Enfin, l'effi-

cacité de certaines de ces méthodes a été comparée (Donoghue, Holland et Thayer, 1993 ; Kim et Cohen, 1995 ; Narayanan et Swaminathan, 1994 ; Rogers et Swaminathan, 1993). Cependant, aucune étude globale de l'efficacité de ces six méthodes, en les comparant toutes en même temps à l'aide de simulations ou de données réelles, n'a été effectuée à ce jour.

La plupart des études antérieures ont pointé le fait que les différentes méthodes confrontées se comportent de façon similaire dans la détection des items fonctionnant différemment. Leurs avantages respectifs relèvent plus de leurs caractéristiques propres que d'une éventuelle supériorité technique sur les autres méthodes. Certaines méthodes sont aussi plus intéressantes de par leur généralisation possible au contexte du fonctionnement différentiel polytomique ou du fonctionnement différentiel non uniforme, comme on l'a vu précédemment.

En conclusion, le choix d'une méthode semble surtout reposer sur des critères techniques (disponibilité de logiciels pour effectuer les analyses) ou sur des préférences personnelles.

1.3. Objectifs de recherche

Compte tenu de ce qui précède, nous nous proposons de réaliser une étude comparative simultanée des six méthodes par la simulation de données. L'originalité de cette analyse est double. Tout d'abord, l'étude simultanée des six méthodes permettra de les comparer sur une même base de travail. Ensuite, le schéma de simulations que nous allons utiliser n'est pas classique dans ce type d'étude empirique, bien qu'il ait été déjà appliqué dans certains travaux et qu'il repose sur une modélisation de la réponse à l'item parfaitement cohérente. Nous espérons ainsi que les résultats empiriques de nos simulations conforteront les connaissances généralement acceptées sur les différentes méthodes de détection du fonctionnement différentiel d'items, tout en provenant d'une étude globale avec un même schéma d'expérience.

2. MÉTHODOLOGIE DE RECHERCHE

Nous décrivons à présent le plan de simulations que nous avons choisi. Nous définissons tout d'abord les différents paramètres choisis pour l'exploitation des données. Ensuite, nous expliquons les outils statistiques que nous utiliserons et qui nous permettront de comparer les méthodes et vérifier leur efficacité.

2.1. Cadre général

Le schéma suivant a été appliqué pour la génération des données. Tout d'abord, chaque base de données contient 1 000 patrons de réponses, la moitié provenant de chacun des groupes de sujets. Ensuite, les tests générés sont de deux longueurs différentes : il s'agit soit de tests relativement courts (avec 20 items), soit de tests de taille moyenne (avec 40 items). La taille du test sera notée *C*. De plus, deux cas sont considérés : soit le test ne comporte aucun item fonctionnant différemment, soit il en contient exactement quatre. Ce dernier nombre a été choisi, car il mène à un pourcentage acceptable (10 % ou 20 %) d'items fonctionnant différemment à l'intérieur du test (Narayanan et Swaminathan, 1994). Lorsqu'il y a présence d'items fonctionnant différemment deux cas sont encore à distinguer : l'effet de ces items est soit symétrique (certains items étant plus difficiles et d'autres, plus faciles) ou asymétrique (l'effet du fonctionnement différentiel va dans le même sens pour tous les items).

Notons enfin que, concernant les groupes de sujets, deux cas de figure sont également à envisager : soit le niveau d'habileté moyen est identique dans les deux groupes, soit le groupe focal a un niveau d'habileté moyen plus élevé. Cela permet ainsi d'étudier l'efficacité des méthodes compte tenu de cet effet appelé l'impact des items (Clauser et Mazor, 1998).

2.2. Paramètres de sujet et d'items et patrons de réponses

Examinons à présent les détails techniques et numériques. Une version modifiée du modèle de Rasch est appliquée pour l'exploitation des données :

$$\text{logit}\,[\Pr(Y_{ijg} \mid \theta_{ig}, \beta_{jg})] = \theta_{ig} - \beta_{jg}$$

où Y_{ijg} est la réponse à l'item j du sujet i provenant du groupe g ; θ_{ig} est l'habileté du sujet i dans le groupe g ; et β_{jg} est la difficulté de l'item j dans le groupe g.

Les habiletés θ_{ig} sont définies comme suit :

$$\theta_{ig} \approx \begin{cases} N(0,1) & \text{(groupe de référence)} \\ N(\gamma,1) & \text{(groupe focal)} \end{cases}$$

Le paramètre γ prend soit la valeur 0 (et il n'y a donc pas de différence de niveau moyen d'habileté), soit la valeur 1 (Kim et Cohen, 1995).

Pour établir les paramètres d'items, nous procédons comme suit. Tout d'abord, les paramètres β_{jR} du groupe de référence sont définis selon une loi N(0,1). Ensuite, les paramètres β_{jF} du groupe focal sont obtenus par la relation $\beta_{jF} = \beta_{jR} + \delta_j$, où l'incrément δ_j caractérise l'effet du fonctionnement différentiel d'items. Bien entendu, pour les items ne fonctionnant pas différemment, cet incrément est égal à 0, de sorte que les difficultés sont identiques dans les deux groupes de sujets. Par contre, en présence des quatre items fonctionnant différemment, les incréments δ_j prennent les valeurs suivantes : $(\delta_1, \delta_2, \delta_3, \delta_4,) = (0{,}5\,;\, 0{,}5\,;\, 1\,;\, 1)$ lorsque l'effet est asymétrique (tous les items sont plus difficiles pour le groupe focal), ou $(\delta_1; \delta_2; \delta_3; \delta_4) = (0{,}5; -0{,}5; 1; -1)$ lorsque l'effet est symétrique. L'étendue des valeurs de l'effet du fonctionnement différentiel est conforme aux choix habituels pour des simulations à ce sujet (par exemple, Narayanan et Swaminathan, 1994). Notons aussi que nous supposons que les items fonctionnant différemment sont les quatre premiers du test. Cela n'est toutefois qu'un artifice de construction dans les simulations et ne soulève pas d'autres problèmes.

Finalement, une fois les paramètres de sujet et d'items établis aléatoirement, la réponse Y_{ijg} de chaque sujet à chaque item est fournie selon une loi de Bernoulli dont la probabilité de succès est donnée par le modèle de Rasch modifié.

2.3. Méthodes d'analyse des résultats

Le modèle de simulations que nous proposons ici regroupe donc 12 situations différentes, selon le choix de la longueur de test C (20 ou 40 items), le nombre d'items fonctionnant différemment (0 ou 4 avec effet asymétrique, 4 avec effet symétrique) et le niveau d'habileté moyen γ du groupe focal (0 ou 1). Dans chacune des situations, 50 ensembles de données ont été générés aléatoirement.

Dans chaque ensemble de données, nous classons les items selon qu'ils fonctionnent différemment ou non, à l'aide des six méthodes présentées à la section 1. Pour ce faire, nous avons utilisé une technique personnelle dans le logiciel R (le code est disponible sur demande). De plus, nous avons pris la règle $|P\text{-}DIF| > 0{,}05$ comme critère de classification des items par standardisation et nous avons fixé le risque de première espèce α à 5 % pour les autres méthodes. Enfin, l'efficacité de chaque approche est déterminée en calculant trois quantités empiriques : l'erreur de type I, la puissance de la méthode et le taux d'erreurs total.

L'erreur de type I est la probabilité de reconnaître un item comme fonctionnant différemment, alors qu'il n'en est rien. On peut donc estimer cette valeur en calculant la proportion de faux positifs dans les simulations (c'est-à-dire le nombre d'items ne fonctionnant pas différemment, mais néanmoins classés comme tels). La puissance du test, quant à elle, est égale à 1 moins la probabilité de classer un item comme ne fonctionnant pas différemment alors que c'est le cas (cette dernière n'est rien d'autre que l'erreur de type II). Par un raisonnement similaire, la puissance d'une méthode est donc estimée par 1 moins le pourcentage de faux négatifs (items classés comme ne fonctionnant pas différemment alors que ce n'est pas le cas). Enfin, le taux d'erreurs total consiste à déterminer le pourcentage d'items qui n'ont pas été classés dans la bonne catégorie. Il s'agit donc d'une moyenne pondérée entre les faux positifs et les faux négatifs (la pondération étant obtenue par le nombre réel d'items présentant un fonctionnement différentiel par rapport au nombre total d'items).

La comparaison des méthodes s'effectue donc sur la base de ces trois quantités empiriques. Elles sont regroupées par tableaux pour simplifier le travail.

3. RÉSULTATS

La présentation des résultats est divisée en deux parties distinctes : le cas où les données ne contiennent pas d'items fonctionnant différemment est traité séparément de la situation où quatre items présentant un fonctionnement différentiel se trouvent dans les données simulées.

3.1. Simulations sans fonctionnement différentiel d'items

Le tableau 3.2 reprend les résultats des simulations lorsque les données ne contiennent aucun item fonctionnant différemment. Les méthodes utilisées sont les suivantes : la méthode Mantel-Haenszel (MH), la standardisation (P-DIF), SIBTEST, la régression logistique (Régr. Logit), Raju et le test du rapport de vraisemblance (Test RV). Notons que, comme le fonctionnement différentiel d'items ne se trouve pas dans cette partie des simulations, les puissances ne peuvent être déterminées (vu qu'il est impossible d'obtenir des faux négatifs). Le tableau 3.2 ne reprend donc que les erreurs de type I (qui correspondent aux taux d'erreurs dans ce cas particulier).

Tableau 3.2
Erreurs de type I empiriques pour les six méthodes en l'absence d'items fonctionnant différemment

Méthode	Habiletés moyennes			
	Égales		Différentes	
	20 items	40 items	20 items	40 items
MH	4,0%	4,4%	3,5%	4,0%
P-DIF	6,0%	6,9%	7,1%	12,1%
SIBTEST	5,8%	4,9%	6,8%	5,6%
Rég. Logit	5,0%	5,3%	5,8%	6,4%
Raju	2,9%	3,5%	3,4%	2,6%
Test RV	5,2%	5,4%	6,0%	6,2%

Nous remarquons tout d'abord que les erreurs de type I empiriques ne diffèrent pas beaucoup entre les méthodes, à l'exception de la standardisation (P-DIF), qui produit des valeurs supérieures à toutes les autres, et la méthode de Raju, qui fournit les valeurs minimales. De plus, ces valeurs sont majoritairement proches de 5% qui est le risque de première espèce fixé *a priori*.

Lorsque la taille du test augmente, l'erreur de type I de la plupart des méthodes tend à se rapprocher de ce risque de 5%, les exceptions notoires étant la standardisation, la régression logistique et le test du rapport de vraisemblance. Remarquons cependant que, hormis pour le cas de la standardisation, ces variations ne sont pas très importantes.

Enfin, lorsque les niveaux d'habileté moyens des groupes de sujets sont différents, les erreurs de type I ont tendance à augmenter légèrement (par rapport au cas de l'égalité des niveaux d'habileté moyens) pour une taille de test fixée. Seule la méthode Mantel-Haenszel fournit des valeurs légèrement inférieures, l'écart n'étant pas plus grand que 0,5%.

3.2. Simulations avec fonctionnement différentiel d'items

Passons à présent aux résultats des simulations lorsque quatre items présentent un fonctionnement différentiel. Le tableau 3.3 reprend les erreurs de type I (α), les puissances (1-β) et les taux d'erreurs (*T.E.*) pour chaque méthode et dans chaque situation.

Tableau 3.3
Erreurs de type I (α), puissances (1-β) et taux d'erreurs (T.E.) empiriques des six méthodes en présence de quatre items fonctionnant différemment

Groupes	Items	Méthode	Effet asymétrique			Effet symétrique		
			a	*1-b*	*T.E.*	*a*	*1-b*	*T.E.*
Habiletés moyennes égales	20	MH	14,5 %	81,0 %	15,4 %	5,1 %	93,5 %	5,4 %
		P-DIF	19,3 %	82,5 %	18,9 %	7,3 %	92,0 %	7,4 %
		SIBTEST	17,5 %	77,0 %	18,6 %	7,1 %	92,0 %	7,3 %
		Rég. Logit	13,6 %	75,5 %	15,8 %	4,9 %	92,5 %	5,4 %
		Raju	11,6 %	78,0 %	13,7 %	3,3 %	93,0 %	4,0 %
		Test RV	16,6 %	85,0 %	16,3 %	6,0 %	94,5 %	5,9 %
	40	MH	6,7 %	88,5 %	7,2 %	4,8 %	94,0 %	5,0 %
		P-DIF	10,8 %	88,5 %	10,9 %	7,9 %	94,0 %	7,8 %
		SIBTEST	7,2 %	84,5 %	8,1 %	5,4 %	93,0 %	5,6 %
		Rég. Logit	6,6 %	85,5 %	7,4 %	5,1 %	93,5 %	5,2 %
		Raju	5,6 %	87,5 %	6,3 %	3,6 %	92,5 %	4,0 %
		Test RV	7,7 %	91,5 %	7,8 %	5,8 %	95,0 %	5,7 %
Habiletés moyennes différentes	20	MH	14,5 %	76,5 %	16,3 %	4,0 %	94,5 %	4,3 %
		P-DIF	20,1 %	77,5 %	20,6 %	9,5 %	88,5 %	9,9 %
		SIBTEST	15,8 %	74,5 %	17,7 %	5,8 %	92,5 %	6,1 %
		Rég. Logit	14,0 %	72,5 %	16,7 %	8,4 %	92,0 %	8,3 %
		Raju	12,1 %	78,0 %	14,1 %	2,8 %	93,0 %	3,6 %
		Test RV	18,0 %	79,0 %	18,6 %	4,9 %	95,5 %	4,8 %
	40	MH	6,4 %	84,5 %	7,3 %	4,9 %	95,5 %	4,9 %
		P-DIF	13,7 %	83,5 %	14,0 %	11,8 %	93,0 %	11,3 %
		SIBTEST	9,9 %	72,0 %	11,7 %	5,8 %	86,0 %	6,7 %
		Rég. Logit	8,1 %	84,0 %	8,9 %	6,7 %	94,5 %	6,6 %
		Raju	4,9 %	86,0 %	5,8 %	4,2 %	97,0 %	4,1 %
		Test RV	8,1 %	88,5 %	8,4 %	6,2 %	97,5 %	5,9 %

En guise de première conclusion, on voit que les méthodes sont nettement mieux adaptées pour la détection du fonctionnement différentiel symétrique. En effet, les erreurs de type I sont nettement plus grandes et les puissances nettement plus petites dans le cas du fonctionnement différentiel asymétrique. En ce qui concerne les erreurs de type I, nous retrouvons des tendances similaires au cas précédent. D'abord, ces valeurs se rapprochent du risque de première espèce de 5 % lorsque la taille du test augmente (excepté pour la standardisation dans le cas symétrique). Ensuite, la méthode de Raju fournit les valeurs les plus petites et la standardisation mène aux erreurs de type I les plus élevées. Enfin, une différence entre les habiletés moyennes a un impact sur les erreurs de type I quoique relativement modéré.

Sans surprise, la puissance des méthodes augmente avec la taille du test. De plus, conformément aux conclusions précédentes sur les erreurs de type I, les méthodes sont plus puissantes lorsque l'effet du fonctionnement différentiel d'items est symétrique plutôt qu'asymétrique. Notons aussi que la puissance diminue lorsqu'une différence entre les niveaux d'habileté moyens des sujets se produit. Il y a cependant très peu de variations au regard de la puissance entre les méthodes. Le test du rapport de vraisemblance est légèrement plus puissant que les autres méthodes bien que l'écart entre les différentes valeurs soit relativement faible.

Terminons en examinant les taux d'erreurs des méthodes. Ceux-ci ont tendance à diminuer lorsque : *a*) la taille du test augmente ; *b*) l'effet du fonctionnement différentiel est symétrique plutôt qu'asymétrique ; *c*) il n'y a pas de différence entre les niveaux d'habileté moyens des sujets. La méthode de Raju semble donner les meilleures classifications des items, tandis que la standardisation fournit les taux d'erreurs les plus élevés. Ces différences sont cependant relativement modérées.

4. DISCUSSION

Le fait que les six méthodes étudiées se comportent de manières pratiquement identiques confirme les résultats obtenus dans les études précédentes mentionnées dans la section 1.2. Il est toutefois apparu deux tendances surprenantes concernant la méthode de Raju et la standardisation qu'il convient d'examiner plus en détail.

Premièrement, la méthode de Raju semble particulièrement bien adaptée dans ce contexte puisqu'elle représente les plus petits taux d'erreurs. Un tel avantage sur les autres méthodes classiques n'apparaît dans aucune autre étude comparative. Néanmoins, la supériorité apparente de la méthode de Raju n'est probablement qu'un artefact dû à l'utilisation du modèle de Rasch, tant pour l'établissement de nos données que pour le sens du fonctionnement différentiel d'items. Il est en effet attendu que cette méthode soit très efficace puisqu'elle décrit parfaitement le modèle sous-jacent aux données simulées. Il est fort probable que dans un contexte plus général cet avantage tende à diminuer.

Deuxièmement, les résultats relatifs à la standardisation, quoique se situant dans la lignée des autres méthodes, ne semblent pas tendre en sa faveur, notamment en termes d'erreurs de type I et de taux d'erreurs totaux. L'explication de ce phénomène (qui est inattendu en vertu des études antérieures) est probablement liée au choix du seuil de détection pour la statistique *P-DIF* que nous avions fixé à 0,05 pour

des raisons de facilité. Une étude centrée sur l'adéquation d'un tel seuil de détection est en cours. Il apparaît que des valeurs légèrement supérieures (de l'ordre de 0,07 ou 0,08) permettraient de corriger les classifications des items de manière importante. Les résultats de cette étude annexe n'étant pas encore entièrement connus, nous nous sommes donc contentés de suivre une règle usuelle de classification des items avec la standardisation, ce qui peut expliquer un tel comportement sous-optimal.

CONCLUSION

Dans ce chapitre, nous avons proposé une étude globale de six méthodes classiques de détection du fonctionnement différentiel des items. Malgré une abondante littérature relative à ces techniques, il apparaît qu'une telle étude comparative n'a jamais été réalisée. Les conclusions de nos simulations vont dans le sens commun : aucune méthode ne semble se révéler comme étant meilleure que les autres.

Ce chapitre a mis l'accent sur les méthodes couramment utilisées pour détecter le fonctionnement différentiel d'items. Cependant, des approches novatrices ont été récemment proposées. Ces nouvelles méthodes reposent sur des modélisations complexes du fonctionnement différentiel des items et nécessitent le recours à des programmes avancés de calcul numérique. Parmi ces techniques récentes, mentionnons les approches par les modèles linéaires hiérarchisés (Kim, 2003), les modèles de mixture (Bolt et Cohen, 2005) et les modèles à coefficients aléatoires (De Boeck, 2008). Toutes ces techniques devraient permettre d'éviter les inconvénients habituellement attribués aux méthodes classiques, le plus important étant le recours à des items (ne fonctionnant pas différemment) déterminés *a priori*.

RÉFÉRENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement, 29*(1), 67-91.

Ankenmann, R. D., Witt, E. A. et Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of educational measurement, 36*(4), 277-300.

Bolt, D. M. et Cohen, A. S. (2005). A mixture model analysis of differential item functioning. *Journal of educational measurement, 42*(2), 133-148.

Camilli, G. et Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, Californie : Sage.

Chang, H.-H., Mazzeo, J. et Roussos, L. (1996). Detecting DIF for polytomously scored items : an adaptation of the SIBTEST procedure. *Journal of educational measurement, 33*(3), 333-353.

Clauser, B. E et Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational measurement : issues and practice, 17*(1), 31-44.

Cohen, A. S., Kim, S.-H. et Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning under the graded response model. *Applied psychological measurement, 20*(1), 15-26.

Cook, L. L. et Eignor, D. R. (1991). IRT equating methods. *Educational measurement : issues and practice, 10*, 37-45.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533-559.

Donoghue, J. R., Holland, P. W. et Thayer, D. T. (1993). A Monte-Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. Dans P. W. Holland et H. Wainer (dir.), *Differential item functioning*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning. Standardization and the Mantel-Haenszel method. *Applied measurement in education, 2*(3), 217-233.

Dorans, N. J. et Kullick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of educational measurement, 23*(4), 355-368.

Finch, W. H. et French, B. (2007). Detection of crossing differential item functioning : a comparison of four methods. *Educational and psychological measurement, 67*(4), 565-582.

Hambleton, R. K. et Rogers, H. J. (1989). Detecting potentially biased test items : comparison of IRT area and Mantel-Haenszel methods. *Applied measurement in education, 2*(4), 313-334.

Holland, P. W. et Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. Dans H. Wainer et H. I. Braun (dir.), *Test validity*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Ironson, G. H. et Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. *Journal of educational measurement, 16*(4), 209-225.

Jiang, H. et Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of educational and behavioral statistics, 23*(4), 291-322.

Jodoin, M. G. et Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied measurement in education, 14*(4), 329-349.

Kim, S.-H. et Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied measurement in education, 8*(4), 291-312.

Kim, S.-H., Cohen, A. S. et Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of educational measurement, 32*(3), 261-276.

Kim, W. (2003). *Development of a differential item functioning (DIF) procedure using the hierarchical generalized linear model : a comparison study with logistic regression procedure*. Thèse de doctorat inédite. Pennsylvania State University.

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

Mantel, N. et Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute, 22*, 719-748.

Mazor, K. M., Clauser, B. E. et Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and psychological measurement, 54*(2), 284-291.

Millsap, R. E. et Everson, H. T. (1993). Methodology review : statistical approaches for assessing measurement bias. *Applied psychological measurement, 17*(4), 297-334.

Narayanan, P. et Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied psychological measurement, 18*(4), 315-328.

Narayanan, P. et Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied psychological measurement, 20*(3), 257-274.

Penfield, R. D. (2003). Application of the Breslow-Day test of trend in odds ratio heterogeneity to the detection of nonuniform DIF. *Alberta journal of educational research, 49*(3), 231-243.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups : a comparison of three Mantel-Haenszel procedures. *Applied measurement in education, 14*, 235-259.

Potenza, M. T. et Dorans, N. J. (1995). DIF assessment for polytomously scored items : a framework for classification and evaluation. *Applied psychological measurement, 19*(1), 23-37.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied psychological measurement, 14*(2), 197-207.

Raju, N. S., Bode, R. K. et Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic to detect differential item functioning. *Applied measurement in education, 2*, 1-13.

Rogers, H. J. et Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied psychological measurement, 17*(2), 105-116.

Rudner, L. M., Getson, P. R. et Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of educational measurement, 17*(1), 1-10.

Shealy, R. et Stout, W. (1993) A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Shepard, L. A., Camilli, G. et Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of educational statistics, 6*(4), 317-375.

Swaminathan, H. et Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of educational measurement, 27*(4), 361-370.

Thissen, D., Steinberg, L. et Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. Dans H. Wainer et H. I. Braun (dir.), *Test validity*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Uttaro, T. et Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied psychological measurement, 18*(1), 15-25.

Chapitre 4

Variables de prédiction du niveau de difficulté de tâches d'évaluation comportant des équations du premier degré en mathématiques et en sciences au secondaire[1]

Martin Riopel, Fadia Sakr, Gilles Raîche,
Patrice Potvin et Valérie Léocadie Djédjé

La production automatisée de tâches d'évaluation est un courant de recherche qui vise à développer des modèles afin de prédire des paramètres comme le niveau de difficulté de tâches d'évaluation à partir d'autres caractéristiques fondamentales de ces tâches. Une fois validés, ces modèles permettraient de produire automatiquement de nouvelles tâches adaptées aux besoins lors de l'évaluation des élèves. Telles que présentées dans Irvine et Kyllonen (2002), les recherches récentes sur la production automatisée de tâches d'évaluation se sont surtout intéressées à des modèles cognitifs généraux ne concernant pas spécifiquement les compétences visées en sciences et en mathématiques dans les écoles. Ce chapitre présente les résultats d'une recherche sur la production automatisée des tâches d'évaluation comportant des équations du premier degré en mathématiques et en sciences au secondaire. Plus précisément, cette recherche a étudié une centaine de tâches provenant de la Banque d'instruments de mesure de la société Gestion du réseau informatique des commissions scolaires (GRICS) qui ont été proposées à 6 910 élèves âgés de 14 à 15 ans et

1. Cette recherche a été rendue possible grâce à une subvention (#410-2007-2357) du Conseil de recherches en sciences humaines du Canada (CRSH) et une subvention (#119232) du Fonds québécois de la recherche sur la société et la culture (FQRSC).

provenant de 22 commissions scolaires francophones du Québec entre 1996 et 2003. Le modèle proposé repose sur neuf variables permettant de prédire le niveau de difficulté avec un coefficient de corrélation de 0,78. Le modèle a été explicitement développé pour permettre la classification de tâches existantes, mais aussi pour permettre la production automatisée de nouvelles tâches. À titre d'exemple d'application, le modèle a été utilisé pour produire 864 nouvelles tâches différentes dont le taux de réussite prédit varie de 0,04 % (tâche très difficile) à 98,7 % (tâche très facile). Ce modèle pourrait être utilisé en ligne par des enseignants et des chercheurs pour soutenir la production de tâches d'évaluation ainsi que dans des environnements informatisés d'évaluation adaptative.

L'évaluation des apprentissages revêt une importance capitale en éducation ; élaborer des tâches d'évaluation constitue à la fois un art et une science (Cronbach et Shapiro, 1982). Plus spécifiquement, l'évaluation adéquate des apprentissages en mathématiques nécessite des enseignants l'utilisation d'une grande diversité de stratégies et d'outils d'évaluation ainsi que l'agencement de ces stratégies et de ces outils de concert avec les cheminements des résultats d'apprentissage et l'équivalence en ce qui a trait à la fois à la mise en application d'appréciation et de notation (Ministère de l'Éducation, du Loisir et du Sport, 2006). Cependant, des recherches (Boucher, Marsolais, Legendre, Scallon, Francœur, Jobin, St-Pierre, Lussier, Lemay, Jalbert et Munn, 2001 ; Dutrenit, 2006 ; Fortin, 2008) ont montré que les enseignants ne disposent pas toujours du temps requis pour élaborer des tâches ou des instruments de mesure très rigoureux au plan docimologique. De plus, il est toujours difficile pour eux de définir et de prédire le niveau de difficulté des items d'évaluation.

La production automatisée de tâches d'évaluation nécessite la construction de modèles pour prédire le niveau de difficulté de nouvelles tâches. Une fois le modèle validé, de nouvelles tâches adaptées pourraient être créées automatiquement pour répondre au besoin de l'évaluation de sujets. Ce procédé comporte des avantages quant au coût d'élaboration de nouvelles tâches et de la validité des tests produits. La production automatisée des tâches est soutenue par les modélisations de la théorie de la réponse à l'item et convient aux évaluations adaptatives et informatisées. Cependant, elle peut être appliquée aussi lors des évaluations classiques en classe (Riopel, Raîche, Pilote et Potvin, 2009). Dans leurs études, Lane (1991a), Sheehan et Mislevy (1994), Sebrechts, Enright, Bennett et Martin (1996), Enright

et Sheehan (2002) ainsi que Schulz, Lee et Mullen (2005) ont identifié des variables prédictives du niveau de difficulté des items en mathématiques.

L'objectif de la recherche a été de construire un modèle permettant de prédire le niveau de difficulté des items à partir de ces variables. Nous exposons la problématique du langage mathématique, de la compétence en algèbre et de l'équivalence des évaluations. La définition des notions de fonctions et d'équations ainsi que cinq recherches sur la production des tâches d'évaluation servent de cadre de référence. Au plan méthodologique, nous décrivons le processus de construction du modèle permettant de prédire le niveau de difficulté des items à partir des variables identifiées dans les études antérieures. L'analyse du modèle obtenu s'ouvre sur des recommandations et des perspectives de recherche.

1. LANGAGE MATHÉMATIQUE ET COMPÉTENCES EN ALGÈBRE

Le langage mathématique semble complexe. Il requiert à la fois une connaissance de l'algèbre dans la résolution des équations et une connaissance de la langue pour la compréhension des problèmes énoncés (Programme de formation de l'école québécoise, 2001). MacGregor et Price (1999) notent que les recherches sur le vocabulaire et l'apprentissage des mathématiques visent à vérifier la compréhension par les élèves des informations mathématiques formelles exprimées en langage naturel. Cela au travers de la signification des mots dans un contexte mathématique et la familiarisation avec le modèle de conversation ou de discours utilisé à l'école pour parler et écrire dans le langage mathématique.

Ainsi, pour Perrin-Glorian (1994), considérer les expressions ou les énoncés en mathématiques d'un point de vue langagier oblige les élèves à les analyser pour mieux les comprendre, formuler des résultats ou des questions et les interpréter. Toutefois, les recherches de Mayer, Larkin et Kadane (1984) indiquent que les élèves éprouvent des difficultés à résoudre un simple problème d'algèbre. De plus, les travaux de De Serre et Groleau (1997) indiquent que les difficultés langagières sont présentes en mathématiques, dans l'utilisation du langage naturel, graphique et symbolique. En outre, De Serres et Groleau (1997) et Laborde (1983) ajoutent qu'abstraire les objets mathématiques de leur contexte représente une difficulté langagière lors de l'utilisation imbriquée des langages symbolique, naturel et graphique. Il en résulte des difficultés pour les enseignants à élaborer des tâches d'évaluation

qui puissent tenir compte de tous ces éléments. Cependant, dans la prochaine section, on verra que l'équivalence des évaluations comporte elle aussi des difficultés.

2. ÉQUIVALENCE DES ÉVALUATIONS ET INSTRUMENTS DE MESURE

Les décisions importantes, comme le classement des élèves, la sélection à l'entrée des études supérieures ou la sanction des études, ne peuvent se baser sur une évaluation approximative, ni sur un classement qui relève de l'arbitraire (Barnes, 1998 ; Bouvette, 1997). Or, les recherches de Barnes (1998) montrent que les méthodes d'évaluation actuelles des difficultés des élèves en mathématique sont critiquées par les éducateurs. Elles montrent également que le problème de l'équivalence des multiples tâches d'évaluation utilisées par les établissements secondaires au Québec se pose avec acuité.

Par ailleurs, en ce qui concerne la mesure, les épreuves du ministère de l'Éducation, du Loisir et du Sport sont paramétrées dans le cadre de la théorie classique des scores (TCS) et s'intéressent beaucoup plus aux groupes qu'aux capacités de l'individu afin de dégager les caractéristiques de l'instrument (Coutu, 1996). Par conséquent, pour les mêmes sujets, un test contenant des items faciles renvoie à une image de sujets compétents tandis qu'un test comprenant des questions difficiles mène à une représentation de sujets incompétents. Dans ces conditions, il s'avère complexe de composer l'item idéal parce que l'enseignant ne connaît pas d'avance les paramètres des questions. Or, le facteur essentiel pour assurer la validité de l'évaluation projetée selon Aschbacher (1991) réside dans la connaissance au préalable du niveau d'habileté approximatif de l'élève.

La production automatisée des tâches d'évaluation basées sur les modélisations de la théorie de la réponse à l'item offre la possibilité de construire des modèles permettant de prédire le niveau de difficulté des items et de mesurer le niveau d'habileté de chaque sujet indépendamment de son groupe d'appartenance. L'importance de s'intéresser à l'évaluation de l'apprentissage des élèves afin de diagnostiquer et de mieux comprendre leurs difficultés d'apprentissage en mathématique est ainsi justifiée. Les sections qui suivent présentent les notions de fonction et d'équation ainsi que celles de la production des tâches d'évaluation.

3. CADRE DE RÉFÉRENCE

Le champ des fonctions et des équations joue un rôle considérable dans les mathématiques et dans d'autres disciplines, telles que l'économie et la physique.

3.1. Notions de fonction et d'équation

Dans le domaine de la fonction linéaire, l'équation du premier degré offre deux perspectives. Elle est considérée comme un processus ou une action lorsqu'elle lie la valeur x à la valeur y (Breidenbach, Dubinsky, Nichols et Hawks, 1992 ; Even, 1990). Par contre, elle représente un objet quand on l'identifie comme un ensemble (Moschkovich, Schoenfeld et Arcavi, 1993 ; Schwarz et Yerushalmy, 1992 ; Sfard, 1992). Pour Moschkovich (2004), travailler avec les fonctions peut être interprété comme une perspective traitant les droites comme objet. En effet, faire correspondre la droite à son équation ($y = mx + b$) est considéré comme une action. L'ordonnée à l'origine est b, la pente m et leurs signes soit 0, positif ou négatif représentent les caractéristiques de l'équation qui peut avoir comme caractéristiques des données du problème l'une des possibilités suivantes :

- l'ordonnée à l'origine avec la pente ;
- une coordonnée avec la pente ;
- plusieurs coordonnées avec la pente ;
- l'ordonnée à l'origine avec une coordonnée ;
- l'ordonnée à l'origine avec deux ou plusieurs coordonnées ;
- deux ou plusieurs coordonnées.

Selon Breton et Smith (1987), l'on procède au calcul de la pente (le taux de variation) ou au calcul de l'ordonnée à l'origine à l'aide de deux coordonnées, ou encore en remplaçant dans l'équation $y = mx + b$, l'une des coordonnées pour trouver la valeur de b, si cette valeur n'est pas donnée dans l'énoncé du problème. La pente et l'ordonnée à l'origine représentent le cas le plus facile à résoudre par les élèves tandis que le cas le plus difficile se présente dans la résolution de deux ou de plusieurs coordonnées. Dans cette perspective, identifier l'équation reviendrait à calculer la pente à l'aide de deux coordonnées. Il s'agira ensuite de remplacer l'une des coordonnées dans l'équation pour trouver l'ordonnée à l'origine. De la bipolarisation de la notion de fonction découlent les interactions entre les différents modes de représentation de l'équation. Dans ce sens, pour Stump (2001), les connaissances conceptuelles de la pente désignent la compréhension des relations entre les représentations variées de la pente algébrique

et géométrique tandis que les connaissances d'application de la pente supposent une familiarité avec la relation de la pente typiquement symbolique (formule algébrique : $m = (y_2 - y_1) / (x_2 - x_1)$) et les conditions d'application de cette formule. Nous avons classé les caractéristiques des données du problème dans le tableau 4.1 selon leur niveau de difficulté et selon leur présence ou leur absence.

Tableau 4.1
Niveau de difficulté des caractéristiques des données

Niveau de difficulté			Pente		Ordonnée à l'origine		Une coordonnée		Deux ou plusieurs coordonnées	
F	**M**	**D**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**
1				1		1	1		1	
	1			1	1			1	1	
	1			1	1		1			1
		1	1		1		1			1
1			1			1	1			1
1			1			1		1	1	

Source : D'après Breton et Smith, 1987.
Notes : F : facile, M : moyen, D : difficile.
A : absent, P : présent.

De plus, Stump (2001) indique que les items nécessitant une connaissance conceptuelle ont généralement un niveau de difficulté moins élevé que les items exigeant une connaissance d'application étant donné que les relations entre les représentations exigent moins d'effort de la part des élèves que les connaissances d'application. Ces connaissances sont connues sous le vocable d'habiletés cognitives. En outre, De Mars (1998) indique que les items qui offrent des réponses à choix multiples semblent plus faciles à résoudre que les items qui offrent des réponses élaborées. Le tableau 4.2 présente la classification des habiletés et du type de réponse aux items selon leur degré de complexité.

Tableau 4.2
Niveau de difficulté de l'habileté cognitive et du type de réponse

Habileté cognitive			Type de réponse		
Conceptualisation	Application	Résolution	Choix multiples	Courte	Élaborée
Facile	Moyen	Difficile	Facile	Moyen	Difficile

3.1.1. *Différents modes de représentation des fonctions*

Des chercheurs ont trouvé une diversité de conceptions sur les représentations de la fonction et présenté les avantages qu'elles offrent aux élèves. Pour Lobato et Siebert (2002) et le *Group for the Psychology of Mathematics Education* (2004), comprendre les fonctions, c'est avoir l'habileté de pouvoir coordonner les différents modes de représentation de la fonction telles les représentations conventionnelles de tableaux, d'équations et de graphiques. Par conséquent, un processus métacognitif se développe lorsque l'élève utilise différentes représentations pour résoudre les équations linéaires. Yerushalmy (2000) et le *Group for Psychology of Mathematics Education* (2004) notent que l'approche fonctionnelle permet aux élèves de développer des habiletés leur permettant de résoudre les équations du premier degré.

En outre, des auteurs tels Moschkovich (1992), Yerushalmy et Chazan (2002) notent que les modes de représentation sous forme de tableau et sous forme de graphique semblent plus faciles à élaborer, car ils simplifient les connaissances requises pour la résolution d'un problème. Par contre, De Serres et Groleau (1997), Nadot (1993), Dreyfus et Mazouz (1993) estiment que les modes de représentation sous forme naturelle et sous forme d'équation semblent plus ardus à réaliser, les élèves trouvant complexe la résolution des équations et la traduction du code algébrique du texte. Par ailleurs, Carpenter, Corbitt, Kepner, Lindquist et Reys (1980) ont remarqué que les élèves résolvent mieux les problèmes ou répondent mieux à la tâche quand ils sont présentés dans un contexte purement mathématique que lorsqu'ils comportent un énoncé. Il en ressort que le mode de représentation sous la forme numérique apparaît plus facile à concevoir que celle sous la forme naturelle. De plus, Duval (1993) et Bloch (2003) indiquent que les interactions entre les différentes représentations de l'objet mathématique devraient être considérées comme nécessaires afin de construire le concept visé (voir tableau 4.3).

3.1.2. *Interactions de représentation : données et réponses d'items*

Selon Confrey et Smith (1994), Bloch (2003), De Serres et Groleau (2003), les principales difficultés dans les interactions entre les représentations des fonctions peuvent être classées en diverses catégories :

- langage symbolique versus langage naturel ;
- langage symbolique versus langage graphique ;
- langage graphique versus langage symbolique ;
- langage numérique (tables) versus langage graphique.

Langage symbolique versus langage naturel

Vergnaud, Cortes et Favre-Artigue (1988), De Serres et Groleau (1997), Rojano (2002), De Serres et collab. (2003) ainsi que le Group for the Psychology of Mathematics Education (2004) mentionnent que les élèves éprouvent des difficultés dans la traduction des équations ou les fonctions linéaires du langage naturel au langage symbolique. En effet, le langage naturel en mathématiques est composé de termes usuels et scientifiques propres à la discipline et le langage symbolique, d'un ensemble de symboles et de règles régissant leur agencement (De Serres et Groleau, 1997). Ces auteurs signalent que les élèves trouvent compliqué la traduction des problèmes comportant des énoncés du code algébrique au texte (naturel au symbolique pour les connaissances conceptuelles) et la résolution de l'équation correspondante.

Langage symbolique versus langage graphique

Afin de pouvoir interpréter un graphique de façon spontanée, l'élève devrait posséder des connaissances préalables, car le graphique illustre un processus composé de codes représentant des concepts (De Serres et Groleau, 1997 ; Dreyfus et Mazouz, 1993 ; Nadot, 1993). Bien que l'élève possède les connaissances nécessaires, il n'utilise pas toujours la représentation graphique de façon automatique. De plus, De Serres et collab. (2003) et Bloch (2003) indiquent que l'articulation entre le registre des représentations graphiques et celui des équations n'est pas bien établie chez les élèves. Ceux-ci éprouvent des difficultés à trouver l'équation à partir d'un graphique parce qu'ils perçoivent les équations d'une façon plutôt algébrique (symbolique) que visuelle (graphique). En outre, les élèves ne réussissent pas à relier les informations issues de différents contextes, comme par exemple relier une équation à un graphique, leurs connaissances étant compartimentées. Les travaux de Schwarz et Yerushalmy (1992), Moschkovich (1992) et de Moschkovich et collab. (1993) ont montré que la représentation symbolique de la fonction rend la nature de son processus intelligible, tandis que la représentation graphique supprime le caractère actif de la fonction et aide ainsi à la présenter plutôt comme une entité. Une compréhension adéquate de l'algèbre requiert de l'élève une familiarité avec les deux aspects de la fonction, étant donné que les deux représentations, équation et graphique, permettent à l'élève de développer des généralisations de la droite.

Langage graphique versus langage symbolique

Stump (2001), Lobato et Siebert (2002) mentionnent que les élèves ont de la difficulté à interpréter et à appliquer la notion de pente (ou taux de variation) dans les fonctions linéaires présentées graphiquement.

Les élèves ne comprennent pas la pente sous les formes fonctionnelle (algébrique) et physique (géométrique : mesure de la pente). Les recherches sur l'apprentissage des fonctions et de leur graphique par ces auteurs montrent les difficultés persistantes qui se posent devant ces différents systèmes de notation.

Langage numérique (tables) versus langage graphique

Pour Moschkovich (1992), Yerushalmy et Chazan (2002) et le *Group for Psychology of Mathematics Education* (2004), les équations présentées sous forme d'exemples, avec des représentations variées et reliées entre elles, supportent la pensée algébrique, c'est-à-dire la transposition d'un contexte situationnel à un contexte mathématique. Selon ces auteurs, la table de valeurs permet à l'élève de développer ses arguments et de se livrer à des conjectures. Quant au graphique, il constitue un moyen productif de développer une représentation élaborée des fonctions chez l'élève. Ces deux éléments, table de valeurs et graphique, visent à réduire la difficulté d'interaction cognitive avec certains aspects symboliques des mathématiques. Il en résulte que le niveau de difficulté des problèmes comportant des représentations sous forme de table versus graphique ou graphique versus table semble moins élevé que ceux comportant des équations versus graphique et vice versa.

Pour conclure sur les difficultés principales dans les différentes formes d'interaction entre les représentations données-réponse, Moschkovich (1992), Yerushalmy et Chazan (2002) notent que le niveau de difficulté de l'interaction tableau-graphique et l'interaction graphique-tableau est faible. Vergnaud, Cortes, Favre-Artigue (1988), De Serres et Groleau (1997), Rojano (2002) et De Serres et collab. (2003) indiquent que le niveau de difficulté de l'interaction naturelle-équation et de l'interaction équation-naturelle est moyen. Schwarz et Yerushalmy (1992), Moschkovich (1992) et Moschkovich et collab. (1993) mentionnent que le niveau de difficulté de l'interaction naturelle-graphique est faible. De Serres et collab. (2003) et Bloch (2003) soulignent que le degré de difficulté de l'interaction équation-graphique et de l'interaction graphique-équation est moyen. Breton et Smith (1987) estiment que l'interaction équation-nombre et l'interaction tableau-équation et vice versa représentent un niveau de difficulté faible. Nous avons classifié les interactions de représentation données-réponse au tableau 4.3 selon leur niveau de difficulté.

Tableau 4.3
Niveau de difficulté de l'interaction de la représentation données-réponse

Représentation des données				Représentation de la réponse				Niveau de difficulté de l'interaction de la représentation données-réponse	
Tab(F)	Gr(F)	Nat(M)	Eq(M)	Tab(F)	Gr(F)	Nat(M)	Eq(M)	F	M
1					1			1	
5							5	5	
	1			1				1	
	4						4		4
		2					2		2
		3			3			3	
			2			2			2
			4		4				4
			5			5		5	
			5	5				5	

Notes : Tab : tableau, Nat : naturel, Eq : équation, Gr : graphique.
F : facile, M : moyen.
1 : Moschkovich, 1992 ; Yerushalmy et Chazan, 2002.
2 : De Serres et Groleau, 1997 ; De Serres et collab., 2003 ; Rojano, 2002 ; Vergnaud, Cortes et Favre-Artigue, 1988.
3 : Moschkovich, 1992 ; Moschkovich et collab., 1993 ; Schwarz et Yerushalmy, 1992.
4 : Bloch, 2003 ; De Serres et collab., 2003.
5 : Breton et Smith, 1987.

Le tableau 4.4 illustre le niveau de difficulté de la combinaison des données et de la représentation de la réponse selon les points de vue des auteurs concernés.

À la lumière de ce qui précède sur les notions de fonction et d'équation, nous avons trouvé que les variables, telles que les caractéristiques des paramètres de l'équation et les caractéristiques des données du problème (présence de pente, d'une ordonnée à l'origine, d'une coordonnée et de deux ou plusieurs coordonnées au tableau 4.1), les habiletés cognitives (de conceptualisation, d'application et de résolution au tableau 4.2), les types de réponses (choix de réponse, réponse courte, réponse élaborée au tableau 4.2), la difficulté de l'interaction des représentations données-réponse, la représentation des données, la représentation de la réponse (tableau 4.3), permettent de prédire le niveau de difficulté dans la résolution des équations du premier degré. La section suivante traite de la production automatisée des tâches d'évaluation et des conclusions de quelques recherches dans ce domaine.

Tableau 4.4
Niveau de difficulté de la combinaison des données et de la représentation de la réponse

Représentation des données					**Représentation de la réponse**					**Niveau de difficulté de l'interaction de la représentation données-réponse**		
Tab(F)	Gr(F)	Nom(F)	Nat(M)	Eq(M)	Tab(F)	Gr(F)	Nom(F)	Nat(M)	Eq(M)	F	M	D
1						1				1		
6					6						6	
6							6			6		
6								6			6	
5									5	5		
	1				1					1		
	4								4		4	
	6						6			6		
	6							6			6	
	6					6					6	
		6			6					6		
		6				6				6		
		6					6				6	
		6						6			6	
		6							6		6	
			2						2		2	
			3			3				3		
			6		6						6	
			6				6				6	
			6					6				6
				2				2			2	
				4		4					4	
				5	5					5		
				5			5			5		
				6					6			6

Notes : Tab : tableau, Nat : naturel, Nom : nombre, Eq : équation, Gr : graphique
F : facile, M : moyen, D : difficile.
1 : Moschkovich, 1992 ; Yerushalmy et Chazan, 2002.
2 : De Serres et Groleau, 1997 ; De Serres et collab., 2003 ; Rojano, 2002 ; Vergnaud, Cortes et Favre-Artigue, 1988.
3 : Moschkovich, 1992 ; Moschkovich et collab., 1993 ; Schwarz et Yerushalmy, 1992.
4 : Bloch, 2003 ; De Serres et collab., 2003.
5 : Breton et Smith, 1987.
6 : Basés sur la classification de chacune des représentations par les auteurs précédents.

3.2. Production automatisée des tâches d'évaluation : bref historique

Irvine (2002) note que les travaux traitant de la production automatisée ont permis d'entrevoir le rapprochement de deux principales sphères de recherche en psychologie expérimentale telles que présentées par Cronbach (1957) : la première sphère contient les résultats d'expérimentation basés sur des modèles complexes cognitifs explicites et un nombre de sujets relativement restreint ; la deuxième sphère porte sur la modélisation des différences individuelles observées qui nécessitent un nombre de sujets plus élevé. Ces deux domaines sont associés aux deux principales théories concernant la production automatisée de tâches d'évaluation soutenues par Bejar, Morley et Benett (2003), à savoir la théorie forte fondée sur des modèles cognitifs pour élaborer des tâches d'évaluation standardisées et la théorie faible fondée sur l'analyse des résultats de banques de tâches pour synthétiser des modèles. Ces auteurs signalent qu'actuellement l'analyse cognitive s'avère plus complexe et par conséquent la théorie faible est généralement appliquée. Celle-ci tend à s'appuyer sur un ensemble de tâches existantes dont les paramètres d'items sont connus. De cet ensemble, des tâches sont sélectionnées pour servir à construire des modèles généraux d'items. Dans le contexte de la théorie faible, au lieu des principes psychologiques, ce sont plutôt les connaissances dans une discipline donnée qui dirigent la conception de modèles de tâches (Riopel, Raîche, Pilote et Potvin, 2009). Fisher (1995) indique que Scheiblechner (1971) fut le premier à manifester un intérêt pour la production automatisée de tâches d'évaluation dans le contexte des modélisations issues de la théorie des réponses à l'item. Il s'est inspiré d'un modèle de régression linéaire pour prédire la valeur du paramètre de difficulté des tâches de compréhension de propositions logiques présentées graphiquement en fonction de trois opérations cognitives : la négation, la disjonction et l'asymétrie. Cinq études plus récentes portant sur les paramètres ou variables qui influencent le niveau de difficulté de tâches de résolution de problèmes en mathématique sont présentées dans la section suivante.

3.2.1. *Des travaux récents*

Lane (1991a) a tenté de vérifier le degré de difficulté et la constance de la pente d'un item où les processus cognitifs variaient en utilisant la théorie de la réponse à l'item et en examinant les connaissances et les méthodes nécessaires pour résoudre des problèmes (algèbre, intérêt et surface). Ainsi, une douzaine d'ensembles comportant deux paires d'items ont été construits. Ces items ont été soumis à 597 élèves inscrits

à un cours d'algèbre pour débutants. Chacun de ces problèmes dépendait d'un type distinct d'habileté, telles la conceptualisation et l'application. L'auteure a aussi identifié quatre étapes de base pour résoudre ces problèmes, soit la traduction, la compréhension, la planification et l'exécution.

Les classifications de la difficulté et de l'uniformité de la pente pour les items ont été vérifiées par une régression multiple linéaire appliquée aux paramètres de difficulté d'items. Le premier ensemble de modèles de comparaison vérifiait l'égalité des paramètres de difficulté et de discrimination pour chacun des ensembles des deux items. Le deuxième ensemble de modèles de comparaison vérifiait l'uniformité de la pente dans les problèmes de distance-taux-temps. Le troisième ensemble de modèles de comparaison examinait dans quelle mesure la familiarité du contexte affectait la difficulté de l'item pour les deux types de problèmes complexes de distance-taux-temps. Quant au dernier ensemble de modèles de comparaison, il vérifiait le classement de la difficulté et l'uniformité de la pente pour les items. Lane a obtenu un coefficient de détermination de 0,60 pour les trois types de problèmes (algèbre, intérêt et surface). En outre, le niveau de difficulté n'a pas été influencé par l'exploitation de la valeur inconnue pour répondre au problème posé. La statistique G2 obtenue était égale à 6,0. Le modèle élaboré s'est avéré incapable de détecter les relations préalables entre les habiletés de conceptualisation et d'application. Toutefois, nous avons identifié les habiletés cognitives comme étant des variables qui peuvent permettre de prédire le niveau de difficulté d'une tâche d'évaluation.

L'étude de Sheehan et Mislevy (1994), quant à elle, visait à déterminer le degré de prédiction des paramètres d'items d'un test de base en mathématique. Deux types d'attributs d'items ont été considérés : 1) les caractéristiques apparentes des items afin de vérifier si le contenu de l'item incluait ou non une équation et 2) l'aspect du processus de la solution afin de vérifier si elle requérait ou non l'application d'une formule standard et le format de la réponse (élaborée ou à choix multiples). Cinq cent dix items ont été classés dans des sous-ensembles tels que les opérations et les nombres, les relations mathématiques, les interprétations des données, la géométrie et les mesures ainsi que le raisonnement puis remis à 900 sujets pour un classement sur une échelle de difficulté de 1 à 5. Le traitement des données a été fait à partir de l'analyse de régression multiple afin d'évaluer la capacité de prédire la valeur du paramètre de difficulté.

Trente variables ont été utilisées dans l'analyse, dont huit présentaient un niveau de signification de 0,15. En utilisant ces huit variables, Sheehan et Mislevy (1994) ont obtenu au modèle 1, un coefficient de détermination ajusté de 0,28. En ajoutant le taux de difficulté moyen des items, ils ont obtenu au modèle 2 un coefficient de détermination ajusté de 0,21. Quand ils ont inclus les items soumis à la comparaison quantitative, le coefficient de détermination ajusté a atteint 0,29 au modèle 3. Au modèle 4, ils ont ajouté quatre autres variables afin de préciser leurs résultats. Ces variables représentaient les paramètres suivants : application d'algorithme standard, histogramme, traduction de mots en symboles et nombre ainsi qu'interprétation des données du type ordre et combinaison. Avec l'ajout de ces paramètres, le coefficient de détermination ajusté du modèle a atteint 0,36. Le paramètre de difficulté au modèle 4, expliquant 36 % de variabilité, correspondait donc au meilleur modèle. De plus, ils ont découvert que les items qui offrent des réponses à choix multiples sont plus faciles à traiter que les items qui demandent des réponses élaborées. Nous avons donc identifié le type de réponse en tant que variable pouvant aider à prédire le niveau de difficulté d'une tâche d'évaluation. Le tableau 4.5 résume les résultats de l'étude de Sheehan et Mislevy.

Puisque le nombre d'items disponibles était limité, les modèles développés ne pouvaient pas avoir une validation croisée. Dans ce contexte, d'autres recherches devaient être menées dans le but de valider la structure du modèle et d'étudier la stabilité des paramètres estimés.

Sebrechts, Enright, Bennett et Martin (1996) ont cherché à évaluer la performance d'un indicateur permettant de mesurer quantitativement l'habileté de raisonnement. Ils ont examiné les relations entre les attributs, les erreurs et les difficultés des problèmes. Ces relations ont été évaluées sur la base de quatre activités cognitives, soit la traduction, l'intégration, l'évaluation pour planifier et l'exécution. Ces chercheurs ont utilisé un ensemble de problèmes avec énoncés algébriques issus des mesures d'évaluation standardisée connu sous le nom de *Graduate Record Examination* (GRE).

L'échantillon était composé de 75 problèmes qui ont été classés en catégories d'équations de probabilité, d'intérêt, de travaux et de distance et administrés à 51 étudiants. Vingt problèmes ont été administrés dans un format de réponses à choix multiples et les autres dans un format de réponses ouvertes. Les approches pour résoudre les problèmes étaient le calcul, le raisonnement, l'estimation et le remplacement des valeurs. Ces auteurs ont trouvé 0,39 comme valeur du coefficient de détermination en algèbre pour les réponses élaborées et 0,47

pour les réponses multiples. Les caractéristiques des problèmes avec énoncés, comme la nécessité d'appliquer des concepts algébriques, la complexité et le contenu, ont constitué des éléments importants pour prédire le niveau de difficulté de la mesure de l'habileté de raisonnement. Ainsi, ces résultats représentent surtout un bon indicateur de difficulté pour les problèmes à réponses élaborées.

Tableau 4.5
Sommaire des résultats des paramètres de difficulté : coefficients de régression estimés et les valeurs de R^2

Paramètres	Modèle 1	Modèle 2	Modèle 3	Modèle 4
Ordonnée à l'origine	–0,16	–2,15	–2,50	–1,90
Taux de difficulté		0,48	0,54	0,50
Comparaison quantitative	0,40		0,71	0,56
Application d'algorithme standard	0,55			0,44
Histogramme	0,97			–0,84
Ordre et combinaison	1,19			
Traduction de mots en symboles				0,405
BDE (application non standard)	0,48			
BDE (application de raisonnement multiple)	0,53			
AC (ordre et combinaison)	–1,67			–0,60
AC (reconnaissance seulement ou rappel)	–0,69			
Degrés de liberté	(8,11)	(1,11)	(2,11)	(6,10)
R^2	0,33	0,22	0,30	0,39
R^2 ajusté	0,28	0,21	0,29	0,36

Source : Adapté de Sheehan et Mislevy, 1994.
Notes : Tous les coefficients de régression étaient significatifs à alpha = 0,15.
Le R^2 ajusté a été corrigé pour le nombre de variables dans le modèle.
AC : domaine de contenu = nombres et opérations et interprétation des données.
BDE : domaine de contenu = relations en mathématiques, géométrie, mesures et raisonnement.

Quant à Enright et Sheehan (2002), leur objectif était d'analyser le niveau de difficulté des problèmes de résolution. Pour ce faire, ces auteurs se sont basés sur des mécanismes comme les informations, les processus, les stratégies et les connaissances emmagasinées (*knowledge stores*) qu'Embretson (1983) avait identifiés pour répondre à des items. Vingt problèmes de mesure quantitative du *Graduate Record Examination* issus des recherches de Sebrechts, Enright, Bennett et Martin (1996) ont été sélectionnés et administrés à 50 collégiens. Ils ont porté

sur la complexité mathématique (le nombre d'opérations, le nombre de contraintes et le nombre de niveaux de parenthèses), le contexte (le temps, l'argent ou la distance) et la situation algébrique.

Trois niveaux d'habileté cognitive ont été définis : l'habileté d'application, l'habileté conceptuelle et l'habileté de résolution. Les résultats ont montré qu'au niveau de l'habileté cognitive, les problèmes d'application étaient les plus faciles à résoudre (65 % comme niveau de difficulté). Ces chercheurs ont obtenu 0,32 comme valeur du coefficient de détermination pour les 368 items en algèbre et 0,33 comme valeur du coefficient de détermination pour les 339 items en complexité mathématique. Nous avons donc identifié les habiletés cognitives et le niveau des élèves en tant que variables qui permettent de prédire le degré de difficulté d'une tâche d'évaluation. Le tableau 4.6 présente la régression de la classification des caractéristiques des items et du niveau cognitif.

Tableau 4.6
Prédiction du paramètre de difficulté pour deux échantillons d'items

	Coefficients de régression linéaire	
Effet	échantillon CP5 (n = 339)	échantillon CP6 (n = 368)
Ordonnée à l'origine	0,36**	0,32**
Démarche	–0,81***	–0,51***
Degré élevé	0,60***	0,86***
Application de démarche	–0,52*	–1,22***
Algèbre	0,33*	0,13
R^2	0,36	0,37
Validation croisée R^2	0,32	0,32

Source : Adapté d'Enright et Sheehan, 2002.
Notes : ***$p < 0{,}001$; **$p < 0{,}01$; *$p < 0{,}05$.

La plupart des travaux d'Enright et Sheehan (2002) sur la prédiction des valeurs des paramètres d'items ont été réalisés à partir de tâches d'évaluation dans un contexte expérimental très précis. Par conséquent, les résultats de ces travaux sont moins applicables aux divers contextes éducatifs.

Schulz, Lee et Mullen (2005) ont utilisé le niveau de la 8e année en mathématiques du contenu du *National Assessment of Educational Progress* (NAEP) pour vérifier si la performance était associée à la maîtrise de plusieurs habiletés et la performance des étudiants. Deux cent quatre-vingt-quatorze items issus du *National assessment of educational progress* portant sur les nombres, les mesures, la géométrie, la

statistique et l'algèbre et requérant trois habiletés cognitives (conceptualisation, application et résolution) ont été utilisés. La modélisation logistique issue de la théorie des réponses à l'item a été appliquée. De plus, la corrélation variait de 0,86 à 0,93 pour les nombres, la statistique et l'algèbre. L'échelle du National Assessment of Educational Progress pour la moyenne des scores pondérés se situait entre 100 et 500 pour les sujets. Des résultats, il se dégage que les items qui relevaient des habiletés de résolution de problèmes semblaient plus difficiles et complexes que ceux relevant des habiletés d'application et de conceptualisation.

Le tableau 4.7 montre les résultats de l'étude. Le critère hypothétique était de 65 % et plus pour le niveau de performance d'un domaine donné. Les items du domaine A-1 ont eu une moyenne de difficulté $b = -0{,}85$ et semblaient plus faciles à résoudre que les items des autres scores en algèbre. Le plus petit écart entre les moyennes des scores de la discipline de l'algèbre était de 1,34 pour les domaines A-2 (0,49) et A-1 (–0,85) ; et le plus grand écart était de 1,96 pour les domaines A-3 (1,11) et A-1 (–0,85).

Tableau 4.7
Sommaire statistique des scores de la discipline

Discipline	Domaine d'enseignement	Nombre d'items	Score des domaines	Moyenne du niveau de difficulté des items (*b*)	Limite du niveau de la performance du domaine introduit	Limite du niveau de la performance du domaine maitrisé
Algèbre	A-1	8	A-1	–0,85	66 %	93 %
	A-2	8				
	A-3	10	A-2	0,49	35 %	79 %
	A-4	4				
	A-5	6	A-3	1,11	22 %	62 %
	A-6	2				
	Total	38				

Source : Adapté de Schulz, Lee et Mullen, 2005.
Notes : A-1 : opération de base.
A-2 : logique, modèles simples, raisonnement algébrique.
A-3 : système de coordonnées.
A-4 : utilisation de variables dans des expressions.
A-5 : résolutions de problèmes avec équations et inéquations.
A-6 : modèles complexes.

Dans cette étude, Schulz, Lee et Mullen (2005) ont constaté que très peu de domaines présentaient le niveau de difficulté prévu. Ils ont démontré que l'accroissement de la performance signifiait l'augmentation du pourcentage acceptable du score dans une discipline et la

maîtrise de celle-ci dans une séquence égalait le niveau de difficulté prévu. Nous avons donc pris les habiletés cognitives et le niveau des élèves comme variables permettant de prédire le niveau de difficulté d'une tâche d'évaluation.

De l'analyse des résultats des cinq précédentes études et des conclusions d'autres auteurs rapportées dans la section sur la notion des fonctions et des équations, nous avons dégagé les neuf neuf variables suivantes qui permettent de prédire le niveau de difficulté des tâches d'évaluation d'équations du premier degré :

- les caractéristiques des paramètres de l'équation ;
- les caractéristiques des données du problème (présence de pente, d'une ordonnée à l'origine, d'une coordonnée ou de deux ou de plusieurs coordonnées) ;
- les habiletés cognitives de conceptualisation, d'application et de résolution ;
- la représentation des données du problème ;
- la représentation de la réponse ;
- la difficulté de l'interaction des représentations données-réponse ;
- les types de réponses ;
- la difficulté des données ;
- le nombre d'équations et le niveau des élèves.

3.3. Objectif de la recherche

L'objectif de la recherche était, à partir des neuf variables retenues, de construire un modèle qui permet de prédire le niveau de difficulté dans les tâches d'évaluation d'équations du premier degré d'une centaine d'items issus de la banque d'instruments de mesure (BIM) du réseau secondaire au Québec.

4. MÉTHODOLOGIE

Pour atteindre l'objectif, une procédure méthodologique descriptive et empirique a été mise en place. Nous avons eu accès aux épreuves de la banque d'instruments de mesure (BIM) de 1997 à 2003. Ces épreuves contiennent des questionnaires à propos des nombres, de l'algèbre, des mesures et de la géométrie. Au total, 100 items de ces épreuves ont été retenus et administrés à 6 910 élèves du niveau secondaire 3 et 4 dont l'âge moyen varie entre 14 et 15 ans provenant de 22 commissions scolaires francophones du Québec.

Les items retenus englobent des questions propres à l'algèbre, en particulier les équations du premier degré à une ou à deux inconnues. Ils ont été rédigés et révisés par des spécialistes de l'éducation de plusieurs commissions scolaires dont celles de Bersimis, Fermont, Littoral, Manicouagan, la Moyenne-Côte-Nord, Port-Cartier, Sept-Îles, Tadoussac, Pierre-Neveu, Le Gardeur, Chomedey-de-Laval, Berthier-Nord-Joli, Académie Sainte-Thérèse, la région de l'Outaouais, Affluents, Seigneurie-des-Mille-Îles, la Rivière-du-Nord, des Laurentides et Laval avec la collaboration de la société GRICS. Le traitement informatisé des données a été réalisé avec le logiciel *SPSS 15.0*.

5. RÉSULTATS

Les items de chaque épreuve ont été classifiés selon 4 catégories :

- le domaine (nombres, algèbre, géométrie, statistiques, mesures) ;
- l'habileté cognitive (concept, application, résolution) ;
- le type de réponses (réponse à choix multiples, réponse courte, réponse élaborée) ;
- le niveau de difficulté des items (facile, moyen, difficile).

Nous avons obtenu les neuf variables identifiées dans les précédentes études.

Tableau 4.8
Neuf variables qui permettent de prédire le niveau de difficulté d'une tâche d'évaluation comportant une équation à une ou deux inconnues

Description	Symbole
Taux de réussite	IR
Caractéristiques des données du problème : présence de l'ordonnée à l'origine	CDPor
Caractéristiques des paramètres de l'équation : signe de la pente – ou 0 ou +	CPEp2
Nombre d'équations (1 ou 2)	Neq
Niveau des élèves (item 314, 416, 426 ou 436)	NEL
Habileté cognitive, soit concept ou application ou résolution	HAB2
Type de réponse (choix multiples, courte ou élaborée)	TR
Difficulté de l'interaction des représentations données-réponse	DC2
Caractéristiques des données du problème : présence de la pente	CPDp
Caractéristiques des paramètres de l'équation : signe de l'ordonnée à l'origine – ou 0 ou +	CPEor2

5.1. Prédiction du taux de réussite des items

Une analyse de régression linéaire a été appliquée aux 100 items. Elle a donné un coefficient de régression de 0,93. La valeur obtenue au test *F* est 6,87 avec une signification statistique de 0,0001. Nous avons calculé la valeur du taux de réussite modélisé pour chaque item, ce qui nous a amené à éliminer tous les items dont la différence entre le taux de réussite réel (IR) et le taux de réussite modélisé (IRmod) est plus grand que la limite arbitraire (0,25)**.** Nous avons constaté que seuls les items 13 et 21 correspondaient à cette condition. L'item 13 présente une différence de 0,27 et l'item 21 une différence de 0,26. Quatre-vingt-dix-huit items ont ainsi été analysés en rapport avec la variable dépendante qui est le taux de réussite (IR).

Nous avons imposé une limite arbitraire à la valeur du coefficient de corrélation de Pearson pour diminuer le nombre des variables indépendantes. Cette limite a été fixée à |0,18|. Une valeur inférieure à cette limite est donc considérée comme faible et n'influence pas vraiment le taux de réussite. Avec cette nouvelle valeur comme limite, nous avons obtenu des résultats préliminaires supérieurs à 0,9 et une valeur de 416,44 comme résultats au test *F*. Le taux de réussite de 43 items parmi les 98 items utilisés n'étant pas connu, nous avons analysé 55 items pour lesquels nous connaissions le taux de réussite correspondant.

Après avoir trouvé un coefficient de régression préliminaire supérieur à 0,9 ; nous avons choisi de vérifier la qualité du modèle de la régression linéaire par l'utilisation de la variable indépendante qui a le plus grand coefficient de corrélation de Pearson. Cette variable est la présence de l'ordonnée à l'origine (une des caractéristiques des données du problème) et est la plus influente sur le taux de réussite. Nous avons également ajouté les trois autres variables qui affichent les plus grands coefficients de Pearson qui sont le signe de la pente (CPEp2) (une des caractéristiques des paramètres de l'équation), le nombre d'équations (Neq) et le niveau des élèves (NEL). Nous remarquons que le coefficient de régression est passé de 0,38 à 0,68. D'autres variables comme la présence de la pente (CPDp) (une des caractéristiques des données du problème) et le signe de l'ordonnée à l'origine (CPEor2) (une des caractéristiques des paramètres de l'équation). Nous constatons que le coefficient de régression a atteint 0,70 avec ces six variables.

De plus, nous avons considéré la variable *difficulté de l'interaction des représentations données-réponse* (DC2) et les variables qui influencent de façon significative le taux de réussite, notamment *l'habileté cognitive* (HAB2) et le *type de réponse* (TR). Nous avons ainsi neuf variables et obtenons un coefficient de régression (corrélation) de 0,78. Nous avons obtenu un coefficient de détermination (corrélation), par l'analyse de

régression linéaire, de 0,61 et un coefficient de détermination ajusté de 0,497. Nous pouvons en déduire que ces neuf variables (tableau 4.9) influencent le niveau de réussite des items et qu'elles représentent le minimum requis pour construire des items d'une équation ou de deux équations du premier degré :

- la présence de l'ordonnée à l'origine (caractéristiques des données du problème) ;
- la présence de la pente (caractéristiques des données du problème) ;
- le signe de la pente (caractéristiques des paramètres de l'équation) ;
- le signe de l'ordonnée à l'origine (caractéristiques des paramètres de l'équation) ;
- le nombre d'équations ;
- le niveau des élèves ;
- l'habileté cognitive ;
- le type de réponse ;
- la difficulté de l'interaction des représentations données-réponse.

Ces résultats présentés sont supérieurs à ceux obtenus par Enright et Sheehan (2002) qui sont de 0,33 pour 339 items et de 0,32 pour 368 items en algèbre. Le tableau 4.9 présente le modèle 9 qui a été retenu après les précédents calculs.

Tableau 4.9
Sommaire de la corrélation (qualité) du modèle de la régression linéaire (9 variables et 45 items, pente et ordonnée à l'origine)

Modèle	R	R^2	R^2 ajusté	Estimation de l'erreur type	Variables prédictives pour le taux de réussite
9	0,78	0,61	0,50	0,14	IR, NEL, DC2, CPEp2, CDPor, CPEor2, CPDp, Neq, HAB2

Le tableau 4.10 présente les statistiques descriptives pour les 45 items et les 9 variables.

Tableau 4.10
Statistiques descriptives pour les 45 items et les 9 variables

Moyenne	Valeur minimale	Valeur maximale	Écart type	N	Corrélation de Pearson	Description	Symbole
na	na	na	na	na	na	Taux de réussite	IR
0,76	0	1	0,43	98	0,39	Caractéristiques des données : présence de l'ordonnée à l'origine	CDPor
2,74	0	4	0,68	92	0,38	Caractéristiques des paramètres de l'équation : signe de la pente – ou 0 ou +	CPEp2
1,50	1	2	0,56	98	0,35	Nombre d'équations (1 ou 2)	Neq
1,79	1	4	0,87	98	0,35	Niveau des élèves (314, 416, 426 ou 436)	NEL
1,90	1	3	0,78	98	–0,27	Habileté cognitive, soit concept ou application ou résolution	HAB2
1,73	1	3	0,86	98	–0,26	Type de réponse (choix multiples, courte ou élaborée)	TR
1,64	1	3	0,48	98	–0,17	Difficulté de la combinaison des représentations données-réponse	DC2
0,68	0	1	0,47	98	0,16	Caractéristiques de l'interaction desdonnées : présence de la pente	CPDp
2,83	0	4	0,50	93	–0,01	Caractéristiques des paramètres de l'équation : signe de l'ordonnée à l'origine – ou 0 ou +	CPEor2

Les données du tableau 4.11 représentent les coefficients des neufs variables qui seront utilisés pour la construction du modèle dans la section 5.3.

Tableau 4.11
Coefficients pour le calcul du taux de réussite des neuf variables

Description	Coefficient			
Constante	0,15			
1) Ordonnée à l'origine (caractéristiques des données du problème)	Présence		Absence	
	0,17		0,00	
2) Pente (caractéristiques des données du problème)	Présence		Absence	
	–0,06	0,00		
3) Signe de l'ordonnée à l'origine (caractéristiques des paramètres de l'équation)	Rien	Négatif	Nul	Positif
	0,00	0,03	0,07	0,10
4) Signe de la pente (caractéristiques des données du problème)	Rien	Négatif	Nul	Positif
	0,00	0,13	0,26	0,40
5) Habileté	Concept	Appli-cation	Résolution	
	–0,01	–0,02	–0,03	
6) Type de réponse	Multiple	Courte	Élaborée	
	–0,05	–0,09	–0,14	
7) Nombre d'équations	Une		Deux	
	0,09		0,18	
8) Niveau des élèves	314	416	426	436
	0,07	0,14	0,22	0,29
9) Difficulté de l'interaction des représentations données-réponse	Facile	Moyenne	Difficile	
	–0,10	–0,20	–0,30	

5.2. Discussion

C'est pour tenter de connaître *a priori* le niveau de difficultés des tâches d'évaluation des équations du premier degré en algèbre afin de résoudre le problème de l'équivalence des tâches d'évaluation que cette étude a été entreprise. Les résultats ont montré, conformément aux recherches dans le cadre de référence de cette étude, que neuf variables prédictives expliquent le taux de réussite des tâches d'évaluation des équations du premier degré.

En outre, nous avons trouvé un coefficient de détermination de 0,61, comme résultat pour ces 9 variables et les 45 items analysés. Ce coefficient est pratiquement égal à celui obtenu par l'étude de Lane (1991) soit 0,60. De plus, il dépasse largement celui trouvé par Sheehan, Kathleen, Mislevy et Robert (1994) : 0,36. Il est également supérieur à celui obtenu par Sebrechts, Enright, Bennett et Martin (1996) : 0,39 comme coefficient de détermination en algèbre pour les réponses élaborées et 0,47 pour les réponses à choix multiples et représente plus du double du coefficient de détermination de Enright et Sheehan (2002) : 0,32. Nous pouvons en conclure que les résultats de cette recherche sont au moins comparables à ceux des chercheurs précédents et expliquent le niveau de difficulté des items dans les tâches d'évaluation des équations du premier degré en mathématiques au secondaire. De plus, les items utilisés ont été rédigés par des personnes provenant d'horizons professionnels et géographiques divers, dont des conseillers pédagogiques et enseignants de plusieurs commissions scolaires, des représentants du ministère de l'Éducation au Québec et de la société GRICS. Cette diversité de sources confère au modèle une certaine pertinence parce qu'il ne dépend pas des tâches d'une seule personne.

Il est à noter que nous avons essayé de poursuivre les calculs et que nous sommes parvenus à l'identification de 21 variables pouvant influencer le niveau de difficulté d'une tâche d'évaluation comportant une équation, tentative qui semble donc prometteuse pour d'autres variables. En plus d'analyser ces résultats, nous proposons un modèle pour construire des items.

5.3. Application du modèle

Afin de construire un item facile, c'est-à-dire ayant un taux de réussite de 75 % ou plus, nous avons procédé en trois étapes : 1) utiliser le choix de combinaisons présenté au tableau 4.12 ; 2) choisir parmi les huit possibilités des signes de la pente et de l'ordonnée à l'origine au tableau 4.13 ; 3) finalement, choisir parmi les dix-huit possibilités présentées au tableau 4.14.

Tableau 4.12
Combinaison : interaction des représentations, données, présence de la pente et de l'ordonnée à l'origine

Possibilité	Difficulté de l'interaction des représentations données-réponse	Difficulté des données	Présence de la pente	Présence de l'ordonnée
1	Facile	Facile	Oui	Oui

Tableau 4.13
Combinaison : signe de la pente et de l'ordonnée à l'origine

Possibilité	Signe de la pente	Signe de l'ordonnée à l'origine
1	Positif	Positif
2	Négatif	Positif
3	Négatif	Négatif
4	Les deux	Les deux
5	Nul	Positif
6	Nul	Négatif
7	Positif	Nul
8	Négatif	Nul

Tableau 4.14
Combinaisons : habileté, type de réponse et nombre d'équations

Possibilité	Habileté cognitive	Type de réponse	Nombre d'équations
1	Concept	Choix multiples	Une
2	Concept	Choix multiples	Deux
3	Concept	Courte	Une
4	Concept	Courte	Deux
5	Concept	Élaborée	Une
6	Concept	Élaborée	Deux
7	Application	Choix multiples	Une
8	Application	Choix multiples	Deux
9	Application	Courte	Une
10	Application	Courte	Deux
11	Application	Élaborée	Une
12	Application	Élaborée	Deux
13	Résolution	Choix multiples	Une
14	Résolution	Choix multiples	Deux
15	Résolution	Courte	Une
16	Résolution	Courte	Deux
17	Résolution	Élaborée	Une
18	Résolution	Élaborée	Deux

Nous avons constaté que le nombre de possibilités de combinaisons des choix était de : $1 \times 8 \times 18 = 144$ possibilités pour construire des items faciles. En outre, le calcul des taux de réussite s'est effectué à partir du tableau 4.11 en utilisant les coefficients correspondant au calcul du taux de réussite des neuf variables retenues.

Quelques exemples de calculs (tableau 4.11) – Exemple 1

Item de niveau 314 : 0,153 + –0,101 + –0,058 + 0 + 0,099 + 0,396 + –0,02 + –0,094 + 0,182 + 0,072 = 62,9 %

Item de niveau 416 : 0,153 + –0,101 + –0,058 + 0 + 0,099 + 0,396 + –0,02 + –0,094 + 0,182 + 0,144 = 70,1 %

Item de niveau 426 : 0,153 + –0,101 + –0,058 + 0 + 0,099 + 0,396 + –0,02 + –0,094 + 0,182 + 0,216 = 77,3 %

Item de niveau 436 : 0,153 + –0,101 + –0,058 + 0 + 0,099 + 0,396 + –0,02 + –0,094 + 0,182 + 0,288 = 84,5 %

Cette combinaison présente un niveau de difficulté facile pour les niveaux 426 et 436. Par contre, elle représente un niveau de difficulté moyenne pour les niveaux 314 et 416, car le taux de réussite est inférieur à 75 %. Néanmoins, on remarque que le taux de réussite augmente avec le niveau de l'élève.

Un exemple d'item de niveau de difficulté facile – Exemple 2

(Pour des élèves de niveau 416 avec un taux de réussite prédit de 92 %)

a) Données du problème : Les tableaux ci-dessous représentent les coûts des cours de danse suivis dans deux écoles différentes selon le nombre d'heures (h). C1 est le coût du premier cours et C2 est le coût du deuxième cours.

Tableau 1

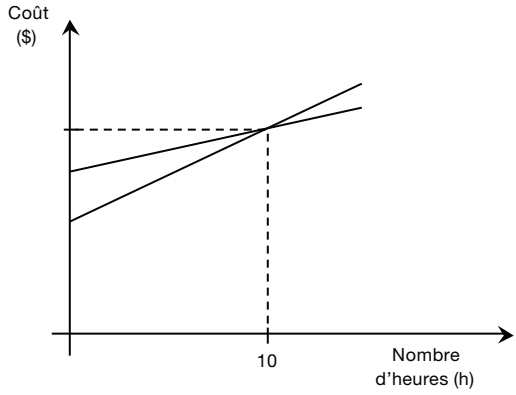
Temps (h)	0	1	2
Coût ($)	30	35	40

Tableau 2

Temps (h)	0	1	2
Coût ($)	20	26	32

b) Question : Représenter graphiquement les coûts des deux cours de danse.

c) Réponse attendue :



Résultat : On obtient 87,7 % comme taux de réussite pour le niveau 416.

Afin de construire un item avec un niveau de difficulté moyen et un taux de réussite se situant entre 50 % et 74 %, nous avons procédé en trois étapes : 1) choisir parmi les trois possibilités présentées au tableau 4.15 ; 2) sélectionner parmi les huit possibilités du signe de la pente et de l'ordonnée à l'origine au tableau 4.13 ; 3) choisir parmi les dix-huit possibilités du tableau 4.14.

Tableau 4.15
Combinaison : interaction des représentations, données, présence de la pente et de l'ordonnée à l'origine

Possibilité	Difficulté de l'interaction des représentations données-réponse	Difficulté des données	Présence de la pente	Présence de l'ordonnée
1	Facile	Difficile	Non	Non
2	Moyen	Moyen	Oui	Non
3	Difficile	Facile	Oui	Oui

Le nombre de possibilités de combinaisons des choix était de : $3 \times 8 \times 18 = 432$ cas pour produire des items ayant un niveau de difficulté moyen. Le calcul des taux de réussite a été fait à partir des coefficients correspondant au calcul du taux de réussite des neuf variables (tableau 4.11).

Exemples de calculs pour quelques items (tableau 4.11) – Exemple 3

Item de niveau 314 : 0,153 + –0,101 + –0,058 + 0,099 + 0,396 + –0,010 + –0,47 + 0,091 + 0,072 = 59,5 %

Item de niveau 416 : 0,153 + –0,101 + –0,058 + 0,099 + 0,396 + –0,010 + –0,47 + 0,091 + 0,144 = 66,7 %

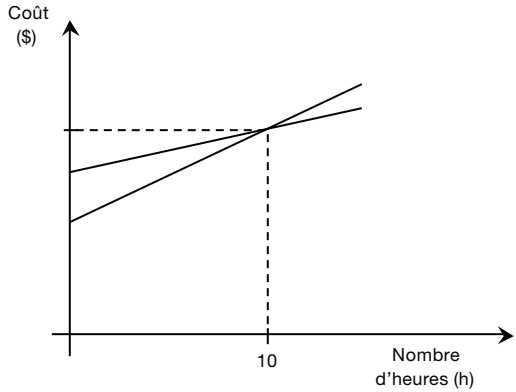
Item de niveau 426 : 0,153 + –0,101 + –0,058 + 0,099 + 0,396 + –0,010 + –0,47 + 0,091 + 0,216 = 73,9 %

Item de niveau 436 : 0,153 + –0,101 + –0,058 + 0,099 + 0,396 + –0,010 + –0,47 + 0,091 + 0,288 = 81 %

Cette combinaison produit un niveau de difficulté moyen pour les niveaux 314, 416 et 426. Par contre, elle présente un niveau de difficulté faible pour le niveau 436.

Un exemple d'item de niveau de difficulté moyen – Exemple 4

(Pour des élèves de niveau 426 avec un taux de réussite de 69,7 %)
a) Données du problème Le système de relations ci-dessous représente les coûts de cours de danse suivis dans deux écoles différentes selon le nombre d'heures (h). C1 est le coût du premier cours et C2 est le coût du deuxième cours.
   1re équation : C1 = 5h + 30
   2e équation : C2 = 6h + 20
b) Question : Représenter graphiquement le système de relations.
c) Réponse attendue :



Résultat : On obtient 70,6 % comme taux de réussite pour le niveau 426.

En outre, un item difficile, c'est-à-dire dont le taux de réussite est inférieur à 50 % a été élaboré selon les deux étapes suivantes : 1) choisir parmi les deux possibilités présentées au tableau 4.16 ; 2) sélectionner parmi les huit possibilités de signe de la pente et de l'ordonnée à l'origine au tableau 4.13 ; 3) choisir parmi les dix-huit possibilités présentées au tableau 4.14.

Tableau 4.16
Combinaison : interaction des représentations, données, présence de la pente et de l'ordonnée à l'origine

Possibilité	Difficulté de l'interaction des représentations données-réponse	Difficulté des données	Présence de la pente	Présence de l'ordonnée
1	Moyenne	Difficile	Non	Non
2	Difficile	Difficile	Non	Non

Le nombre de possibilités de combinaisons des choix était de : $2 \times 8 \times 18 = 288$ cas pour construire des items avec un niveau de difficulté élevé. Le calcul des taux de réussite a été fait à partir des coefficients correspondant au calcul du taux de réussite des neuf variables (tableau 4.11).

Quelques exemples de calculs pour quelques items (tableau 4.11) : Exemple 5

Item de niveau 314 : 0,153 + –0,303 + 0 + 0 + 0,033 + 0,132 + –0,03 + –0,094 + 0,091 + 0,72 = 5 %

Item de niveau 416 : 0,153 + –0,303 + 0 + 0 + 0,033 + 0,132 + –0,03 + –0,094 + 0,091 + 0,144 = 12,6 %

Item de niveau 426 : 0,153 + –0,303 + 0 + 0 + 0,033 + 0,132 + –0,03 + –0,094 + 0,091 + 0,216 = 19,8 %

Item de niveau 436 : 0,153 + –0,303 + 0 + 0 + 0,033 + 0,132 + –0,03 + –0,094 + 0,091 + 0,288 = 27 %

Nous pouvons conclure que cette combinaison présente un degré de difficulté élevé pour tous les niveaux.

Ainsi, la description du processus a mené à la construction d'un modèle de 864 items différents qu'on pourrait multiplier sans modifier le niveau de difficulté avec un taux de réussite prédit de 0,04 % à 98,7 %. Pour les tâches d'évaluation, 144 items ont un niveau de difficulté faible (taux de réussite entre 75 % et 100 %), 432 items ont un niveau de difficulté moyen (taux de réussite entre 50 % et 75 %) et 288 items ont un niveau de difficulté élevé (taux de réussite inférieur à 50 %). De plus, ce modèle a expliqué le taux de réussite avec un coefficient de détermination de 0,61. Ainsi, neuf variables prédictives, soit : la présence de l'ordonnée à l'origine dans les données du problème, la présence de la pente dans les données du problème et le signe de la pente, le signe de l'ordonnée à l'origine, le nombre d'équations, le niveau des élèves, l'habileté cognitive, le type de réponse, la difficulté de l'interaction des représentations données-réponse ont permis de développer ce modèle qui a pu produire 864 items différents avec un taux de réussite prédit de 0,04 % à 98,7 % et de conclure que l'on a pu prédire efficacement le niveau de difficulté des équations du premier degré.

CONCLUSION

Suite à cette étude plus exploratoire que prescriptive, mais qui décrit le processus de construction d'items en utilisant les variables prédictives, nous pouvons conclure qu'il a été utile d'expliquer ce processus

qui a contribué à construire des items avec leur niveau de difficulté prévu. De celui-ci, un modèle a émergé, lequel pourrait servir dans d'autres disciplines telles que les sciences. De plus, il pourrait également être informatisé et utilisé par les enseignants et concepteurs de programmes d'enseignement afin de faciliter l'évaluation des élèves en mathématiques. Il pourrait aussi servir à construire des systèmes informatisés d'évaluation adaptative.

Toutefois, cette étude n'a pas la prétention d'avoir fait le tour de la problématique des équations parce que 21 variables mériteraient encore d'être analysées. Son ouverture sur d'autres pistes de recherche sur l'équation du deuxième degré ou en modélisation logistique à trois paramètres, telles que les difficultés, la discrimination et la pseudo-chance sur les tâches d'évaluation et à l'utilisation d'enquêtes internationales (TIMSS et PIRS) apparaît souhaitable. Ces avenues de recherche représentent des possibilités d'enrichissement de la compréhension des facteurs influençant les tâches d'évaluation au secondaire en mathématique.

RÉFÉRENCES

Aschbacher, P. R. (1991). Performance assessment : State activity, interest and concerns. *Applied measurement in education, 4*(4), 275-288.

Barnes, L. B. (1998). The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied measurement in education, 11*(2), 179-194.

Bejar, I. I., Morley, M. E. et Benett, R. E. (2003). A feasibility study of on-the-fly-item generation in adaptive testing. *The Journal of technology, learning and assessment*, *2*(3), 1-29.

Bloch, I. (2003). Teaching functions in a graphic milieu : what forms of knowledge enable students to conjecture and prove ? *Educational studies in mathematics, 52*(1), 3-28.

Bouvette, M. (1997). *Typologie et paramètres des items dans le paradigme de la théorie des réponses aux items.* Mémoire de maîtrise inédit. Université du Québec à Montréal, Montréal, Québec.

Breidenbach, D., Dubinsky, E., Nichols, D. et Hawks, J. (1992). Development of the process conception of function. *Educational studies in mathematics*, *23*(3), 247-285.

Breton, G. et Smith, J-G. (1987). *Mathématique au secondaire. BMS 4 : cahier d'exercices :programme 414.* Montréal, Québec : Éditions HRW.

Boucher, M., Marsolais, A., Legendre, M.-F., Scallon, G., Francœur, P., Jobin, M., St-Pierre, M., Lussier, D., Lemay, V., Jalbert, P. et Munn, J. (2001). L'évaluation des apprentissages : un sens à trouver. *Vie pédagogique, 120.* Québec : Ministère de l'Éducation du Québec.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., Lindquist, M. M. et Reys, R. E. (1980). Solving verbal problems : results and implications from national assessment. *The arithmetic teacher, 28*(1), 8-12.

Confrey, J. et Smith, E. (1994). Exponential functions, rates of change, and the multiplicative unit. *Educational studies in mathematics. 26*, 135-164.

Coutu, M. (1996). *La typologie des items comme facteur pouvant en influencer les paramètres : une étude portant sur les épreuves uniques du MEQ en histoire 414 entre 1988 et 1993*. Mémoire de maîtrise inédit. Université du Québec à Montréal, Montréal, Québec.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American psychologist, 12,* 671-684.

Cronbach, L. J. et Shapiro, K. (1982*). Designing evaluations of educational and social programs*. San Francisco, Californie *:* Jossey Bass.

De Mars, C. E. (1998). *The impact of test consequences and response format on performance*. Thèse de doctorat inédite. Michigan State University.

De Serres, M. et Groleau (1997). *Mathématiques et langages*. Montréal, Québec : Collège Jean-de-Brébeuf, Direction pédagogique, service de la recherche.

De Serres, M., Bélanger, M., Piché, M. C., Riopel, M., Staub, C. et de Grandpré, C. (2003). *Intervenir sur les langages en mathématiques et en sciences*. Montréal, Québec : Modulo.

Dreyfus, A. et Mazouz, Y. (1993). L'utilisation judicieuse du langage des graphiques par des élèves de seconde dans le domaine de la biologie. *Les Sciences de l'éducation – Pour l'Ère nouvelle, 1*(3), 245-266.

Dutrenit, J.-M. (2006). Évaluation modes d'emploi : évaluer et stimuler les élèves. *La Nouvelle revue de l'AIS, 32, 9-18*.

Duval, R. (1993). Graphiques et équations, Les représentations graphiques dans l'enseignement et la formation. *Les Sciences de l'éducation – Pour l'Ère nouvelle, 1*(3), 57-72.

Embretson, S. (1983). Construct validity : construct representation versus nomothetic span. *Psychological bulletin, 93*(1), 179-197.

Enright, M. K. et Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items : Implications for item generation. Dans S. H. Irvine et P. C. Kyllonen (dir.), *Item generation for test development*. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Even, R. (1990). Subject matter knowledge for teaching and the case of function. *Educational studies in mathematics*, *21*(6), 521-544.

Fischer, G. H. (1995). The Linear logistic test model. Dans G. H. Fischer et I. W. Molenaar (dir.), *Rasch models. Foundations, recent developments, and applications*. New York, New York : Springer-Verlag.

Fortin, P. (2008). Des phénix ou des cancres ? *L'actualité, 33*(13).

Irvine, S. H. (2002). The foundation of item generation for mass testing. Dans S. H. Irvine et P. C. Kyllonen (dir.), *Item generation for test development*. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Irvine, S. H. et Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Lane, S. (1991a). Use of restricted item response models for examining item difficulty ordering and slope uniformity. *Journal of educational measurement, 28*(4), 295-309.

Lane, S. (1991b). Implications of cognitive psychology for measurement and testing : assessing students' knowledge structures. *Educational measurement : Issues and practice, 10*(1), 31-36.

Lobato, J. et Siebert, D. (2002). Quantitative reasoning in a reconceived view of transfer. *Journal of mathematical behavior, 21*(1), 87-116.

MacGregor, M. et Price, E. (1999). An exploration of aspects of language proficiency and algebra learning. *Journal for research in mathematics education*, *30*(4), 449-467.

Mayer, R. E., Larkin, J. H. et Kadane, J. B. (1984). A cognitive analysis of mathematical problem-solving ability. Dans R. J. Sternberg (dir.), *Advances in the psychology of human intelligence, 2*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Mesa, V. (2004). Characterizing practices associated with functions in middle school textbooks : an empirical approach. *Educational studies in mathematics*, *56*(2-3), 255-286.

Moschkovich, J. (1992). *Making sense of linear equations and raphs : An analysis of students' conceptions and language use*. Thèse de doctorat inédite. University of California at Berkeley.

Moschkovich, J., Schoenfeld, A. et Arcavi, A. (1993). Aspects of understanding : on multiple perspectives and representation of linear relations, and connections among them. Dans T. A. Romberg, E. Fennema et T. P. Carpenter (dir.), *Integrating research on the graphical representation of function*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Moschkovich, J. N. (2004). Appropriating mathematical practices : a case study of learning to use and explore functions through interaction with a tutor. *Educational studies in mathematics, 55*(1-3), 49-80.

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

Nadot, S. (1993). Les représentations graphiques des fonctions. *Les Sciences de l'éducation – Pour l'Ère nouvelle, 1*(3), 137-158.

Perrin-Glorian, M.-J. (1994). Contraintes de fonctionnement des enseignants au collège : ce que nous apprend l'étude de « classes faibles ». *Petit x*, *35*, 5-40.

Riopel, M., Raîche, G., Pilote, M. et Potvin, P. (2009). La production automatisée de tâches d'évaluation. Dans J.-G. Blais (dir.), *Éducation et TIC en éducation/formation : enjeux, modèles et applications*. Québec, Québec : Presses de l'Université Laval.

Rojano, T. (2002). Mathematics learning in the junior secondary school : students' access to significant mathematical ideas. Dans L. D. English (dir.), *Handbook of international research in mathematics education*. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Schulz, E. M., Lee, W. C. et Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of educational measurement*, *42*(1), 1-26.

Schwarz, J. et Yerushalmy, M. (1992). Getting students to function in and with algebra. Dans G. Harel and E. Dubinsky (dir.), *Learning the concept of function : aspects of epistemology and pedagogy*. Washington, Columbia : Mathematical association of America (MAA).

Sebrechts, M. M., Enright, M., Bennett, R. E. et Martin, K. (1996). Using algebra word-problems to assess quantitative ability : attributes, strategies, and errors. *Cognition and instruction*, *14*(3), 285-343.

Sfard, A. (1992). Operational origins of mathematical objects and the quandary of reification – the case of function. Dans G. Harel and E. Dubinsky (dir.), *Learning the concept of function : aspects of epistemology and pedagogy*. Washington, Columbia : Mathematical association of America (MAA).

Sheehan, K. et Mislevy, J. (1994). *A tree-based analysis of items from an assessment of basic mathematics skills*. Educational testing service, Princeton, New Jersey. Reports – Evaluative/feasibility (142).

Stump, S. L. (2001). Developing preservice teachers' pedagogical content knowledge of slope. *The journal of mathematical behavior*, *20*(2), 207-227.

Vergnaud, G., Cortes, A. et Favre-Artigue, P. (1988). *Introduction de l'algèbre auprès de débutants faibles. Problèmes épistémologiques et didactiques*. Actes du colloque de Sévres, France : Didactique et acquisition des concepts scientifiques.

Yerushalmy, M. (2000). Problem solving strategies and mathematical resources : a longitudinal view on problem solving in a functional based approach to algebra. *Educational Studies in Mathematics, 43,* 125-147.

Yerushalmy, M. et Chazan D. (2002). Algebra : curricular, graphing, and research. Flux in school algebra : graphing technology, and research on student learning and teacher knowledge. Dans L. D. English (dir.), *Handbook of international research in mathematics education*. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Chapitre 5

Identification des patrons de réponses inappropriés à un test à partir des stratégies qui sous-tendent les comportements des répondants

Patricia Brassard, Sébastien Béland et Gilles Raîche

Nous déterminerons les stratégies qui sous-tendent les comportements des étudiants qui tentent de sous-performer à un test. Une analyse de protocole de rapports écrits nous permettra de catégoriser ces stratégies. La concordance des patrons de réponses avec ces catégories sera vérifiée par une régression logistique pour données nominales ainsi qu'une validation croisée. Nous identifierons des catégories de comportements réels associés à des patrons de réponses inappropriés. Des modèles informatisés d'étudiants qui tentent de se sous-classer intentionnellement, plus fiables quant à la validité de l'interprétation de la mesure, pourront alors être créés.

Le phénomène du surclassement intentionnel est bien connu en milieu scolaire. En effet, diverses stratégies sont mises en place afin d'enrayer le plagiat ou le copiage. Nous rencontrons aussi des situations de sous-classement intentionnel, quoique moins documentées, dans la littérature spécialisée, et elles constituent un problème majeur pour certains. En effet, nous avons remarqué que des étudiants tentent de se sous-classer intentionnellement au test de classement en anglais langue seconde, le TCALS II. Le sous-classement intentionnel à une évaluation est donc présent en milieu scolaire (Dodeen, 2003 ; Fournier, 1992 ; Meijer, 1998 ; Raîche, 2002), mais on le trouve également ailleurs. Nous le

relevons, entre autres, lors d'évaluations professionnelles pour l'obtention d'un emploi (Meijer, 1998 ; Zickar et Drasgow, 1996, p. 71), lors d'évaluations psychologiques (Zickar, Drasgow, 1996, p. 71) et lors d'évaluations juridiques (Bénézech, 2007, p. 361-362). Ce constat conduit à s'interroger davantage sur cette réalité afin de la combattre.

Plusieurs auteurs se sont penchés sur la problématique des patrons de réponses inappropriés (Dodeen, 2003 ; Fournier, 1992 ; Hendrawan, Glas et Meijer, 2005 ; Karabatsos, 2003 ; Levine et Rubin, 1979 ; Meijer, 1996, 1998 ; Meijer et Sijtsma, 1995 ; Mischel, 2004 ; Nering et Meijer, 1998 ; Raîche, 2002 ; Seol, 1998 ; Sijtsma et Meijer, 2001 ; Van der Flier, 1982). On a donc ressenti le besoin d'identifier ces patrons de réponses. De fait, plusieurs approches sont proposées. En effet, Meijer et Sijtsma (1995), Levine et Rubin (1979), Van der Flier (1982), Dodeen (2003), Karabatsos (2003), Reise et Flannery (1996) proposent une approche statistique (*person fit*). D'autres préconisent plutôt une approche qui étudie les courbes de réponse des personnes (*person response function*) (Nering et Meijer, 1998 ; Sijtsma et Meijer, 2001). Mischel (2004), quant à lui, favorise davantage une approche qui tient compte de l'impact du contexte situationnel. Certains (Bertrand et Blais, 2004 ; Laveault et Grégoire, 2002 ; Mislevy et Verhelst, 1990 ; Raîche, 2002) se sont plutôt intéressés au contexte psychométrique qui repose sur la théorie de réponse à l'item. On y trouve une réalité dichotomique, où une bonne réponse est identifiée par 1 et une mauvaise réponse par 0. On y propose divers indices de détection de patrons de réponses inappropriés qui permettent une modélisation informatisée (Karabatsos, 2003 ; Meijer, 1998 ; Meijer et Sijstma, 1995). Toutefois, ces indices génériques ne sont pas propres à la sous-performance intentionnelle. Nous remarquons également que ces approches mettent en évidence des biais d'items, où l'accent est mis sur la construction de l'item. Il y a aussi des biais de personne, dont la cause est attribuée aux comportements du répondant (Bertrand et Blais, 2004, p. 279-313 ; Seol, 1998, p. 2). Par conséquent, devant la variété de ces indices, des chercheurs effectuent présentement des travaux afin d'évaluer l'efficacité de ces indices de détection en fonction de différents types de comportements qui peuvent créer des patrons de réponses inappropriés. Des simulations informatisées de patrons de réponses inappropriés sont généralement effectuées à partir de comportements qui ne sont pas pragmatiques. L'analyse de la méthodologie de ces recherches nous permet d'affirmer que ces patrons de réponses reflètent mal la stratégie mise de l'avant par un répondant pour tenter de tricher à un test.

D'ailleurs, comme le signalent quelques auteurs tels que Raîche (2002), la simulation d'un patron de réponses qui consiste à répondre au hasard n'est pas représentative du processus de réflexion du

répondant et de son comportement qui sont beaucoup plus complexes que cela. Dans la recension d'écrits sur le sujet, on ne trouve aucun cas où une validation des comportements générateurs de patrons de réponses inappropriés est effectuée directement auprès des personnes qui tentent de déjouer le test. La complexité des facteurs, dont les comportements qui déterminent le patron de réponses inapproprié, explique la nécessité d'une investigation afin de recueillir plus d'informations sur les répondants et les items concernés. Il s'avère alors utile d'interroger les répondants désireux d'être sous-classés (Meijer, 1998, p. 157-158).

Compte tenu de ces intérêts pratiques et scientifiques, un aspect déterminant de l'identification des comportements réels et des patrons de réponses correspondants n'a pas encore été étudié et devient la pierre angulaire de notre recherche. Mais comment identifier un patron de réponses inapproprié à partir des comportements et des stratégies du répondant qui tente de se sous-classer intentionnellement ?

Nous allons d'abord délimiter le cadre théorique de notre recherche. Nous allons y définir les termes spécifiques de notre problématique : le patron de réponses, le processus question-réponse, les catégorisations de patrons de réponses inappropriés, le comportement et la stratégie. Nous y identifierons les types de patrons de réponses inappropriés. Puis, nous analyserons quelques recherches et leur méthodologie, telles que celles de Cronbach (1946), Johnson (1998), Meijer (1996), Mislevy et Verhelst (1990), Raîche (2002) ainsi que Baumgartner, Steenkamp et Jan-Benedict (2001), afin de détecter les patrons de réponses inappropriés. Nous déterminerons ensuite notre propre méthodologie avant de présenter notre analyse des résultats. Enfin, nous mettrons en perspective nos résultats par rapport aux recherches antérieures avant de formuler notre conclusion.

1. CONTEXTE THÉORIQUE

1.1. Opérationnalisation des termes de la recherche

Le patron de réponses est un élément déterminant de notre recherche. Avant de situer ce terme à l'intérieur de notre contexte de recherche, nous allons le définir en regard d'autres auteurs et de définitions de concepts sous-jacents au patron de réponses inapproprié. Ces concepts psychométriques sont d'abord liés aux grands courants qui ont influencé les méthodes de détection de patrons de réponses inappropriés : les modèles de mesure dans le cadre de la théorie classique et les

modèles en théorie de la réponse aux items (Bertrand et Blais, 2004). Il existe d'autres types de modèles présentés à l'intérieur de l'ouvrage de Bertrand et Blais (2004) ainsi que celui de Laveault et Grégoire (2002).

Thurstone (1929) a été le premier à montrer l'intérêt d'identifier des patrons de réponses inappropriés occasionnés par un manque d'intérêt ou la nonchalance de la part du répondant. Il suggère de relever des caractéristiques précises de patrons de réponses inappropriés afin de les identifier et de les éliminer dans la mesure du possible. Puis, Cannell, Miller et Oksenberg (1981, p. 393) ont établi un diagramme qui décrit le processus question-réponse des répondants selon sept étapes. Nous nous intéressons particulièrement aux facteurs qui peuvent influencer le processus de prise de décision de la réponse donnée et aux biais qui s'introduisent dans ce processus. Ces deux étapes du processus question-réponse du répondant peuvent toutes deux donner des patrons de réponses inappropriés. Ensuite, Smith (1982, p. 126) précise que ces patrons de réponses inappropriés peuvent être classés en deux catégories. D'abord, il y a les caractéristiques propres à la personne qui fournit des patrons de réponses inappropriés qui sont indépendants de l'item. Puis, il y a les patrons de réponses inappropriés qui peuvent résulter d'une interaction entre la personne et l'item. Il distingue des stratégies et des comportements propres à chacune de ces sources d'erreurs. Les auteurs soulignent que cette catégorisation n'est pas exclusive et que d'autres catégories peuvent exister. Enfin, Ro (2001, p. 6-7) exprime une vision plus large que Smith (1982). Il procède à une recension d'écrits d'auteurs qui ont procédé à une catégorisation des stratégies, des comportements et des facteurs qui créent des patrons de réponses inappropriés, en cinq catégories : 1) utilisation d'une stratégie par le répondant (*sources by test-taking strategies*) ; 2) comportement des répondants (*sources by examinee behavior*) ; 3) facteurs culturels du répondant (*sources by examinee external conditions*) ; 4) facteurs internes du répondant (*sources by exminee internal conditions*) et 5) déficience au sein du test (*sources by faultyness on the test or scoring*). Nous avons retenu cette catégorisation, car elle semble être l'une des plus exhaustives. Elle permet d'avoir une vue d'ensemble.

Le comportement et la stratégie sont deux autres éléments déterminants de notre recherche. Le sujet soumis à un test y répond. Pour ce faire, il écrit sur sa feuille- réponse. Ceci illustre son comportement. Afin de poser un geste, le répondant a réfléchi. Il a mis de l'avant certaines stratégies. Selon Rey et Rey-Debove (1991, p. 351), un comportement est l'ensemble des réactions objectivement observables. Legendre (2005, p. 259) spécifie qu'en psychologie humaniste, un comportement est l'activité d'une personne qui se divise en deux caté-

gories : l'aspect externe et l'aspect interne. Le premier est observable et mesurable. Le second est accessible par introspection. De Landsheere (1992) ajoute que cette dernière catégorie d'entités non observables comprend la conscience, les attitudes et les besoins. Nous ajoutons à cette liste les stratégies cognitives telles que le comportement intériorisé. En effet, selon Legendre (2005, p. 1261), une stratégie cognitive est une technique ou une procédure intellectuelle qu'une personne juge propice à la résolution d'un problème. De plus, ce procédé est mis en œuvre alors qu'une personne fait intervenir un processus métacognitif. En effet, elle doit connaître les manières de faire pour accomplir la tâche, ainsi que d'autres paramètres tels que ses ressources personnelles, la nature et les objectifs de la tâche. De même, afin d'activer une stratégie pour atteindre un but spécifique, les principaux facteurs de la situation doivent être connus de la personne. Un autre lien entre la métacognition et les stratégies cognitives est déterminé par l'aboutissement des deux activités. À la fin de la mise en action de son processus métacognitif, l'individu fait face à une réussite ou à un échec. Il procède à la validation de son jugement par l'analyse de la rétroaction externe et interne présente pendant et à la fin du processus. Cela s'apparente au résultat de la mise en place d'une stratégie. Pour atteindre un but précis, l'individu choisit des opérations ingénieuses et ordonnées afin de favoriser, au meilleur de ses connaissances, sa réussite.

La catégorisation de ces comportements qui donnent des patrons de réponses inappropriés nous semble tributaire du jugement qui guide le chercheur tout au long de son processus de recherche. Meijer (1996, p. 7) souligne que même si un patron de réponses est reconnu inapproprié chez un sujet, les chercheurs ne peuvent être absolument certains du type de biais correspondant. En effet, les diverses formes de comportements qui donnent des patrons de réponses inappropriés peuvent aboutir à la même forme de patron de réponses à l'item. La distinction n'est donc pas évidente. Enfin, jusqu'à maintenant, les méthodes de mesure appropriées sont confrontées à ce problème de réponses inappropriées au test. Des recherches récentes tentent d'appliquer des théories psychologiques au processus de détection de patrons de réponses inappropriés (Meijer, 1996, p. 8).

Nous aborderons ces concepts en suivant certaines lignes directrices. D'abord, nous nous intéresserons plus particulièrement à la théorie de réponse à l'item et à ses propriétés liées aux habiletés et aux paramètres d'item. Nous allons examiner le modèle logistique à un paramètre qui ne tient compte que du paramètre de difficulté de l'item. Cette modélisation nous permettra de ne pas tenir compte des autres paramètres difficilement contrôlables que sont les paramètres de discrimination et de pseudo-chance (Raîche, 2002, p. 59). Puis, nous

étudierons le comportement observable et non observable. Le comportement observable est l'ensemble des actions qu'un étudiant met en place pour faire ses choix de rédaction. Cela comprend ce qu'il écrit et où il l'écrit. Le comportement non observable est lui aussi subdivisé en sous-ensembles. Parmi ceux-ci il y a les stratégies, spécifiquement les stratégies cognitives, qui font partie du processus métacognitif. L'atteinte du but, dans un cas de désir de sous-classement intentionnel à un test se manifeste par l'échec au test. La notion de réussite est ici relative au contexte, mais fait appel aux mêmes procédés. En somme, un patron de réponses inapproprié est identifiable lorsque le patron de réponses est improbable étant donné les caractéristiques du modèle de la théorie de réponse à l'item. Au sens psychométrique, un patron de réponses est inapproprié lorsque l'étudiant ne répond pas en faisant appel à toute son habileté.

1.2. Types de patrons de réponses inappropriés

On trouve divers patrons de réponses inappropriés dans la littérature spécialisée. Nous avons déjà vu que la catégorisation de patrons de réponses était tributaire de la théorie qui guide le chercheur. Un des éléments qui encadre cette analyse est le contexte dans lequel se trouve le patron de réponses inapproprié. Nous identifions trois contextes différents possibles. Il y a le contexte d'incompréhension, de copiage ou de plagiat, et finalement le contexte de sous-performance intentionnelle. Nous identifions des types de comportements et de stratégies qui donnent des patrons de réponses inappropriés dans chacun de ces contextes.

En premier lieu, des chercheurs se sont intéressés aux patrons de réponses inappropriés provoqués par l'incompréhension, une incapacité ou une difficulté. Le sujet soumis au test peut alors adopter diverses stratégies. Nous trouvons, par exemple, la stratégie qui consiste à faire semblant (*faking*) (Alliger et Dwight, 2000 ; Meijer, 1996 ; Peeters et Lievens, 2005 ; Reise et Waller, 1993 ; Van der Flier, 1982 ; Zickar et Drasgow, 1996). Alliger et Dwight (2000, p. 69) indiquent que, dans le cadre de tests qui mesurent des habiletés qui ne sont pas cognitives, il est possible que des sujets prétendent être ce qu'ils ne sont pas de manière à redorer leur image (*faking good*). Meijer (1996) relève un autre comportement : le comportement de la personne méthodique et lente *(plodding behavior)* qui désire faire de son mieux en prenant le temps nécessaire pour répondre à chaque item, mais qui ne possède pas la capacité intellectuelle pour répondre correctement à tous les items. Elle n'est donc pas efficace dans l'espace de temps dont elle dispose.

Ensuite, le contexte de copiage ou de plagiat se définit par l'intention d'un étudiant qui n'a pas les aptitudes requises pour réussir le test à faire appel à de l'aide extérieure afin de surperformer. Plusieurs auteurs se sont intéressés à cette réalité (Brezina, 2000 ; Cizek, 1999, 2003 ; Hutton, 2006 ; Kuehn, Stanwyck et Holland, 1990 ; Levine et Drasgow, 1988 ; Meijer, 1996 ; Meijer et Sijtsma, 1995 ; Nathanson, Paulhus et Williams, 2006 ; Van der Ark, Emons et Sijtsma, 2008). Cizek (1999, p. 1-2) définit cette réalité propre au contexte scolaire, comme toute action qui viole les règlements de l'établissement scolaire qui ont trait à l'administration d'un test ; tout comportement qui octroie un privilège à un étudiant auquel les autres n'ont pas droit ; toute action qui diminue la valeur du jugement porté sur la performance de l'étudiant au test. Nathanson, Paulhus et Williams (2006, p. 113 et 116) ont tenté de relever des comportements qui peuvent être des indicateurs de tricherie. Ils indiquent que les différences de personnalité et le bagage de connaissances jouent un rôle dans la tricherie. Ces personnes ne montrent pas d'empathie pour l'ardeur au travail préparatoire manifestée par les autres étudiants pour réussir le test.

Enfin, des auteurs examinent le contexte de sous-performance intentionnelle (Berkowitz, Cicchelli, 2004 ; Fournier, 1992 ; Hendrawan, Glas et Meijer, 2005 ; McCoach et DelSiegle, 2003 ; Nurmi, Onatsu et Haavisto, 1995 ; Raîche, 2002). Comme le soulèvent Raîche (2002, p. 2) et Fournier (1992, p. 21), soumis à un test qui vérifie les habiletés langagières en anglais, langue seconde, certains étudiants parviennent volontairement à un résultat inférieur à ce qu'ils sont capables d'obtenir. Ces élèves désireraient être classés dans un cours facile afin d'obtenir de meilleurs résultats avec peu d'effort. Nous avons trouvé peu d'écrits scientifiques traitant de ce sujet. Toutefois, nous relevons un sujet connexe qui est davantage étudié : les surdoués qui sous-performent *(gifted underachievers)*. McCoach et Siegle (2003, p. 150-152) ont déterminé les caractéristiques propres aux étudiants surdoués et à ceux qui en plus d'être surdoués sous-performent. De manière plus générale, Hendrawan, Glas et Meijer (2005, p. 32) soulignent l'impact qu'a la stratégie de réponse au hasard chez un étudiant qui possède une grande capacité intellectuelle. Ces derniers répondent incorrectement à des items faciles. Ces étudiants obtiendront alors un score inférieur à ce qu'ils seraient en mesure d'obtenir. L'aptitude évaluée sera jugée inférieure au niveau réel.

Nous nous intéressons spécifiquement au contexte de sous-performance intentionnelle et nous définissons un étudiant qui fait preuve de sous-performance intentionnelle alors qu'il livre un patron de réponses inapproprié de manière volontaire par le biais d'une stratégie expressément choisie.

1.3. Méthodologie des recherches qui traitent du patron de réponses inapproprié

La recherche de Cronbach (1946) identifie des catégories de facteurs étrangers au test qui peuvent influencer le résultat du répondant. Il identifie six catégories qui sont le produit de comportements qui témoignent d'une tendance qu'entretient une personne afin de fournir un patron de réponses différent de celui qu'il aurait fourni si le même contenu avait été présenté sous une autre forme. Il ajoute que certaines différences individuelles de patron de réponses peuvent être reliées entre elles. Ces catégories peuvent être tributaires de l'item, de la situation ou du répondant. Toutefois, il importe de connaître la nature et l'origine des différences individuelles quant à la manière de répondre aux items qui donnent des patrons de réponses inappropriés.

La recherche de Mislevy et Verhelst (1990) propose un modèle qui tient compte de l'utilisation de plusieurs stratégies par le répondant tout au long du test afin d'analyser les stratégies mises de l'avant en éducation et en psychologie. Ce modèle s'accorde avec la théorie de réponse à l'item. Ils suggèrent que le choix d'une stratégie n'est pas directement observé, mais peut être inféré d'après les patrons de réponses et le modèle d'analyse choisi. Il serait alors possible d'évaluer la compétence du sujet en fonction des diverses stratégies possibles pour répondre aux items. Ils déterminent trois catégories de stratégies qui sont utilisées selon la disposition des items d'après leur niveau de difficulté. Pour identifier ces stratégies, ils insistent sur les paramètres de l'item et leurs caractéristiques qui peuvent conduire le répondant à choisir une stratégie. Il serait peut-être intéressant d'utiliser cette méthodologie afin d'étudier les paramètres liés au répondant de manière plus générale. Les auteurs mettent l'accent sur la stratégie de réponse au hasard. Ne serait-il pas intéressant d'extrapoler cette méthodologie à un autre type de patron de réponse ?

La recherche menée par Meijer (1996) propose une revue de la littérature spécialisée afin de déterminer des catégories de comportements qui donnent des patrons de réponses inappropriés. Il identifie sept catégories. Il propose une interprétation possible de patron de réponses associé à chaque catégorie en regard de l'erreur de Guttman. Il procède à cette analyse dans un contexte fictif. Les choix de Meijer sont totalement intuitifs, basés sur des théories antérieures. Il est difficile d'associer un patron de réponses, grâce à ses caractéristiques propres, à un comportement donné.

La recherche de Johnson (1998) propose une taxonomie des facteurs qui relèvent de la psychologie affective. Ces facteurs peuvent donner des patrons de réponses inappropriés. Afin de valider cette

taxonomie, il s'appuie non seulement sur une revue de la littérature scientifique sur le sujet, mais aussi sur un comité d'experts. Ce dernier évalue la taxonomie une fois constituée. Il est intéressant que Jonhson ait utilisé une méthodologie qui mette à contribution l'avis d'un comité d'experts de manière à valider sa taxonomie. Cela a permis certains ajustements ou questionnements. Entre autres, l'auteure a été amenée à se questionner sur la catégorisation de cette taxonomie en regard de l'individu et des situations extérieures. La méthodologie utilisée ici, bien que plus complète que celles des auteurs qui la précèdent grâce à une validation sur le terrain, ne permet pas de valider cette taxonomie par des données réelles.

La recherche de Raîche (2002), quant à elle, s'intéresse au développement d'indices d'ajustement inadéquat qui servent spécifiquement à dépister le sous-classement intentionnel chez des étudiants de niveau collégial au test de classement en anglais langue seconde, le TCALS II. Afin d'y parvenir, le chercheur a interrogé des étudiants pour déterminer des catégories de stratégies qu'ils mettraient de l'avant s'ils désiraient se sous-classer. Il a trouvé sept catégories. N'aurait-il pas été intéressant de distribuer au préalable le test afin d'observer ces stratégies sur un patron de réponses ? Nous nous demandons également comment cette méthode exploratoire qui a amené l'auteur à une catégorisation peut être optimisée. Il serait intéressant de relever ces stratégies chez des sujets qui tentent vraiment de se sous-classer.

La recherche de Baumgartner, Steenkamp et Jean-Benedict (2001) procède à une étude corrélationnelle. Ils s'intéressent aux différents types de réponses des sujets à un test qui peuvent contaminer les résultats, la validité et les conclusions du test. Dans ce contexte, les auteurs s'intéressent à cinq types de réponses. Ils montrent que des patrons inappropriés affectent négativement les conclusions d'un test. À cet effet, ils portent une attention particulière à l'impact des échelles de mesure et à leurs interrelations selon le type de réponse reçue. Cette étude tente de sensibiliser les chercheurs qui effectuent des études de marché à cette réalité. Nous nous demandons si cette étude corrélationnelle ne serait pas encore plus valable si elle s'appuyait sur des catégories déterminées à partir de données réelles.

Dans toutes ces recherches, il semble qu'il y ait un aspect méthodologique qui n'a pas encore été considéré jusqu'à présent afin d'identifier des catégories de stratégies, de comportements ou de facteurs qui peuvent donner un patron de réponses inapproprié : une étude sur le terrain auprès de sujets qui présentent un véritable patron de réponses inapproprié. En effet, il serait intéressant de faire davantage de recherche pour vérifier si cette brèche, inexploitée jusqu'à présent,

ne permettrait pas de mettre plus fortement en rapport les patrons de réponses inappropriés avec la catégorie respective de leur source. De fait, il importe de préciser cette source par des conditions clairement établies chez le sujet. D'autant plus que nous n'avons relevé qu'une étude qui traite véritablement du phénomène de sous-classement intentionnel. Nous voulons donc identifier les stratégies utilisées par les étudiants pour sous-performer intentionnellement à un test.

2. MÉTHODOLOGIE

Nous procéderons selon deux phases complémentaires. La première phase est exploratoire. Une analyse qualitative des réponses des étudiants sera faite. Des catégories seront déterminées par émergence (Ericsson et Simon, 1993 ; Paillé et Mucchielli, 2003). Nous centrerons notre attention sur les tâches élémentaires d'extraction des unités des courts textes, de repérage des mots signifiants et de formation de classes par le regroupement sémantique, comme c'est le cas dans l'ensemble des procédures spécifiques propres à l'instrumentation en analyse de textes (Landry, Bhanji-Pitman et Auger, 2005, p. 70). La deuxième phase de notre analyse est confirmatoire. Nous procéderons alors à une régression logistique pour données nominales, puis à une validation croisée. Cette méthode nous permettra de détecter les biais de notre analyse exploratoire.

2.1. Sujets

Pour procéder à cette recherche, nous avons bénéficié de la collaboration volontaire de 151 répondants. Deux d'entre eux ont remis un patron de réponses qui n'a pas été pris en compte, car ils ont été égarés. Au total, nous disposons de 149 patrons de réponses pour procéder à notre analyse. Quatre-vingt-six d'entre eux proviennent de cinq groupes distincts du cégep du Vieux-Montréal et les 63 autres proviennent de sept groupes distincts du cégep André-Laurendeau. Tous ces étudiants sont en première année. Les étudiants du cégep du Vieux-Montréal ont été classés au niveau 3 en anglais, langue seconde. Ils sont alors presque tous bilingues (français-anglais). Les étudiants du cégep André-Laurendeau ont été classés au niveau 2. Ils sont de niveau intermédiaire en anglais, langue seconde. Tous les étudiants sont inscrits à des programmes différents et suivent le même cours d'anglais, langue seconde, obligatoire. Enfin, tous les étudiants ont déjà été soumis au TCALS II lors de leur classement, en anglais, langue seconde.

2.2. Instrumentation

Deux instruments ont été utilisés lors de la cueillette des données : le TCALS II et un questionnaire composé d'un item à réponse élaborée. Le TCALS II est constitué de sept sections : les phrases, les dialogues, le miniexposé, le vocabulaire, la grammaire, l'analyse d'erreurs et la lecture. Pour le remplir, les répondants disposent de 90 minutes. Les étudiants répondent à des items à choix multiples. Le TCALS II est constitué d'items dont le niveau de difficulté est relativement faible. Malgré cela, il semble demeurer l'instrument adéquat afin d'évaluer le niveau d'habileté d'une étudiante ou d'un étudiant en anglais, langue seconde (Raîche, 2002, p. 78). Le questionnaire, pour sa part, était composé de l'item suivant : *décrivez la stratégie que vous avez tenté d'utiliser pour vous sous-classer au test.* Ils ont disposé de 15 minutes pour répondre.

2.3. Déroulement

Les étudiants des deux cégeps ont été soumis au TCALS II en deux temps : en mai 2006 au cégep du Vieux-Montréal et en septembre ainsi qu'en octobre 2006 au cégep André-Laurendeau. Trois professeurs d'anglais, langue seconde ont procédé à l'administration de ce test. Les étudiants ont été interviewés par les chercheurs lors d'un cours d'anglais régulier,en l'absence du titulaire. L'assistante de recherche a expliqué le contexte ainsi que les objectifs de cette recherche et comment se ferait la cueillette des données. Les étudiants ont dû répondre au TCALS II et tenter de s'y sous-classer. Ils ont ensuite répondu au questionnaire. Afin d'être certaine que les étudiants comprennent bien la tâche qui leur était assignée, l'assistante de recherche a donné un exemple de stratégie possible : répondre au hasard à toutes les questions du test. Aucune autre piste de réponse n'a été proposée afin d'éviter que les répondants reproduisent ces dernières dans leurs propres réponses. Les répondants ont disposé du temps nécessaire pour répondre à cette question. Enfin, l'assistante de recherche a répondu aux questions des étudiants, à savoir quelles étaient les stratégies déjà identifiées par les chercheurs.

2.4. Méthode d'analyse des données

Dans un premier temps, pour une étude exploratoire, les protocoles des répondants sont analysés de manière à faire ressortir des catégories de stratégies utilisées par les étudiants afin de se sous-classer intentionnellement au test. Le sens attribué à chaque catégorie est déterminé par les réponses des étudiants au questionnaire. Pour ce faire, on surligne

les descripteurs de chaque catégorie. Ensuite, on regroupe les stratégies qui ont un sens commun. Puis, à l'aide d'une grille d'analyse qualitative, nous identifions les stratégies utilisées par chaque étudiant. Nous nous demandons à nouveau si la première analyse qualitative était juste. Nous relevons ensuite la catégorie la plus fréquente pour chacun. Enfin, nous tentons de relever des comportements qui pourraient être observables quant au patron de réponses de chaque catégorie de stratégies.

Dans un second temps, afin de vérifier les résultats obtenus, nous tentons de prédire l'appartenance des patrons de réponses à chacune des catégories par une régression logistique et une validation croisée à l'aide du logiciel R. Nous calculons le pourcentage de cas bien classés par la régression logistique, la fréquence de chaque stratégie, la fréquence de classement des diverses stratégies, le pourcentage du classement des diverses stratégies, le nombre d'observations utilisées pour la validation croisée, et enfin le pourcentage de cas bien classés par la régression logistique en validation croisée.

2.5. Considérations éthiques

L'assistante de recherche a préconisé une approche participative auprès des étudiants lors de sa présentation. Les participants ont été soumis au TCALS II seulement sur une base volontaire. Ceux qui ont refusé de se plier au test ont quitté le local. La confidentialité a été assurée. Les objectifs et le protocole de recherche ont été présentés intégralement. L'assistante de recherche a fait signer un formulaire de consentement à chaque participant. Elle leur a remis une copie dudit formulaire. Ce document leur fournissait de l'information sur la recherche à laquelle il participait, entre autres, les coordonnées du site Web où il est possible de suivre les travaux en cours, l'évolution de la recherche et ses résultats.

3. RÉSULTATS

Dans un premier temps, l'analyse exploratoire nous a permis de déterminer les catégories de stratégies présentées au tableau 5.1, adoptées par les étudiants afin de se sous-classer intentionnellement au test de classement en anglais, langue seconde, le TCALS II. Nous distinguons sept catégories (identifiées de 1 à 7) qui regroupent 21 stratégies (identifiées de A à U). Il est à noter que le répondant 99 n'a pas déclaré la stratégie utilisée afin de se sous-classer intentionnellement au test. La stratégie 2, soit de ne pas répondre, ne sera toutefois pas étudiée ici et sera retirée des analyses.

Tableau 5.1
Catégories de stratégies adoptées par les étudiants afin de se sous-classer intentionnellement

1. Stratégies de hasard
   - A. Choisir une réponse au hasard (bonne ou mauvaise)
   - B. Répondre et illustrer un motif sur la feuille-réponse
   - C. Choisir une réponse d'après une émotion
   - D. Répondre de façon systématique
   - E. Choisir les réponses d'après ses chiffres préférés
2. Stratégie d'absence de réponse
   - F. Ne pas répondre
3. Stratégies uniques
   - G. Choisir la mauvaise réponse
   - H. Choisir la bonne réponse
   - I. Inattention
   - J. Erreur de lecture
4. Stratégies doubles
   - K. Faire des choix en tenant compte du niveau de difficulté de l'item (facile, difficile)
   - L. Association de mots lus ou entendus à l'intérieur des questions et des réponses
   - M. Modifier le sens d'un mot volontairement
   - N. Alterner les bonnes et les mauvaises réponses
   - O. Sélectionner aléatoirement des questions à réussir ou à échouer
5. Stratégie multiple
   - P. Diversifier les stratégies (plus de 3 stratégies utilisées)
6. Stratégies de premier plan avec une intention claire
   - Q. Copier sur un pair/Consulter un pair
   - R. Choisir ses réponses en visant un niveau de maîtrise de la langue
   - S. Tricherie
7. Stratégies de second plan
   - T. Camouflage
   - U. Calculer le nombre de bonnes et mauvaises réponses

Il nous est difficile de dégager une séquence pour chaque catégorie de stratégies dans les patrons de réponses. Nous remarquons que tous les patrons de réponses sont hétérogènes. De plus, il y a rarement des répondants qui omettent de répondre volontairement à certaines questions. Par ailleurs, nous avons tenté d'observer la séquence de la stratégie utilisée dans le patron de réponses. Par exemple, certains étudiants utilisent une stratégie de second plan telle que calculer le nombre de bonnes et mauvaises réponses. Toutefois, il ne nous a pas été possible d'en arriver à des résultats concluants. En effet, il n'y a que 33 répondants qui ont utilisé une seule stratégie. C'est trop peu d'étudiants par rapport aux 21 stratégies relevées pour que nous soyons

en mesure d'établir une quelconque séquence du patron de réponses. Il n'est pas indiqué qu'un étudiant qui dit vouloir choisir la mauvaise réponse y parvient. Enfin, lorsque des étudiants utilisent plus d'une stratégie, il peut être difficile de les identifier.

Dans un deuxième temps, nous avons procédé à l'analyse confirmatoire, spécifiquement pour les sept catégories retenues. Nous avons utilisé 50 % des observations pour procéder à la régression logistique, soit 75 observations. Le tableau 5.2 indique la fréquence de chacune des catégories de stratégies. Nous remarquons que la classe 1, stratégies de hasard, est la plus fréquente avec 69 observations. À l'opposé, la catégorie 5, stratégie multiple, n'est relevée qu'une seule fois.

Tableau 5.2
Fréquence de chacune des sept catégories de stratégies

Catégorie	1	3	4	5	6	7
Pourcentage	0,47	0,25	0,14	0,01	0,08	0,05

Note : La stratégie 2 (ne pas répondre) ne figure pas dans l'analyse.

Pour effectuer la validation croisée, 50 % des observations ont été utilisées, soit 75 observations. Le tableau 5.3 indique la fréquence du classement de catégories de stratégies par la régression logistique. Nous observons que la catégorie 1, stratégies de hasard, est bien classée dans 18 cas et que la catégorie 4, stratégies doubles, est bien classée dans 5 cas. Ces dernières sont les deux catégories qui ont été les mieux classées par la régression logistique.

Tableau 5.3
Fréquence du classement de catégories de stratégies par la régression logistique

Catégories	1	3	4	6	7
1	18	8	5	2	5
3	8	4	4	1	1
4	3	2	5	0	0
5	0	1	0	0	0
6	2	1	1	0	0
7	2	0	0	1	0

Note : La stratégie 2 (ne pas répondre) ne figure pas dans l'analyse.

Le pourcentage des cas bien classés par la régression logistique en validation croisée est de 36,5 %. Le tableau 5.4 précise le pourcentage du classement des diverses stratégies. Les données indiquent à quel point le patron de réponses prédit l'appartenance à une des

catégories de stratégies. Nous remarquons que la catégorie 1, stratégies de hasard, est clairement identifiable à 47 % des cas. De même, la catégorie 4, stratégies doubles, est bien identifiée à 50 % des cas. Nous pouvons affirmer que les patrons de réponses appartiennent surtout à ces deux catégories. À l'opposé, il est intéressant de remarquer que le classement de la catégorie 3, stratégies uniques, correspond avec 44 % à la catégorie 1, stratégies de hasard. De plus, la catégorie 5, stratégies multiples, correspond à 100 % des cas à la catégorie 3, stratégies uniques. De même, la catégorie 6, stratégies de premier plan avec une intention claire, et la catégorie 7, stratégies de second plan, ont un classement respectif qui correspond à la catégorie 1, choisir une réponse au hasard, avec 50 % et 67 %. De manière générale, la validation croisée nous indique qu'il est très difficile de prédire correctement la stratégie que l'étudiant aurait utilisée.

Tableau 5.4
Proportion du classement des sept catégories de stratégies par la régression logistique en validation croisée

Catégories	1	3	4	6	7
1	0,47	0,21	0,13	0,05	0,13
3	0,44	0,22	0,50	0,00	0,00
4	0,30	0,20	0,50	0,00	0,00
5	0,00	1,00	0,00	0,00	0,00
6	0,50	0,25	0,25	0,00	0,00
7	0,67	0,00	0,00	0,33	0,00

Note : La stratégie 2 (ne pas répondre) ne figure pas dans l'analyse.

4. DISCUSSION

Nous avons tenté de déterminer la nature et l'origine des différences individuelles qui poussent une personne à donner un patron de réponses inapproprié tel que le décrit Cronbach (1946). Nous nous sommes intéressés spécifiquement aux stratégies choisies par le répondant pour se sous-classer intentionnellement. Toutefois, comme Meijer (1996, p. 7), nous constatons que, même si un patron de réponses est désigné comme étant inapproprié chez un sujet, nous ne pouvons être absolument assurés du type de biais correspondant. En effet, chaque comportement peut donner un patron de réponses inapproprié qui présente des similitudes avec d'autres. D'ailleurs, il est rare qu'un répondant n'utilise qu'une seule stratégie de sous-classement intentionnel.

Lors de notre analyse exploratoire, nous avons été d'abord étonnés de remarquer que, dans le sous-classement intentionnel, des étudiants ont utilisé une stratégie propre à d'autres types de patron de réponses inapproprié. En effet, nous avons relevé le copiage comme une stratégie de sous-classement intentionnel. Des étudiants affirment, par exemple, copier sur un étudiant qu'ils estiment plus faible qu'eux. Ces résultats correspondent davantage à la définition de Cizek (1999, p. 1-2) comme étant toute action qui diminue la qualité du jugement issu de la performance de l'étudiant au test. De même, nous trouvons des étudiants qui adoptent la stratégie de tromperie en se mettant dans la peau de quelqu'un de plus faible qu'eux. Cette stratégie semble correspondre à celle qui consiste à faire semblant (*faking*) (Alliger, Dwight, 2000 ; Ford, 1996 ; Meijer, 1996 ; Peeters et Lievens, 2005 ; Reise et Waller, 1993 ; Van der Flier, 1982 ; Zickar et Drasgow, 1996). Cette stratégie était adoptée jusqu'alors par des répondants confrontés à l'incompréhension, à une incapacité ou à une difficulté (Alliger et Dwight, 2000, p. 69). Ensuite, nous avons tenté de pallier certaines lacunes des recherches de Meijer (1996), de Johnson (1998) et de Raîche (2002). En effet, nous avons proposé une taxonomie des stratégies qui donnent des patrons de réponses inappropriés à partir de réponses réelles alors que des étudiants étaient amenés à se sous-classer intentionnellement. Comme Johnson (1998), nous nous sommes intéressés aux écrits scientifiques sur le sujet. Puis, tout comme Mislevy et Verhelst (1990) l'indiquent, nous constatons que le choix d'une stratégie n'est pas directement observé. En effet, il nous a été difficile de trouver une séquence du patron de réponses directement en rapport avec la stratégie choisie par le répondant.

Lors de l'analyse confirmatoire, comme Mislevy et Verhelst (1990), nous avons tenté d'inférer le choix de la stratégie par le répondant des patrons de réponses et du modèle d'analyse confirmatoire choisi, soit la régression logistique pour données nominales suivie d'une validation croisée. Nous constatons que deux catégories de stratégies semblent être plus faciles à reconnaître : les stratégies de hasard et les stratégies doubles.

CONCLUSION

Nous avons essayé de repérer les répondants qui tentent de se sous-classer intentionnellement au test de classement en anglais, langue seconde au collégial, le TCALS II. Pour ce faire, nous avons spécifiquement porté notre analyse sur les stratégies qui sous-tendent le comportement observable sur le patron de réponses des répondants qui désirent se sous-classer intentionnellement à un test. À la lumière

de la littérature spécialisée, une étude sur le terrain s'est imposée afin d'identifier des catégories de stratégies réelles associées à des patrons de réponses inappropriés. Des étudiants de cégep ont passé le TCALS II avec la volonté de se sous-classer intentionnellement. Ils ont ensuite répondu à un questionnaire qui nous permettait de procéder à l'analyse de protocoles de rapports écrits. L'analyse des résultats, d'abord exploratoire, nous a permis de classer ces 21 stratégies en sept catégories. Nous avons aussi examiné les patrons de réponses. Ensuite, une analyse confirmatoire, par une régression logistique pour des données nominales et une validation croisée, nous a permis de constater la très faible concordance des patrons de réponses avec les catégories 1, stratégies de hasard, et 4, stratégies doubles. Plusieurs autres patrons de réponses correspondent à la stratégie 1, stratégies de hasard, par la nature de leur patron de réponses. Ainsi, la prédiction de la stratégie utilisée à partir de l'analyse du patron de réponses s'est avérée peu efficace.

Notre recherche comporte toutefois ses limites :

1. La petite taille de notre échantillon, 149 sujets, impose des limites à notre analyse et à nos conclusions. En effet, le grand nombre de stratégies relevées par rapport aux nombres d'observations pour chacune d'elle ne nous permet pas de généraliser nos résultats, ni d'obtenir des résultats vraiment concluants.
2. Il était difficile d'analyser la stratégie adoptée dans le patron de réponses correspondant compte tenu de la variété des stratégies mises de l'avant par plusieurs répondants.
3. L'analyse exploratoire s'est faite, en grande partie, à partir de données qualitatives. Le jugement des chercheurs a certainement influencé les choix des stratégies et des catégories relevées. La terminologie utilisée pour désigner les stratégies et les catégories a pu nuire aussi au classement des stratégies. Il existe peut-être des termes plus justes pour décrire ces dernières.

Compte tenu de toutes ces limites, nous recommandons que, lors de recherches ultérieures, on utilise un échantillon plus représentatif. De plus, il serait intéressant d'analyser les patrons de réponses selon les parties du test, le TCALS II. En effet, plusieurs étudiants ont indiqué avoir utilisé plus d'une stratégie selon la partie spécifique où chaque stratégie a été mise de l'avant. Il serait aussi profitable de procéder à une analyse terminologique des stratégies et des catégories identifiées afin de rendre plus solide l'analyse exploratoire. De plus, nous recommandons qu'une comparaison entre les stratégies relevées dans la littérature spécialisée et celles de la présente recherche soit faite afin d'assurer une plus grande validité à la catégorisation des stratégies.

Enfin, nous croyons qu'il pourrait être avantageux de mettre de l'avant une méthodologie plus proche de celle que Mislevy et Verhelst (1990) ont utilisée. Par exemple, une stratégie en particulier pourrait être étudiée en profondeur.

En définitive, cette recherche sur le sous-classement intentionnel s'ajoute aux autres. Elle nous permet de constater qu'une étude de données réelles est possible et valable afin d'identifier des patrons de réponses inappropriés. Les projets de recherches connexes pourront peut-être, à la lumière de recommandations que nous faisons, permettre à d'autres chercheurs de créer des modèles informatisés d'étudiants qui tentent de se sous-classer intentionnellement, plus fiables quant à la validité de l'interprétation de la mesure.

RÉFÉRENCES

Alliger, G. M., Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and psychological measurement, 60*(1), 59-72.

Baumgartner, H. et Steenkamp, J. B. E. M. (2001). Response styles in marketing research : a cross-national investigation. *Journal of marketing research, 38*(2), 143-156.

Bénézech. M. (2007). Vérité et mensonge : l'évaluation de la crédibilité en psychiatrie légale et en pratique judiciaire. *Annales médico- psychologiques, 165*(5), 351-364

Berkowitz, E. et Cicchelli, T. (2004). Metacognitive strategy use in reading of gifted high achieving and gifted underachieving middle school students in New York City. *Educational and urban society*, 37(1), 37-57.

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

Bouchard, S. et Cyr, C. (dir.) (2006). *Recherche psychosociale : pour harmoniser recherche et pratique*. Sainte-Foy, Québec : Presses de l'Université du Québec.

Brezina, T. (2000). Are deviants different from the rest of us ? Using student accounts of accademic cheating to explore a popular myth. *Teaching sociology, 28*, 71-78.

Cannell, F., Miller, P. V. et Oksenberg, L. (1981). Research on interviewing techniques. *Sociological methodology, 12*, 389-437.

Cizek G. J. (1999). Cheating on tests : how to do it, detect it, and prevent it. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Cronbach, L. J. (1946). Response set and test validity. *Educational and psychological measurement, 6*, 475-94.

De Landsheere, G. (1992). *Dictionnaire de l'évaluation et de la recherche en éducation* (2e édition). Paris, France : Presses universitaires de France.

Dodeen, H. (2003). The use of person-fit statistics to analyse placement tests. Paper presented at the 2003 Annual Meeting of the American Educational Research Association. Chicago, Illinois : AERA.

Ericsson, K. A. et Simon, H. A. (1993). *Protocol analysis. Verbal reports as data* (2e édition). Cambridge, Massachusetts : MIT Press.

Fournier, P. (1992). *Pour un test incontestable : rapport de recherche sur les tests de classement en anglais (langue seconde) au collégial*. Québec, Québec : Ministère de l'Éducation, Direction générale de l'enseignement collégial.

Hendrawan, I., Glas, A. W. et Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied psychological measurement, 29*(1), 26-44.

Hutton, P.A. (2006). Understanding student cheating and what educators can do about it. *College teaching, 54*(1), 171-176.

Johnson, E. M. (1998). *A taxonomy of person misfit on affective measures*. Thèse de doctorat inédite, Denver, Colorado : Université de Denver.

Johnson, J. A. (1981). The « self-disclosure » and « self-presentation » views of item response dynamics and personality scale validity. *Journal of personality and social psychology, 40*, 761-769.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied measurement in education, 16*(4), 277-298.

Kuehn, P., Stanwyck, D. J. et Holland, C. L. (1990). Attitudes toward « Cheating » behaviors in the ESL classroom. *TESOL Quarterly, 24*(2), 313-317.

Landry, N., Bhanji-Pitman, S. et Auger, R. (2005). Comparaison d'un mode de sélection par le chercheur et d'un mode d'extraction automatisée de données textuelles. Actes du colloque 2004 de l'Association pour la recherche qualitative. Trois-Rivières, Canada. Dans C. Royer, J. Moreau et F. Guillemette (dir.), *Recherches qualitatives. L'instrumentation dans la collecte des données : choix et pertinence* (Hors-Série, no 2).

Laveault, D. et Grégoire, J. (2002). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (2e édition). Bruxelles, Belgique : De Boeck.

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

Levine, M. V. et Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*(2), 161-176.

Levine, M. V. et Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of educational statistics, 4*(4), 269-290.

McCoach, D. B. et Siegle, D. (2003). Factors that differentiate underachieving gifted students from high-achieving gifted students. *Gifted child quarterly, 47*(2), 144-154.

Meijer, R. R. (1996). Person-fit research : an introduction. *Applied measurement in education*, *9*(1), 3-8.

Meijer, R. R. (1998). Consistency of test bahaviour and individual difference in precision of prediction. *Journal of occupational and organizational psychology, 71*(2), 147-160.

Meijer, R. R. et Sijtsma, K. (1995). Detection of aberrant item score patterns : a review of recent developments. *Applied measurement in education, 8*(3), 261-272.

Mischel, W. (2004). Toward an integrative science of the person. *Annual review of psychology, 55*, 1-22.

Mislevy, R. J. et Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195-215.

Nathanson, C., Paulhus, D. L. et Williams, K. M. (2006). Predictors of a behavioral measure of scholastic cheating : personality and competence but not demographics. *Contemporary educational psychology, 31*(1), 97-122.

Nering, M. L. et Meijer, R. R. (1998). A comparison of the person response function and the l(z) person-fit statistic. *Applied psychological measurement, 22*(1), 53-69.

Nurmi, J. E., Onatsu, T. et Haavisto, T. (1995). Underachievers' cognitive and behavioral strategies. Self- handicapping at school. *Contemporary educational psychology, 20*(2), 188-200.

Paillé, P. et Mucchielli, A. (2003). *L'analyse qualitative en sciences humaines et sociales*. Paris, France : Armand Colin.

Peeters, H. et Lievens, F. (2005). Situational judgment tests ans their predictiveness of college students' success : the influence of faking. *Educational and psychological measurement, 65*(1), 70-89.

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

Reise, S. P. et Flannery, W. P. (1996). Assessing person-fit on measures of typical performance. *Applied measurement in education, 9*(1), 9-26.

Reise, S. P. et Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of personality and social psychology, 65*(1), 143-151.

Rey, A. et Rey-Debove, J. (1991). *Dictionnaire alphabétique et analogique de la langue française*. Paris, France : Dictionnaires Le Robert.

Ro, S. (2001). *Characteristics of a likelihood-based person-fit index under the graded response model*. Thèse de doctorat inédite. University of Minnesota.

Robert, P., Rey, A. et Rey-Debove, J. (1967-1992). *Dictionnaire alphabétique et analogique de la langue française*. Paris, France : Le Robert Paris.

Seol, H. (1998). *Sensitivity of five rash-model-based fit indices to selected person and item aberrances : a simulation study.* Thèse de doctorat inédite. Ohio State University.

Sijtsma, K. et Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*(2), 191-207.

Smith, R. M. (1982). *Detecting measurement disturbances with the Rash model*. Thèse de doctorat inédite. University of Chicago, Chicago, Illinois.

Thurstone, L. L. (1929). Theory of attitude measurement. *Psychological review, 36*(3), 222-241.

Van der Ark, L. A., Emons, W. H. M. et Sijtsma, K. (2008). Detecting answer copying using alternate test forms and seat locations in small-scale examinations. *Journal of educational measurement, 45*(2), 99-117.

Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of cross-cultural psychology, 13*(3), 267-298.

Zickar, M. J. et Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied psychological measurement, 20*(1), 71-87.

Chapitre 6

Étude du comportement de 15 indices de détection de patrons de réponses inappropriés paramétriques et non paramétriques à partir d'une analyse par corrélations canoniques

Sébastien Béland, Patricia Brassard et Gilles Raîche

Lors de l'administration d'instruments de mesure du niveau d'habileté d'un étudiant, on assiste souvent à des tentatives de tricherie. Des indices ont été développés pour détecter les patrons de réponses inappropriés créés par ces tentatives. Or, la comparaison entre ces divers indices se limite généralement au calcul des corrélations entre eux. À ce titre, il serait plus indiqué d'utiliser l'analyse factorielle ou l'analyse par corrélation canonique afin d'obtenir une meilleure synthèse de ces corrélations. Dans le cadre de cette recherche, nous utilisons une démarche de type Monte Carlo pour étudier la relation qui existe entre certains indices de détection paramétriques et non paramétriques.

Lors de l'administration d'instruments visant à mesurer le niveau d'habileté d'un étudiant, on assiste souvent à des tentatives de tricherie. Ainsi, le Centre pour l'intégrité académique (*Center for Academic Integrity*) rapportait que plus de 75 % des étudiants ont admis avoir déjà triché à un test (Hutton, 2006, p. 171). De même, une étude de Laurier, Froio, Pearo et Fournier (1998) montrait que 78 % des enseignants du niveau collégial croient que les étudiants ne répondent pas correctement lors des épreuves scolaires. Conséquemment, ce type de comportement peut avoir des effets pernicieux sur la validité d'un test. En effet, l'observation d'un biais systématique dans la façon de répondre d'un étudiant peut causer

d'importants problèmes d'interprétation des scores. Ainsi, quelles voies s'offrent à nous pour tenter de détecter les individus qui adoptent de tels comportements lorsqu'ils passent un test ?

Zickar et Drasgow (1996) recensent plusieurs approches qui permettent d'éviter les comportements inappropriés. Notons, à titre d'exemple, que de nombreux auteurs conseillent aux évaluateurs de proposer des items qui sont difficiles à mésinterpréter. Par exemple, Edwards (1970) a suggéré de favoriser les questions ambiguës et moins explicites que ce que les évaluateurs souhaitent mesurer. Pour la mesure, quelques chercheurs suggèrent plutôt d'utiliser des indices de détection de patrons de réponses inappropriés tels que ceux développés par Tatsuoka et Tatsuoka (1982) ou Drasgow, Levine et Williams (1985). Dans le cadre de ce chapitre, nous nous concentrerons plus particulièrement sur cette dernière stratégie. Ainsi, nous focaliserons notre démarche sur l'objectif général suivant : analyser la relation existant entre différents indices de détection de patrons de réponses inappropriés.

1. CONTEXTE THÉORIQUE

1.1. Cadre conceptuel

Il existe plus d'une cinquantaine d'indices de détection de patrons de réponses inappropriés qui peuvent se diviser en deux grandes catégories (Karabatsos, 2003 ; Meijer et Sijtsma, 2001) : les indices paramétriques et les indices non paramétriques. Dans un premier temps, on constate que les indices non paramétriques découlent de la logique de comparaison en paires d'items (*group-based index*), dont la notion a surtout été élaborée par Guttman (1944, 1950). Au plan pratique, cette approche est la plus facile à appliquer, car les calculs sont directement effectués à partir des données brutes du test. Dans un deuxième temps, on voit que les indices paramétriques proviennent du modèle probabiliste de la théorie de la réponse aux items (TRI). Dans ce cas-ci, le calcul des patrons de réponses est plus fastidieux, car les données sont tributaires des estimations des modèles logistiques de la théorie de la réponse aux items.

Dans le cadre de cette recherche, nous avons sélectionné 15 indices qui ont été regroupés dans le tableau 6.1 :

Tableau 6.1
Classification des 15 indices sélectionnés

Indices non paramétriques		Indices paramétriques	
G	Guttman (1944, 1950)	U	Wright et Stone (1979)
NCI	Tatsuoka et Tatsuoka (1982)	lnU	Wright (1980)
		ZU	Wright et Stone (1979)
		W	Wright (1980)
		LnW	Wright (1980)
		ZU	Wright et Stone (1979)
		Lo	Levine et Rubin (1979)
		Lz	Drasgow, Levine et Williams (1985)
		zeta	Tatsuoka (1996)
		M	Molenaar et Hoijtink (1990)
		s	Ferrando (2004)
		c et d	Blais, Raîche et Magis (2009)

1.1.1. *Les indices non paramétriques du type Guttman*

Les indices non paramétriques suivent tous la même logique de base : le respect du vecteur de réponses parfait de Guttman (*Guttman perfect pattern*). Ainsi, les patrons de réponses possibles devraient arborer une des trois configurations suivantes lorsque tous les items sont classés selon leur degré de difficulté. En premier, le patron de type réussite parfaite : l'étudiant a bien répondu à tous les items du test (1111111111). En second, le patron de type échec parfait : l'étudiant n'a pas obtenu une seule bonne réponse dans le test (0000000000). Enfin, le patron de réponses parfaitement cohérent : si l'étudiant trouve une bonne réponse à un item, il trouvera aussi une bonne réponse à tous les items de moindre difficulté (ex. : 1111100000).

Le premier indice élaboré, l'indice *G* (Guttman, 1944, 1950), mesure le nombre de paires d'items qui dévient du vecteur de réponses parfait de Guttman. Mathématiquement parlant, cette statistique prend la forme suivante :

$$G = \sum_{d,f} X_{id}\,(1 - X_{if}) \qquad (1)$$

où l'on considère la réponse du répondant *j* aux items difficiles *d*, en paires d'items {*d*,*f*}, et sa réponse aux items faciles *f*, en paires d'items {*d*,*f*}. Ainsi, la paire d'items {0,1} présenterait une erreur de Guttman (*Guttman's error*), car l'étudiant a donné une mauvaise réponse à l'item le plus facile et une bonne réponse à l'item le plus difficile. Dans le

cadre de ce chapitre, nous retiendrons aussi le *norm conformity index* (*NCI*), de Tatsuoka et Tatsuoka (1982), qui mesure la conformité d'un patron de réponses au vecteur de réponses parfait de Guttman.

1.1.2. *Les indices paramétriques*

Les indices non paramétriques comportent plusieurs désavantages. Notons, à titre d'exemple, le fait que les estimations de ces indices ne sont pas indépendantes du niveau d'habileté (q) et que leur point de coupure (soit, la valeur numérique à partir de laquelle on considère un patron de réponses comme étant inapproprié) est tributaire de la nature des données traitées (Meijer et Sijtsma, 2001). Pour pallier ces problèmes, des indices découlant de la théorie de la réponse aux items ont été développés. Cette approche statistique (Bertrand et Blais, 2004 ; Hambleton, Swaminathan et Rogers, 1986 ; Hambleton et Zaal, 1991 ; Lord, 1980 ; Rasch, 1980 ; van der Linden et Hambleton, 1997) permet d'estimer la probabilité que l'étudiant j réponde correctement à l'item i :

$$P(X_1 = x \mid \theta) = c_i + \frac{1 - c_i}{1 + e^{a_i(\theta - b_i)}} \tag{2}$$

en tenant compte de l'habileté de l'individu (q), d'un paramètre de difficulté de l'item b_i, d'un paramètre de discrimination de l'item a_i et d'un paramètre de pseudo-chance c_i. Puisque nous traitons des items à réponses dichotomiques (0 ou 1), la probabilité qu'un étudiant obtienne le vecteur de réponses X (0 ou 1) suit la loi de Bernoulli :

$$P(X = x \mid \theta) = \prod_{i=1}^{I} \left\{ P(X_i = 1 \mid \theta)^{X_i} \, P(X_i = 0 \mid \theta)^{(1 - X_i)} \right\} \tag{3}$$

1.1.3. *Les indices paramétriques de vraisemblance*

Proposé par Levine et Rubin (1979), L_o est l'un des indices paramétriques le mieux documenté. Mécaniquement, cette approche calcule le maximum du logarithme de vraisemblance d'un patron de réponses :

$$L_0 = \sum_{i=1}^{I} \left\{ X_i \ln P(X_i = 1 \mid \theta) + (1 - X_i) \ln \left[1 - P(X_i = 0 \mid \theta) \right] \right\} \tag{4}$$

Malheureusement, cet indice est difficile à interpréter, car il est aussi partiellement dépendant du niveau d'habileté de la personne. Pour cette raison, Drasgow, Levine et Williams ont proposé une version standardisée de L_o en 1985 : l'indice *Lz*. Mathématiquement,

$$Lz = \frac{L_0 - E(L_0)}{VAR(L_0)^{1/2}} \tag{5}$$

Puisqu'il se distribue selon une loi normale, l'interprétation des résultats de cette statistique est aisée : $Lz \leq -1{,}96$ indique un patron de réponses inapproprié.

L'indice *M* de Molenaar et Hoijtink (1990) est aussi inspiré de l'indice L_o de Levine et Rubin (1979), mais il calcule tout simplement la somme des scores qu'un individu a obtenu à chacun des items et le degré de difficulté de ceux-ci de la façon suivante :

$$M = -\sum_{i=1}^{I} X_i b_i$$

1.1.4. *Les indices de type carrés moyens*

Deux indices paramétriques présentent une forme sensiblement différente des indices précédents, mais reposent sur la logique des carrés moyens : l'indice *U* et l'indice *W*. Dans un premier temps, Wright et Stone (1979) élaborent l'indice *U* (*outfit mean square*), qui est la moyenne des résidus de toutes les réponses au test élevée au carré. Il est à noter que dans ce chapitre, nous utilisons aussi la version standardisée *ZU* et le logarithme de *U* (*lnU*) (Wright, 1980). Dans un deuxième temps, l'indice *W* (ou *infit mean square*), de Wright et Stone (1979), représente plutôt la moyenne des résidus des réponses aux items pondérée par la somme des variances. Encore ici, nous présentons le logarithme de l'indice (*logW*) et sa version standardisée : *ZW* (Wright et Stone, 1979).

Zêta (Tatsuoka, 1996) est un indice qui a été moins souvent traité dans la littérature spécialisée. À l'instar de Raîche (2002), nous croyons qu'il serait aussi intéressant d'observer comment il se comporte avec d'autres indices. Mathématiquement, le numérateur de *zêta* est la covariance conditionnelle des vecteurs $P(X_i = 1 \mid \theta) - X_i$ et $P(X_i = 1 \mid \theta) - T(X_i \mid \theta)$, où $T(X_i \mid \theta)$ est la moyenne du nombre de bonnes réponses données par un individu. Le dénominateur est tout simplement la déviation standard conditionnelle du numérateur.

1.1.5. *Les indices de variabilité personnelle*

Contrairement à ce qui précède, cette catégorie d'indices ne suppose pas que l'habileté d'un individu demeure la même tout au long du test. Ainsi, Ferrando (2004) a élaboré un indice de discrimination personnelle *s* à l'aide du modèle de Rasch adapté :

$$P(X_i = 1 \mid \theta\ , s, b_i) = \Phi\left(\frac{\theta - b_i}{s}\right) = \Phi(a_i(\theta - b_i)) \tag{6}$$

où $a_i = s^{-1}$ et Φ est la distribution logistique. De leur côté, Blais, Raîche et Magis (2009) font l'analyse des deux autres paramètres de personne en créant un indice de pseudo-chance *C* et un indice d'inattention personnelle *D*.

1.2. Cadre théorique

Historiquement, les recherches produites dans le domaine ont permis de mettre en évidence les trois qualités essentielles que doit posséder un indice de détection de patrons de réponses inappropriés (St-Onge, 2007, p. 10). Dans un premier temps, Harnisch et Linn (1981), Molenaar et Hoijtink (1990) et Meijer et Sijtsma (2001) ont démontré l'importance de rendre l'estimation d'un indice indépendante du niveau d'habileté de l'individu en le standardisant. En effet, cela rendrait les indices plus faciles à interpréter : ils peuvent être directement comparés entre eux et leurs points de coupure sont alors faciles à trouver. Dans un deuxième temps, tout un champ de recherche s'est intéressé à étudier les points de coupure des différents indices. Enfin, plusieurs chercheurs ont tenté de préciser l'effet de l'environnement de simulation sur le taux de détection des indices. Puisque c'est cet aspect que nous développons dans le cadre de ce chapitre, nous lui consacrerons la prochaine section.

1.2.1. *Étude de l'environnement de la simulation sur le taux de détection des indices*

Dans le cas présent, nous avons retenu 27 études qui comparent de 2 à 36 indices entre eux. Afin d'avoir une meilleure vue d'ensemble de ce qui s'est fait, nous avons divisé ce champ d'investigation en quatre grandes catégories. Premièrement, de nombreuses recherches ont proposé d'étudier les indices en utilisant la corrélation de Pearson. Par exemple, Harnisch et Linn (1981), Rudner (1983) et Meijer (1994a) signalent généralement des relations fortes entre certains indices non paramétriques. De son côté, Rudner (1983) met aussi en évidence l'existence d'un coefficient de corrélation fort entre certains indices paramétriques. Comme on pouvait s'y attendre, Birenbaum (1985, 1986) ainsi que Li et Olejnik (1997) ont confirmé le fait que les indices standardisés sont fortement corrélés entre eux et faiblement corrélés avec le score total.

Deuxièmement, le modèle de simulation choisi n'est pas le même dans toutes les études. Ainsi, nous remarquons que la majorité des chercheurs font leurs simulations à l'aide du modèle de Rasch (Drasgow, 1982 ; Karabatsos, 2003 ; Li et Olejnik, 1997 ; Rudner, 1983 ; Smith, 1985) ou du modèle logistique à trois paramètres (Kogut, 1987 ; Meijer, Muijtjens et van der Vleuten, 1996 ; Raîche, 2002 ; Rogers et Hattie, 1987 ; Rudner, 1983). Il est à noter que seulement une étude utilise le modèle à quatre paramètres (Emons, Glas, Meijer et Sijtsma, 2003).

Troisièmement, certains auteurs montrent que plus le test est long (en nombre d'items), plus plus on détecte facilement l'étudiant au comportement inapproprié (Karabatsos, 2003 ; Li et Olejnik, 1997 ; Meijer, 1994a ; Rudner, 1983). Enfin, de nombreuses recherches ont étudié l'effet des réponses inappropriées sur le taux de détection des indices. Ainsi, Rudner (1983), Meijer, Muijtjens et van der Vleuten (1996) et Karabatsos (2003) ont démontré que les données faussement faibles (*spurliously low*) ont un taux de détection plus élevé que les données faussement élevées (*spurliously high*). Toutefois, Li et Olejnik (1997) démontrent que, dans le cadre du modèle de Rasch à une dimension, les données faussement faibles et faussement élevées ont à peu près le même niveau de détection. À l'opposé, il n'y a que Glas et Meijer (2003) qui ont examiné une diminution de la détection avec l'augmentation du niveau de réponses inappropriées dans un test.

1.3. Synthèse

La méthodologie utilisée dans ces études est déficiente en ce qui concerne certains aspects :

1. Les différents travaux utilisent une méthode de simulation de patrons de réponses inappropriés plutôt artisanale qui n'est pas basée sur un modèle mathématique formel (comme ceux issus des modélisations des paramètres personnels).
2. L'étendue des paramètres d'items et des paramètres de personne a été relativement limitée.
3. Ce type de recherche est généralement limité à seulement quelques indices de détection et ne couvre pas un éventail assez large : bien qu'il existe plus d'une cinquantaine d'indices de détection différents, 63 % des études recensées dans ce chapitre comparent seulement deux à quatre indices entre eux. De plus, 67 % des études dans le domaine traitent uniquement des indices paramétriques, alors que 11 % des études comparent uniquement les indices non paramétriques.

4. La comparaison se limite généralement à l'inspection de tableaux de corrélations entre ces indices. Il serait plus approprié d'effectuer des analyses qui permettent de mieux faire la synthèse des informations contenues dans ces tableaux de corrélations, par exemple, en utilisant une analyse factorielle (Mulaik, 1972) ou la corrélation canonique (Tabachnik et Fidell, 2001).
5. Ces études ne permettent pas de juger de l'importance de l'explication (la variance expliquée) du caractère inapproprié des patrons de réponses par l'ensemble de ces indices.

1.4. Objectif spécifique

Dans le cadre de ce chapitre, nous analysons les corrélations canoniques entre 15 indices de détection de patrons de réponses inappropriés et les trois paramètres de personne : *S* (fluctuation), *C* (pseudo-chance) et *D* (inattention personnelle), proposés par Blais, Raîche et Magis (2009).

2. MÉTHODOLOGIE

Pour conduire cette recherche de type empirique, nous utilisons le logiciel de programmation statistique R (version 2.8.1) afin de produire aléatoirement des patrons de réponses. À la limite, cela aura le net avantage de permettre de contrôler l'information simulée et de vérifier comment les indices se comportent les uns par rapport aux autres.

2.1. Simulations

Les simulations se font selon la modélisation logistique à quatre paramètres à laquelle on ajoutera quatre paramètres de personnel, proposés par Blais, Raîche et Magis. Les paramètres d'items et les paramètres personnels sont introduits au hasard selon les distributions de probabilité suivantes :

1. *b* et *q :* N (0,1) ;
2. *a* et *A* : uniforme (variant entre 0 et l'infini) ;
3. *c*, *C*, *d* et *D* : uniforme (variant entre 0 et 1).

Ainsi, les simulations se feront en deux étapes. Dans l'étude de simulation 1, nous produirons 1 000 patrons de réponses de 5, 10, 20 et 40 items pour chacune des conditions de simulation. Dans l'étude de simulation 2, nous tenterons de vérifier les résultats de la simulation 1 en produisant aléatoirement 5 000 patrons de réponses de 40 items.

2.2. Méthodes d'analyse des données

Dans le cadre de cette recherche, trois méthodes d'analyse sont mises de l'avant :

- Un tableau de corrélations de Pearson sera dressé pour les valeurs des indices de détection et les valeurs des paramètres personnels.
- Pour vérifier si l'ensemble des divers indices de détection permettent de détecter l'ensemble des comportements simulés par la modélisation à quatre paramètres personnels, les corrélations canoniques (Tabachnik et Fidell, 2001) entre les quatre paramètres personnels et les divers indices de détection seront calculées.
- Cette même analyse permettra aussi de synthétiser la relation qui existe entre ces différents indices et leur proximité entre eux. À cette fin, les saturations sur les axes canoniques seront analysées aussi bien en observant leurs valeurs que par leur représentation graphique.

3. RÉSULTATS

3.1. Corrélation de Pearson

Quelques constats intéressants peuvent être faits en ce qui concerne les corrélations présentées au tableau 6.2. On constate tout d'abord qu'il y a des coefficients de corrélation situés entre –0,01 et 0,03 pour les paramètres de personne *q* et *S*. Par contre, les paramètres *C* et *D* présentent des coefficients de corrélation respectifs de 0,64 et –0,65 avec le score total. Enfin, tous les indices et tous les paramètres de personne sont faiblement corrélés avec le calcul de la probabilité (entre –0,13 et 0,06).

Dans un deuxième temps, nous allons examiner les indices de détection. À l'exception de *c* et *d*, les indices sont tous faiblement corrélés avec le score total. D'une part, les deux indices non paramétriques, *NCI* et *G*, présentent une corrélation de –0,69. Il est à noter que ces indices sont aussi bien corrélés à *zêta*. De son côté, *G* présente une corrélation de 0,81 avec l'indice *s* de Ferrando. D'autre part, les indices paramétriques *U*, *ZU*, *lnU*, L_o, L_z, *W*, *ZW* et *lnW* sont généralement bien corrélés. Comme on pouvait s'y attendre, les indices *c* et *d* présentent un $r = 0{,}66$ avec leurs paramètres de personne respectifs *C* et *D*. De son côté, l'indice *M* est très peu corrélé avec tous les indices sélectionnés.

Tableau 6.2
Tableau de corrélations

	thêta	*S*	*C*	*D*	*Total*	*prob*	L_o	L_z	*zêta*	W	ZW	lnW	U	ZU	lnU	M	NCI	G	s	c	d
thêta	1	0,02	0,01	0,03	0,02	-0,03	-0,02	-0,02	0	0,02	0,02	0,02	0	0,01	0	0	0,01	0	-0,01	0,01	-0,01
S	0,02	1	-0,01	0,01	-0,02	-0,04	-0,05	-0,05	0,06	0,04	0,07	0,05	0,02	0,08	0,03	0,01	-0,06	0,07	0,1	0,02	0,05
C	0,01	-0,01	1	0	0,64	-0,04	-0,15	-0,14	0,2	0,13	0,16	0,14	0,09	0,16	0,09	0,01	-0,19	0,15	0,09	0,66	-0,56
D	0,03	0,01	0	1	-0,65	0,03	-0,14	-0,13	0,19	0,12	0,15	0,13	0,08	0,16	0,08	-0,01	-0,19	0,16	0,1	-0,59	0,66
total	0,02	-0,02	0,64	-0,65	1	-0,05	0,01	0,01	0,01	-0,01	-0,01	-0,01	-0,01	-0,01	-0,01	0,01	0,01	-0,01	-0,01	0,95	-0,94
prob	-0,03	-0,04	-0,04	0,03	-0,05	1	0,06	0,05	-0,01	-0,05	-0,1	-0,07	-0,03	-0,1	-0,03	0,01	0,01	-0,1	-0,13	-0,05	0,06
L_o	-0,02	-0,05	-0,15	-0,14	0,01	0,06	1	0,99	-0,27	-0,97	-0,97	-0,98	-0,84	-0,86	-0,84	0,01	0,38	-0,31	-0,24	-0,07	-0,09
zêta	0	0,06	0,2	0,19	0,01	-0,01	-0,27	-0,22	1	0,2	0,32	0,21	0,04	0,37	0,05	0	-0,87	0,64	0,39	0,18	0,18
W	0,02	0,04	0,13	0,12	-0,01	-0,05	-0,97	-0,99	0,2	1	0,92	1	0,91	0,76	0,91	-0,01	-0,31	0,25	0,2	0,06	0,08
ZW	0,02	0,07	0,16	0,15	-0,01	-0,1	-0,97	-0,95	0,32	0,92	1	0,94	0,73	0,89	0,74	-0,01	-0,41	0,39	0,34	0,08	0,11
lnW	0,02	0,05	0,14	0,13	-0,01	-0,07	-0,98	-0,99	0,21	1	0,94	1	0,9	0,78	0,9	-0,01	-0,32	0,28	0,24	0,06	0,08
U	0	0,02	0,09	0,08	-0,01	-0,03	-0,84	-0,89	0,04	0,91	0,73	0,9	1	0,6	1	0	-0,23	0,18	0,14	0,04	0,06
ZU	0,01	0,08	0,16	0,16	-0,01	-0,1	-0,86	-0,82	0,37	0,76	0,89	0,78	0,6	1	0,62	-0,01	-0,52	0,49	0,41	0,1	0,13
lnU	0	0,03	0,09	0,08	-0,01	-0,03	-0,84	-0,9	0,05	0,91	0,74	0,9	1	0,62	1	0	-0,24	0,19	0,16	0,04	0,06
M	0	0,01	0,01	-0,01	0,01	0,01	0,01	0,01	0	-0,01	-0,01	-0,01	0	-0,01	0	1	0	-0,01	-0,01	0,01	-0,01
NCI	0,01	-0,06	-0,19	-0,19	0,01	0,01	0,38	0,35	-0,87	-0,31	-0,41	-0,32	-0,23	-0,52	-0,24	0	1	-0,69	-0,45	-0,19	-0,22
G	0	0,07	0,15	0,16	-0,01	-0,1	-0,31	-0,28	0,64	0,25	0,39	0,28	0,18	0,49	0,19	-0,01	-0,69	1	0,81	0,12	0,15
s	-0,01	0,1	0,09	0,1	-0,01	-0,13	-0,24	-0,22	0,39	0,2	0,34	0,24	0,14	0,41	0,16	-0,01	-0,45	0,81	1	0,14	0,18
c	0,01	0,02	0,66	-0,59	0,95	-0,05	-0,07	-0,06	0,18	0,06	0,08	0,06	0,04	0,1	0,04	0,01	-0,19	0,12	0,14	1	-0,84
d	-0,01	0,05	-0,56	0,66	-0,94	0,06	-0,09	-0,09	0,18	0,08	0,11	0,08	0,06	0,13	0,06	-0,01	-0,22	0,15	0,18	-0,84	1

3.2. Corrélations canoniques

La corrélation canonique est utilisée pour étudier la relation existant entre deux groupes de variables, où chaque groupe est composé d'au moins deux variables. Dans le cadre de cette étude, nous tentons de savoir si les paramètres personnels *S*, *C* et *D* (les variables indépendantes) prédisent bien les 15 indices sélectionnés préalablement (les variables dépendantes).

3.2.1. *Simulation 1*

Dans un premier temps, nous avons procédé à des corrélations conditionnelles à quatre longueurs de test : 5, 10, 20 et 40 items. Ainsi, les résultats obtenus à chacune des itérations laissent entrevoir une structure à un facteur prédominant. Dans le cas de l'axe 1, le pourcentage de la variance de la variable canonique dépendante expliquée par la variable canonique indépendante est de 0,84 pour le test de 20 items et 0,91 pour un test de 40 items. À cet effet, le tableau 6.3 montre que la corrélation canonique augmente lorsque le test simulé est plus long

Tableau 6.3
Corrélations canoniques (1 000 patrons de réponses)

Nombre d'items	Corrélations canoniques		
	I	II	III
5	0,52	0,22	0,14
10	0,73	0,24	0,15
20	0,84	0,31	0,13
40	0,91	0,36	0,18

Le tableau 6.4 nous permet d'en savoir un peu plus sur la nature de cette dimension prédominante. Ici, les variables dépendantes (les indices de détection) semblent toutes liées au paramètre de personne *S* : en effet, nous remarquons que les coefficients estimés de *S* sont toujours le centre de gravité de la structure des données. Cela est d'autant plus vrai lorsque le nombre d'items augmente.

Un autre élément intéressant nous permet aussi d'en savoir un peu plus sur la structure des données : comme les données de l'axe 1 l'indiquent, les coefficients de corrélation *C* et *D* tendent à s'opposer. Ainsi, pour un test de 40 items, le coefficient *C* est égal à –0,08, alors qu'il est égal à 0,09 pour *D*.

Enfin, l'analyse des coefficients relatifs aux indices de détection nous permet de préciser encore plus notre argumentation. À part *C* (–0,08) et *D* (0,09), tous les indices de détection sont alignés sur l'axe 1 des données simulées lorsque le test comporte 40 items.

Tableau 6.4
Coefficients estimés des paramètres de personne (1 000 patrons de réponses)

Nombres d'items	Paramètres de personne	Coefficients estimés		
		I	II	III
5	*S*	0,00	0,00	0,00
	C	0,10	0,06	–0,07
	D	–0,10	0,08	–0,06
10	*S*	0,00	–0,00	0,01
	C	–0,09	–0,09	–0,01
	D	0,10	–0,08	–0,02
20	*S*	0,00	0,00	0,01
	C	–0,09	–0,08	–0,02
	D	0,09	–0,08	–0,03
40	*S*	0,00	0,00	0,01
	C	–0,08	–0,08	–0,03
	D	0,09	–0,08	–0,04

3.2.2. *Simulation 2*

Deuxièmement, nous avons simulé 5 000 patrons de réponses de 40 items afin de valider les résultats présentés plus haut. Ici, nous avons obtenu une corrélation canonique de 0,91 sur le premier axe, de 0,33 sur le deuxième axe et de 0,13 sur le troisième axe. Cette structure des données concorde bien avec celles que nous avons obtenues dans le cadre des simulations avec 1 000 patrons de réponses. Même constat pour les données du tableau 6.5 : les coefficients *C* (–0,04) et *D* (0,04) se distribuent aussi de part et d'autre du centre de gravité des données *S* (0,00).

Tableau 6.5
Coefficients estimés des paramètres de personne (5 000 patrons de réponses)

Paramètres de personne	Coefficients estimés		
	I	II	III
S	0,00	0,00	0,00
C	–0,04	–0,04	0,01
D	0,04	–0,04	0,01

Enfin, la figure 1 illustre l'alignement des différents indices sur le paramètre de personne *S*. Ainsi, il n'y a que les indices estimés *c* et *d* et les paramètres de personne *C* et *D* qui ne sont pas expliqués par ce premier axe.

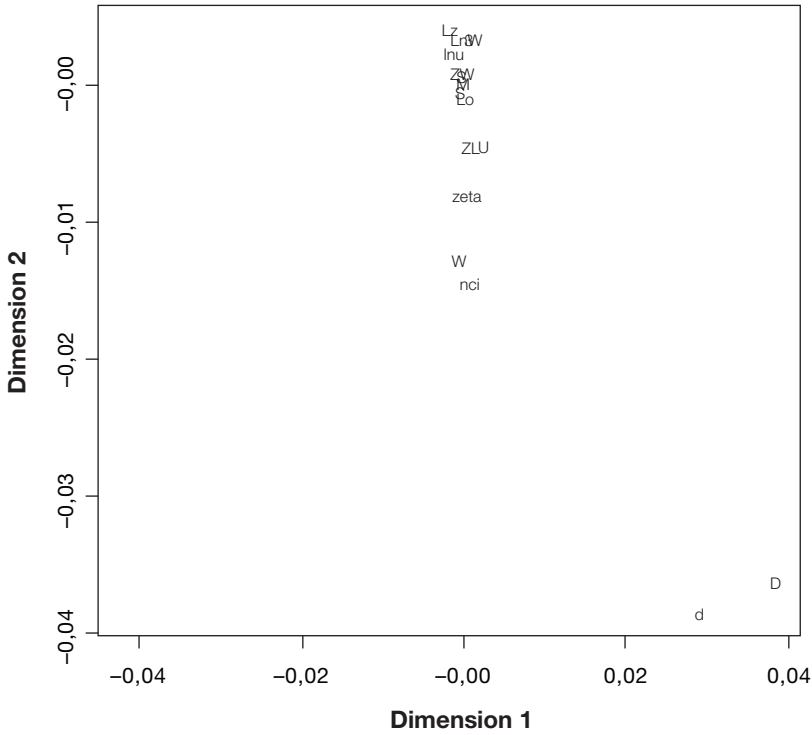


Figure 6.1
Coefficients estimés des indices et des paramètres de personnes (5 000 patrons de réponses)

4. DISCUSSION

Le tableau de corrélations que nous avons obtenu plus haut confirme les résultats de plusieurs autres recherches. Comme Harnisch et Linn (1981) et Rudner (1983) l'ont montré pour d'autres indices non paramétriques, nous avons aussi trouvé un coefficient de corrélation relativement élevé entre *NCI* et *G*. Néanmoins, nous avons obtenu un résultat différent de celui de Harnisch et Linn (1981) lorsque ceux-ci vérifient la corrélation de *NCI* avec le score total : leur résultat est de –0,54, alors qu'il est de 0,01 dans le cadre de cette recherche.

Concernant les corrélations obtenues pour *W*, L_o et L_z, nos résultats sont aussi conformes à ceux de Harnisch et Tatsuoka (1983). Par contre, nous trouvons un indice paramétrique *U* plus fortement corrélé à *W*, L_o et L_z que ces derniers. En ce qui a trait aux corrélations entre les indices et les scores totaux, nos résultats rejoignent ceux de Rogers

et Hattie (1987) et de Li et Olejnik (1997) : nous avons aussi obtenu des corrélations faibles entre le score total et les indices standardisés L_z, *ZU* et *ZW*.

Pour les paramètres de personne, notre étude reste purement exploratoire. Néanmoins, les résultats présentés ici soutiennent la démarche entreprise par Ferrando (2004) et Blais, Raîche et Magis (2009) puisque les corrélations canoniques ont mis en évidence l'importance de la fluctuation personnelle dans la structure des données, ainsi que de la pseudo-chance et de l'inattention personnelle.

CONCLUSION

À cause de sa visée uniquement exploratoire, cette recherche comporte plusieurs limites importantes. À la lumière des résultats obtenus, nous proposons deux autres pistes de réflexion qui pourraient élargir la portée de cette analyse. Premièrement, il faudrait intégrer d'autres indices de détection aux calculs que nous avons effectués : les indices non paramétriques ont été sous-représentés et il serait intéressant d'inclure certains indices paramétriques plus récents comme l'indice L_z* de Snijders. Enfin, il serait pertinent de calculer de nouvelles corrélations canoniques en simulant des patrons de réponses selon différents niveaux de *S*, *C* et *D*.

RÉFÉRENCES

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

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and psychological measurement, 45*(3), 523-534.

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied psychological measurement, 10*(2), 167-174.

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 technologies de l'information et de la communication. Enjeux, applications et modèles de mesure*. Sainte-Foy, Québec : Presses de l'Université Laval.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied psychological measurement*, *6*(3), 297-308.

Drasgow, F., Levine, M. V. et Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British journal of mathematical and statistical psychology, 38,* 67-86.

Edwards, A. L. (1970). *The measurement of personality traits by scales and inventories*. New York, New York : Holt, Rinehart and Winston.

Emons, W. H. M., Glas, C. A. W., Meijer, R. R. et Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement*, *27*(6), 459-478

Ferrando, P. J. (2004). Person reliability in personality measurement : an item response theory analysis. *Applied psychological measurement*, *9*(1), 47-64.

Glas, C. A. W. et Meijer, R. R. (2003). A Bayesian approach to person-fit analysis in item response theory models. *Applied psychological measurement*, *27*(3), 217-233.

Guttman, L. (1944). A basis for scaling qualitative data. *American sociological review, 9,* 139-150.

Guttman, L. (1950). The basis for scalogram analysis. Dans S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star et J. A. Clausen (dir.), *Measurement and Prediction*. Princeton, New Jersey : Princeton University Press.

Hambleton, R. K., Swaminathan, H. et Rogers, H. J. (1985). *Fundamentals of item response theory*. Measurement methods for the social sciences Series. Newbury Park, Californie : Sage.

Hambleton, R. K. et Zaal, J. N. (1991) *Advances in educational and psychological testing : theory and application*. Dordrecht, Pays-Bas : Kluwer academic publishers.

Harnisch, D. L. et Linn, R. L. (1981). Analysis of item response patterns : Questionable test data and dissimilar curriculum practices. *Journal of educational measurement*, *18,* 133-146.

Harnisch, D. L., et Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. Dans R. Hambleton (dir.), *Applications of item response theory* (p. 104-122). Vancouver, Colombie-Britannique : Educational Research Institute of British Columbia.

Hutton, P. A. (2006). Understanding student cheating and what educators can do about it. *College teaching*, *54*(1), 171-176.

Karabatsos, G. (2003). Comparing the aberrant response detectionp of thirty-six person-fit statistics. *Applied measurement in education*, *16*, 277-298.

Kogut, J. (1987). *Detecting aberrant item response patterns in the Rasch model* (Research Report 87-3). Enschede, Pays-Bas : University of Twente, Department of Education.

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.

Levine, M. V. et Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of educational and behavioral statistics, 4(4),* 269-290.

Li, M. F. et Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied psychological measurement*, *21*(3), 215-231.

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

Meijer, R. R. (1994a). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied psychological measurement*, *18*(4),311-314.

Meijer, R. R. et Sijtsma, K. (2001). Methodology review : evaluating person fit. *Applied psychological measurement*, *25*(2), 107-135.

Meijer, R. R., Muijtjens, A. M. M. et van der Vleuten, C. P. M. (1996). Nonparametric person-fit research : some theoretical issues and an empirical example. *Applied measurement in education, 9(1),* 77-89.

Molenaar, I. W. et Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55(1),* 75-106.

Mulaik, S. A. (1972). *The foundation of factor analysis*. New York, New York : McGraw-Hill.

R Development Core Team (2009). R : a language and environment for statistical computing. Vienne, Autriche : R Foundation for statistical computing.

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, Illinois : The University of Chicago Press.

Rogers, H. J. et Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied psychological measurement*, *11*(1), 47-57.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of educational measurement*, *20*(3), 207-219.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and psychological measurement*, *45*(3), 433-444.

St-Onge, C. (2007). La vraisemblance de patrons de réponses : étude de la précision des indices d'ajustement des scores individuels, de leurs points critiques et du taux optimal d'aberrance. Thèse de doctorat inédite. Université Laval, Québec.

Tabachnik, B. G. et Fidell, L. S. (2001). *Using multivariate statistics* (4e édition). Needham Heights, Massachusetts : Allyn and Bacon.

Tatsuoka, K. et Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of educational statistics, 7*(3), 215-231.

Tatsuoka, K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied measurement in education*, *9*(1), 65-75.

Van der Linden, W. J. et Hambleton, R. K. (dir.) (1997). *Handbook of modern item response theory*. New York, New York : Springer-Verlag.

Wright, B. D. (1980). Afterword. Dans G. Rasch (dir.) : *Probabilistic models for some intelligence and attainment tests : with foreword and afterword by Benjamin D. Wright*. Chicago, Illinois : MESA Press.

Wright, B. D. et Stone, M. H. (1979). *Best test design Rash measurement*. Chicago, Illinois : MESA Press.

Zickar, M. J. et Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied psychological measurement*, *20*(1), 71-87.

Chapitre 7

Utilisation de la théorie des ensembles flous pour valider une épreuve

Paul Martin et Jean-Guy Blais

La théorie des ensembles flous a été utilisée pour valider une épreuve de mathématique administrée aux étudiants de l'École Polytechnique de Montréal lors de leur admission. Un modèle a été établi pour poser un diagnostic sur la capacité d'un étudiant à réussir ses études. Les résultats nous montrent que le modèle est valide pour prédire la réussite future mais non l'échec. En d'autres mots, l'épreuve est moins valide dans un contexte d'orientation que de sélection.

En ce début de XXIe siècle, plusieurs croient qu'il faut désormais piloter les systèmes d'éducation dans le but d'assurer la réussite du plus grand nombre possible d'individus. Cependant, en corollaire à cette visée, il faut mesurer la capacité de l'étudiant au début du processus afin de l'orienter, au besoin, vers une formation d'appoint. Pour ce faire, on doit s'assurer de la validité prédictive de l'instrument qui mesure cette capacité, mais en prédisant autant l'échec que la réussite. Dans un tel contexte, la validité est souvent difficile à démontrer à l'aide des théories de la mesure existantes.

Une nouvelle piste à cet effet, soit la théorie des ensembles flous, a donc été envisagée et a fait l'objet d'une application au niveau universitaire. Plus précisément, l'utilisation de cette théorie devrait nous permettre d'élaborer un modèle pour traduire le score sous la forme d'un diagnostic afin de répondre à la question

suivante : est-ce qu'un diagnostic de capacité établit un pronostic de réussite et, parallèlement, est-ce qu'un diagnostic d'incapacité établit un pronostic d'échec ? La comparaison avec les méthodes traditionnelles de corrélation va nous permettre de distinguer les forces de cette nouvelle approche.

1. CADRE CONCEPTUEL

Par capacité, nous entendons un comportement attendu (De Landsheere, 1979) ou encore une compétence (Bloom, 1988) mesurée au moment de l'entrée dans un cours ou un programme d'études. À l'inverse, la réussite est plutôt le comportement attendu ou la compétence reconnue à la sortie du cours ou du programme d'études. Il s'agit donc de mesurer la capacité des sujets afin d'émettre un diagnostic avec un instrument de mesure éprouvé du point de vue de sa validité pronostique en se référent au critère de réussite. En d'autres mots, un diagnostic de capacité valide se doit donc d'être en même temps un pronostic fiable de la réussite ou de l'échec. Toutefois, la capacité doit être reliée à la nature du programme d'enseignement. Dans la situation étudiée, nous allons diagnostiquer la capacité en mathématique puisque les sujets s'engagent dans des programmes d'études en génie.

Les théories des tests existantes ne nous sont pas d'une grande utilité pour vérifier la validité pronostique d'un tel instrument. En effet, selon Brown (1980), il existe trois manières d'interpréter les scores obtenus avec un instrument de mesure : en se référant à une norme, à un contenu ou à un critère. Or, le but principal de la théorie classique des tests est d'expliquer les scores en fonction d'une norme alors que celui de la théorie de la réponse à l'item est plutôt de faciliter l'interprétation des scores en se rapportant principalement à une habileté envers un contenu. Cependant, ces théories/modèles n'apportent pas d'explications suffisantes lorsqu'il s'agit d'interpréter les scores par rapport à un critère et c'est pourquoi la théorie des ensembles flous nous semble prometteuse à cet égard (Zadeh, 1965).

Il existe des situations où il est difficile de classer avec certitude un élément dans un ensemble ou son complément, qui est spécifié par l'attribut contraire. En effet, un sujet peut réussir ou échouer, mais il se peut aussi que l'on ne soit pas sûr de son classement et que l'on ne puisse affirmer avec certitude s'il réussit ou s'il échoue. De même, on peut diagnostiquer un sujet comme ayant la capacité ou l'incapacité d'entreprendre des études, mais il se peut aussi que le diagnostic soit incertain. Dans ces derniers cas, on devra recourir aux ensembles flous pour indiquer l'incertitude dans le classement.

1.1. L'appartenance à un ensemble

Le concept d'appartenance a été proposé par Zadeh (1965), le précurseur de la théorie des ensembles flous, et il s'opérationnalise par un indice spécifiant le degré de *vérité* de l'appartenance d'un élément à un ensemble (équation 1).

$$m_S(x) \rightarrow U\colon [0,1] \tag{1}$$

L'indice, $m_S(x)$, représente le degré d'appartenance d'un élément x à l'ensemble S et se définit dans l'univers, U, par un intervalle fermé de nombres réels compris entre 0 et 1. Une valeur de 0 pour cet indice signifie qu'il est certain que l'élément x n'appartient pas à l'ensemble S et une valeur de 1 qu'il est certain que cet élément appartient à ce même ensemble. Cependant, la particularité d'un ensemble flou est de présenter plusieurs possibilités d'appartenance. Ainsi, une valeur de 0,5 pour cet indice représente un maximum d'ambiguïté quant au classement de l'élément x, à savoir qu'il n'est ni vrai, ni faux que cet élément appartienne à l'ensemble S. L'ensemble classique défini par une appartenance certaine constitue un cas particulier d'un ensemble flou.

Dans le contexte de l'étude présentée, nous avons un ensemble flou X dont l'attribut est la capacité en mathématique et un ensemble flou Y représentant le critère à prédire. Plus particulièrement, cet ensemble flou Y a comme attribut la réussite dans le premier cours de mathématique.

1.2. La fonction d'appartenance discrète

Avec des ensembles flous, il devient nécessaire d'avoir une fonction qui détermine la valeur de l'indice d'appartenance à partir d'une donnée. Cette fonction d'appartenance joue un rôle fondamental lors du classement des données et doit être élaborée à partir de points d'ancrage dans la réalité. Or, comme le dit Kosko (1993), le monde est en gris, mais la science est en noir et blanc. Il faut donc proposer une fonction d'appartenance possédant différents tons de gris pour pouvoir décrire le plus fidèlement possible la réalité.

Une fonction d'appartenance discrète est souvent représentée sous la forme d'un tableau (voir tableau 7.1) où un repère linguistique est proposé pour guider l'affectation d'une valeur de l'indice d'appartenance lors du classement d'une donnée.

Ragin (2000) préconise une méthode pour étalonner une fonction d'appartenance où il faut spécifier trois points d'ancrage qualitatifs importants. D'abord, il faut spécifier le point auquel il y aurait certitude d'appartenance (m_S = 1,00), ensuite le point auquel il y aurait

certitude de non-appartenance ($m_S = 0{,}00$) et, finalement, le point où il y aurait le maximum d'ambiguïté ($m_S = 0{,}50$). Par la suite, l'affectation des valeurs pour l'indice d'appartenance peut se poursuivre en progressant entre les intervalles.

Tableau 7.1
Une fonction d'appartenance discrète (Ragin, 2000)

$\mu_S(x)$	**repère linguistique**
0,00	certainement faux
0,17	sensiblement faux
0,25	plus faux que vrai
0,33	plus ou moins faux
0,50	ni vrai, ni faux (ambiguë)
0,67	plus ou moins vrai
0,75	plus vrai que faux
0,83	vraisemblable
1,00	certainement vrai

Ainsi, l'ensemble flou *Y* faisant office de critère est caractérisé par une fonction d'appartenance discrète selon les cotes obtenues par les sujets, c'est-à-dire F, D, D+, C, C+, B, B+, A, A+. Comme postulat, nous posons qu'il est certainement vrai qu'un sujet réussisse avec une cote supérieure ou égale à B+, qu'il est certainement faux qu'il réussisse s'il obtient la cote F, que la cote C représente le maximum d'ambiguïté. Les valeurs 0,67 et 0,83 correspondent aux cotes C+ et B respectivement et les valeurs 0,33 et 0,17 correspondent aux cotes D+ et D. Les cotes A et A+ représentent aussi une réussite certaine avec une valeur de 1,00.

1.3. La fonction d'appartenance continue

Dans certaines situations, il est préférable de recourir à une fonction d'appartenance continue où chaque élément se caractérise par un indice d'appartenance distinct. C'est notamment le cas lorsqu'il s'agit du score d'un sujet à un test comme par exemple dans le cas de l'ensemble flou *X* de notre étude. Traditionnellement, une interprétation critériée du score est faite lorsque celui-ci est comparé à un score de césure pour décider de la réussite ou de l'échec. On peut comparer cette situation à un classement dans un ensemble classique. En effet, un score de césure permet de prendre une décision quant à l'échec ou la réussite ; toutefois, cette décision peut être erronée si le score brut du sujet se situe tout près du score de césure (l'impact de l'erreur de mesure). Aussi, Smithson (1989) propose une fonction d'appartenance

continue qui épouse une courbe en S comme modèle pour traduire le score *x* à une épreuve en un degré d'appartenance à un ensemble flou. Il propose une fonction logistique à deux paramètres telle que décrite par l'équation 2 et représentée par la courbe de la figure 7.1.

$$\mu_s(x) = \frac{1}{1 + e^{-\alpha(x-\beta)}} \tag{2}$$

où *x* est le score brut, α un paramètre spécifiant la plage des scores où le classement est incertain et β un paramètre associé au score de césure.

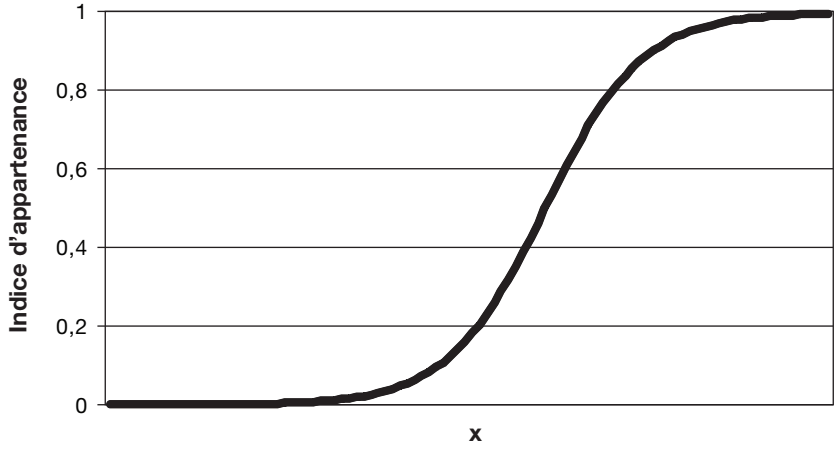


Figure 7.1
Graphique du modèle à deux paramètres de l'équation 2

Lorsque le score brut, *x*, se confond avec le score de césure β, il y a ambiguïté maximum dans le classement puisque l'indice d'appartenance vaut 0,50 en ce point. De plus, une zone est établie autour de ce score de césure pour indiquer que le diagnostic est incertain. La valeur de l'indice d'appartenance est 0,95 lorsque le score brut *x* se situe à une distance $+3/\alpha$ du score de césure β et de 0,05 lorsque le score brut se situe sous le standard de contre-performance, soit à une valeur de $-3/\alpha$ inférieure à β. Par conséquent, nous pouvons émettre avec certitude un diagnostic de capacité pour un score supérieur au standard de performance et un diagnostic d'incapacité pour un score inférieur au standard de contre-performance.

1.4. Le complément d'un ensemble

Chaque attribut devrait normalement avoir son contraire. Ainsi, l'incertitude pour l'indice d'appartenance à un ensemble quelconque, *A*, se reflète pour l'indice d'appartenance au complément de cet ensemble, *A*', par l'équation 3.

$$\mu_{A'}(x) = 1 - \mu_A(x) \tag{3}$$

Cette équation fait en sorte que le cas classique devient un cas particulier puisque s'il est certainement vrai qu'un sujet appartient à l'ensemble *A* ($\mu_A = 1{,}00$) alors il est certainement faux qu'il appartient à *A*' ($\mu_{A'} = 0{,}00$). Cependant, l'équation 3 permet d'apporter des nuances pour les cas incertains. Ainsi, le degré d'incapacité d'un sujet est égal à 1 moins son degré de capacité.

1.5. L'intersection

Avec l'algèbre de Boole qui régit la théorie classique des ensembles, nous retrouvons l'opération d'intersection entre plusieurs ensembles. Pour appartenir à l'intersection de ces ensembles, un élément doit appartenir à tous les ensembles qui composent cette intersection. L'algèbre flou vient nuancer à l'aide de l'opérateur minimum comme le décrit l'équation 4 pour le cas de l'intersection de deux ensembles quelconques, *A* et *B*:

$$\mu_{A\cap B}(x) = \text{Min}\ (\mu_A(x), \mu_B(x)). \tag{4}$$

Autrement dit, l'indice d'appartenance d'un sujet à l'intersection de deux ensembles est le minimum entre ces indices d'appartenance particuliers à chacun de ces ensembles. Ces opérateurs logiques vont permettre de traiter les données sans pour autant assumer qu'elles proviennent d'une mesure d'intervalle.

1.6. Le cardinal scalaire

Nous pouvons obtenir la mesure d'un ensemble flou quelconque, *A*, en prenant son cardinal scalaire, c'est-à-dire en additionnant, comme le montre l'équation 5, les indices d'appartenance à ce même ensemble *A,* et ce, pour tous les éléments *x* de l'univers *U* (les barres verticales qui encadrent l'ensemble *S* indiquent qu'il s'agit du cardinal scalaire):

$$|A| = \Sigma\ \mu_A(x). \tag{5}$$

Ainsi, on peut interpréter le cardinal scalaire, $|A|$, comme étant une statistique pour décrire l'ensemble flou.

1.7. L'inférence des valeurs vraies pour les prémisses

Avec la théorie des ensembles flous, on ne calcule pas un coefficient de corrélation ; on raisonne plutôt en recourant à la méthode de raisonnement du syllogisme appliqué avec des données. Rappelons qu'un syllogisme comprend trois parties : la prémisse, la situation d'un cas particulier et la conclusion. Or, la prémisse représente l'inclusion d'un ensemble flou quelconque *A* dans un autre ensemble flou *B*.

En effet, si un élément *x* appartient à un ensemble *A* (cas particulier) et que cet ensemble est inclus à son tour dans un ensemble *B* (prémisse), alors il est certain que cet élément *x* appartient aussi à l'ensemble *B* (conclusion). D'une manière générale, nous pouvons dire que tous les éléments appartenant à *A* appartiennent aussi à *B*. Cependant, avec les ensembles flous il faut nuancer et c'est pourquoi Klir et Yuan (1995, p. 28) décrivent le degré d'inclusion d'un sous-ensemble flou *A* dans un ensemble flou *B* par l'équation 6 :

$$\left\| A \subseteq B \right\| = \frac{\left| A \cap B \right|}{\left| A \right|} \tag{6}$$

Smithson (2005, p. 443) ajoute que l'idée générale est de comparer les degrés d'appartenance à l'intersection des ensembles flous *A* et *B* avec les degrés d'appartenance pour l'ensemble *A* qui est inclus. Il ajoute que l'opérateur minimum, ne requérant pas de conditions strictes d'utilisation d'une échelle d'intervalles, permet beaucoup d'applications en sciences sociales. L'expression $||A \subseteq B||$ signifie *la valeur vraie que l'ensemble A soit inclus dans l'ensemble B*. Cette valeur vraie vaut 1,00 lorsque l'ensemble *A* est entièrement inclus dans l'ensemble *B* et vaut 0,00 lorsque l'ensemble *A* en est complètement exclu. Enfin, la valeur vraie d'une prémisse permet d'inférer une règle d'implication logique du type : *si x appartient à A alors x appartient à B*.

Dans notre recherche, nous avons les ensembles *X* et *Y* qui forment quatre partitions possibles dans l'univers : $X' \cap Y'$, $X' \cap Y$, $X \cap Y'$, $X \cap Y$. Or, seules les partitions $X' \cap Y'$ et $X \cap Y$ sont cohérentes alors que les deux autres ne le sont pas. En effet, il est cohérent de dire qu'un sujet pour lequel on a posé un diagnostic de capacité réussira, c'est-à-dire $X \cap Y$, et il est aussi cohérent d'affirmer qu'un étudiant présentant un diagnostic d'incapacité échouera, c'est-à-dire $X' \cap Y'$. S'il existe une relation causale entre la mesure de la capacité *X* et le critère de réussite *Y* alors nous pouvons affirmer que l'instrument de mesure possède une validité prédictive. En ce sens, de Landsheere (1979, p. 17) définit une cause comme étant un antécédent nécessaire, c'est-à-dire une condition qui précède toujours l'apparition d'un phénomène donné et en l'absence de laquelle le phénomène ne se produit jamais.

Cette définition nous amène à rechercher non seulement l'explication de la réussite, mais aussi celle de l'échec et l'équation 7 indique comment faire ce calcul des valeurs vraies pour les prémisses :

$$\underline{\|X \subseteq Y\|} = \frac{|X \cap Y|}{|X|}$$
$$\underline{\|X' \subseteq Y'\|} = \frac{|X' \cap Y'|}{|X'|} \qquad (7)$$

Or, la valeur vraie de la prémisse est aussi la valeur vraie d'inclusion. Par exemple, pour une valeur vraie d'inclusion significative de $\|X' \subseteq Y'\|$, on peut établir une règle du type : *si un diagnostic d'incapacité est posé envers un sujet alors il échouera* (Schneider et Kandel, 1992, p. 34).

2. MÉTHODOLOGIE

Le processus de validation est effectué en trois temps : la calibration du modèle, l'inférence des valeurs vraies pour les prémisses et enfin, une étude de corrélation.

2.1. Sujets

La méthode développée précédemment a été appliquée à une épreuve de mathématique administrée aux étudiants de l'École Polytechnique de Montréal lors de leur admission. Plus particulièrement, nous utilisons des données en provenance des cohortes de 1997 et 1999. Seules les données pour les sujets ayant poursuivi leurs études dans cette institution ont été retenues.

2.2. Instrumentation

Pour chacune de ces cohortes, les données utilisées sont le score à l'épreuve de mathématique (X) et la cote au cours de mathématique Calcul I (Y). Chacune des épreuves de mathématique comporte 60 items à réponse choisie qui se répartissent en six domaines. Elles sont produites à partir d'une banque de plus de 900 items. Par conséquent, on assume que les épreuves de mathématique pour les cohortes de 1997 et 1999 sont équivalentes.

2.3. Déroulement

La première cohorte de 1997 a servi à calibrer le modèle à deux paramètres pour déterminer la certitude dans le diagnostic de capacité (voir équation 2 et figure 7.1). On utilise la méthode des groupes contraires (Nedelsky, 1954) pour calculer le score de césure d'un instrument critérié qui, pour nous, se confond avec le paramètre β du modèle. Les groupes contraires sont déterminés en fonction des cotes obtenues au cours Calcul I. Ainsi, on établit un groupe fort avec les étudiants qui ont obtenu une cote supérieure ou égale à C+ et un groupe faible avec ceux qui ont obtenu une cote inférieure ou égale à D+. Pour chacun de ces groupes, on calcule la moyenne à l'épreuve de mathématique et la moyenne de ces deux moyennes nous donne le score de césure β.

Par la suite, on utilise l'indice de Livingston pour estimer la fidélité de l'épreuve en appliquant la méthode moitié-moitié. Ce coefficient a la particularité de recourir à une variance qui tient compte de la dispersion des scores par rapport au score de césure plutôt que par rapport à la moyenne ; ainsi, le calcul s'effectue à partir du paramètre β trouvé à l'étape précédente. Avec cet indice, on établit l'erreur-type de mesure et l'intervalle de confiance où se situe le score vrai de césure. De plus, un intervalle au seuil de confiance de 95 % correspond à la zone d'incertitude du modèle (voir figure 7.1), soit $\beta \pm 3/\alpha$. De là, on peut déduire le second paramètre, α.

2.4. Considérations éthiques

Les données ont été fournies par l'École Polytechnique de Montréal et aucun renseignement nominatif ne permettait d'identifier les sujets pour les cohortes de 1997 et 1999.

2.5. Méthode d'analyse des résultats

En appliquant le modèle calibré aux scores obtenus à l'épreuve de mathématique pour les étudiants de la cohorte de 1999, on obtient leur appartenance à l'ensemble flou *X*. De plus, pour obtenir un ensemble flou *Y* représentant la réussite des étudiants de la cohorte de 1999 au cours Calcul I, on utilise une fonction d'appartenance discrète où les points d'ancrage sont : certitude de réussite à B+ ($\mu_Y = 1{,}00$), certitude d'échec à F ($\mu_Y = 0{,}00$) et maximum d'ambiguïté à C ($\mu_Y = 0{,}50$).

À partir des indices d'appartenance, μ_X et μ_Y pour chacun des sujets aux ensembles *X* et *Y,* on a calculé les valeurs vraies d'inclusion : $||X \subseteq Y||$ et $||X' \subseteq Y'||$ qui constituent les prémisses pour les règles d'inférence. Un test statistique *z* sur les proportions est utilisé pour vérifier

si la valeur vraie est significative, c'est-à-dire si elle est supérieure à l'ambiguïté de 0,50. L'interprétation de ces valeurs vraies s'effectue en comparant les valeurs numériques obtenues aux repères linguistiques établis dans le tableau 7.1.

L'étude de corrélation est effectuée selon deux méthodes avec les données de la cohorte de 1999. L'ensemble des scores à l'épreuve de mathématique représente la variable *X* alors que l'ensemble des cotes au cours Calcul I est associée à la variable critère *Y*. Il s'agit de décider à l'aide d'un test statistique unilatéral s'il y a une corrélation positive et significative entre les deux variables *X* et *Y*.

D'une part, on calcule le coefficient de corrélation de Pearson, r_{XY}, en assumant que les données proviennent de mesures sur une échelle d'intervalles. Les cotes pour le cours Calcul I sont traduites en nombres selon la suite : F = 0,00 ; D = 0,50 ; D+ = 1,00 ; … A+ = 4,00. On utilise le test *t* pour vérifier la signification de la corrélation positive ($\rho > 0$).

D'autre part, on calcule le coefficient phi, r_ϕ, en assumant que les données sont dichotomiques (échec/réussite). La variable *X* est l'ensemble des décisions échec/réussite issu de la comparaison des scores bruts à l'épreuve avec le score de césure β trouvé lors de la calibration du modèle. De même, la variable *Y* est l'ensemble des décisions issu de la comparaison des cotes au cours Calcul I avec la cote C. On utilise le test du chi-carré, χ^2, pour vérifier la signification de la corrélation positive ($\rho > 0$).

3. RÉSULTATS

La cohorte de 1997 comportait 344 sujets pour lesquels nous avions des données. La moyenne et l'écart-type obtenus à l'épreuve ont été respectivement 38/60 (63 %) et 8,93. Par contre, nous avions 389 sujets pour la cohorte de 1999 ; la moyenne et l'écart-type étaient respectivement 34/60 (57 %) et 8,42.

3.1. Calibration du modèle

La moyenne du groupe faible au (prétest) de mathématique était 32,5 alors que celle du groupe fort égalait 42,5. Ainsi, le score de césure qui se confond au paramètre β a donné 37,5. Le calcul de l'indice de fidélité de Livingston, r_{Xq}, a été effectué en divisant l'épreuve en deux parties égales de 30 items : les items pairs et impairs. La césure à utiliser est la moitié de la valeur du paramètre β, soit 18,75. Le calcul de cet indice de fidélité, r_{Xq}, nous a donné une valeur de 0,637 ; ainsi on a pu établir

l'erreur-type de mesure et l'intervalle de confiance où se situe le score de césure vraie qui correspond à la zone d'incertitude du modèle. Par la suite, on a pu déduire la valeur 0,278 du paramètre α.

3.2. L'inférence des valeurs vraies pour les prémisses

Le tableau 7.2 montre les valeurs vraies pour les prémisses $||X \subseteq Y||$ et $||X' \subseteq Y'||$. Les calculs pour un test statistique de différence de proportions sont effectués, la valeur de 0,50 étant la proportion de référence.

Tableau 7.2
Valeurs vraies des prémisses pour la cohorte de 1999 (*N* = 389)

Diagnostic	Intersection	Prémisse	H_0	Calcul	Signification (α = 0,001)	Décision
$\|X\|$ = 114,80	$\|X \cap Y\|$ = 115,60	$\|\|X \subseteq Y\|\|$ = 0,80	≤ 0,50	z = 7,20	> 3,0902	Rejetée
$\|X'\|$ = 244,20	$\|X' \cap Y'\|$ = 156,50	$\|\|X' \subseteq Y'\|\|$ = 0,64	≤ 0,50	z = 4,37	> 3,0902	Rejetée

3.2.1. *Les corrélations*

On retrouve dans le tableau 7.3 les coefficients de corrélation ainsi que les résultats aux tests statistiques pour décider de leur signification. Rappelons que les tests statistiques sont unilatéraux, et ce, afin de décider s'il y a une corrélation positive ($\rho > 0$) entre l'attribut *X* mesuré par l'épreuve de mathématique et le critère *Y*.

Tableau 7.3
Corrélations *X*-*Y* pour la cohorte de 1999 (*N* = 389)

Données	Coefficient	H_0	Calcul	Signification (α = 0,001)	Décision
Continues	$r_{XY} = 0,57$	$\rho \leq 0$	$t = 13,64$	> 3,12	Rejetée
Dichotomiques	$r_{\varphi} = 0,40$	$\rho \leq 0$	$\chi^2 = 62,24$	> 10,83	Rejetée

4. DISCUSSION DES RÉSULTATS

Rappelons le sujet de la recherche : est-ce qu'un diagnostic de capacité établit un pronostic de réussite et, parallèlement, est-ce qu'un diagnostic d'incapacité établit un pronostic d'échec ? Or, la calibration du modèle avec la cohorte de 1997 vient nous signifier que le standard de performance se situe à 49 (82 %) et le standard de contre-performance à 26 (43 %). Par conséquent, on posera avec certitude un diagnostic de capacité pour un sujet ayant obtenu un score supérieur à

49/60 et un diagnostic d'incapacité pour celui dont le score est inférieur à 26/60. De plus, le faible écart entre les moyennes à l'épreuve pour les deux cohortes nous permet d'affirmer qu'il n'y a pas de différences significatives et que le modèle s'applique pour la cohorte de 1999.

À partir des résultats obtenus au tableau 7.2, nous vérifions la validité pronostique de l'épreuve. Toutefois, en se référant aux repères linguistiques du tableau 7.1, nous pouvons nuancer cette validité par la règle suivante : *il est plus vrai que faux d'affirmer que si un sujet a été diagnostiqué certainement capable alors il réussira le cours Calcul I*. Par contre, on ne peut rien affirmer sur le pronostic d'échec d'un étudiant qui aurait reçu un diagnostic certain d'incapacité. Pour un cas où le diagnostic est incertain, le degré de certitude dans la conclusion est le minimum entre le degré de certitude de la prémisse de 0,80 et celui déterminé par le modèle à partir du score obtenu.

Les résultats montrés dans le tableau 7.3 pour les coefficients de corrélation viennent aussi vérifier la validité pronostique. En effet, Violato, McDougall et Marini (1992) affirment qu'il est rare que le coefficient de corrélation de Pearson dépasse 0,60. De plus, Zwick (2002) va plus loin en affirmant que la corrélation entre le score au test d'admission SAT dans les collèges américains et la note à la fin de la première année se situe entre 0,30 et 0,40. Toutefois, ici, on ne peut répondre adéquatement à la question posée dans la recherche puisque nous assumons au départ une symétrie dans les résultats, c'est-à-dire que le même coefficient de corrélation s'applique autant pour les sujets faibles que pour les sujets forts.

CONCLUSION

Rappelons que l'objectif de la recherche était d'élaborer un modèle pour traduire le score à une épreuve d'admission en un diagnostic de capacité à réussir les études. Les résultats issus de l'application de ce modèle ont été comparés aux résultats obtenus avec les méthodes traditionnelles de corrélation. La validité pronostique de l'épreuve ne peut être mise en doute puisqu'il est possible d'établir un pronostic de réussite à partir du score obtenu. Cependant, on ne peut établir un pronostic d'échec et cela rend difficile l'orientation de l'étudiant vers une formation d'appoint. Étant donné que le score obtenu à l'épreuve de mathématique ne compte pas dans l'évaluation de l'étudiant, il est probable que l'absence de motivation à répondre correctement puisse sous-estimer sa capacité réelle et cela expliquerait, en partie du moins,

l'asymétrie dans les résultats obtenus. D'autres données qui n'étaient pas disponibles lors de la recherche pourraient sans doute contribuer à établir un pronostic significatif de l'échec.

Par contre, cette recherche a démontré qu'il est possible d'utiliser la théorie des ensembles flous pour prédire un critère en éducation. Cela vient faciliter grandement la démarche de validation puisqu'il n'est plus nécessaire d'assumer que nous avons des mesures d'intervalle. Or, nous avons utilisé une analyse avec une variable indépendante seulement ; il faudrait, dans une prochaine recherche, recourir à d'autres données pour effectuer une analyse avec plusieurs variables indépendantes.

Le débat public récent sur la forme du bulletin scolaire au Québec vient illustrer l'urgence de reconsidérer l'évaluation des apprentissages. On ne peut prétendre que la théorie des ensembles flous puisse répondre à toutes les questions, mais elle peut certainement contribuer au débat. En effet, cette théorie permet de raisonner l'évaluation plutôt que de la calculer, soit un pas manifestement important dans la bonne direction.

RÉFÉRENCES

Bloom, B. S. dans Legendre, R. (1988). *Dictionnaire actuel de l'éducation*. Paris, France : Larousse.

Brown, F. G. (1980). *Guidelines for test use : a commentary on the standards for educational and psychological tests*. Washington, Colombia : National council on measurement in education.

De Landsheere, G. (1979). *Dictionnaire de l'évaluation et de la recherche en éducation*. Paris, France : Presses universitaires de France.

Klir, G. J. et Yuan, B. (1995). *Fuzzy sets and fuzzy logic : theory and applications*. Englewood Cliffs, New Jersey : Prentice Hall.

Kosko, B. (1993). *Fuzzy thinking, the new science of fuzzy logic*. New York, New York : Hyperion.

Legendre, R. (1988). *Dictionnaire actuel de l'éducation*. Paris, France : Larousse.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and psychological measurement, 14*(1),3-19.

Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago, Illinois : The University of Chicago Press.

Schneider, M. et Kandel, A. (1992). General purpose fuzzy expert systems. Dans A. Kandel (dir.), *Fuzzy expert systems*. Boca Raton, Floride : CRC Press.

Smithson, M. (1989). *Ignorance and uncertainty. Emerging paradigms*. New York, New York : Springer-Verlag.

Smithson, M. (2005). Fuzzy set inclusion. Linking fuzzy set methods with mainstream techniques. *Sociological methods and research, 33*(4), 431-461.

Violato, C., McDougall, D. et Marini, A. (1992). *Educational measurement and evalauation*. Dubuque, Iowa : Kendall-Hunt.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control, 8*, 338-353.

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. New York, New York : Routledge Falmer.

LISTE DES CONTRIBUTEURS

Réjean AUGER
Optimisation et gestion de devis, données, analyses statistiques et textuelles (OgDDASt), Waterloo, Québec, Canada
eval@oddas.ca

Sébastien BÉLAND
Université du Québec à Montréal, Québec, Canada
sebastien.beland.1@hotmail.com

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

Patricia BRASSARD
Université du Québec à Montréal, Québec, Canada
brisederose@hotmail.com

Patrick CHARLES
Université de Montréal, Québec, Canada
patrick.charles@umontreal.ca

Paul DE BOECK
University of Amsterdam, Pays-Bas
paul.deboeck@uva.nl

Valérie Léocadie DJÉDJÉ
Université du Québec à Montréal, Québec, Canada
djedje.valerie@uqam.ca

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

Paul MARTIN
Cégep de Sorel-Tracy, Québec, Canada
paul.martin@cegepst.qc.ca

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

Patrice POTVIN
Université du Québec à Montréal, Québec, Canada
potvin.patrice@uqam.ca

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

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

Fadia SAKR
Université du Québec à Montréal, Québec, Canada
fadia_sakr@sympatico.ca

Serge P. SÉGUIN
Université du Québec à Montréal, Québec, Canada
seguin.serge_p@uqam.ca

Theodore A. WALLS
University of Rhode Island, Kinston, Rhode Island, États-Unis
walls@uri.edu

RÉSUMÉS EN ANGLAIS

CHAPITRE 1

Une solution numérique au test de Cattell pour déterminer le nombre de composantes principales à retenir

Gilles Raîche, David Magis, Theodore A. Walls, Martin Riopel et Jean-Guy Blais

Determining how many primary components should be retained has always been a difficult task for psychometricians. Cattell suggests a graphical and subjective approach. It should be possible to develop a numerical solution to Cattell's screen test. Anindex is developed for this purpose based on the secondary derivative associated with each primary component, the so-called acceleration factor. Two applications of this index are also presented.

CHAPITRE 2

Validité de la précision de la mesure d'une stratégie de testing adaptatif informatisé (TAI) sous trois conditions de représentativité du domaine

Patrick Charles, Réjean Auger, Jean-Guy Blais et Serge P. Séguin

This research focuses on the validity of measurement in education. More specifically, it studies the precision of results obtained bycomputer adaptive testing (CAT) under three conditions of domain representativeness. Data from a conventional paper and pencil test of 156 students, of which 110 have been chosen at random for simulation purposes. The precision of scores has been measured usingthe root-mean-square error (RMSE) and the systematic error or bias. The results indicate that the measurements' precision increases substantially and that the biais decreases, when the number of items increases. Nevertheless, to improve the validity of the interpretation of test results, we question the relevance of invariably increasing the number of items in a test once domain representativeness has been established.

CHAPITRE 3

Comparaison empirique des méthodes classiques de détection du fonctionnement différentiel d'items en psychométrie

David Magis, Paul De Boeck et Gilles Raîche

A major issue in modern psychometrics is the detection of test items whose difficulty levels vary across several subgroups of subjects. Several methods have been proposed to detect those differentially functioning items; however, they were not yet investigated in a simultaneous comparative manner. The purpose of this paper is to perform a simulation study in order to bring additional information to that framework. The study findings suggest that no single method is more effective than the others, but significant methodological differences can be pointed out.

CHAPITRE 4

Variables de prédiction du niveau de difficulté de tâches d'évaluation comportant des équations du premier degré en mathématiques et en sciences au secondaire

Martin Riopel, Fadia Sakr, Gilles Raîche, Patrice Potvin et Valérie Léocadie Djédjé

Item generation is a research effort aimed at developing general models for predicting evaluation tasks parameters such as difficulty, based on other fundamental characteristics of these tasks. Once validated, these models could then be used to automatically generate adaptive new tasks (or items) to evaluate students. As presented in Irvine and Kyllonen (2002), most of the research on item generation focused on general cognitive tasks not specific to mathematics and sciences proficiency in schools. This paper presents the results of a study on predictive variables about evaluation tasks related to linear functions for secondary level students. Namely, the study focused on 100 items from the Banque d'instruments de mesure of the Gestion du réseau informatique des commissions scolaires (GRICS) society that were administered to 6910 students aged between 14 and 15 in 22 school boards in the Province of Quebec between 1996 and 2003. The proposed model is based on nine variables and can predict success rates with a correlation coefficient of 0.78. The model has been explicitly designed to be used for classifying existing items but also for generating new items. As a proof of concept, the proposed model has then been used to generate 864 different items with predicted success rates ranging from 0.04% (very difficult task) to 98.7% (very easy task). It could be used online by teachers or researchers to generate evaluation tasks or by computerized systems for adaptive evaluation.

CHAPITRE 5

Identification des patrons de réponses inappropriés à un test à partir des stratégies qui sous-tendent les comportements des répondants

Patricia Brassard, Sébastien Béland et Gilles Raîche

We will determine the underlying behavioural strategies in students who attempt to underachieve on a test. A protocol analysis of written reports will enable us to divide them into categories. Matches between response patterns and these categories will be verified using logistic regression for nominal data as well as cross-validation. We will identify categories of actual behaviours associated with inappropriate response patterns. Computer models of students who attempt to deliberately underachieve, that are more solid in terms of measure interpretation validity, can then be created.

CHAPITRE 6

Étude du comportement de 15 indices de détection de patrons de réponses inappropriés paramétriques et non paramétriques à partir d'une analyse par corrélations canoniques

Sébastien Béland, Patricia Brassard et Gilles Raîche

Some examinees behave inappropriately in testsituations. In these cases, the use of person-fit indices seems to be one of the best options because of their high detection rate. The comparison of these indices is generally limited to measuring the correlation between them. It would in fact be more appropriate to use factor analysis or canonical correlation analysis to obtain a better synthesis of these correlations. In this research, we compute the canonical correlations between the person parameters and some parametric and non-parametric person-fit statistics.

CHAPITRE 7

Utilisation de la théorie des ensembles flous pour valider une épreuve

Paul Martin et Jean-Guy Blais

The Fuzzy Set Theory was used to validate a test in mathematics that is part of the entrance exam at the École Polytechnique de Montréal. A model was designed to foresee the students' ability to succeed. The results demonstrate that this model is valid for predicting success, but not failure. In other words, this test is valid for selection but not for counselling purposes.

Collection Mesure et Evaluation

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

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

ONT COLLABORÉ À CET OUVRAGE

Réjean Auger
Sébastien Béland
Jean-Guy Blais
Patricia Brassard
Patrick Charles
Paul De Boeck
Valérie L. Djédjé
David Magis
Paul Martin
Karine Paquette-Côté
Patrice Potvin
Gilles Raîche
Martin Riopel
Fadia Sakr
Serge P. Séguin
Theodore A. Walls

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

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

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

www.puq.ca

9 782760 526853
ISBN 978-2-7605-2685-3