Le dépistage du sous-classement aux tests de classement en anglais, langue seconde, au collégial

Gilles Raîche

Le dépistage du sous-classement aux tests de classement en anglais, langue seconde, au collégial

Gilles Raîche

Collège de l'Outaouais

Septembre 2002

Ce projet de recherche a obtenu le soutien du programme d'aide à la recherche sur l'enseignement et l'apprentissage (PAREA) du ministère de l'Éducation du Québec.

Une copie électronique de ce document peut être obtenue à l'adresse suivante :
http //www.umoncton.ca/raicheg/ publications/

Dépôt Légal : .....
Bibliothèque nationale du Canada, 2002
Bibliothèque nationale du Québec, 2002
ISBN 2-9803398-2-2

Remerciements

Ce projet a été financé par le programme de recherche sur l'enseignement et l'apprentissage (PAREA) du ministère de l'Éducation du Québec, alors que l'auteur était conseiller pédagogique au collège de l'Outaouais. Plus spécifiquement, nous remercions Guy Demers, directeur du service de la recherche et du développement, et François Hardy, responsable du programme PAREA, pour avoir appuyé sa réalisation.

Nous tenons à remercier Martin Dandenault et Kevin Johnston, coordonnateurs du département des langues secondes au collège de l'Outaouais, les professeurs du département des langues secondes du collège de l'Outaouais, ainsi que les étudiants du groupe cours qui ont participé à la rencontre, pour leur précieuse collaboration.

Les travaux n'auraient toutefois pu être effectués sans la précieuse collaboration et le constant soutien de Pierrette Pelletier, technicienne en informatique au collège de l'Outaouais. Grâce à elle, à chaque année, nous avons pu obtenir et conserver les résultats individuels au TCALS II.

La collaboration de Richard Fillion, alors directeur des études par intérim au collège de l'Outaouais, a aussi été précieuse. Son constant soutien administratif a été fort apprécié.

Nous ne perdons pas de vue, aussi, que ce projet n'aurait pu naître sans les travaux préliminaires de création du test TCALS II effectués par Michel Laurier, de l'Université de Montréal, de Charles Pearo, du collège André-Laurendeau, de Lydia Froio, du collège de Maisonneuve, ainsi que de Michel Fournier, étudiant au doctorat en éducation à l'Université de Montréal. La qualité de ces travaux a permis l'élaboration d'un test qui tient encore parfaitement la route, après tant d'années, et dont les caractéristiques métrologiques ont sûrement été un gage de réussite quant à l'efficacité de la procédure de détection proposée.

Enfin, nous tenons à remercier Jean-Guy Blais, professeur à la Faculté des sciences de l'éducation de l'Université de Montréal, pour sa participation à la diffusion des résultats, Christian Démoré, étudiant à la maîtrise à l'École de psychologie de l'Université de Moncton, pour sa collaboration à l'analyse des caractéristiques métrologiques du TCALS II, François Vigneau, professeur à l'École de psychologie de l'Université de Moncton, et Douglas Bors, professeur à la Faculté de psychologie de l'Université de Toronto, pour les échanges fructueux, ainsi qu'Andrée Cantin, du cégep régional de Lanaudière, et Michel Coutu, de Mémotexte, pour leurs commentaires et conseils quant à la rédaction du rapport.

Gilles Raîche
Professeur, Université de Moncton
Faculté des sciences de l'éducation

Sommaire

Ce rapport présente les résultats d'une recherche ayant pour objectif de développer des indices d'ajustement inadéquat spécifiques au dépistage du sous-classement intentionnel par les étudiantes et les étudiants du réseau collégial au test de classement en anglais, langue seconde, le TCALS II. La recherche a été menée au collège de l'Outaouais au cours de l'année 2001-2002 et s'est déroulée en plusieurs étapes.

Dans un premier temps, considérant que le test avait été élaboré plusieurs années auparavant, il a été jugé nécessaire de vérifier de nouveau ses caractéristiques métrologiques et d'évaluer s'il pouvait être considéré adapté à une clientèle d'étudiantes et d'étudiants dont le niveau d'habileté en anglais, langue seconde, est plutôt élevé, comme c'est le cas au collège de l'Outaouais. Le TCALS II semble, en général, efficace avec cette clientèle étudiante. Toutefois, aux niveaux supérieurs d'habileté langagière, le TCALS II est beaucoup moins précis. En fait, il n'est constitué d'aucun item dont le niveau de difficulté est suffisamment élevé pour mesurer la compétence langagière de ces étudiantes et de ces étudiants. Pour cette raison, nous recommandons l'ajout de questions d'un niveau de difficulté plus élevé au test.

Dans un second temps, nous avons vérifié directement auprès des étudiantes et des étudiants quelles stratégies ils utiliseraient éventuellement s'ils désiraient se sous-classer au TCALS II. Quatre stratégies ont été signalées par eux : répondre au hasard, donner une mauvaise réponse, omettre de répondre et répéter le même choix de réponse.

Principalement pour détecter les deux premières stratégies de sous-classement, des indices basés sur des modélisations issues de la théorie de la réponse à l'item ont été, soit choisis, soit élaborés. Les indices choisis ont été l_z, provenant des travaux de Drasgow, Levine et Rubin (Drasgow et Levine, 1986; Levine et Rubin, 1979), *W*, utilisé principalement avec le modèle de Rasch (Wright et Stone, 1982), *Zeta*, développé par Tatsuoka (1996) dans un contexte d'évaluation diagnostique, et χ_B^2, indice intégré au logiciel Bilog. Nous avons aussi élaboré des indices qui permettraient, en

théorie, de détecter spécifiquement, soit un patron de réponses au hasard, soit un patron de réponses intentionnellement fausses : I_{hasard}, $I_{inversé}$ et I_{sous}.

Pour évaluer l'efficacité de ces indices, nous avons procédé, dans un troisième temps, à des simulations d'étudiantes ou d'étudiants qui auraient tenté de se sous-classer au TCALS II selon une de ces deux stratégies. Nous avons réalisé 1 000 simulations selon 17 niveaux d'habileté différents variant entre –2,00 et 2,00 par sauts de 0,25 et selon quatre situations différentes : répondre au hasard à 10 % des premières questions, répondre au hasard à 20 % des premières questions, donner une mauvaise réponse à 10 % des premières questions et donner une mauvaise réponse à 20 % des premières questions. De façon globale, l'indice l_z a été jugé le plus efficace de tous. L'identification des étudiantes et des étudiants des niveaux de classement 3 et 4, soit les plus forts, est presque parfaite dans tous les cas. Il permet même une identification relativement adéquate aux niveaux 1 et 2. Au collège de l'Outaouais, l'indice l_z a permis l'identification d'environ 10 % des étudiantes et des étudiants de la cohorte 2002-2003 admis et inscrits au premier et second tours du Service régional d'admission du Montréal métropolitain (SRAM).

Cette recherche nous amène à recommander l'utilisation de l'indice l_z pour identifier les étudiantes et les étudiants qui auraient éventuellement tenté de se sous-classer au test. Cette identification doit toutefois être suivie de mécanismes de vérification. Par exemple, les étudiantes et les étudiants dont le résultat au test est douteux pourraient être convoqués à une seconde séance d'évaluation de leur compétence langagière. C'est ce qui a été fait au collège de l'Outaouais. Nous recommandons aussi aux collèges de mettre en place un mécanisme simple pour identifier les étudiantes et les étudiants qui omettent de répondre ou répètent consécutivement un même choix de réponse à plus de 10 % des questions du test.

Table des matières

Chapitre 1

Alerte! Des étudiants cherchent à rater leur examen!

Depuis plusieurs années, le personnel enseignant du département des langues des collèges québécois cherche à améliorer les procédures de classement des étudiantes et des étudiants aux divers niveaux de formation en anglais et en espagnol, langues secondes. Cette quête de la validité des procédures de classement en langues secondes a pris une plus grande importance avec le renouvellement de la formation générale dans le réseau collégial québécois qui a introduit deux cours obligatoires (ministère de l'Éducation, 1993). Dorénavant, pour obtenir leur diplôme d'études collégiales, tous les étudiantes et les étudiants doivent avoir réussi, en langues secondes, un cours de formation générale commune ainsi qu'un cours de formation générale propre à leur programme d'études. De plus, certaines étudiantes et certains étudiants choisissent un ou des cours de langues secondes à l'intérieur de la formation spécifique à leur programme d'études. C'est le cas notamment du programme *Sciences, lettres et arts*, ainsi que des variations institutionnelles du profil *Lettres* ou du profil *Langues* du programme *Arts et lettres*. La problématique du classement des élèves aux divers niveaux de formation en langues secondes touche donc tous les étudiantes et étudiants du réseau collégial, tous les programmes d'études et, conséquemment, tous les collèges publics et privés.

Il va sans dire que la détermination du niveau d'habileté dans une langue seconde spécifique est importante pour permettre à l'étudiante ou à l'étudiant de recevoir un enseignement approprié à son niveau de connaissances. Elle est aussi cruciale pour permettre de constituer des groupes classes homogènes quant au niveau d'habileté langagière des étudiantes ou des étudiants. À ce moment, l'enseignant peut adapter ses interventions pédagogiques en fonction des besoins d'apprentissage spécifiques des étudiantes ou des étudiants et ces derniers sont alors appelés à fournir des efforts similaires, à percevoir un degré de difficulté du cours comparable et à manifester un niveau d'attention semblable de l'un à l'autre.

1.1 Le sous-classement

Toutefois, selon Fournier (1992) et Raîche (2000), certaines étudiantes et certains étudiants ratent intentionnellement des questions au test de classement dans le but de se retrouver dans un cours plus facile et ainsi, soit se la couler douce, soit obtenir de meilleurs résultats. C'est ce que nous désignons sous le terme de sous-classement. Il faut souligner que, puisque aucune mesure de la manifestation de ce comportement n'a été développée, l'ampleur du phénomène est toutefois difficile à évaluer et ne repose que sur des appréciations subjectives de la part du personnel des collèges. Lorsque des étudiantes et des étudiants réussissent à se sous-classer à un test de classement en langue seconde, le groupe classe ne peut plus être considéré comme homogène et le climat de classe peut en être affecté. Des problèmes d'inattention surgissent et certaines étudiantes ou certains étudiants adoptent des comportements perturbateurs; les notions abordées dans le cours étant trop faciles pour eux, ils affectent éventuellement le déroulement normal du cours.

Fait à noter, selon certains, ce problème de sous-classement intentionnel aux tests de langues serait aussi présent à l'intérieur des collèges communautaires du Nouveau-Brunswick ainsi qu'à l'Université de Moncton (Raîche, 2002), tout comme à l'École de langues du Gouvernement fédéral canadien. Dans un tout autre contexte, On nous a aussi signalé qu'un tel problème de sous-classement existe aussi en Hollande lorsque vient le moment d'administrer des tests d'admission à l'armée. Il est aussi possible de retrouver ce comportement chez la clientèle bénéficiaire de l'assurance-emploi canadienne lorsque celle-ci suit des cours de mise à niveau au secondaire : certains considéreraient alors qu'ils ont tout avantage à demeurer le plus longtemps possible en formation. Le problème du sous-classement pourrait être éventuellement beaucoup plus fréquent qu'on pourrait le croire à prime abord et pourrait être associé à bien d'autres contextes que celui de l'évaluation de la compétence langagière dans les collèges.

1.2 Le TCALS II, une solution?

Le personnel enseignant des départements de langues, qui doit composer avec ce problème, cherche depuis longtemps des moyens simples et efficaces pour, sinon contrer, du moins amenuiser ce comportement de la part des étudiantes et des étudiants. Plusieurs enseignants ont cru que la solution au problème du sous-classement résidait dans l'élaboration d'un test de classement plus efficace dans l'appréciation du niveau d'habileté de l'étudiante ou de l'étudiant. Malheureusement, comme le démontrent les résultats consécutifs à l'élaboration par Laurier, Froio, Pearo et Fournier (1998) d'un test de classement amélioré en anglais, langue seconde, soit le TCALS II, les étudiantes et les étudiants du réseau collégial continuent à tenter de se sous-classer à ce test. D'autres solutions doivent donc être envisagées.

1.3 Des solutions prometteuses

Tenant compte du grand nombre d'administration des tests de classement en anglais, langue seconde, dans chacun des collèges, soit plus de mille par an à l'intérieur d'un collège de taille moyenne, il est nécessaire d'utiliser une formule simple pour dépister les tentatives de sous-classement. Est-ce que l'analyse de la cohérence des réponses à chacune des questions d'un test de classement ne permettrait pas de détecter des patrons de réponses plutôt suspects et ainsi peu probables? Certains auteurs se sont intéressés à cette problématique et ont proposé des indices numériques de scores déviants (*appropriateness measurement, person fit, caution indices*) à un test (Birenbaum, 1985, 1986; Bracey and Rudner, 1992; Drasgow, 1982; Reise et Flannery, 1996). Selon nous, c'est l'une des pistes les plus prometteuse à envisager pour pallier ce problème. Maintes mesures d'ajustement inadéquat ont été proposées. Ces mesures ne s'adressent toutefois pas au problème spécifique du sous-classement. Habituellement, elles visent plutôt à déterminer de façon globale le mauvais ajustement des réponses au test sans se soucier de la nature du problème : copie, réponse au hasard, mauvaise utilisation de la feuille de réponses, etc. Il est donc nécessaire de développer un indice, ou des indices, d'ajustement inadéquat spécifiques au dépistage du sous-classement. C'est l'objectif de cette recherche.

1.4 Déroulement de la recherche

La première étape de la recherche est réalisée auprès de toutes les étudiantes et tous les étudiants inscrits au collège de l'Outaouais au cours de l'année scolaire 2001-2002 à qui le test de classement en anglais, langue seconde, est administré. Selon les données de 2000-2001, cette population devait compter environ 1300 étudiantes et étudiants. Les résultats actuels dénombrent en fait 1415 étudiantes et étudiants pour l'année 2001-2002. La moitié de ces étudiantes et ces étudiants suivent leur premier cours d'anglais, langue seconde, à l'automne 2001, tandis que l'autre moitié le suivent à l'hiver 2002. Cette première étape de notre recherche a pour objectif de déterminer les paramètres de chacune des questions du TCALS II à partir d'une modélisation issue de la théorie de la réponse à l'item (TRI), soit le modèle à un paramètre de Rasch (1960). La connaissance de ces paramètres associés à chacune des questions nous permet d'étudier les qualités métrologiques du test, d'évaluer son adéquation à une population qui devrait être plus habile en anglais et, enfin, de mettre en oeuvre une première tentative de dépistage des étudiantes et des étudiants qui cherchent à se sous-classer au test. Le chapitre deux présente cette démarche.

La seconde étape du projet de recherche, décrite au chapitre trois, consiste à vérifier directement auprès des étudiantes et des étudiants quelles sont les stratégies qu'ils utiliseraient s'ils désiraient se sous-classer au test. À cette fin, un groupe d'étudiants est rencontré au collège de l'Outaouais et nous leur demandons, autant verbalement que par écrit, de nous décrire les stratégies qu'ils pourraient éventuellement utiliser pour se sous-classer. Il s'agit d'une étape exploratoire de la recherche qui devrait nous éviter de ne pas tenir compte de stratégies de sous-classement importantes.

Le chapitre quatre, portrait de la troisième étape de la recherche, s'adresse plus spécifiquement à l'élaboration d'indices qui devraient permettre de dépister les étudiantes et les étudiants qui affichent un comportement de sous-classement à un test. Nous abordons le sujet par la démonstration des difficultés rencontrées par le personnel enseignant quant il cherche à détecter directement en classe les étudiantes et les étudiants qui ont sous-performé au test de classement. Nous arrivons à la même conclusion que celle retrouvée dans la littérature sur la sous-performance des surdoués : le personnel

enseignant est un des plus mauvais juges pour dépister de tels étudiantes et étudiants. Il nous semble donc encore plus pertinent de développer des indices de dépistage performants. Le chapitre quatre est toutefois assez technique pour le lecteur peu enclin aux formulations mathématiques et il peut décider de remettre sa lecture à plus tard. Il pourra y revenir par la suite pour mieux saisir le fonctionnement des stratégies de détection. Suite à une consultation de la littérature, les indices retenus sont tous tributaires des modélisations issues de la théorie de la réponse à l'item. Pour cette raison, les modélisations les plus fréquentes sont présentées. Elles sont accompagnées de leur formulation mathématique ainsi que des méthodes utilisées pour estimer le niveau d'habileté de l'étudiante ou de l'étudiant. Les indices de détection et les équations associées terminent ce chapitre.

Au chapitre cinq, les indices de dépistage du sous-classement sont mis à l'épreuve. Premièrement, des étudiantes et des étudiants qui répondent de façon honnête au test sont simulés dans le but de déterminer la valeur critère à appliquer à chacun des indices retenus pour détecter le comportement de sous-classement. Ces valeurs sont ensuite mises à l'épreuve sur des simulations d'étudiantes et d'étudiants qui utilisent des stratégies pour se sous-classer au test. Nous pouvons ainsi évaluer le pourcentage d'étudiantes et d'étudiants que ces indices nous permettent de dépister correctement. Nous serons alors en mesure d'évaluer la validité de la procédure de dépistage des étudiantes et des étudiants qui ont tenté de se sous-classer au test de classement en anglais, langue seconde. Le chapitre cinq se termine sur l'application de ces résultats au dépistage du comportement de sous-classement pour la cohorte des étudiants des 1er et 2^{e} tours du SRAM qui sont inscrits au Collège de l'Outaouais pour l'année scolaire 2002-2003.

Nous tenons à préciser que, malgré le risque d'alourdir le texte, nous avons, autant que possible, présenté le détail des équations utilisées. De cette façon, il sera non seulement plus facile de reproduire la recherche, mais aussi plus simple d'appliquer et d'informatiser les procédures statistiques proposées.

Enfin, nous terminons ce rapport par de nouvelles pistes à explorer ainsi que par des recommandations quant aux stratégies à utiliser pour dépister les étudiantes et les étudiants qui

cherchent à se sous-classer lors de l'administration du TCALS II. Ces recommandations tiennent aussi compte des actions à envisager lorsqu'une étudiante ou un étudiant potentiellement sous-performant a été détecté.

Chapitre 2

Le test de classement en anglais, langue seconde, au collégial

Au collège de l'Outaouais, la version longue du TCALS II est utilisée depuis 1998-1999. Elle a été élaborée par Laurier, Froio, Pearo et Founier (1998) grâce à l'appui du programme d'aide à la recherche sur l'enseignement et l'apprentissage (PAREA) de la Direction générale de l'enseignement collégial du ministère de l'Éducation du Québec. Auparavant, les procédures de classement en anglais, langue seconde, variaient considérablement d'un collège à un autre rendant totalement impossible l'équivalence inter-institutionnelle du classement.

L'élaboration a débuté à l'automne 1995 et l'instrument a été disponible dans les collèges à partir du printemps 1997. Il a été construit en fonction des compétences langagières visées par la formation générale telles que redéfinies par le ministère de l'Éducation en 1993. Son élaboration a de plus succédé à une vaste consultation auprès des intervenants du réseau collégial québécois : directions des études, coordinations départementales en langues secondes et personnel enseignant. Depuis, le test est à l'essai et aucune recherche n'a encore été réalisée pour vérifier l'adéquation de son application. L'étude actuelle correspond donc, aussi, à une évaluation partielle de l'application du test, cinq ans après son introduction à l'intérieur du réseau collégial québécois.

Cette version longue du TCALS II est composée de 85 items à quatre choix de réponses et de 8 sous-tests présentés séquentiellement lors de l'administration du test. Ces sous-tests sont eux-mêmes divisés en deux catégories. La première de ces catégories est celle de la compréhension auditive où l'étudiante ou l'étudiant doit donner une réponse suite à l'écoute d'un exposé oral pré-enregistré. La seconde catégorie est celle de la compréhension écrite de l'anglais : l'étudiante ou l'étudiant doit donner sa réponse suite à la lecture de la question.

Le tableau 2.1 illustre la composition du test, dont le nombre de questions pour chacun des sous-tests. Nous y remarquons une plus grande couverture du domaine de la compréhension écrite, soit

52 questions pour la compréhension écrite contre 33 questions pour la compréhension auditive. La qualité des conditions d'écoute des exposés oraux pouvant être éventuellement affectée par des facteurs divers qui ne sont pas toujours facilement contrôlables, il nous semble opportun d'utiliser plus de questions en compréhension écrite qu'en compréhension auditive.

Tableau 2.1
Composition de la version longue du test de classement en anglais, langue seconde, au collégial (TACLS II)

Sous-tests	Nombre de questions	Pourcentage du test*
Compréhension auditive	33	39 %
Phrases	12	14 %
Dialogues	13	15 %
Mini-exposés	8	9 %
Compréhension écrite	52	61 %
Vocabulaire	13	15 %
Grammaire	17	20 %
Analyse d'erreurs	7	8 %
Lecture 1	7	8 %
Lecture 2	8	9 %

* L'arrondissement des valeurs peut faire en sorte que la somme des pourcentages ne corresponde pas exactement aux totaux

La fonction du test est strictement de classer les étudiantes et les étudiants à un niveau approprié d'apprentissage en anglais, langue seconde. De cette façon, on peut s'attendre à créer des groupes classes plutôt homogènes où les étudiantes et les étudiants pourront développer leurs habiletés langagières de façon optimale. L'étudiante ou l'étudiant peut être classé selon cinq niveaux différents de compétence en anglais, langue seconde, du plus facile, soit la mise à niveau, au plus difficile, soit le niveau 4. Laurier et ses collaborateurs ont aussi ajusté les scores de césures pour chacun des niveaux de façon à ce que la distribution des étudiantes et des étudiants soit convenable pour l'ensemble des collèges. Suite à des consultations à l'intérieur du réseau collégial, ces scores ont toutefois dû être ajustés de nouveau en 1998-1999. Le tableau 2.2 indique la valeur des scores de césures sur la base du nombre total de bonnes réponses au test (total minimal et total maximal). Il s'agit de l'échelle de mesure utilisée dans les collèges depuis

1998-1999. Les scores de césure sont aussi indiqués sur la base d'une échelle de mesure que nous utilisons à l'intérieur de la recherche ($\hat{\theta}$ minimal et $\hat{\theta}$ maximal). Cette échelle est bâtie autour d'une modélisation plus adaptée aux travaux de recherche que nous présenterons au chapitre quatre. Il s'agit de la modélisation logistique à un paramètre de Rasch. On peut remarquer que cette nouvelle échelle de mesure n'est plus bornée par 0 et 85, mais plutôt par $-\infty$ et ∞. Le tableau 2.2 nous permet ainsi d'assurer la correspondance entre ces deux échelles de mesure tout au long du texte.

Tableau 2.2
Scores de césure retenus pour la version longue du test de classement en anglais, langue seconde, au collégial (TACLS II)

Niveau	Total minimal	Total maximal	$\hat{\theta}$ minimal	$\hat{\theta}$ maximal
Mise à niveau	0	37	$-\infty$	-1,26
1	38	48	-1,20	-0,59
2	49	66	-0,53	0,49
3	67	79	0,55	1,27
4	80	85	1,33	∞

2.1 Les paramètres d'items

La calibration des items est réalisée à partir des résultats obtenus par les étudiantes et les étudiants des quatre tours du Service régional d'admission du Montréal métropolitain (SRAM) de la cohorte inscrite au collège de l'Outaouais en 1998-1999. Le niveau de difficulté moyen des items est de –1,14 avec un écart-type de 0,65. Le niveau de difficulté des items de la section relative à la compréhension auditive affiche une moyenne de –1,10, tandis que celui de la section relative à la compréhension écrite affiche une moyenne de –1,14. Le niveau de difficulté des deux sections du test est donc comparable. Il en est aussi de même en ce qui a trait à l'écart-type : 0,69 et 0,63 respectivement. C'est la sous-section *Lecture 1* qui est constituée des items les plus

difficiles, soit –0,75 en moyenne et représentative du niveau 1, tandis qu'à la sous-section *Vocabulaire* on retrouve les items les plus faciles, soit une moyenne de –1,63 qui est représentative de la mise à niveau. Nous avons aussi vérifié la stabilité des paramètres d'items au cours des années, soit de 1998-1999 à 2002-2003. Les moyennes et écarts-types sont similaires tandis que la corrélation de Pearson entre eux est toujours supérieure à 0,96. Il nous semble donc sensé de considérer ces paramètres stables dans le temps.

Les items peuvent être considérés très faciles pour la plupart des étudiantes et des étudiants de cette cohorte. En se référant au tableau 2.2, on peut ainsi constater que la moyenne globale (-1,14) correspond à des items qui permettent de classer des étudiantes et des étudiants au niveau 1. C'est à ce niveau que le test serait le plus utile, ce que confirmera, un peu plus loin, l'analyse de l'information donnée par le test à différents niveaux d'habileté en anglais, langue seconde.

Les tableaux 2.3 et 2.4 présentent la valeur du paramètre de difficulté pour chacun des 85 items du TCALS II. Le tableau 2.3 présente ces valeurs pour la section de la compréhension auditive, tandis que le tableau 2.4 s'adresse à la compréhension écrite. Nous y remarquons qu'aucun item n'affiche un niveau de difficulté supérieur à 0,49 et qu'ainsi aucun item ne correspond à un niveau de difficulté qui soit supérieur au niveau de classement 2. Les niveaux 3 et 4 ne sont donc pas du tout couverts par les items du test. À ces niveaux, la précision du test est ainsi affectée. Il semble donc qu'il serait nécessaire d'ajouter des items plus difficiles au test. La figure 2.1 nous permet d'ailleurs d'observer la fréquence avec laquelle chaque classe du niveau de difficulté de l'item est couverte.

Tableau 2.3
Valeur du paramètre de difficulté pour chacun des 33 items de la section de compréhension auditive de la version longue du test de classement en anglais, langue seconde, au collégial (TACLS II)

Item	b	Item	b	Item	b	Item	b
Phrases (-1,26)*		10	-0,41	19	-1,34	28	-1,14
1	-2,23	11	-0,16	20	-1,33	29	-1,26
2	-2,07	12	-0,29	21	-0,68	30	-0,72
4	-2,40	Dialogues (-1,15)		22	-0,95	31	-0,41
4	-1,82	13	-1,62	23	-0,76	32	-0,52
5	-1,42	14	-2,40	24	-0,29	33	-1,19
6	-1,48	15	-1,51	25	0,42		
7	-1,42	16	-1,63	Mini-exposés (-0,76)			
8	-0,78	17	-1,49	26	-0,79		
9	-0,66	18	-1,41	27	-0,07		

* La valeur entre parenthèses représente la moyenne de la sous-section

Tableau 2.4
Valeur du paramètre de difficulté pour chacun des 52 items de la section de compréhension écrite de la version longue du test de classement en anglais, langue seconde, au collégial (TACLS II)

Item	b	Item	b	Item	b	Item	b
Vocabulaire (-1,63)*		Grammaire (-1,14)		60	-0,66	72	-1,06
34	-2,40	47	-1,68	61	-0,48	73	-0,80
35	-1,98	48	-1,39	62	-0,28	74	-0,56
36	-2,23	49	-1,86	63	0,00	75	-0,63
37	-2,10	50	-1,66	Analyse d'erreurs (-1,16)		76	-0,55
						77	0,34
38	-1,68	51	-1,26	64	-1,94	78	-1,28
39	-1,67	52	-1,23	65	-1,79	Lecture 2 (-0,84)	
40	-1,72	53	-1,44	66	-1,55	79	-1,22
41	-1,34	54	-1,46	67	-1,09	80	0,54
42	-2,15	55	-1,26	68	-0,83	81	-0,62
43	-1,16	56	-1,47	69	-0,53	82	-1,01
44	-1,23	57	-1,11	70	-0,36	83	-1,40
45	-0,75	58	-1,48	Lecture 1 (-0,75)		84	-0,89
46	-0,84	59	-0,70	71	-1,48	85	-1,26

* La valeur entre parenthèses représente la moyenne de la sous-section

On remarquera aussi que dans plusieurs des sous-tests, le niveau de difficulté des items est plus bas au début de la section. C'est une caractéristique intéressante pour l'ensemble du test, puisque, selon plusieurs, il est préférable d'administrer des items plus faciles au début d'un test pour maintenir la motivation et l'intérêt de l'étudiante ou l'étudiant. Toutefois, est-ce que cette stratégie d'élaboration d'un test doit être appliquée à chacun de ses sous-tests? Ne risque-t-on pas plutôt de perdre alors cet effet éventuellement bénéfique de l'augmentation graduelle du niveau de difficulté des items dans le test?

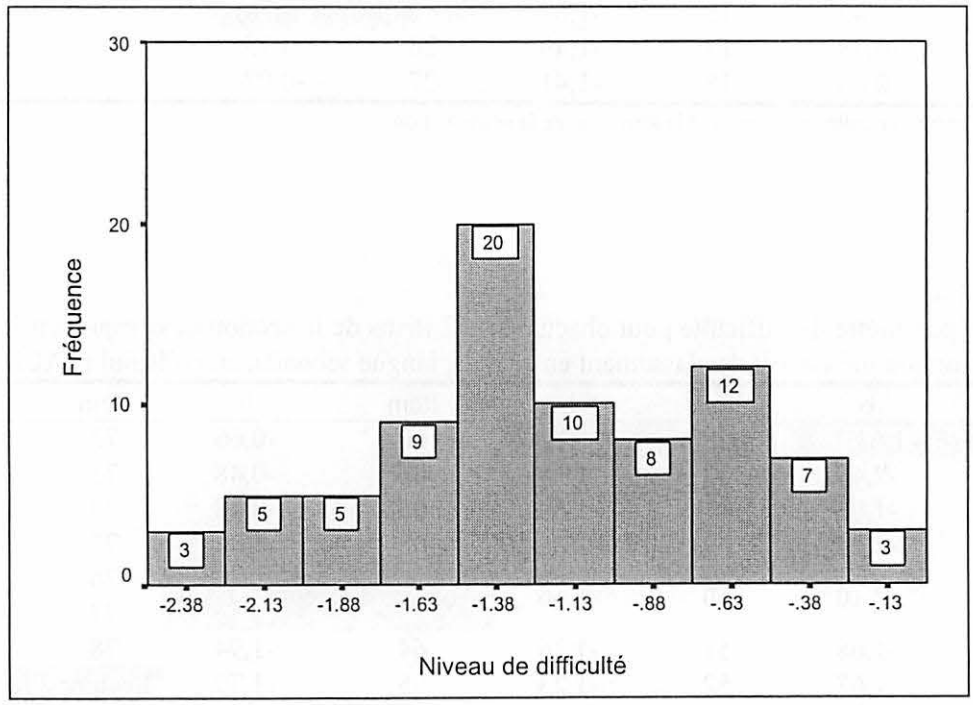

Figure 2.1 Fréquence avec laquelle chaque classe du niveau de difficulté de l'item est couverte dans le test de classement en anglais, langue seconde, au collégial (TCALS II)

2.2 La distribution de probabilité de l'estimateur du niveau d'habileté

Nous voulions vérifier si le TCALS II, malgré ses problèmes de représentativité des items aux niveaux supérieurs d'habileté, pouvait tout de même permettre d'estimer le niveau d'habileté en anglais, langue seconde, avec suffisamment de précision. Il était impossible de réaliser cette vérification à partir de données réelles car nous ne connaissons malheureusement pas à ce moment le niveau d'habileté réel de l'étudiante ou de l'étudiant : nous ne pouvons que l'estimer. C'est pourquoi, à partir d'une simulation, nous avons aussi analysée la distribution de la probabilité de chaque valeur de l'estimateur du niveau d'habileté, $\hat{\theta}$, en fonction du niveau d'habileté, θ, d'une étudiante ou d'un étudiant. Pour 17 niveaux d'habileté, variant entre –2,00 et 2,00 par saut de 0,25, nous avons effectué 1 000 simulations de l'administration du TCALS II. Ces valeurs du niveau d'habileté couvrent bien tous les niveaux d'habileté visés par le test. Nous avons utilisé un simulateur de tests maintes fois éprouvé à l'intérieur de recherches antérieures sur le testing adaptatif (Blais et Raîche, 2002a, 2002b ; Raîche, 2001a, 2001b, 2002b ; Raîche et Blais, 2002a). La programmation a été faite à partir du logiciel SAS dans sa version 6.0 (1996). Il est aussi important de signaler que nous avons utilisé une méthode d'estimation du niveau d'habileté récemment développée qui permet d'obtenir un biais presque nul de cet estimateur quel que soit le niveau d'habileté dans un contexte de testing adaptatif (Raîche et Blais, 2002b, 2002c). Son application dans le contexte d'un test fixe est toutefois nouvelle.

Pour décrire la distribution de probabilité de l'estimateur du niveau d'habileté en fonction du niveau d'habileté, nous avons retenu la moyenne, l'écart-type, les minimums et maximums ainsi que les coefficients d'asymétrie et de kurtose. Notons que les valeurs de ces statistiques ont été obtenues empiriquement à partir des simulations et non pas selon des estimations asymptotiques fréquemment utilisées dans les travaux sur les modélisations issues de la théorie de la réponse à l'item.

Les résultats de cette simulation sont présentés au tableau 2.5. On peut y noter que lorsque le niveau d'habileté, θ, est inférieur à 0,50, la moyenne de l'estimateur du niveau d'habileté, $\hat{\theta}$, est toujours à peu près égale, sinon égale, au niveau d'habileté : en fait, la différence est alors d'au plus 0,06. Elle devient beaucoup plus importante par la suite, quoique tout de même assez raisonnable. Ainsi, entre des valeurs de 0,75 et 1,75 du niveau d'habileté, la différence peut atteindre jusqu'à 0,19. Les valeurs obtenues de la différence entre le niveau d'habileté et son estimateur démontrent encore une fois l'efficacité de la méthode d'estimation développée par Raîche et Blais (2002b, 2002c), principalement aux valeurs extrêmes du niveau d'habileté.

Tableau 2.5

Distribution de l'estimateur du niveau d'habileté ($\hat{\theta}$) en fonction du niveau d'habileté (θ) au test de classement en anglais, langue seconde, au collégial (basée sur les paramètres d'items de la cohorte 1998-1999)

θ (Total)*	$\hat{\theta}$	$S_{\hat{\theta}}$	$Min_{\hat{\theta}}$	$Max_{\hat{\theta}}$	$Asymétrie_{\hat{\theta}}$	$Kurtose_{\hat{\theta}}$
-2,00 (26)	-2,03	0,17	-2,66	-1,56	-0,31	0,23
-1,75 (30)	-1,76	0,16	-2,38	-1,31	-0,21	0,10
-1,50 (34)	-1,50	0,15	-1,91	-0,99	0,10	0,29
-1,25 (38)	-1,26	0,15	-1,75	-0,85	-0,04	-0,19
-1,00 (42)	-0,99	0,15	-1,45	-0,32	0,17	0,48
-0,75 (46)	-0,75	0,15	-1,24	-0,18	0,10	0,52
-0,50 (49)	-0,49	0,15	-0,96	-0,02	0,17	-0,06
-0,25 (53)	-0,23	0,17	-0,70	0,32	0,11	-0,01
0,00 (57)	0,03	0,20	-0,54	0,71	0,30	0,18
0,25 (61)	0,29	0,22	-0,28	1,22	0,40	0,56
0,50 (65)	0,56	0,27	-0,18	1,92	0,80	1,23
0,75 (69)	0,88	0,35	0,10	2,30	0,79	0,95
1,00 (73)	1,16	0,38	0,32	2,30	0,62	1,18
1,25 (77)	1,43	0,42	0,49	2,30	0,37	-0,61
1,50 (80)	1,69	0,41	0,71	2,30	-0,03	-0,99
1,75 (84)	1,87	0,37	0,71	2,30	-0,50	-0,61
2,00 (85)	2,03	0,32	1,01	2,30	-0,90	0,01

* La valeur entre parenthèses correspond au score total équivalent

Autre indicateur de la précision de l'estimateur du niveau d'habileté, l'erreur-type de l'estimateur du niveau d'habileté, $S_{\hat{\theta}}$, est inférieure à 0,30 lorsque le niveau d'habileté varie entre –2,00 et 0,50. Dans bien des cas l'erreur-type est inférieure à 0,20. Par la suite l'erreur-type, varie entre 0,32 et 0,42. Ce fait n'est pas surprenant puisque nous avions remarqué plus haut que le TCALS II est constitué d'items dont le niveau de difficulté est d'au plus 0,49. C'est donc dire qu'au-dessus d'un niveau d'habileté de 0,50, le TCALS II est définitivement moins précis. Nous tenons à souligner que pour les fins de cette recherche, c'est justement à ces niveaux supérieurs que nous souhaiterions réaliser une estimation précise du niveau d'habileté en anglais, langue seconde, de l'étudiante ou de l'étudiant.

Quant aux minimums et maximums de l'estimateur du niveau d'habileté, on remarque que leur étendue augmente généralement avec l'augmentation de la valeur de l'erreur-type correspondante. Les valeurs des coefficients d'asymétrie et de kurtose sont à peu près toujours inférieurs à 1,00 en valeur absolue. Dans le cas du coefficient d'asymétrie, cette valeur critique de 1,00 en valeur absolue n'est jamais dépassée. On peut donc dire qu'en général, sur la base des ces coefficients, le postulat de normalité de ces distributions peut être généralement maintenu. Le seul niveau d'habileté où le coefficient de kurtose affiche une valeur qui pourrait poser des problèmes sérieux d'interprétation est 1,50 : le coefficient de kurtose est alors presque égal à -1,00. La distribution d'échantillonnage de l'estimateur du niveau d'habileté est donc fortement aplatie et l'intervalle de confiance associé à l'erreur-type, déjà importante (0,41), est donc considérablement sous-estimé.

2.3 La fidélité et l'unidimensionnalité du test

Selon les travaux de Laurier, Froio, Pearo et Fournier (1998), dans sa version longue, le TCALS II, ainsi que chacun de ses sous-tests, démontrent un niveau de fidélité très satisfaisant. Le coefficient *alpha* de Cronbach est égal à 0,96 pour la globalité du test. Cette valeur est assez importante et dénote de la précision globale du test quant à l'estimation du niveau d'habileté. Le

coefficient *alpha* varie de 0,48 à 0,81 selon les sous-tests concernés du test. C'est le sous-test Lecture 1 qui affiche la valeur la plus faible, soit 0,48, tandis que c'est le sous-test Grammaire qui présente la valeur la plus élevée, soit 0,84. Ces constatations étaient prédictibles puisque la valeur du coefficient *alpha* de Cronbach est tributaire du nombre d'items. Le sous-test Lecture 1 est constitué de seulement 7 items, contre 17 pour le sous-test Grammaire.

Le coefficient *alpha* de Cronbach étant une mesure globale de la précision du test, nous ne pouvons pas l'utiliser pour analyser la précision du test en fonction du niveau d'habileté de l'étudiante ou de l'étudiant. C'est à partir de la fonction d'information à chaque niveau d'habileté que nous pouvons faire cette analyse. La figure 2.2 présente la variation de l'information donnée par le TCALS II en fonction du niveau d'habileté. On peut y remarquer que l'information est maximale entre des valeurs approximatives du niveau d'habileté de –1,75 et de –0,50. Nous avions d'ailleurs souligné, plus haut, que le TCALS offre de meilleures performances dans l'estimation du niveau d'habileté lorsque le niveau d'habileté de l'étudiante ou de l'étudiant était inférieur ou égal à 0,50. À titre indicatif, on peut obtenir une approximation de l'information donnée par un test à un niveau d'habileté fixé par une fonction utilisant l'erreur-type de l'estimateur du niveau d'habileté, soit :

$$I = \frac{1}{s_\theta^2}. \qquad \textbf{(Équation 2.1)}$$

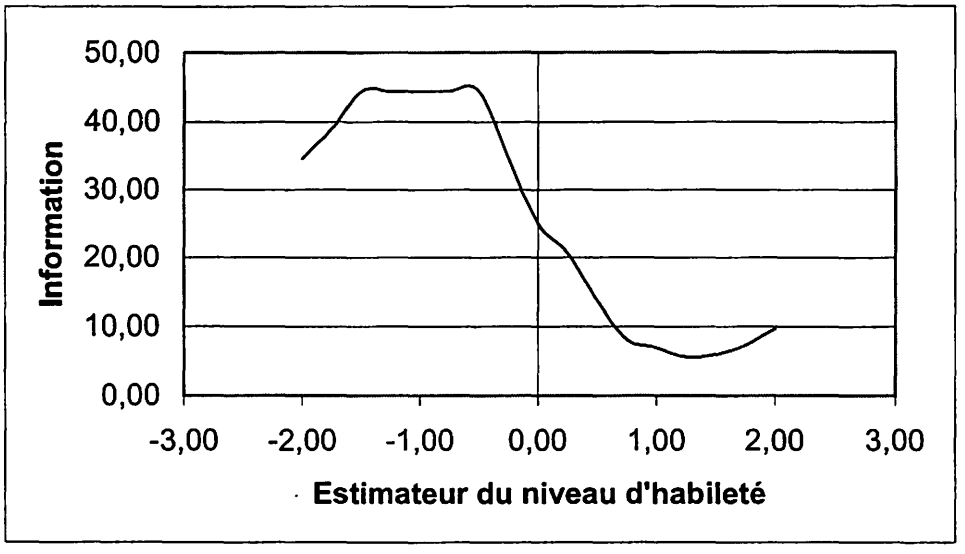

Figure 2.2 Information à chaque niveau d'habileté donnée par le test de classement en anglais, langue seconde, au collégial (TCALS II)

Il est aussi possible de faire un lien entre l'information donnée par un test à chaque niveau d'habileté et le coefficient *alpha* de Cronbach. Ainsi, on peut obtenir une approximation de l'*alpha* de Cronbach par :

$$\alpha_c = 1 - s_\theta^2 . \qquad \textbf{(Équation 2.2)}$$

À titre d'exemple, nous avons indiqué au tableau 2.4 que lorsque le niveau d'habileté est égal à -1,00, l'erreur-type de l'estimateur du niveau d'habileté est de 0,15. Le coefficient *alpha* approximé par l'équation 2.2 est alors égal à 0,98. Il s'agit donc d'une valeur qui se rapproche de celle obtenue par Laurier et ses collaborateurs. Toutefois, lorsque le niveau d'habileté est égal à -1,25, l'erreur-type est égale à 0,42 et la valeur obtenue par approximation du coefficient *alpha* est de 0,82; valeur relativement peu importante pour un tel test, mais tout de même satisfaisante.

Enfin, puisque le résultat fourni par le TCALS II correspond à une valeur unique pour chaque étudiante ou étudiant à qui il est administré, il est donc considéré comme unidimensionnel. Selon cette interprétation, seulement le niveau d'habileté en anglais, langue seconde, serait mesuré par le test. D'autres dimensions, telles que les connaissances historiques ou les valeurs culturelles, ne viendraient pas contaminer le résultat au test. Les travaux de Laurier et de ses collaborateurs confirment l'unidimensionnalité du test. Nous avons analysé de nouveau la dimensionnalité du TCALS II à partir d'une analyse en composantes principales appliquées aux coefficients de corrélation tétrachoriques. Comme le démontre la figure 2.3, plus de 25 % de la variance est expliquée par une première composante principale. La seconde composante explique environ 2 % de la variance seulement. Nous réitérons la conclusion de Laurier et de ses collaborateurs selon laquelle le TCALS II ne mesure essentiellement que l'habileté en anglais, langue seconde.

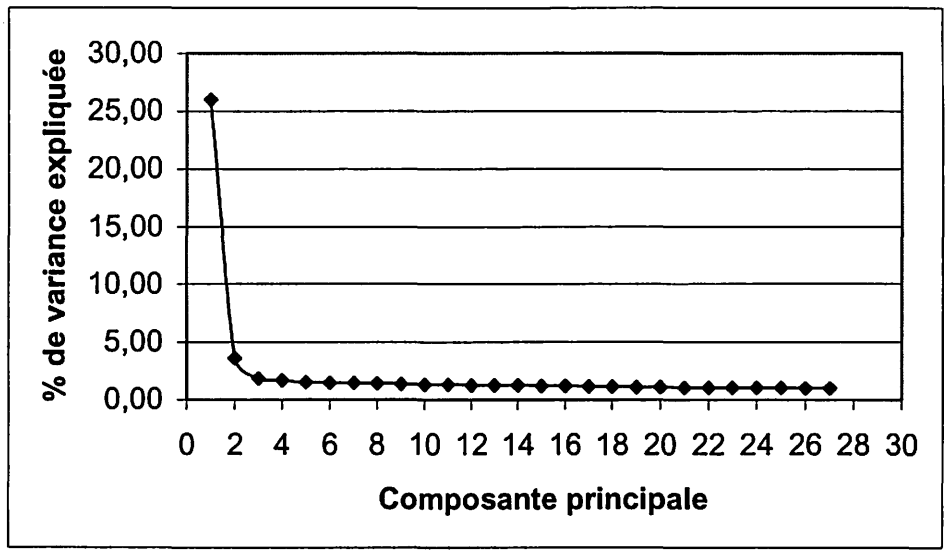

Figure 2.3 Valeurs propres supérieures à 1,00 de chacune des composantes principales du test TCALS II (cohorte 1998-1999)

2.4 Conclusion

Le test de classement en anglais, langue seconde, au collégial est assez précis en autant que le niveau d'habileté estimé soit inférieur ou égal à 0,50; un niveau d'habileté qui correspond à la mise à niveau ainsi qu'aux niveaux 1 et 2. Lorsque les niveaux 3 et 4 sont considérés, le niveau d'habileté langagière est mesuré avec moins de précision. Ceci est dû au fait qu'aucun item dont le niveau de difficulté est supérieur à 0,50 n'est présent dans le test. Des items plus difficiles seraient à ajouter au test. Cette constatation a pour conséquence de rendre éventuellement plus difficiles la détection d'étudiantes ou d'étudiants qui cherchent à se sous-classer, principalement les étudiantes et les étudiants dont le niveau d'habileté en anglais, langue seconde, est plus élevé : ceux et celles que nous voulons précisément détecter. En fait, les indices de détection de patrons de réponses suspects mis à l'essai à l'intérieur de la littérature s'avèrent moins efficaces lorsque le niveau d'habileté est estimé avec moins de précision.

Chapitre 3

Les stratégies utilisées par les étudiantes et les étudiants pour se sous-classer

Une des étapes que nous avons jugée nécessaire à la réalisation de ce projet a été d'identifier les stratégies utilisées par les étudiantes et les étudiants pour tenter de se sous-classer au test de classement en anglais, langue seconde. Nous avions, bien sûr, des hypothèses, au départ, quant aux stratégies adoptées. Ces hypothèses étaient-elles appropriées? Pouvait-il y en avoir d'autres? Y avait-il des variantes et des particularités à ces stratégies?

Nous avons donc décidé d'aller vérifier directement auprès des étudiantes et des étudiants quelles stratégies ils appliqueraient s'ils désiraient sous-performer au TCALS II. Il s'agit d'une étape exploratoire de ce projet de recherche qui nous permettra par la suite de proposer des indices appropriés de détection du comportement de sous-classement. Nous en présentons la méthodologie, suivie des résultats obtenus.

3.1 Méthodologie

3.1.1 Sujets

Seize étudiantes et étudiants inscrits à un cours d'anglais, langue seconde, de niveau 102 au collège de l'Outaouais (604-EWG-03) du même groupe classe ont participé à la rencontre. Il s'agissait d'un cours adapté à la famille des programmes rattachés aux sciences sociales et aux techniques humaines. De ces 16 étudiantes et étudiants, six étaient de sexe masculin et dix de sexe féminin. Six étaient inscrits dans un programme technique, soit Techniques administratives (410.12), tandis que les dix autres étaient inscrits dans un programme pré-universitaire, soit Sciences humaines (300.12 et 300.13). Trois étudiantes et étudiants inscrits en sciences humaines se sont particulièrement montré intéressés par la recherche et ont d'ailleurs signalé leur intérêt à

se diriger plus tard en éducation. Enfin, quelques étudiantes et étudiants ont indiqué qu'ils n'ont pas participé à l'administration du TCALS en 2001-2002.

3.1.2 Déroulement

La rencontre avec les étudiantes et les étudiants s'est déroulée de 11 h 00 à 12 h 00, à la fin de la période d'un cours de niveau 102, le jeudi 14 mars 2002, la semaine suivant la semaine de relâche au collège de l'Outaouais.

La titulaire du cours a présenté le chercheur et a offert ses services pour la prise des notes pendant que celui-ci animait la rencontre. Il a pris le temps de bien expliquer le contexte et les objectifs de la recherche. Il s'est assuré que les étudiantes et les étudiants se souvenaient bien du type de questions dont est composé le TCALS. Il a répondu aux questions des étudiantes et des étudiants pour ensuite présenter le déroulement de la cueillette des informations. Cette cueillette consistait à obtenir, dans un premier temps, par écrit, la réponse à la question suivante : si vous désiriez vous sous-classer au test de classement en anglais, langue seconde, quelle stratégie, ou quelles stratégies, utiliseriez-vous ?

Le chercheur a bien insisté sur le caractère confidentiel des commentaires écrits. Pour s'assurer que les étudiantes et les étudiants comprennent bien la tâche, le chercheur a donné un exemple de stratégie : répondre au hasard à toutes les questions du test. Le chercheur a évité de donner d'autres pistes de réponses pour que les étudiantes et les étudiants ne les reproduisent éventuellement pas intégralement. Les étudiantes et les étudiants avaient au plus 10 minutes pour répondre par écrit à cette question.

Le chercheur a ensuite recueilli les commentaires écrits des étudiantes et des étudiants. Pour conclure, le chercheur a continué la discussion avec les étudiantes et les étudiants, principalement quant aux stratégies éventuelles de sous-classement volontaire.

3.1.3 Considérations éthiques

Au départ, la titulaire du cours a sollicité la participation des étudiantes et des étudiants du groupe classe. Seulement ceux et celles qui le désiraient ont participé à la rencontre. Lors de la rencontre, la confidentialité a été assurée et des explications complètes quant aux objectifs et au protocole de la recherche ont été présentées. Le chercheur a aussi jugé important de prendre le temps de discuter avec les étudiantes et les étudiants du bien-fondé de l'étude et des impacts de celle-ci. Enfin, le chercheur a signalé son intention d'informer les étudiantes et les étudiants des résultats de la recherche et, si possible, de l'évolution de celle-ci.

3.1.4 Méthode d'analyse des résultats

Considérant la nature exploratoire de cette étape de la recherche, l'analyse des résultats se limite au recueil et à l'interprétation des commentaires reçus de la part des étudiantes et des étudiants. La fréquence avec laquelle chacune des stratégies est proposée est aussi indiquée. Toutefois, les commentaires retenus ne se limitent pas aux seules stratégies proposées pour se sous-classer au test de classement en anglais, langue seconde. Tout autre commentaire pertinent est retenu et intégré à l'analyse des résultats.

3.2 Résultats

Des 16 étudiantes et étudiants, 14 ont indiqué qu'ils pourraient éventuellement utiliser une stratégie associée au hasard. Cette stratégie a t-elle été proposée plus fréquemment parce que le chercheur l'a donné en exemple? Il faudrait reproduire ce type de consultation en donnant un autre exemple pour le vérifier.

La seconde stratégie, qui serait éventuellement la plus utilisée, consiste à donner une mauvaise réponse aux questions du test. Selon diverses variantes encore, les étudiantes et les étudiants nous

ont proposé 10 fois cette méthode. Toutefois, la plupart nous ont indiqué qu'ils utiliseraient la stratégie s'il connaissent la bonne réponse. Est-ce à dire que lorsqu'ils ne sont pas certains de la bonne réponse, ils adoptent une autre stratégie? Ou encore, ne s'agit-il que d'une façon de s'exprimer qui signifie que quand l'étudiante ou l'étudiant ne connaît pas la bonne réponse il lui est difficile de choisir la mauvaise réponse?

Six étudiantes et étudiants ont indiqué que la stratégie adoptée serait d'omettre de répondre à certaines questions. À prime abord, l'utilisation de cette stratégie est quelque peu surprenante et dénote de la faiblesse des mécanismes de détection des patrons de réponses suspects dans les collèges. D'une certaine façon ces mécanismes de détection seraient eux-mêmes assez aberrants car à aucune occasion un des intervenants dans le processus ne vérifie la qualité des réponses au test. Dans ce cas, il suffirait de reproduire sur papier les réponses de chacun des étudiantes et des étudiants et de vérifier si certains ont adopté cette stratégie : une opération d'au plus dix minutes. Nous avons fait l'exercice avec les résultats des étudiantes et étudiants des deux premiers tour du SRAM de la cohorte inscrite au collège de l'Outaouais en 2002-2003. Sur les 1 361 patrons de réponses, nous avons retracé 14 cas où il y a huit omissions ou plus. De ces 14 cas, neuf présentent des omissions uniquement à la fin du test. Cela pourrait tout simplement indiquer un manque de temps pour compléter le test. Seulement cinq étudiantes ou étudiants présentent huit omissions ou plus indistinctement à l'intérieur du test et pourraient éventuellement correspondre à des comportements de sous-classement. Dans tous les cas, il ne faudrait pas utiliser le résultat au test lorsque huit questions ou plus n'ont pas reçu de réponse. Dans la même veine, une étudiante ou un étudiant propose de donner la même réponse à plusieurs questions. Plus difficile à inspecter visuellement, cette stratégie serait très facile à détecter par une simple routine informatique.

Trois étudiantes ou étudiants utiliseraient des stratégies plus complexes. Par exemple, une personne nous a dit qu'elle placerait en catégories les questions (facile, moyenne, difficile, etc.) et répondrait correctement aux questions du test selon le niveau voulu. La proposition de cette stratégie nous amène à nous demander si les étudiantes et les étudiants connaissent les scores qu'ils doivent obtenir à chaque niveau. Un autre a indiqué qu'il donnerait la bonne réponse aux

questions vraiment faciles et la mauvaise réponse aux questions un peu plus difficiles. Les stratégies qui ont été proposées par les étudiantes et les étudiants sont présentées au tableau 3.1.

Tableau 3.1
Stratégies que les étudiantes et les étudiants utiliseraient éventuellement pour se sous-classer au test de classement en anglais, langue seconde, au collégial (TCALS II)

Stratégie	Fréquence
Répondre au hasard, cocher au hasard. Répondre au hasard sans lire les questions. Faire du *Wild guess*.	14
Choisir la mauvaise réponse dans les choix (en sachant la mauvaise réponse). Connaître la réponse mais cocher la mauvaise. Cocher la mauvaise réponse par exprès, si nous savons la bonne réponse. Répondre aux questions par une mauvaise réponse si on connaît la bonne réponse. Si je connais la réponse, je peux répondre volontairement la mauvaise réponse. Donner des fausses réponses sachant qu'elles sont fausses. Répondre les mauvaises réponses. Se forcer à donner une mauvaise réponse pour avoir une note inférieure (c'est d'ailleurs ce que j'ai fait lors de mon test pour travailler moins fort et avoir une meilleure note). Répondre l'opposé de ce qu'on croit être la bonne réponse. Connaissant la réponse, en cocher une autre ayant la certitude que ce ne sera pas la bonne réponse. Alterner entre des fausses et des bonnes réponses.	10
Répondre à un certain nombre de questions, ne pas répondre à certaines questions. Omettre de répondre à certaines questions. Ne rien écrire du tout à l'examen. Sauter plusieurs questions. Répondre à une seule question.	6
Catégoriser les questions (faible, moyen, fort, etc.) et répondre correctement aux questions du test selon le niveau voulu. Répondre la bonne réponse aux questions vraiment évidentes et la mauvaise réponse aux questions un peu plus difficiles. Répondre faussement à des questions qui vous paraissent plus faciles.	3
Répondre en calculant le nombre de bonnes réponses accumulées et en répondant au hasard pour le reste.	1
Donner le même choix de réponse tout au long de l'examen.	1
Répondre aux questions les plus faciles et laisser toutes les plus ardues sans réponses.	1

3.3 Conclusion

Certaines stratégies identifiées ne nécessitent pas l'application d'un algorithme complexe pour dépister le comportement de sous-classement. C'est le cas notamment des stratégies où l'étudiante ou l'étudiant omet sciemment de répondre aux questions ou lorsqu'il répète la même réponse à plusieurs items consécutivement. Seules quelques recommandations aux collèges, faciles à appliquer, sont alors nécessaires. Toutefois, les stratégies qui ont été le plus fréquemment identifiées, telles que de répondre au hasard ou de choisir la mauvaise réponse, nécessitent la mise en œuvre de procédures de dépistage plus sophistiquées.

Le prochain chapitre aborde spécifiquement ce problème. La solution proposée consistera à élaborer des procédures de dépistage de ces deux dernières stratégies basées sur des modélisations issues de la théorie de la réponse à l'item.

Chapitre 4

Le dépistage des étudiants qui cherchent à se sous-classer

Ce chapitre s'adresse plus spécifiquement à l'élaboration de mécanismes qui devraient permettre de dépister les étudiantes et les étudiants qui affichent un comportement de sous-classement au TCALS II, principalement ceux et celles qui ont utilisé les stratégies de réponses au hasard et de réponses inversées. Avant tout, cependant, nous soulignons les difficultés rencontrées par le personnel enseignant quant il cherche à détecter, directement en classe, les étudiantes et les étudiants qui ont réussi à se sous-classer au TCALS II. Nous arrivons à la même conclusion que celle retrouvée dans la littérature sur la sous-performance des surdoués : le personnel enseignant est un des plus mauvais juges pour dépister de tels étudiantes et étudiants. Il nous semble donc encore plus pertinent de développer des indices de dépistages performants.

Différents indices retenus, tous tributaires des modélisations issues de la théorie de la réponse à l'item (TRI), seront ensuite présentés. Pour cette raison, auparavant, les modélisations les plus fréquentes issues de la théorie de la réponse à l'item sont décrites. Elles sont accompagnées de leur formulation mathématique, ainsi que des méthodes habituellement utilisées pour estimer le niveau d'habileté de l'étudiante ou de l'étudiant. Les indices de détection et les formulations mathématiques associées terminent ce chapitre.

4.1 L'enseignant, un mauvais juge

Comme nous l'avons souligné au premier chapitre, depuis l'introduction au Québec de la réforme de l'enseignement collégial en 1993, le personnel enseignant des collèges se préoccupe beaucoup de la problématique du sous-classement intentionnel au test de classement en anglais, langue seconde. Selon Laurier, Froio, Pearo et Fournier (1998), 78 % du personnel enseignant des

collèges affirment que les étudiantes et les étudiants ne répondent pas correctement aux questions du test de façon intentionnelle. Il va de soi, que pour se permettre d'affirmer une telle idée, le personnel enseignant spécialisé dans l'enseignement des langues secondes doit pouvoir juger correctement qu'une étudiante ou un étudiant a bien réussi à se sous-classer. Nous avons voulu le vérifier directement auprès des enseignantes et des enseignants du département des langues secondes du collège de l'Outaouais.

Dans un premier temps, avant même l'élaboration des indices de dépistage du sous-classement, nous avons tenté, de façon exploratoire, de dépister les étudiantes et les étudiants qui auraient pu tenter de se sous-classer. À cette fin, nous avons utilisé un indice de détection de patrons de réponses aberrants déjà intégré à un logiciel reconnu et spécialisé dans l'analyse des items d'un test, soit Bilog. Les résultats de la cohorte des 1 415 étudiantes et des étudiants des quatre tours du SRAM admis et inscrits au collège de l'Outaouais en 2001-2002 en main, nous avons utilisé le logiciel Bilog. L'indice de détection appliqué par Bilog, χ_B^2, est décrit plus bas à la fin de la section 4.5.1. Il est important de souligner que cet indice n'est pas destiné spécifiquement à la détection d'un comportement de sous-classement et que, de ce fait, les étudiantes et étudiants qu'il permet d'identifier pourraient, pour bien d'autres raisons, obtenir un patron de réponses peu normal.

De ces 1 415 étudiantes et étudiants, Bilog a permis d'identifier 105 cas potentiels de sous-classement, dont 22 inscrits à leur premier cours d'anglais, langue seconde, à l'automne 2001. Cette opération s'étant déroulée en octobre 2001, nous avons pu vérifier si ces étudiantes et ces étudiants avaient bel et bien été identifiés par le personnel enseignant du département des langues secondes du collège de l'Outaouais lorsqu'ils étaient inscrits à leur premier cours d'anglais, langue seconde, à l'automne 2001. En fait, sur ces 22 cas, seulement une étudiante a été déplacée vers un cours d'un niveau supérieur.

Restaient 83 étudiantes et étudiants qui allaient être inscrits à leur premier cours d'anglais, langue seconde, plus tard dans leur cheminement scolaire. Pour plusieurs ce serait à l'hiver 2002, principalement pour ceux et celles inscrits dans un programme pré-universitaire. Pour plusieurs

autres, surtout pour ceux et celles inscrits dans un programme technique, ce serait au plus tôt à l'automne 2002. Nous avons fourni la liste de ces 83 étudiantes et étudiants au personnel enseignant du département des langues secondes en leur suggérant de surveiller de près ces étudiantes et ces étudiants lors de leurs premières semaines de cours. Nous leur avons aussi indiqué qu'une rencontre de groupe serait réalisée avec eux après la semaine de relâche au mois de mars 2002 dans le but de vérifier l'adéquation entre les résultats de détection obtenus par Bilog et leur appréciation de ces résultats. Fait à noter, cette vérification devait s'effectuer tout de suite après la rencontre évoquée au chapitre précédent avec un groupe d'étudiantes et d'étudiants visant à la cueillette d'informations quant aux stratégies utilisées par eux pour se sous-classer.

Lors de la rencontre de groupe du personnel enseignant du mois de mars, cinq enseignantes et enseignants sur sept du département des langues secondes du collège de l'Outaouais étaient présents. Ces enseignantes et enseignants dispensent de la formation aux quatre niveaux de cours en anglais, langue seconde. Certains enseignent aussi des cours relatifs à la langue anglaise à l'Université d'Ottawa et à l'Université du Québec à Hull. D'autres enseignent, ou ont enseigné, l'anglais, langue seconde, au secondaire. L'une d'entre eux dispense aussi de la formation en espagnol, langue seconde. De tous, une seule personne n'a pas encore obtenu sa permanence d'emploi. Il s'agit donc de personnes qui possèdent une excellente expérience de l'enseignement de l'anglais, langue seconde. La plupart ont suivi le dossier du test de classement et ont participé à ses multiples versions au cours des années. Une de ces personnes a d'ailleurs contribué à la création des questions de la dernière version du TCALS, telle qu'elle est administrée aujourd'hui au Collège de l'Outaouais.

Le chercheur a bien expliqué le but principal de la rencontre, soit de vérifier si les étudiantes et les étudiants de la cohorte 2001-2002 qui avaient été identifiés par Bilog avaient aussi été identifiés par eux comme ayant été mal classés par le TCALS II. Dans un premier temps, un rappel des événements antérieurs a été fait. Ensuite, les enseignantes et les enseignants ont vérifié si les noms et les numéros de matricules des étudiantes et des étudiants correspondaient à ceux et celles qu'ils considèrent avoir été mal classés à l'intérieur des cours spécifiques.

Des 83 étudiantes et étudiants identifiés par Bilog qui auraient éventuellement pu être inscrits à leur premier cours d'anglais, langue seconde, à l'hiver 2002, seulement quatre avaient aussi été identifiés par le personnel enseignant, tandis que six n'auraient pas été mal classés et qu'une étudiante aurait été déclassée. Il ne faut pas oublier que Bilog n'est pas destiné uniquement à la détection d'un patron de réponses relatif à un comportement de sous-classement. Somme toute, il semble qu'il n'y a pas concordance entre le verdict des enseignantes et des enseignants et les indications obtenues à partir de Bilog.

La rencontre s'est poursuivie par une discussion sur les raisons de la distance entre leur verdict et celui de Bilog, ainsi que sur les actions subséquentes à entreprendre. Les enseignantes et les enseignants ont indiqué qu'il leur était fort probablement difficile d'identifier les étudiantes et les étudiants sous-classés car ces derniers prennent un profil langagier bas lors du premier cours. L'enseignante qui a assisté le chercheur lors de la rencontre préalable des étudiantes et des étudiants a signalé la surprise qu'elle a eue quant elle a réalisé que les étudiantes et les étudiants qu'elle avait cru sous-classés n'étaient pas ceux et celles qu'elle avait identifiés. Pourtant, celle-ci est la plus expérimentée parmi les enseignantes et les enseignants de ce département. C'est elle qui a participé à la création de questions pour le test. C'est elle qui s'est le plus impliquée au collège de l'Outaouais au fil des ans dans la recherche de solutions au problème de sous-classement. Il s'agit donc d'une personne, non seulement très compétente, mais qui était sûrement parmi les plus outillée pour identifier les étudiantes et les étudiants sous-performants. Et pourtant, il semble bien que les étudiantes et les étudiants sous-classés aient réussi à lui camoufler leurs réelles compétences langagières.

L'enseignante ou l'enseignant ne serait peut-être pas le meilleur juge pour détecter les étudiantes et les étudiants qui auraient réussi à se sous-classer. Ceci semblerait aller strictement dans le même sens que les résultats observés à l'intérieur de multiples recherches réalisées sur la sous-performance (*underachievement*) des surdoués (*gifted*). Selon ces recherches (Ciha, Harris, Hoffman et Potter, 1974; Coleman et Cross, 2001, p. 91-96; Feldhusen et Jarwan, 2000, p. 275; Gagné, 1994; Pegnato et Birch, 1959; Rimm, 1995; Whitmore, 1980, p. 85-87), les enseignantes

et les enseignants seraient les moins bons juges pour détecter les étudiantes et les étudiants sous-performants.

Selon nous, il est possible que les étudiantes et les étudiants qui ont réussi à se sous-classer cherchent à ne pas se faire remarquer, comme l'a suggéré le personnel enseignant. Toujours selon nous, les étudiantes et les étudiants qui, selon la perception des enseignantes et des enseignants, auraient réussi à se sous-classer seraient en réalité ceux et celles les plus visibles dans la salle de classe. Ce seraient ceux et celles qui parleraient le plus ou encore qui auraient un comportement plutôt perturbateur. D'ailleurs, le fait que les cours d'anglais, langue seconde, depuis la réforme de l'enseignement collégial de 1993, soient devenus obligatoires a eu pour conséquence de forcer des étudiantes et des étudiants, non nécessairement motivés, à suivre ces cours. Pour certains d'entre eux, ces cours d'anglais, langue seconde, seraient moins intéressants et ils seraient alors moins bien disposés à participer aux activités de classe.

L'enseignante ou l'enseignant pourrait éventuellement se laisser berner par ses perceptions. Si notre hypothèse était vraie, cela signifierait que, sans nier l'existence du problème du sous-classement, la source du problème de gestion de classe vécu par le personnel enseignant en anglais, langue seconde, serait tout autre. Les étudiantes et les étudiants perturbateurs ne seraient pas ceux et celles qui ont réussi à se sous-classer. Pour détecter ces étudiantes et étudiants, l'utilisation d'un indice de dépistage du comportement de sous-classement lors de l'administration du test de classement devient donc encore plus nécessaire.

Enfin, la rencontre avec le personnel enseignant s'est terminée sur les suites à donner aux travaux. Le chercheur et les enseignantes et les enseignants se sont entendus pour aller plus loin dans l'évaluation de l'indice de dépistage intégré au logiciel Bilog. Il a été décidé que les prochaines étudiantes et les prochains étudiants de la cohorte 2002-2003, qui seraient identifiés par Bilog comme ayant éventuellement adopté un comportement suspect au test, seraient convoqués à un examen écrit supplémentaire. Selon le personnel enseignant du collège de l'Outaouais, une composition écrite sur un sujet personnel où l'étudiant aurait à utiliser des temps de verbes différents, dont le passé, serait très efficace. Les consignes spécifiques données à

l'étudiante ou à l'étudiant sont présentées à l'annexe 1. Les résultats obtenus permettront de vérifier d'une façon beaucoup plus précise l'adéquation de l'identification obtenue à partir de Bilog.

Nous enchaînons maintenant avec la description du modèle théorique, soit la théorie de la réponse à l'item, et les modélisations associées qui guident la formulation mathématique de tous les indices de dépistages de patrons de réponses aberrants que nous avons adoptés à l'intérieur de cette recherche.

4.2 La théorie de la réponse à l'item

La théorie de la réponse à l'item postule qu'il est possible de spécifier une fonction mathématique reliant la probabilité d'une réponse à un item au niveau d'habileté du répondant (Goldstein, 1994a, p. 366; Goldstein, 1994b, p. 109; Laveault et Grégoire, 1997, p. 291; van der Linden et Hambleton, 1997, p. v). Certaines des modélisations proposées dépendent du type de réponses aux items : réponses dichotomiques (Lord, 1952; Lord et Novick, 1968, p. 365), réponses polytomiques (Baker, 1992, p. 251-288; Bock, 1997; Roberts, Donoghue et Laughlin, 2000; Thissen, 1988), réponses ordonnées (Samejima, 1997), réponses polytomiques partiellement ordonnées (Wilson, 1992) et réponses continues (Samejima, 1973b). D'autres modélisations dépendent de l'échelle de mesure postulée pour le niveau d'habileté : catégorielle, classes latentes (Gitomer et Rock, 1993); continue, trait latent (Samejima, 1973b, 1974); ou hybride, classes latentes et trait latent (Yamamoto, 1995; Yamamoto et Everson, 1995; Yamamoto et Gitomer, 1993). Dans d'autres cas, des habiletés différentes contribuent à produire la réponse à l'item : les modèles sont alors multidimensionnels (Ackerman, 1994; Goldstein et Wood, 1989, p. 160-162; Luecht, 1996; McDonald, 1982, p. 381-384, 1997, 2000; Reckase, 1985, 1997). Des modèles non paramétriques existent aussi (Junker et Sijtsma, 2000; Mokken,

1997; Mokken et Lewis, 1982; Molenaar, 1997; Ramsay, 1991, 1993a, 1993b, 1997; Stout, 1990).

Dans la plupart des tests utilisés actuellement, le type de réponses aux items est dichotomique (bonne ou mauvaise réponse) et l'unidimensionnalité de l'habileté sur une échelle continue est supposée. En fonction de ces caractéristiques, trois modèles ont été privilégiés (Hambleton, Swaminathan et Rogers, 1991, p. 12-18; Wainer et Mislevy, 1990, p. 68-72). Ils ne diffèrent que par le nombre de paramètres impliqués dans la fonction modélisant la probabilité d'obtenir une bonne réponse à un item selon le niveau d'habileté. Ce sont les modèles à un, à deux et à trois paramètres. Ces trois modèles sont présentés ainsi que celui, moins fréquent, à quatre paramètres.

Dans le modèle à un paramètre, seul le niveau de difficulté de l'item, b_g, est considéré. Le modèle à deux paramètres ajoute un paramètre de discrimination, a_g, qui correspond à la pente maximale de la fonction. Le modèle à trois paramètres intègre de plus un paramètre de pseudo-chance, c_g. Dans ce modèle, il est postulé que la probabilité d'une bonne réponse à un item n'est pas nécessairement nulle lorsque le niveau d'habileté est très faible. De façon similaire, le modèle à quatre paramètres incorpore un paramètre, γ_g, qui correspond à la valeur asymptotique supérieure de la probabilité d'une bonne réponse à l'item. Il y est postulé que, même si le niveau d'habileté est très élevé, la probabilité d'une bonne réponse à l'item n'est pas nécessairement égale à 1, donc certaine. Ce dernier modèle, pourtant intéressant au plan théorique, ne semble cependant pas utilisé dans la pratique. Barton et Lord ont été incapables de lui trouver des avantages significatifs (1981 : voir Hambleton et Swaminathan, 1987, p. 49).

4.3 Les modélisations issues de la théorie de la réponse à l'item

4.3.1 La modélisation logistique à un paramètre

Comme le souligne Blais (1987, p. 24), le modèle logistique à un paramètre a été développé par Georges Rasch (1960) indépendamment des travaux de Lord (1952) et de Birnbaum (1968), et presque simultanément à ces travaux. Il est souvent nommé le modèle de Rasch (Rasch 1960 : voir Molenaar, 1995, p. 15), quoique Rasch ait formulé ce modèle de manière différente, mais mathématiquement équivalente. On reconnaît le modèle à un paramètre par l'expression de la probabilité d'une bonne réponse à un item, $P(X_g = 1|\theta)$, à partir de la fonction logistique suivante :

$$P(X_g = 1|\theta) = \frac{1}{1+e^{-1.7(\theta - b_g)}}, \qquad \textbf{(Équation 4.1)}$$

où X_g est la réponse à l'item, θ est le niveau d'habileté et b_g le niveau de difficulté de l'item. La valeur de X_g est égale à 1 pour une bonne réponse et à 0 pour une mauvaise réponse tandis que θ et b_g, par une transformation en scores standardisés (scores z) peuvent prendre des valeurs comprises dans l'intervalle $\langle -\infty, \infty \rangle$.

La figure 4.1 présente une courbe caractéristique d'item du modèle à un paramètre. Plus la valeur de b_g est élevée, plus le niveau de difficulté de l'item correspondant est élevé. Cette courbe permet d'observer la probabilité d'obtenir une bonne réponse à un item en fonction du niveau d'habileté. Selon le modèle à un paramètre, la probabilité d'obtenir une bonne réponse est nulle lorsque le niveau d'habileté est très faible. Elle est certaine lorsque le niveau d'habileté est élevé.

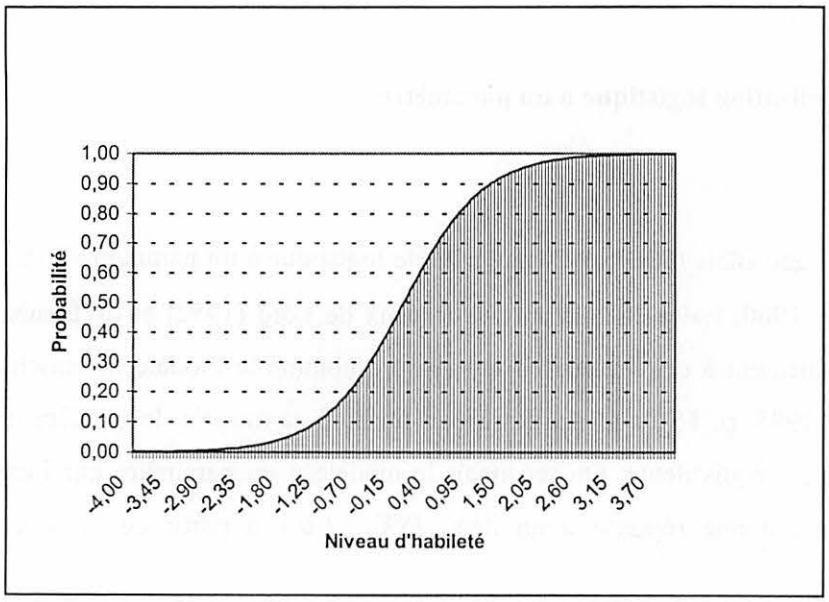

Figure 4.1 Courbe caractéristique d'item de la modélisation logistique à un paramètre

Puisque

$$P(X = x \mid \theta) = \prod_{g=1}^{n} P(X_g = 1 \mid \theta)^{X_g} P(X_g = 0 \mid \theta)^{1-X_g}, \qquad \textbf{(Équation 4.2)}$$

la probabilité d'obtenir un patron de réponses spécifique est égale à

$$P(X = x \mid \theta) = \prod_{g=1}^{n} \left[\frac{1}{1+e^{-1,7(\theta - b_g)}} \right]^{X_g} \left[1 - \frac{1}{1+e^{-1,7(\theta - b_g)}} \right]^{1-X_g}. \qquad \textbf{(Équation 4.3)}$$

4.3.2 La modélisation logistique à deux paramètres

Le modèle logistique à deux paramètres propose d'ajouter au niveau de difficulté, b_g, un second paramètre, la discrimination. Le paramètre de discrimination, a_g, peut prendre des valeurs qui varient entre $-\infty$ et ∞. Il est toutefois inhabituel d'obtenir des valeurs supérieures à 2 et, lorsque la valeur est négative, l'item devrait être rejeté (Hambleton, Swaminathan et Rogers, 1991, p.15). Selon Blais (1987, p. 24), l'intervalle raisonnable du paramètre de discrimination varie entre 0,50 et 2,00. Dans ce modèle, les items ne partagent pas tous le même pouvoir de discrimination lorsque le niveau d'habileté est égal au niveau de difficulté. L'indice de discrimination est proportionnel à la pente de la courbe caractéristique de l'item lorsque le niveau d'habileté est égal au niveau de difficulté de l'item. Un item dont le paramètre de discrimination affiche une valeur négative indique que la probabilité d'une bonne réponse à l'item diminue avec le niveau d'habileté. Si le modèle est monotone croissant, alors la valeur du paramètre de discrimination est fixée par cette condition et alors $a_g > 0$. L'équation de la fonction logistique à deux paramètres prend la forme suivante :

$$P(X_g = 1 \mid \theta) = \frac{1}{1 + e^{-1,7a_g(\theta - b_g)}} \qquad \textbf{(Équation 4.4)}$$

et

$$P(X = x \mid \theta) = \prod_{g=1}^{n} \left[\frac{1}{1 + e^{-1,7a_g(\theta - b_g)}} \right]^{X_g} \left[1 - \frac{1}{1 + e^{-1,7a_g(\theta - b_g)}} \right]^{1 - X_g}. \qquad \textbf{(Équation 4.5)}$$

La figure 4.2 illustre la courbe caractéristique d'un item selon la modélisation logistique à deux paramètres.

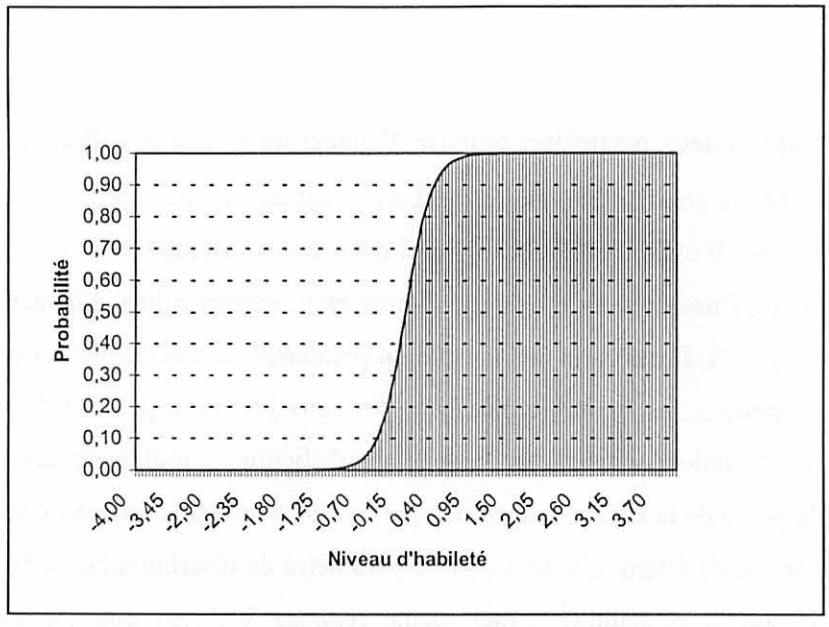

Figure 4.2 Courbe caractéristique d'item de la modélisation logistique à deux paramètres

4.3.3 La modélisation logistique à trois paramètres

Le modèle logistique à trois paramètres ajoute un troisième élément aux deux modèles précédents : l'indice de pseudo-chance, c_g. Selon ce modèle, la probabilité d'une bonne réponse à un item n'est pas nécessairement nulle lorsque le niveau d'habileté est très faible. Des facteurs externes au niveau d'habileté peuvent affecter la probabilité d'une bonne réponse. Par exemple, dans un choix de réponses *vrai ou faux,* la réponse *vrai* pourrait être naturellement préférée par les individus dont le niveau d'habileté est très faible. Quoique le paramètre de pseudo-chance puisse prendre des valeurs comprises entre 0,00 et 1,00, sa valeur ne devrait pas dépasser celle d'une réponse au hasard et serait proportionnelle au nombre de choix de réponses (Blais, 1987, p. 24; Hambleton, Swaminathan et Rogers, 1991, p. 17). Selon Laveault et Grégoire (1997,

p. 294), la valeur du paramètre de pseudo-chance est généralement inférieure à celle qui correspondrait à un choix de réponses complètement au hasard. La figure 4.3 montre une courbe caractéristique d'items selon le modèle à trois paramètres. Selon Wainer et Mislevy (1990, p. 2), ce modèle est largement utilisé dans les applications de testing à grande échelle. L'équation du modèle logistique à trois paramètres est la suivante :

$$P(X_g = 1 \mid \theta) = c_g + \frac{1 - c_g}{1 + e^{-1,7(\theta - b_g)}} \qquad \textbf{(Équation 4.6)}$$

et

$$P(X = x \mid \theta) = \prod_{g=1}^{n} \left[c_g + \frac{1 - c_g}{1 + e^{-1,7a_g(\theta - b_g)}} \right]^{X_g} \left[1 - c_g + \frac{1 - c_g}{1 + e^{-1,7a_g(\theta - b_g)}} \right]^{1 - X_g} . \qquad \textbf{(Équation 4.7)}$$

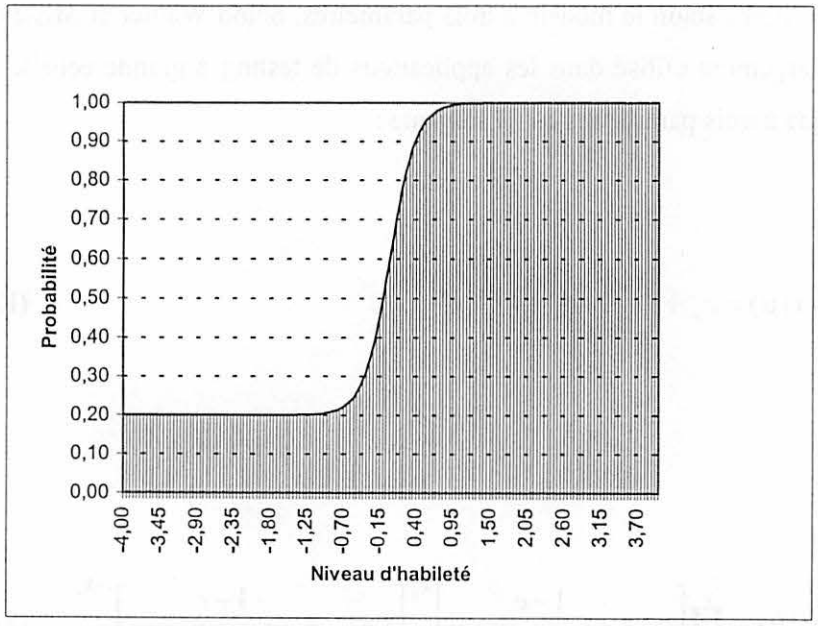

Figure 4.3 Courbe caractéristique d'item de la modélisation logistique à trois paramètres

4.3.4 La modélisation logistique à quatre paramètres

Dans le modèle logistique à quatre paramètres, on suppose que la probabilité d'une bonne réponse à un item n'est pas nécessairement égale à 1,00 lorsque le niveau d'habileté est très élevé. On incorpore alors le paramètre γ_g à la fonction logistique pour en tenir compte, et celui-ci varie à l'intérieur de l'intervalle compris entre 0,00 et 1,00. La fonction logistique à quatre paramètres semble toutefois n'être jamais utilisée, puisque des problèmes d'estimation numérique lui sont associés lorsque les méthodes d'estimation par vraisemblance maximale sont appliquées. La figure 4.4 présente la courbe caractéristique d'un item obtenue selon cette modélisation. La fonction est la suivante (Hambleton et Swaminathan, 1987, p. 49) :

$$P(X_g = 1 \mid \theta) = c_g + \frac{\gamma_g - c_g}{1 + e^{-1,7(\theta - b_g)}} \qquad \textbf{(Équation 4.8)}$$

et

$$P(X = x \mid \theta) = \prod_{g=1}^{n} \left[c_g + \frac{\gamma_g - c_g}{1 + e^{-1,7a_g(\theta - b_g)}} \right]^{X_g} \left[1 - c_g + \frac{\gamma_g - c_g}{1 + e^{-1,7a_g(\theta - b_g)}} \right]^{1-X_g} \qquad \textbf{(Équation 4.9)}$$

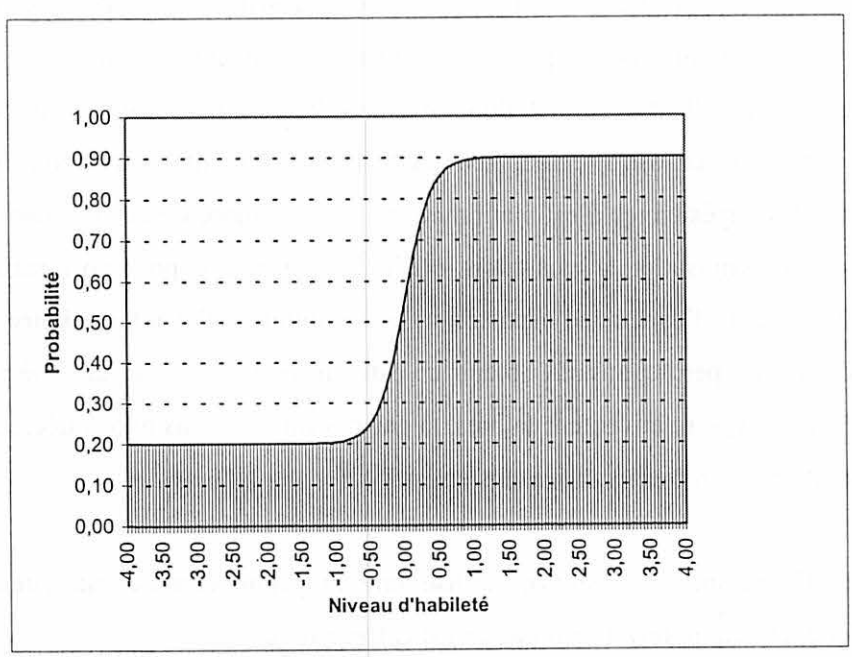

Figure 4.4 Courbe caractéristique d'item de la modélisation logistique à quatre paramètres

Un tel modèle a été proposé par McDonald (1967, p. 67) et par Burton et Lord (1981 : voir Hambleton et Swaminathan, 1987, p. 50) dans le contexte d'une modélisation basée sur la loi

normale. Ils l'ont cependant abandonné puisqu'ils ne lui ont pas trouvé d'avantages appréciables dans le gain de précision des mesures.

4.4 L'estimation du niveau d'habileté

Lorsque la valeur du paramètre de difficulté est fixée pour tous les items, il est possible d'estimer le niveau d'habileté à partir de différentes méthodes. Les méthodes d'estimation les plus fréquemment employées sont les méthodes de vraisemblance maximale (*maximum likelihood*, ML), de maximisation à posteriori (*maximization a posteriori*, MAP) et de l'espérance a posteriori (*expected a posteriori*, EAP). La méthode de vraisemblance maximale, la première à avoir été proposée, ne tient pas compte de l'information a priori contenue dans les résultats obtenus préalablement. Elle ne permet donc pas d'utiliser toute l'information connue et la précision de l'estimation en est affectée. Pour cette raison, les auteurs préfèrent généralement utiliser des méthodes bayésiennes d'estimation qui tiennent compte de ces informations a priori. Les méthodes de maximisation a posteriori et de l'espérance a posteriori satisfont à cette exigence. La méthode de l'espérance a posteriori offre toutefois l'avantage d'être une des plus aisée à manipuler et elle permet aussi d'obtenir un estimateur du niveau d'habileté parmi les plus précis (biais et erreur-type relativement faibles). Pour ces raisons nous avons décidé d'adopter la méthode d'estimation selon l'espérance a posteriori.

La méthode de l'espérance a posteriori utilise une moyenne comme estimateur du niveau d'habileté (Raîche, 2001a, p. 66). L'estimateur est calculé selon :

$$\hat{\theta} = \frac{\int_{\theta=-\infty}^{\infty} \theta f(\theta) P(X_g = 1 \mid \theta)^{X_g} P(X_g = 0 \mid \theta^{1-X_g})}{\int_{\theta=-\infty}^{\infty} f(\theta) P(X_g = 1 \mid \theta)^{X_g} P(X_g = 0 \mid \theta^{1-X_g})} \qquad \textbf{(Équation 4.10)}$$

avec comme variance

$$S_{\hat{\theta}} = \frac{\int_{\theta=-\infty}^{\infty} (\theta - \hat{\theta})^2 f(\theta) P(X_g = 1 | \theta)^{X_g} P(X_g = 0 | \theta^{1-X_g})}{\int_{\theta=-\infty}^{\infty} f(\theta) P(X_g = 1 | \theta)^{X_g} P(X_g = 0 | \theta^{1-X_g})}, \qquad \textbf{(Équation 4.11)}$$

où $f(\theta)$ correspond à la distribution de probabilité a priori du niveau d'habileté. Le calcul est réalisé à partir d'une approximation numérique (Baker, 1992, p. 210-211; Bock et Mislevy, 1982; Raîche, 2001a, p. 130-131).

Généralement, la distribution retenue est normale et centrée réduite (N(0, 1)). Toutefois, nous avons pu constater que le biais de l'estimateur du niveau d'habileté peut être assez important lorsqu'une distribution N(0, 1) est retenue et que le niveau d'habileté s'éloigne de 0. Récemment, Raîche et Blais (2002a; 2002b) ont proposé une amélioration à la méthode de l'espérance a posteriori. Ils ont ainsi suggéré d'utiliser une distribution a priori N(0, 1) seulement pour le premier item administré. Pour les autres items, la moyenne de la distribution a priori est égale à l'estimateur du niveau d'habileté obtenu à l'item précédent, soit $f(\hat{\theta}_{g-1})$.

Cette stratégie, dans un contexte de testing adaptatif, a permis, à toutes fins pratiques, de rendre le biais nul quel que soit le niveau d'habileté du sujet. Nous avons décidé d'adopter cet ajustement adaptatif à la méthode de l'espérance a posteriori. Nous tenons toutefois à souligner que nous avons remarqué, dans cette recherche, que l'estimateur de niveau d'habileté ne maintient plus la propriété de statistique suffisante (Rost, 1995, p.37-38; Wright, 1997, p. 35-36), caractéristique prisée de la modélisation logistique à un paramètre de Rasch. Cette propriété permettait à l'estimateur du niveau d'habileté d'avoir une correspondance parfaite avec le score total au test, soit le nombre de bonnes réponses. Il n'est donc plus certain que l'utilisation de la modélisation logistique à un paramètre de Rasch permet à l'estimateur du niveau d'habileté de se retrouver sur

une échelle d'intervalle, comme le soutiennent certains auteurs (Fischer, 1995, p. 37-38; Michell, 2002, p. 13-14; Wright, 1997, p. 37-38).

4.5 Des indices pour détecter les patrons de réponses aberrants

4.5.1 Des indices non spécifiques à un comportement particulier

Les premiers travaux relatifs à la détection de patrons de réponses aberrants ne s'intéressaient pas au dépistage de comportements spécifiques, encore moins au dépistage du comportement de sous-classement. Ces indices avaient plutôt pour but de détecter toute structure de réponses à un test qui apparaîtrait étrange. Parmi les indices les plus populaires qui ont été proposés nous retrouvons les indices l_z, W et *Zeta*. Le logiciel Bilog, que nous avons utilisé dans les phases préliminaires de la recherche, ainsi que pour comparer l'exactitude de nos calculs, utilise un indice, auquel nous attribuerons le symbole χ^2_B, dont la justification est assez différente de l_z, W et *Zeta*. Nous avons utilisé cet indice au début de nos travaux. Nous présentons maintenant les étapes du calcul de ces quatre indices, ainsi que la façon de les interpréter.

L'indice standardisé l_0 a été proposé par Levine et Rubin en 1979 et a été mis en application à plusieurs reprises par la suite (Drasgow et Guertler, 1987; Drasgow et levine, 1986; Drasgow, Levine et McLaughlin, 1987, 1991; Drasgow, Levine et Williams, 1985; Levine et Drasgow, 1982, 1983). Cet indice était toutefois difficile à interpréter car il est dépendant de la valeur du niveau d'habileté estimé. Pour pallier ce problème, certains auteurs (Drasgow, Levine et Williams, 1985) lui ont apporté une modification et ont ainsi proposé le nouvel indice standardisé l_z. Cet indice serait distribué selon une distribution de probabilité N(0,1), ce qui lui confère des propriétés très intéressantes au point de vue statistique ainsi qu'au plan pratique. Pour un test constitué de n items, lorsque la probabilité d'obtenir une bonne réponse à chacun des items $P(X_g = 1|\theta)$ est obtenue selon l'une des quatre modélisations présentées auparavant, il est égal à :

$$l_z = \frac{l_o - E(l_o)}{\left[Var(l_o)\right]^{1/2}} \qquad \textbf{(Équation 4.12)}$$

où

$$l_o = \sum_{g=1}^{n}\left[X_g \ln P(X_g = 1 \mid \theta) + (1 - X_g) \ln P(X_g = 0 \mid \theta)\right], \qquad \textbf{(Équation 4.13)}$$

$$E(l_o) = \sum_{g=1}^{n}\left[P(X_g = 1 \mid \theta) \ln P(X_g = 1 \mid \theta) + P(X_g = 0 \mid \theta) \ln P(X_g = 0 \mid \theta)\right]. \qquad \textbf{(Équation 4.14)}$$

et

$$Var(l_o) = \sum_{g=1}^{n}\left\{P(X_g = 1 \mid \theta)P(X_g = 0 \mid \theta)\left[\ln \frac{P(X_g = 1 \mid \theta)}{P(X_g = 0 \mid \theta)}\right]\right\}. \qquad \textbf{(Équation 4.15)}$$

Lorsque l_z affiche une valeur inférieure ou égale à –1,65, le patron de réponses obtenu est considéré comme peu probable. Toute valeur supérieure à –1,65 est jugée convenable.

De leur côté, Wright et Stone (1979; Wright et Master, 1982; Bond, et Fox, 2001; Li, Olejnik et Bashaw, 1991), ainsi que Smith (1991, 2002, Smith et Suh, 2002), ont proposé un indice légèrement différent, *W*. Il est égal à

$$W = \frac{\sum_{g=1}^{n}\left[\left[X_g - P(X_g = 1 \mid \theta)\right]^2\right]}{\sum_{g=1}^{n}\left[P(X_g = 1 \mid \theta)P(X_g = 0 \mid \theta)\right]}. \qquad \textbf{(Équation 4.16)}$$

Les logiciels destinés exclusivement à la modélisation logistique à un paramètre utilisent fréquemment l'indice *W*. Lorsque la valeur affichée par *W* est sensiblement supérieure à 1, le patron de réponses est jugé aberrant. La distribution de probabilité de *W* a toutefois été peu étudiée. Les auteurs supposent que la moyenne de *W* est égale à 1 tandis que peu d'entre eux s'entendent sur la variance qui lui est associée. C'est pourquoi il est difficile de statuer à quel moment la valeur affichée par *W* est sensiblement supérieure à 1.

Enfin, Tatsuoka (1996, Birenbaum, Kelly et Tatsuoka, 1992a, 1992b) a proposé un indice de détection des patrons de réponses aberrants dans le but de diagnostiquer des problèmes d'apprentissage chez les étudiants, plus particulièrement des problèmes de conceptions erronées en mathématiques. Cet indice, *Zeta*, s'est toutefois avéré intéressant quant à la détection de toutes formes de patrons de réponses aberrants. Il est calculé comme suit :

$$Zeta = \frac{\sum_{g=1}^{n}\left[P(X_g = 1 \mid \theta) - X_g\right]\left[P(X_g = 1 \mid \theta) - T(\theta)\right]}{\sum_{g=1}^{n}\left\{P(X_g = 1 \mid \theta)P(X_g = 0 \mid \theta)\right]\left[P(X_g = 1 \mid \theta) - T(\theta)\right]^2\right\}^{1/2}}, \qquad \textbf{(Équation 4.17)}$$

où

$$T(\theta) = \frac{\sum_{g=1}^{n} P(X_g = 1 \mid \theta)}{n}. \qquad \textbf{(Équation 4.18)}$$

Zeta afficherait une distribution de probabilité N(0,1) et une valeur sensiblement supérieure à 0 ($Zeta \geq 1{,}65$) indiquerait que le patron de réponses est peu probable.

L'indice χ_B^2, utilisé à l'intérieur du logiciel Bilog (Mislevy et Bock, 1986; Mislevy et Stocking, 1987) repose sur une logique différente. Il vise à déterminer si le patron de réponses d'un individu s'ajuste bien à la modélisation utilisée en comparant cet ajustement avec une modélisation où on a ajouté un paramètre d'item supplémentaire. Dans notre recherche, cet indice compare la probabilité d'obtenir un patron de réponses telle que calculée selon une modélisation logistique à un paramètre à la probabilité d'obtenir de patron de réponses si on avait appliqué la modélisation logistique à deux paramètres. Une valeur trop élevée de χ_B^2 signifierait que l'indice de discrimination d'une étudiante ou d'un étudiant particulier n'est pas constant tout au long du test. L'indice χ_B^2 pourrait ainsi détecter l'étudiante ou l'étudiant qui donne des mauvaises réponses plutôt que des bonnes réponses puisque pour celui-ci l'indice de discrimination ne correspond pas à celui de la modélisation logistique à un paramètre. L'indice χ_B^2 est égal à

$$\chi_B^2 = \left\{-2 * \ln P(X_g = 1 \mid \theta)_{1PL}\right\} - \left\{-2 * \ln P(X_g = 1 \mid \theta)_{2PL}\right\}, \qquad \textbf{(Équation 4.19)}$$

où $P(X_g = 1 \mid \theta)_{1PL}$ correspond à l'équation 4.3 de la modélisation logistique à un paramètre, tandis que $P(X_g = 1 \mid \theta)_{2PL}$ correspond à l'équation 4.5 de la modélisation logistique à deux paramètres. L'indice χ_B^2 se distribuerait selon la distribution de probabilité du χ^2 avec un nombre de degrés de liberté égal au nombre d'items du test, ici 85.

4.5.2 Des indices spécifiques à un comportement particulier

Levine et Drasgow (1982, 1988; Drasgow, Levine et Zickar, 1996) suggèrent que la détection de patrons de réponses aberrants gagnerait à être effectuée en ciblant spécifiquement les comportements qui sont directement la cause du mauvais ajustement du patron de réponses. Il suffirait alors de calculer le rapport entre la vraisemblance marginale d'un patron de réponses fourni selon le comportement aberrant spécifique et la vraisemblance marginale d'un patron de réponses obtenu selon un comportement dit normal au test. L'indice serait alors égal à :

$$I_{aberrant} = \frac{P_{aberrant}}{P_{normal}}. \qquad \textbf{(Équation 4.20)}$$

Si $I_{aberrant}$ est supérieur à une valeur critique, le patron de réponses est alors considéré suspect.

Sachant que la probabilité conditionnelle au niveau d'habileté θ d'obtenir un patron de réponses normal est calculée à partir d'une des modélisations logistiques de la réponse à l'item,

$$P_{normal|\theta} = P(X = x \mid \theta) = \prod_{g=1}^{n} P(X_g = 1 \mid \theta)^{X_g} P(X_g = 0 \mid \theta)^{1-X_g}. \qquad \textbf{(Équation 4.21)}$$

et que la probabilité marginale (non conditionnelle) d'obtenir un patron de réponses normal est calculée selon :

$$P_{normal} = \int_{\theta=-\infty}^{\infty} P_{normal|\theta} d\theta \qquad \textbf{(Équation 4.22)}$$

La probabilité conditionnelle d'obtenir un patron de réponses au hasard dépend du nombre de choix de réponses, k, à chacun des items et est pour sa part égale à :

$$P_{hasard|\theta} = \prod_{g=1}^{n} \left(\frac{1}{k}\right)^{X_g} \left(\frac{k-1}{k}\right)^{1-X_g}, \qquad \textbf{(Équation 4.23)}$$

tandis que la probabilité marginale d'obtenir un patron de réponses au hasard est égale à

$$P_{hasard} = \int_{\theta=-\infty}^{\infty} P_{hasard|\theta} d\theta . \qquad \textbf{(Équation 4.24)}$$

L'indice de détection d'un patron de réponses obtenu au hasard est alors égal à

$$I_{hasard} = \frac{P_{hasard}}{P_{normal}} . \qquad \textbf{(Équation 4.25)}$$

Puisque les étudiants nous ont signalé qu'ils utiliseraient la stratégie de donner la mauvaise réponse quant ils connaissent la bonne réponse, la vraisemblance conditionnelle de production du patron de réponses associé est égale à

$$P_{inverse|\theta} = \prod_{g=1}^{n} P(X_g = 1|\theta)^{1-X_g} P(X_g = 0|\theta)^{X_g}, \qquad \textbf{(Équation 4.26)}$$

tandis que la vraisemblance marginale correspond à

$$P_{inverse} = \int_{\theta=-\infty}^{\infty} P_{inverse|\theta}\, d\theta\,. \qquad \textbf{(Équation 4.27)}$$

L'indice associé correspond alors à

$$I_{inverse} = \frac{P_{inverse}}{P_{normal}}\,. \qquad \textbf{(Équation 4.28)}$$

Enfin, nous avons cru utile d'améliorer la détection par l'utilisation d'un indice de sous-classement global qui permettrait de combiner les stratégies de réponses au hasard et de réponses inversées. En fait, il s'agit de la vraisemblance conditionnelle de la production de réponses au hasard ou de la production de réponses inversées, soit :

$$P_{sous} = \prod_{g=1}^{n} [P_{hasard} + P_{inverse}] = \prod_{g=1}^{n} \left[\left(\frac{1}{k}\right)^{X_g} \left(\frac{k-1}{k}\right)^{1-X_g} P(X_g = 1 \mid \theta)^{1-X_g} P(X_g = 0 \mid \theta)^{X_g} \right].$$

(Équation 4.29)

La vraisemblance marginale est alors égale à

$$P_{sous} = \int_{\theta=-\infty}^{\infty} P_{sous|\theta}\, d\theta\,. \qquad \textbf{(Équation 4.30)}$$

L'indice associé devient alors égal à

$$I_{sous} = \frac{P_{sous}}{P_{normal}}. \qquad \textbf{(Équation 4.31)}$$

4.6 Choix quant aux modélisations et aux indices

Des indices de sous-classement basés sur les modélisations issues de la théorie de la réponse à l'item (TRI) ayant été élaborés, nous devons maintenant vérifier s'ils sont de bons prédicteurs du sous-classement intentionnel des étudiants qui suivent leurs cours d'anglais, langue seconde. Nous avons retenu la modélisation logistique à un paramètre de Rasch. Elle présente l'avantage, non négligeable, de permettre des simulations sans se préoccuper de l'impact, difficilement contrôlable, des paramètres de discrimination et de pseudo-chance. De plus, l'interprétation des résultats obtenus en est simplifiée car seul le paramètre de difficulté de l'item, notion un peu plus habituelle pour une intervenante ou un intervenant en éducation, est tenu en compte. Enfin, la modélisation logistique à un paramètre permettra plus facilement la mise en œuvre éventuelle d'une version informatisée et adaptative du test de classement en anglais, langue seconde, au collégial.

Quant aux indices de dépistage des patrons de réponses aberrants, seuls les indices l_z, *W, Zeta,* I_{hasard}, $I_{inverse}$ et I_{sous} sont retenus pour les simulations ainsi que les décisions finales. L'indice de dépistage, χ^2_B, propre à Bilog n'est utilisé qu'aux étapes préliminaires avant que soit réalisée, plus tard dans le cheminement du projet de recherche, l'informatisation des indices.

Chapitre 5

La mise à l'épreuve de certains indicateurs de patrons de réponses aberrants

À l'intérieur de ce chapitre nous mettons à l'épreuve les indices de détection de patrons de réponses aberrants que nous avons présentés au chapitre quatre : l_z, W, $Zeta$, I_{hasard}, $I_{inverse}$ et I_{sous}. Toutefois, selon plusieurs auteurs, les caractéristiques des distributions de probabilité de ces indices pourraient être fortement tributaires de la composition même du test et varier d'un niveau d'habileté à un autre (Bay et nering, 1998; Bedrick, 1997; Klauer, 1995; Liou, 1993; Meijer et Molenaar et Sijtsma, 1994; Meijer, Muijtjens et van der Vleuten, 1996; Meijer et Nering, 1997, Meijer et Sijtsma, 2001; Meijer et van Krimpen-Stoop, 1998, 2000; Molenaar et Hoijtink, 1990). Entre autres, la moyenne, ainsi que l'écart-type, pourraient s'éloigner considérablement des valeurs suggérées dans la littérature. Des essais préliminaires de notre part appuient ces réserves. C'est pourquoi, avant tout, nous avons jugé nécessaire d'étudier les caractéristiques de la distribution de probabilité de ces indices à partir d'une simulation où nous avons pu contrôler la nature des patrons de réponses aux items, qu'il s'agisse de patrons de réponses normaux, de patrons de réponses au hasard ou de patrons de réponses inversées.

De plus, puisque avec des données réelles nous ne connaissons pas à l'avance qui sont les élèves qui ont cherché à se sous-classer, il est impossible de mettre à l'épreuve les indices de détection du comportement de sous-classement. Ce n'est seulement qu'en simulant les patrons de réponses normaux et aberrants qu'il est possible de vérifier le pouvoir de détection des différents indices. De cette façon, nous pourrons connaître les valeurs critères des différents indices de détection de patrons de réponses aberrants à utiliser spécifiquement avec le TCALS II.

Cette section se terminera par l'application des valeurs critères obtenues à partir des simulations à des données réelles, soit les résultats de l'administration du TCALS II aux étudiantes et aux

étudiants des 1er et 2e tours du Service régional d'admission du Montréal métropolitain (SRAM) de la cohorte 2002-2003 au collège de l'Outaouais.

5.1 La simulation des patrons de réponses

Nous avons simulé 1 000 patrons de réponses dans 85 conditions différentes, soit 85 000 patrons de réponses différents. Pour réaliser cette simulation, un logiciel, destiné à la simulation de tests adaptatifs, élaboré auparavant par l'auteur (Raîche, 2002b), a été modifié. Plus précisément, nous avons généré de façon aléatoire 1000 patrons de réponses à 17 niveaux d'habileté, variant entre -2,00 et 2,00 par saut de 0,25, et selon cinq conditions :

1) patrons de réponses totalement normaux;
2) patrons de réponses avec 10 % de réponses générées au hasard;
3) patrons de réponses avec 20 % de réponses générées au hasard;
4) patrons de réponses avec 10 % de réponses inversées;
5) patrons de réponses avec 20 % de réponses inversées.

Un patron de réponses dit normal est produit en utilisant une procédure courante dans les travaux relatifs à la théorie de la réponse à l'item (Harwell, Stone, Hsu et Kirisci, 1996, p. 116; Raîche, 2001a, p. 128-129). En premier lieu, pour chaque item du test, un nombre aléatoire se distribuant entre 0 et 1 selon une loi de probabilité uniforme $U(0,1)$ est généré. Ce nombre est ensuite comparé à la probabilité théorique d'obtenir une bonne réponse. Les paramètres d'items calibrés à partir des quatre tours du SRAM de la cohorte 1998-1999 au collège de l'Outaouais sont utilisés pour calculer la probabilité d'obtenir une bonne réponse à l'item. Si la probabilité d'obtenir une bonne réponse à l'item affiche une valeur supérieure ou égale au nombre aléatoire généré, alors on note une bonne réponse. Sinon, on note une mauvaise réponse. Ainsi,

$$\text{si } P(X_g = 1 \mid \theta) \geq U(0,1), \text{ alors } X_g = 1; \text{ sinon } X_g = 0. \qquad \textbf{(Équation 5.1)}$$

Un patron de réponse qui aurait été produit au hasard par une étudiante ou un étudiant est généré à partir de la stratégie suivante. Un nombre aléatoire se distribuant entre 0 et 1 selon une loi uniforme $U(0,1)$ est généré. Si l'inverse du nombre du choix de réponses, k, est supérieur ou égal au nombre généré, on note une bonne réponse. Sinon, on note une mauvaise réponse. Ainsi,

$$\text{si } \frac{1}{k} \geq U(0,1), \text{ alors } X_g = 1; \text{ sinon } X_g = 0. \qquad \textbf{(Équation 5.2)}$$

Toutefois, puisqu'il nécessaire de détecter une étudiante ou un étudiant qui cherche à se sous-classer en ne répondant au hasard qu'à quelques questions, seulement 10 % et 20 % des réponses au test sont simulées de cette façon. Les autres questions sont simulées selon un patron de réponses normal. De plus, nous postulons que l'étudiante ou l'étudiant, en cherchant à contrôler le nombre de mauvaises réponses, n'applique la stratégie de réponses au hasard qu'au début du test. Ce sont ainsi les premières questions du test qui sont générées au hasard.

Quant à la stratégie de donner une mauvaise réponse à la question, nous postulons encore que l'étudiante ou l'étudiant n'inversera son choix de réponses qu'au début du test et pour seulement quelques questions. Nous simulons ainsi 10 % et 20 % de réponses inversées aux premières questions du test en notant, cette fois-ci, une mauvaise réponse lorsque la probabilité d'obtenir une bonne réponse à l'item affiche une valeur supérieure ou égale à un nombre aléatoire généré selon une distribution de probabilité $U(0,1)$. Sinon, on note une bonne réponse. Ainsi,

$$\text{si } P(X_g = 1 \mid \theta) \leq U(0,1), \text{ alors } X_g = 1; \text{ sinon } X_g = 0. \qquad \textbf{(Équation 5.3)}$$

À une première étape, nous effectuons la simulation de patrons de réponses normaux. L'analyse de la distribution de probabilité de chacun des indices de détection à chaque niveau d'habileté est réalisée. L'analyse de l'efficacité des indices de détection est par la suite réalisée en utilisant les valeurs critères obtenues aux simulations de patrons de réponses normaux.

5.2 Les résultats des simulations

5.2.1 Les patrons de réponses au hasard

Lorsque l'étudiante ou l'étudiant choisit la réponse au hasard à seulement 10 % des premières questions du test, soit les 9 premières questions, la distribution de probabilité de l'estimateur du niveau d'habileté est significativement affectée. Ainsi, comme l'attestent les résultats présentés au tableau 5.1, plus le niveau d'habileté en anglais, langue seconde, θ, est élevé, plus cette distribution s'éloigne de la distribution de probabilité de l'estimateur du niveau d'habileté, $\hat{\theta}$, que nous avions observée au tableau 2.5. Il est assez évident que la valeur de l'estimateur du niveau d'habileté en vient à s'éloigner rapidement du niveau d'habileté.

Ainsi, lorsque le niveau d'habileté est égal à 2,00, valeur représentative d'une étudiante ou d'un étudiant du niveau de classement 4, l'estimateur du niveau d'habileté est, en moyenne égal, à 0,60, valeur plutôt représentative d'une étudiante ou d'un étudiant au maximum du niveau de classement 2. Au mieux, quelques rares étudiantes ou étudiants seront classés au niveau 4, tel que l'atteste le maximum de l'estimateur du niveau d'habileté de 1,51. En fait, si on ne détecte pas la tentative de sous-classement, presque tous les étudiantes et les étudiants du niveau de classement 4 seront classés aux niveaux de classement 2 ou 3.

L'analyse des maximums et des minimums nous révèle que les étudiantes et les étudiants du niveau 3 (θ = 0,75 à 1,25) sont significativement moins affectés que les étudiantes et les étudiants du niveau de classement 4 puisqu'ils sont au plus déclassés d'un seul niveau. Une plus grande proportion sera d'ailleurs classée correctement au niveau 3. L'impact de la stratégie de réponses au hasard est donc nettement moins efficace pour ces étudiantes et étudiants.

Le niveau de classement 2 couvre une assez large étendue du niveau d'habileté en anglais, langue seconde. Tous ceux et celles qui ont un niveau d'habileté supérieur ou égal à 0,00 ne pourront pas se sous-classer en utilisant cette stratégie. Seulement les étudiantes et les étudiants les plus faibles du niveau 2 (θ = -0,25 et –0,50) pourraient occasionnellement réussir à se sous-classer au niveau 1.

Quant aux étudiantes et aux étudiants du niveau 1, les minimums nous indiquent que quelques étudiantes et étudiants pourraient se sous-classer en mise à niveau si leur niveau d'habileté était inférieur à -0,75.

Ainsi, avec seulement 10 % de choix de réponses au hasard, les étudiantes et les étudiants, principalement les plus forts, peuvent facilement réussir à se sous-classer d'au moins un niveau, sinon de deux niveaux.

Tableau 5.1

Distribution de l'estimateur du niveau d'habileté ($\hat{\theta}$) en fonction du niveau d'habileté (θ) en anglais, langue seconde (réponses au hasard = 10 % et réponses inversées = 0 %)

θ (Total)*	$\hat{\theta}$	$S_{\hat{\theta}}$	$Min_{\hat{\theta}}$	$Max_{\hat{\theta}}$	$Asymétrie_{\hat{\theta}}$	$Kurtose_{\hat{\theta}}$
-2,00 (26)	-2,09	0,16	-2,96	-1,56	-0,37	0,77
-1,75 (30)	-1,85	0,16	-2,59	-1,34	-0,25	0,48
-1,50 (34)	-1,62	0,15	-2,20	-0,99	-0,02	0,59
-1,25 (38)	-1,37	0,15	-1,92	-0,89	-0,03	0,03
-1,00 (42)	-1,15	0,15	-1,63	-0,58	-0,11	0,18
-0,75 (46)	-0,92	0,14	-1,31	-0,46	0,08	-0,15
-0,50 (49)	-0,70	0,15	-1,20	-0,13	0,26	0,26
-0,25 (53)	-0,50	0,15	-0,99	0,04	0,10	0,13
0,00 (57)	-0,29	0,16	-0,74	0,24	0,11	-0,06
0,25 (61)	-0,10	0,16	-0,50	0,59	0,44	0,64
0,50 (65)	0,07	0,16	-0,37	0,71	0,36	0,28
0,75 (69)	0,21	0,17	-0,23	0,84	0,34	0,26
1,00 (73)	0,34	0,18	-0,18	1,01	0,37	0,57
1,25 (77)	0,43	0,17	-0,02	1,01	0,49	0,40
1,50 (80)	0,50	0,17	0,04	1,22	0,39	0,31
1,75 (84)	0,57	0,16	0,12	1,51	0,68	1,63
2,00 (85)	0,60	0,15	0,24	1,51	0,80	1,72

* Le total, entre parenthèses, correspond à une estimation du nombre de bonnes réponses au test

L'analyse du tableau 5.2 nous révèle que lorsque l'étudiante ou l'étudiant utilise la stratégie de réponses au hasard sur le premier 20 % des questions du test, son taux de réussite, si on peut s'exprimer ainsi, est considérablement augmenté. Comme nous l'indiquent les maximums et les minimums, presque tous les étudiantes et les étudiants du niveau 4 réussissent à se sous-classer au niveau 2. Parmi eux, seulement ceux et celles dont le niveau d'habileté est très élevé en anglais, langue seconde, ne réussissent qu'à se sous-classer d'un seul niveau, mais ils sont à la limite du sous-classement au niveau 2.

Les étudiantes et les étudiants du niveau 3 se sous-classent, sans exception, d'un niveau. Cette stratégie est donc bien plus efficace pour eux que celle de ne répondre au hasard qu'à seulement 10 % des questions du test.

Tableau 5.2

Distribution de l'estimateur du niveau d'habileté ($\hat{\theta}$) en fonction du niveau d'habileté (θ) en anglais, langue seconde (réponses au hasard = 20 % et réponses inversées = 0 %)

θ (Total)*	$\hat{\theta}$	$S_{\hat{\theta}}$	$Min_{\hat{\theta}}$	$Max_{\hat{\theta}}$	$Asymétrie_{\hat{\theta}}$	$Kurtose_{\hat{\theta}}$
-2,00 (26)	-2,11	0,18	-2,75	-1,56	-0,42	0,48
-1,75 (30)	-1,91	0,17	-2,51	-1,41	-0,15	-0,10
-1,50 (34)	-1,69	0,15	-2,15	-1,24	-0,14	-0,04
-1,25 (38)	-1,46	0,15	-2,00	-0,99	-0,18	0,05
-1,00 (42)	-1,26	0,15	-1,75	-0,74	0,00	0,21
-0,75 (46)	-1,05	0,14	-1,48	-0,58	0,04	0,03
-0,50 (49)	-0,86	0,13	-1,27	-0,42	0,21	0,23
-0,25 (53)	-0,68	0,14	-1,10	-0,13	0,07	0,11
0,00 (57)	-0,54	0,14	-0,92	0,04	0,11	0,14
0,25 (61)	-0,38	0,14	-0,77	0,10	0,36	0,12
0,50 (65)	-0,28	0,14	-0,70	0,24	0,23	0,41
0,75 (69)	-0,19	0,13	-0,54	0,32	0,32	0,41
1,00 (73)	-0,11	0,13	-0,50	0,40	0,40	0,41
1,25 (77)	-0,50	0,12	-0,42	0,49	0,39	0,50
1,50 (80)	-0,01	0,12	-0,33	0,49	0,48	0,82
1,75 (84)	0,02	0,12	-0,33	0,40	0,29	0,19
2,00 (85)	0,04	0,12	-0,23	0,71	0,75	1,67

* Le total, entre parenthèses, correspond à une estimation du nombre de bonnes réponses au test

En ce qui concerne les étudiantes et les étudiants du niveau 2, plusieurs en arrivent à se sous-classer d'un niveau. Enfin, plus d'étudiantes et d'étudiants du niveau 1 réussissent aussi à se sous-classer vers la mise à niveau.

De manière générale, on remarque que plus une étudiante ou un étudiant répond au hasard aux questions du TCALS II, plus il peut réussir à se sous-classer, et ce de façon importante. De surcroît, plus son niveau d'habileté est élevé, plus la stratégie permet d'atteindre l'objectif de sous-classement. La figure 5.1 permet de comparer l'efficacité des stratégies de réponses au hasard à 10 % et à 20 % avec un comportement dit normal. On y voit bien la diminution de la valeur de l'estimateur du niveau d'habileté en fonction de l'importance du nombre d'items auxquels la réponse est effectuée au hasard.

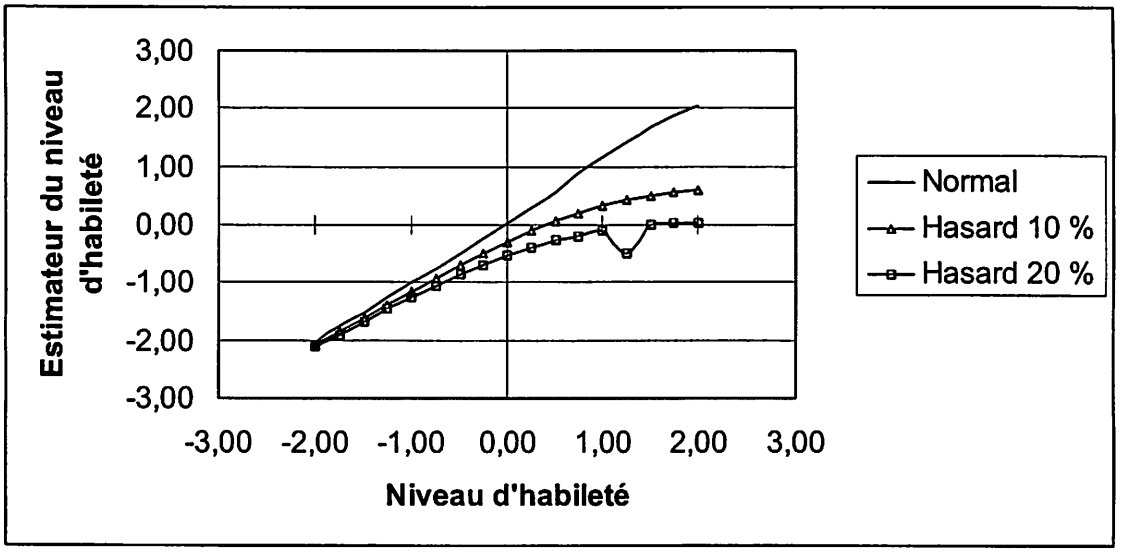

Figure 5.1 Estimateur du niveau d'habileté ($\hat{\theta}$) en fonction du niveau d'habileté (θ) en anglais, langue seconde, selon l'importance du nombre d'items auxquels la réponse est effectuée au hasard : 0, 10 et 20 %

Pour nous permettre de mieux saisir l'importance de la réussite du sous-classement par la stratégie de réponses au hasard, nous avons vérifié le pourcentage d'étudiantes et d'étudiants classés à chaque niveau de classement en fonction de leur niveau de classement réel et du taux de réponses données au hasard aux premières questions du test. Le tableau 5.3 présente les résultats de cette analyse. Pour chaque niveau de classement (niveau réel), le pourcentage d'étudiantes et d'étudiants classés à un niveau spécifique (niveau estimé) par le TCALS II est indiqué. Lorsque aucune réponse au hasard (0 %) n'est donnée aux premières questions du test, très peu de cas de sous-classements sont observés, au plus 17,60 %. De plus, les cas de sous-classements sont d'ailleurs plus fréquents aux niveaux 1, et 2 qu'aux niveaux 3 et 4. En fait, au niveau 3, on retrouve plus de cas de sur-classements (29,37 %) que de sous-classements (10,70 %). La situation est totalement différente lorsque le taux de réponses au hasard est de 10 %. Ainsi, aux niveaux 1 et 2, il y a sous-classement d'un niveau dans respectivement 33,90 % et 31,98 %. Au

niveau 3, le pourcentage de sous-classement d'un niveau atteint 95,00 %. Au niveau 4, la situation est plus dramatique. On y observe 25,03 % de cas de sous-classement d'un niveau et 74,90 % de cas de sous-classement de deux niveaux.

Tableau 5.3
Efficacité pour les étudiantes et les étudiants du sous-classement par la stratégie de réponses au hasard

	Niveau estimé par le résultat au TCALS II				
Niveau réel	Mise à niveau	Niveau 1	Niveau 2	Niveau 3	Niveau 4
Mise à niveau (0 %)	98,57 %	1,43 %	0,00 %	0,00 %	0,00 %
(10 %)	99,63 %	0,37 %	0,00 %	0,00 %	0,00 %
(20 %)	99,93 %	0,07 %	0,00 %	0,00 %	n.d.
Niveau 1 (0 %)	17,60 %	81,20 %	1,20 %	0,00 %	0,00 %
(10 %)	33,90 %	66,07 %	0,03 %	0,00 %	0,00 %
(20 %)	50,03 %	49,97 %	0,00 %	0,00 %	n.d.
Niveau 2 (0 %)	0,00 %	12,40 %	80,24 %	7,14 %	0,22 %
(10 %)	0,00 %	31,98 %	67,98 %	0,04 %	0,00 %
(20 %)	0,02 %	57,60 %	42,38 %	0,00 %	n.d.
Niveau 3 (0 %)	0,00 %	0,00 %	10,70 %	59,93 %	29,37 %
(10 %)	0,00 %	0,00 %	95,00 %	5,00 %	0,00 %
(20 %)	0,00 %	0,17 %	99,83 %	0,00 %	n.d.
Niveau 4 (0 %)	0,00 %	0,00 %	0,00 %	12,53 %	87,47 %
(10 %)	0,00 %	0,00 %	74,90 %	25,03 %	0,07 %
(20 %)	0,00 %	0,00 %	99,97 %	0,03 %	n.d.

Quand le taux de réponses données au hasard aux premières questions du test grimpe à 20 %, les résultats sont encore plus frappants. Ainsi, 50,03 % de cas de sous-classement d'un niveau sont observés au niveau 1. Au niveau 2, on observe même quelques cas de sous-classement de deux niveaux (0,02 %). Au niveau 3, la plupart des cas de sous-classement sont d'un niveau (99,83 %), tandis qu'au niveau 4 la majorité des cas de sous-classement sont de deux niveaux (99,97 %). Nous concluons, une fois de plus, à l'efficacité pour les étudiantes et les étudiants de la stratégie de sous-classement par un choix de réponses au hasard aux premières questions du TCALS II. Un

taux de réponses données au hasard aussi bas que 10 % est extrêmement efficace pour une étudiante ou un étudiant.

Après avoir noté l'ampleur de la réussite des étudiantes et des étudiants à se sous-classer en utilisant une des stratégies de réponses au hasard, nous allons maintenant évaluer notre taux de réussite à les détecter.

Le tableau 5.4 présente, pour chacun des indices de détection non spécifiques, les valeurs à partir desquelles nous pouvons prendre la décision de considérer que l'étudiante ou l'étudiant a adopté une stratégie de sous-classement en acceptant de cibler par erreur dans 5 % des cas une étudiante ou étudiant qui a adopté un comportement honnête au test. En premier, lieu, on remarque que les valeurs critères, quel que soit l'indice de détection, varient selon la valeur de l'estimateur du niveau d'habileté. Cela confirme les observations de plusieurs auteurs (Meijer, 1996, 1997; Meijer et Sijtsma, 2001; Meijer et van Krimpen-Stoop, 1998; Molenaar et Hoijtink, 1990, 1996). Il aurait donc été inapproprié d'utiliser une valeur commune quel que soit l'estimateur du niveau d'habileté. De plus, en ce qui a trait aux indices l_z et *Zeta*, la valeur critère s'éloigne fréquemment de façon assez importante de la valeur critère généralement suggérée de $-1{,}65$ pour l_z et de 1,65 pour *Zeta*. Définitivement, les indices l_z et *Zeta* ne se distribuent pas selon une loi de probabilité N(0,1). Nos résultats rejoignent ainsi ceux des auteurs qui ont analysé de plus près la distribution de probabilité de ces indices : celle de l_z, principalement, n'est pas une distribution normale centrée réduite et elle n'est même pas une distribution normale. De plus, les recherches récentes soutiennent aussi que cette distribution de probabilité varie en fonction du niveau d'habileté, ce que nous observons aussi.

Tableau 5.4
Valeurs critères des indices d'ajustement non optimaux du patron de réponses et proportion de patrons détectés selon le niveau d'habileté estimé lorsque les réponses sont produites au hasard (probabilité de fausse détection = 0,05)

	l_z			W			*Zeta*		
$\hat{\theta}$	Critère	10 %	20 %	Critère	10 %	20 %	Critère	10 %	20 %
-2,00	-1,01	0,66	0,73	1,17	0,45	0,63	1,45	0,49	0,68
-1,75	-1,24	0,21	0,38	1,16	0,20	0,38	1,54	0,20	0,40
-1,50	-1,41	0,19	0,41	1,16	0,19	0,43	1,60	0,19	0,43
-1,25	-1,45	0,30	0,59	1,14	0,31	0,62	1,60	0,28	0,59
-1,00	-1,53	0,38	0,72	1,14	0,38	0,73	1,49	0,38	0,73
-0,75	-1,46	0,50	0,86	1,15	0,44	0,81	1,54	0,44	0,81
-0,50	-1,57	0,54	0,92	1,16	0,49	0,88	1,65	0,47	0,88
-0,25	-1,27	0,82	0,99	1,17	0,70	0,97	1,49	0,71	0,97
0,00	-1,21	0,91	1,00	1,22	0,69	0,97	1,54	0,76	0,99
0,25	-0,87	0,98	1,00	1,25	0,67	0,98	1,10	0,89	1,00
0,50	-0,88	0,99	1.00	1,23	0,84	0,95	1,26	0,92	1,00
0,75	-0,94	0,99	1.00	1,26	0,84	1,00	1,36	0,92	1,00
1,00	-0,94	1,00	(nd.)	1,26	0,85	(nd.)	1,36	0,97	(nd.)
1,25	-0,87	1,00	(nd.)	1,27	0,67	(nd.)	1,14	0,83	(nd.)
1,50	-0,81	1,00	(nd.)	1,27	1,00	(nd.)	1,12	1,00	(nd.)
1,75	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)
2,00	-0,77	(nd.)	(nd.)	1,27	(nd.)	(nd.)	0,97	(nd.)	(nd.)

Le taux de détection des patrons de réponses au hasard est assez important, quel que soit l'indice utilisé et quel que soit le taux de réponses au hasard. Globalement, toutefois, c'est l'indice l_z qui semble le plus efficace. Cette dernière constatation confirme les résultats observés à l'intérieur de la plupart des recherches sur les indices de détection des patrons de réponses aberrants (Meijer et Sijtsma, 2001; Nering et Meijer, 1998). Ainsi, une étudiante ou un étudiant dont l'estimateur du niveau d'habileté serait égal ou supérieur à –0,50, soit représentatif du niveau de classement 2 ou plus, serait détecté dans 92 % des cas par l'indice l_z quand il répond au hasard à 20 % des premières questions. Même les étudiantes et les étudiants dont l'estimateur du niveau d'habileté est représentatif du niveau 1, soit quand $\hat{\theta}$ = –1,25 à –0,75, affichent un taux de détection assez élevé, soit entre 59 % et 86 % quand ils répondent au hasard à 20 % des premières questions. Les

autres indices, *W* et *Zeta*, sont aussi assez efficaces. Toutefois, c'est clairement l'indice l_z qui est le plus utile. En se référant au tableau 5.2, on peut remarquer que les étudiantes et les étudiants des niveaux 3 et 4 qui répondent au hasard à 20 % des premières questions obtiennent une valeur de l'estimateur du niveau d'habileté dont les minimums et maximums varient entre –0,54 et 0,71. Les étudiantes et les étudiants des niveaux 3 et 4 sont ainsi détectés dans 92 % des cas et plus.

Le taux de détection est bien sûr inférieur lorsque l'étudiante ou l'étudiant ne répond au hasard qu'à 10 % des premières questions. Tout de même, avec l'indice l_z, il est possible de détecter au moins 82 % des étudiantes et des étudiants dont l'estimateur du niveau d'habileté est supérieur à –0,50. Lorsque l'estimateur du niveau d'habileté est égal ou supérieur à 0,00, le taux de détection est égal ou supérieur à 98 %. Utilisant le même raisonnement que précédemment, on remarque au tableau 5.1 que les étudiantes et les étudiants des niveaux 3 et 4 qui répondent au hasard à 10 % des premières questions obtiennent une valeur de l'estimateur du niveau d'habileté dont les minimums et maximums varient entre –0,23 et 1,51. Les étudiantes et les étudiants des niveaux de classement 3 et 4 sont ainsi détectés dans 82 % des cas et plus.

L'indice l_z, s'avère donc extrêmement efficace pour détecter les étudiantes et les étudiants qui répondent au hasard à aussi peu que 10 % des premières questions. Ceux et celles qui utilisent cette stratégie avec un taux de réponses au hasard supérieur à 10 % sont détectés presque à coup sûr.

Les résultats de l'analyse des taux de détection selon les indices optimaux, I_{hasard}, $I_{inversée}$ et I_{sous}, sont présentés au tableau 5.5. Ces indices affichent une performance inférieure à ceux de l_z, indice non optimal, ceci quelle que soit le taux de réponses données au hasard. Étrangement, l'indice I_{hasard} est l'indicateur le moins efficace. Lorsque le taux de réponse au hasard n'est que de 10 %, il permet d'obtenir un taux de détection raisonnable seulement si l'estimateur du niveau d'habileté est égal ou supérieur à 1,50. Cela ne représente aucun intérêt pour le dépistage du sous-classement puisque l'étudiante ou l'étudiant se retrouve alors classé au niveau le plus élevé tout de même. Lorsqu'il répond au hasard à 20 % des premières questions, il doit obtenir au moins une valeur de l'estimateur du niveau d'habileté égale à 0,50 pour être détecté dans 91 % des cas.

En se référant aux maximums et aux minimums observés au tableau 5.1, on peut déduire que plusieurs étudiantes et étudiants des niveaux 3 et 4 seraient détectés correctement, mais pas tous cependant. Cette constatation peut paraître surprenante à prime abord. Cependant, en y regardant de plus près, on doit admettre que plusieurs caractéristiques de l'indice I_{hasard} peuvent expliquer ce faible rendement. Premièrement, l'indice est utilisé pour détecter un comportement très spécifique, contrairement aux indices non optimaux. Deuxièmement, l'indice ne repose pas sur une procédure vraiment optimale car plutôt que de s'intéresser à 10 % ou 20 % des premières questions répondues au hasard, il traite toutes les questions du test comme étant répondues au hasard.

Tableau 5.5
Valeurs critères des indices d'ajustement optimaux du patron de réponses et proportion de patrons de réponses détectés selon le niveau d'habileté estimé lorsque les réponses sont produites au hasard (probabilité de fausse détection = 0,05)

	I_{hasard}			$I_{inversée}$			I_{sous}		
$\hat{\theta}$	Critère	10 %	20 %	Critère	10 %	20 %	Critère	10 %	20 %
-2,00	0,258179	0,16	0,30	4,29 E-08	0,44	0,59	0,258179	0,18	0,35
-1,75	5,151367	0,14	0,30	1,25 E-07	0,16	0,30	5,151367	0,14	0,30
-1,50	1,331787	0,12	0,25	2,67 E-09	0,14	0,31	1,331787	0,12	0,25
-1,25	0,029076	0,13	0,23	9,54 E-12	0,25	0,52	0,029076	0,13	0,23
-1,00	0,000088	0,13	0,20	3,75 E-13	0,38	0,71	0,000088	0,13	0,21
-0,75	3,17 E-08	0,11	0,18	3,61 E-13	0,47	0,83	3,67 E-08	0,17	0,43
-0,50	9,18 E-12	0,11	0,12	2,29 E-12	0,55	0,91	6,51 E-11	0,45	0,86
-0,25	1,93 E-16	0,09	0,04	5,11 E-10	0,69	0,98	5,11 E-10	0,69	0,98
0,00	1,33 E-21	0,05	0,05	0,001713	0,55	0,91	0,001713	0,55	0,91
0,25	1,50 E-27	0,06	0,44	0,001758	0,69	0,96	0,001758	0,69	0,96
0,50	1,04 E-34	0,18	0,91	0,001713	0,90	0,95	0,001713	0,90	0,95
0,75	2,37 E-37	0,31	1,00	0,001713	0,94	1,00	0,001713	0,94	1,00
1,00	2,94 E-40	0,60	(nd.)	0,001713	0,98	(nd.)	0,001713	0,98	(nd.)
1,25	3,65 E-42	0,50	(nd.)	0,003540	0,83	(nd.)	0,003540	0,83	(nd.)
1,50	1,38 E-44	1,00	(nd.)	0,006632	1,00	(nd.)	0,006632	1,00	(nd.)
1,75	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)
2,00	1,31 E-48	(nd.)	(nd.)	0,007862	(nd.)	(nd.)	0,007862	(nd.)	(nd.)

Les indices $I_{inversée}$ et I_{sous} sont beaucoup plus performants que l'indice I_{hasard}. Leur efficacité, légèrement inférieure à celle de l_z, est même comparable à celle des indices non optimaux *W* et *Zeta*, indépendamment du taux de réponses données au hasard.

Nous tenons aussi à souligner que l'hypothèse d'indépendance entre l'événement *répondre au hasard* et l'événement *répondre par des questions inversées*, postulée par l'équation 4.29, est douteuse. Quand on vérifie les résultats de détection obtenus au tableau 5.5, on remarque que les taux de détection de I_{sous} sont quelquefois inférieurs à ceux de $I_{inversée}$. L'hypothèse d'indépendance devrait nous mener à ce que le taux de détection de I_{sous} soit du moins le plus important parmi I_{hasard} et $I_{inversée}$.

Pour être en mesure de bien évaluer l'efficacité de l'indice l_z et ainsi nous permettre de détecter une étudiante ou étudiant qui utilise la stratégie de réponses au hasard, nous avons vérifié le pourcentage d'étudiantes et d'étudiants sous-classés d'au moins un niveau qui a été effectivement détecté par l'indice l_z. Cette analyse est réalisée en fonction du niveau réel de l'étudiante ou de l'étudiant et selon le taux de réponses données au hasard aux premières questions du test. Le tableau 5.6 présente les résultats de cette analyse. Pour chaque niveau de classement (niveau réel), le pourcentage d'étudiantes et d'étudiants sous-classés d'au moins un niveau et détectés par l'indice l_z est indiqué.

Lorsque aucune réponse au hasard (0 %) n'est donnée aux premières questions du test, très peu de cas sont faussement détectés, quel que soit le niveau réel de l'étudiante ou de l'étudiant. Le pourcentage de fausses détections varie entre 2,93 %, au niveau de classement 4, et 6,85 %, au niveau de classement 3. Nous aurions tout de même apprécié ne pas dépasser un pourcentage de fausse détection de 5 % à tous les niveaux de classement.

Quand 10 % des premières réponses sont données au hasard, le taux de détection des cas réels de sous-classements est supérieur à 90 % pour les étudiantes et les étudiants des niveaux 3 et 4. Au niveau 2, tout de même, un peu plus de la moitié des cas sont détectés, tandis que seulement 25,76 % des cas peuvent être dépistés au niveau 1. Lorsque 20 % des réponses sont données au

hasard, ces pourcentages augmentent considérablement. Même au niveau 1, plus de la moitié des cas de sous-classement sont détectés, tandis qu'aux niveaux 2, 3 et 4 le taux de détection est supérieur à 86,00 %. L'indice l_z constitue donc un indicateur très efficace pour nous permettre de détecter l'utilisation d'une stratégie de réponses au hasard.

Tableau 5.6
Efficacité du dépistage du sous-classement par la stratégie de réponses au hasard

	Efficacité de la détection de l'indice l_z	
Niveau réel	Non détecté	Détecté
Niveau 1 (0 %)	95,08 %	4,92 %
(10 %)	74,24 %	25,76 %
(20 %)	49,63 %	50,37 %
Niveau 2 (0 %)	95,48 %	4,52 %
(10 %)	48,41 %	51,59 %
(20 %)	13,09 %	86,91 %
Niveau 3 (0 %)	93,15 %	6,85 %
(10 %)	2,95 %	97,05 %
(20 %)	0,30 %	99,70 %
Niveau 4 (0 %)	97,07 %	2,93 %
(10 %)	0,27 %	99,73 %
(20 %)	0,00 %	100,00 %

5.2.2 Les patrons de réponses inversées

L'impact de la stratégie de donner la mauvaise réponse aux premières questions du test est beaucoup plus important que celui que nous avons observé avec la stratégie de répondre au hasard à ces mêmes questions. Ainsi, comme nous l'indiquent les tableaux 5.7 et 5.8, ainsi que la figure 5.2, la distribution de probabilité de l'estimateur du niveau d'habileté est fortement affectée et sa moyenne s'éloigne considérablement de la valeur correspondante du niveau d'habileté.

Tableau 5.7

Distribution de l'estimateur du niveau d'habileté ($\hat{\theta}$) en fonction du niveau d'habileté (θ) en anglais, langue seconde (réponses au hasard = 0 % et réponses inversées = 10 %)

θ (Total)*	$\hat{\theta}$	$S_{\hat{\theta}}$	$Min_{\hat{\theta}}$	$Max_{\hat{\theta}}$	$Asymétrie_{\hat{\theta}}$	$Kurtose_{\hat{\theta}}$
-2,00 (26)	-1,94	0,16	-2,44	-1,48	-0,11	-0,11
-1,75 (30)	-1,75	0,16	-2,38	-1,24	0,00	-0,09
-1,50 (34)	-1,54	0,15	-2,20	-1,06	-0,36	0,19
-1,25 (38)	-1,34	0,14	-1,87	-0,85	-0,04	0,13
-1,00 (42)	-1,14	0,15	-1,56	-0,66	0,15	-0,01
-0,75 (46)	-0,92	0,15	-1,56	-0,46	-0,05	0,20
-0,50 (49)	-0,71	0,14	-1,17	-0,18	0,18	0,19
-0,25 (53)	-0,51	0,14	-0,99	-0,02	0,06	0,09
0,00 (57)	-0,33	0,14	-0,70	0,17	0,14	-0,11
0,25 (61)	-0,15	0,15	-0,58	0,24	0,20	-0,08
0,50 (65)	-0,01	0,14	-0,46	0,59	0,28	0,50
0,75 (69)	0,10	0,14	-0,23	0,59	0,15	-0,05
1,00 (73)	0,22	0,13	-0,13	0,71	0,00	-0,11
1,25 (77)	0,30	0,11	-0,08	0,59	-0,20	-0,14
1,50 (80)	0,36	0,10	-0,08	0,59	-0,33	0,05
1,75 (84)	0,40	0,09	0,04	0,71	-0,61	0,25
2,00 (85)	0,43	0,08	0,10	0,59	-0,86	0,62

* Le total, entre parenthèses, correspond à une estimation du nombre de bonnes réponses au test

Au tableau 5.1, on avait observé que lorsque 10 % des premières réponses étaient données au hasard et que le niveau d'habileté était de 2,00, les minimums et maximums de l'estimateur du niveau d'habileté étaient respectivement de 0,24 et de 1,51. Avec la stratégie de réponses inversées, au même niveau d'habileté, l'estimateur du niveau d'habileté varie maintenant entre 0,10 et 0,59. Les cas de sous-classement de deux niveaux deviennent alors plus fréquents. La stratégie est donc plus efficace lorsque le niveau d'habileté est élevé. Cependant, aux niveaux d'habileté les plus faibles, soient ceux de la mise à niveau et du niveau 1, la stratégie est généralement moins efficace. Il est alors possible d'obtenir une valeur de l'estimateur du niveau d'habileté plus faible à partir de la stratégie de réponses au hasard à 10 % des premières questions du test. Par exemple, la valeur minimale de l'estimateur du niveau d'habileté, lorsque le niveau

d'habileté est de –2,00 et que 10 % des réponses sont données au hasard, est égale à –2,96. Quand la stratégie de réponses inversées est appliquée, le minimum n'est plus que de –2,44.

Tableau 5.8

Distribution de l'estimateur du niveau d'habileté ($\hat{\theta}$) en fonction du niveau d'habileté (θ) en anglais, langue seconde (réponses au hasard = 0 % et réponses inversées = 20 %)

θ (Total)*	$\hat{\theta}$	$S_{\hat{\theta}}$	$Min_{\hat{\theta}}$	$Max_{\hat{\theta}}$	$Asymétrie_{\hat{\theta}}$	$Kurtose_{\hat{\theta}}$
-2,00 (26)	-1,74	0,15	-2,26	-1,34	-0,12	-0,07
-1,75 (30)	-1,62	0,15	-2,10	-1,13	-0,04	-0,13
-1,50 (34)	-1,48	0,14	-2,10	-0,99	-0,13	0,27
-1,25 (38)	-1,32	0,15	-1,79	-0,85	0,03	0,14
-1,00 (42)	-1,15	0,14	-1,63	-0,66	-0,11	0,18
-0,75 (46)	-0,99	0,14	-1,52	-0,46	0,07	0,05
-0,50 (49)	-0,84	0,14	-1,34	-0,23	0,22	0,41
-0,25 (53)	-0,70	0,13	-1,10	-0,28	0,15	-0,09
0,00 (57)	-0,57	0,13	-0,99	-0,18	-0,03	0,02
0,25 (61)	-0,47	0,13	-0,81	-0,08	0,23	-0,04
0,50 (65)	-0,38	0,11	-0,70	-0,02	-0,01	-0,03
0,75 (69)	-0,32	0,09	-0,58	-0,02	-0,04	-0,16
1,00 (73)	-0,26	0,09	-0,58	-0,02	-0,13	-0,02
1,25 (77)	-0,22	0,08	-0,50	-0,02	-0,29	0,05
1,50 (80)	-0,19	0,06	-0,42	-0,02	-0,17	-0,02
1,75 (84)	-0,17	0,05	-0,37	-0,02	-0,45	0,51
2,00 (85)	-0,16	0,04	-0,33	-0,08	-0,63	0,43

* Le total, entre parenthèses, correspond à une estimation du nombre de bonnes réponses au test

Lorsque 20 % des premières réponses au test sont inversées, comme nous en informe le tableau 5.8, nous retrouvons la même tendance générale, mais beaucoup plus marquée. Ainsi, lorsque le niveau d'habileté est élevé, l'estimateur du niveau d'habileté est beaucoup plus influencé par cette stratégie de réponses inversées que par la stratégie de réponses au hasard, tandis qu'on observe la situation opposée lorsque le niveau d'habileté est faible. Il s'agit donc d'une stratégie de sous-classement qui avantage, au sens de la réussite du sous-classement, les plus forts. Les

plus faibles, pour se sous-classer, auraient plutôt avantage à utiliser la stratégie de réponses au hasard. Comme nous l'avions fait lors de l'analyse de la stratégie de réponses au hasard, nous reproduisons graphiquement à la figure 5.2 la relation entre le niveau d'habileté et la moyenne de l'estimateur du niveau d'habileté. La comparaison des figures 5.2 et 5.1 illustre bien les différences quant à l'efficacité de ces deux stratégies.

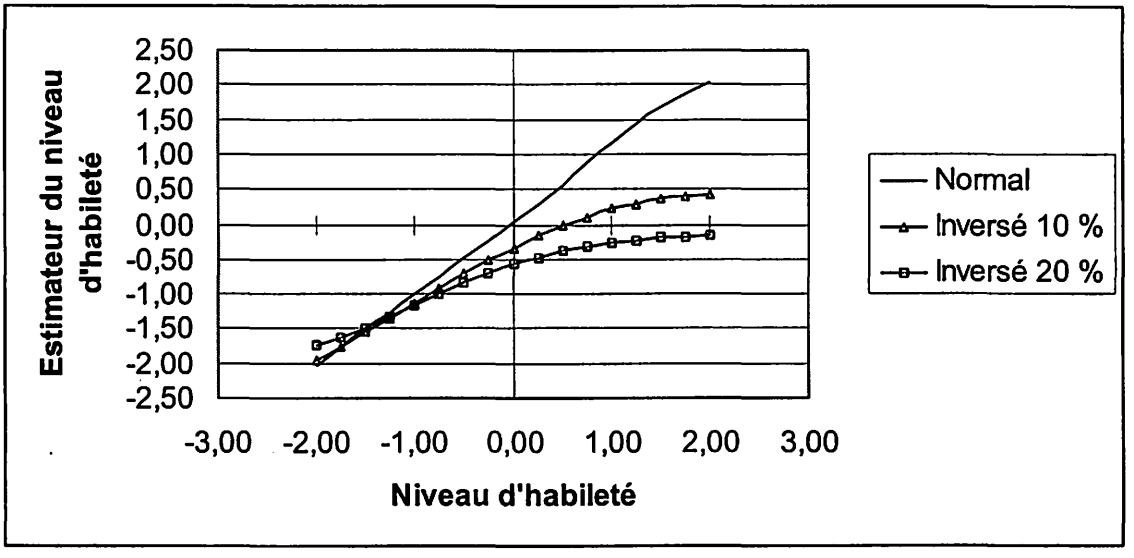

Figure 5.2 Estimateur du niveau d'habileté ($\hat{\theta}$) en fonction du niveau d'habileté (θ) en anglais, langue seconde, selon l'importance du nombre d'items auxquels la réponse est inversée : 0, 10 et 20 %.

Pour chaque niveau de classement (niveau réel), nous vérifions maintenant le pourcentage d'étudiantes et d'étudiants classés à un niveau spécifique (niveau estimé) par le TCALS II. Les résultats de cette analyse sont décrits au tableau 5.9. Les pourcentages de sous-classement lorsque aucune réponse au hasard (0 %) n'est donnée aux premières questions du test sont strictement les mêmes que ceux que nous avons présentés au tableau 5.3. Ils sont reproduits au tableau 5.9 uniquement pour faciliter les comparaisons. Lorsque le taux de réponses inversées est égal à

10 %, on peut observer que 32,37 % des individus du niveau 1 réussissent à se sous-classer vers une mise à niveau et que 33.96 % des individus du niveau 2 arrivent à se sous-classer au niveau 1. Au niveau 3, le pourcentage de sous-classement d'un niveau atteint 99,97 % tandis qu'au niveau 4, seulement 0,03 % se sous-classent d'un niveau, mais 99,97 % affichent un sous-classement de deux niveaux.

Tableau 5.9
Efficacité pour les étudiantes et les étudiants du sous-classement par la stratégie de réponses inversées

	Niveau estimé par le résultat au TCALS II				
Niveau réel	Mise à niveau	Niveau 1	Niveau 2	Niveau 3	Niveau 4
Mise à niveau (0 %)	98,57 %	1,43 %	0,00 %	0,00 %	0,00 %
(10 %)	99,47 %	0,53 %	0,00 %	0,00 %	0,00 %
(20 %)	98,20 %	1,80 %	0,00 %	n.d.	n.d.
Niveau 1 (0 %)	17,60 %	81,20 %	1,20 %	0,00 %	0,00 %
(10 %)	32,37 %	67,00 %	0,03 %	0,00 %	0,00 %
(20 %)	31,27 %	68,70 %	0,03 %	n.d.	n.d.
Niveau 2 (0 %)	0,00 %	12,40 %	80,24 %	7,14 %	0,22 %
(10 %)	0,00 %	33,96 %	66,04 %	0,00 %	0,00 %
(20 %)	0,02 %	66,98 %	33,00 %	n.d.	n.d.
Niveau 3 (0 %)	0,00 %	0,00 %	10,70 %	59,93 %	29,37 %
(10 %)	0,00 %	0,00 %	99,97 %	0,03 %	0,00 %
(20 %)	0,00 %	1,80 %	98,20 %	n.d.	n.d.
Niveau 4 (0 %)	0,00 %	0,00 %	0,00 %	12,53 %	87,47 %
(10 %)	0,00 %	0,00 %	99,97 %	0,03 %	0,00 %
(20 %)	0,00 %	0,00 %	100,00 %	n.d.	n.d.

Quand le taux de réponses inversées est égal à 20 %, 31,27 % de cas de sous-classement d'un niveau sont observés au niveau 1. Au niveau 2, le taux de sous-classement d'un niveau est égal à 66,98 % et quelques cas de sous-classement de deux niveaux (0,02 %) sont observés. Au niveau 3, on remarque 98,20 % de cas de sous-classement d'un seul niveau et 1,80 % de cas de sous-classement de deux niveaux. Enfin, au niveau 4, la totalité des cas de sous-classement sont de

deux niveaux (100,00 %). Nous concluons, une fois de plus, à l'efficacité pour les étudiantes et les étudiants de la stratégie de sous-classement par un choix de réponses inversées aux premières questions du TCALS II. Comme nous l'avons noté auparavant, la stratégie de réponses inversées est toutefois plus appropriée pour les étudiantes et les étudiants des niveaux 3 et 4, tandis que ceux et celles des niveaux 1 et 2 obtiennent de meilleurs résultats de sous-classement avec la stratégie de réponses au hasard.

Qu'en est-il maintenant de l'efficacité des indices l_z, *W* et *Zeta* pour détecter l'utilisation de la stratégie de sous-classement par réponses inversées? À cette fin, le tableau 5.10 nous indique de nouveau que l'indice l_z est généralement supérieur aux indices *W* et *Zeta*. De plus, comme les résultats de sous-classement que nous venons tout juste d'observer nous le font espérer, le taux de détection des cas de sous-classement est supérieur, à taux égal de réponses aberrantes, à celui obtenu à partir de la stratégie de réponses au hasard. On obtient déjà un taux de détection de 95 % lorsque l'estimateur du niveau d'habileté est égal à -0,25 avec aussi peu que 10 % de réponses inversées. Lorsque le taux de réponses inversées est égal à 20 %, le taux de détection est toujours égal ou supérieur à 80 %. L'indice l_z s'avère donc extrêmement efficace pour détecter les patrons de réponses aberrants issus de cette stratégie.

Tableau 5.10

Valeurs critères des indices d'ajustement non optimaux du patron de réponses et proportion de patrons détectés selon le niveau d'habileté estimé lorsque les réponses sont inversées (probabilité de fausse détection = 0,05)

	l_z			W			*Zeta*		
$\hat{\theta}$	Critère	10 %	20 %	Critère	10 %	20 %	Critère	10 %	20 %
-2,00	-1,01	0,75	0,96	1,17	0,51	0,90	1,45	0,58	0,93
-1,75	-1,24	0,34	0,90	1,16	0,27	0,80	1,54	0,28	0,83
-1,50	-1,41	0,19	0,80	1,16	0,20	0,72	1,60	0,20	0,74
-1,25	-1,45	0,31	0,86	1,14	0,33	0,84	1,60	0,31	0,82
-1,00	-1,53	0,42	0,92	1,14	0,43	0,92	1,49	0,43	0,92
-0,75	-1,46	0,58	0,98	1,15	0,51	0,97	1,54	0,51	0,97
-0,50	-1,57	0,75	1,00	1,16	0,99	0,96	1,65	0,68	0,99
-0,25	-1,27	0,95	1,00	1,17	0,85	1,00	1,49	0,85	1,00
0,00	-1,21	0,98	1,00	1,22	0,78	1,00	1,54	0,87	1,00
0,25	-0,87	1,00	(nd.)	1,25	0,88	(nd.)	1,10	0,99	(nd.)
0,50	-0,88	1,00	(nd.)	1,23	0,98	(nd.)	1,26	1,00	(nd.)
0,75	-0,94	1,00	(nd.)	1,26	1,00	(nd.)	1,36	1,00	(nd.)
1,00	-0,94	(nd.)	(nd.)	1,26	(nd.)	(nd.)	1,36	(nd.)	(nd.)
1,25	-0,87	(nd.)	(nd.)	1,27	(nd.)	(nd.)	1,14	(nd.)	(nd.)
1,50	-0,81	(nd.)	(nd.)	1,27	(nd.)	(nd.)	1,12	(nd.)	(nd.)
1,75	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)
2,00	-0,77	(nd.)	(nd.)	1,27	(nd.)	(nd.)	0,97	(nd.)	(nd.)

Comme nous l'avons observé plus haut, l'indice I_{hasard} est très peu efficace pour détecter les patrons de réponses aberrants et comme on pouvait s'en douter, son efficacité pour dépister l'utilisation de la stratégie de réponses inversées est encore moins importante que celle que nous avons observée pour dépister la stratégie de réponses au hasard. Cette remarque est valable quel que soit le taux de réponses inversées. Les indices $I_{inversée}$ et I_{sous} offrent encore une fois de meilleures performances que l'indice I_{hasard}, sans toutefois rivaliser sérieusement avec l'indice l_z. Ils ne rivalisent pas plus avec les indices W et *Zeta* lorsque l'estimateur du niveau d'habileté est inférieur à –1,00.

Tableau 5.11

Valeurs critères des indices d'ajustement optimaux du patron de réponses et proportion de patrons détectés selon le niveau d'habileté estimé lorsque les réponses sont inversées (probabilité de fausse détection = 0,05)

	I_{has}			I_{inv}			I_{sous}		
$\hat{\theta}$	Critère	10 %	20 %	Critère	10 %	20 %	Critère	10 %	20 %
-2,00	0,258179	0,07	0,11	4,29 E-08	0,34	0,79	0,258179	0,09	0,24
-1,75	5,151367	0,07	0,18	1,25 E-07	0,14	0,56	5,151367	0,07	0,22
-1,50	1,331787	0,09	0,30	2,67 E-09	0,10	0,56	1,331787	0,09	0,31
-1,25	0,029076	0,12	0,33	9,54 E-12	0,26	0,79	0,029076	0,12	0,33
-1,00	0,000088	0,11	0,26	3,75 E-13	0,42	0,92	0,000088	0,11	0,35
-0,75	3,17 E-08	0,13	0,23	3,61 E-13	0,55	0,97	3,67 E-08	0,21	0,76
-0,50	9,18 E-12	0,11	0,09	2,29 E-12	0,73	0,99	6,51 E-11	0,66	0,99
-0,25	1,93 E-16	0,06	0,04	5,11 E-10	0,87	1,00	5,11 E-10	0,88	1,00
0,00	1,33 E-21	0,04	0,25	0,001713	0,58	0,97	0,001713	0,58	0,97
0,25	1,50 E-27	0,06	(nd.)	0,001758	0,89	(nd.)	0,001758	0,89	(nd.)
0,50	1,04 E-34	0,28	(nd.)	0,001713	0,98	(nd.)	0,001713	0,98	(nd.)
0,75	2,37 E-37	1,00	(nd.)	0,001713	1,00	(nd.)	0,001713	1,00	(nd.)
1,00	2,94 E-40	(nd.)	(nd.)	0,001713	(nd.)	(nd.)	0,001713	(nd.)	(nd.)
1,25	3,65 E-42	(nd.)	(nd.)	0,003540	(nd.)	(nd.)	0,003540	(nd.)	(nd.)
1,50	1,38 E-44	(nd.)	(nd.)	0,006632	(nd.)	(nd.)	0,006632	(nd.)	(nd.)
1,75	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)	(nd.)
2,00	1,31 E-48	(nd.)	(nd.)	0,007862	(nd.)	(nd.)	0,007862	(nd.)	(nd.)

La figure 5.3 résume les observations que nous avons faites en ce qui a trait à l'efficacité de l'indice l_z, indice généralement le plus performant, en fonction de la valeur obtenue de l'estimateur du niveau d'habileté. On y remarque clairement la supériorité du taux de détection lorsque le taux de réponses aberrantes est égal à 20 %. Il est aussi évident que le taux de détection est plus élevé lorsque la stratégie de réponses inversées est adoptée.

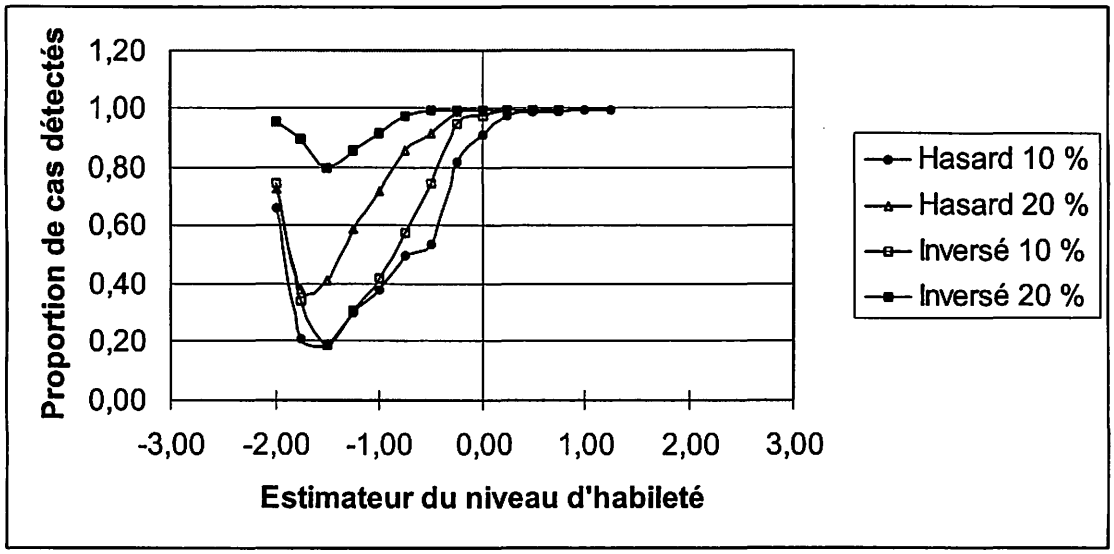

Figure 5.3 Proportion de patrons de réponses aberrants détectés en fonction de la valeur obtenue de l'estimateur du niveau d'habileté et selon le degré et le type de d'aberrance.

Quand nous vérifions le pourcentage d'étudiantes et d'étudiants sous-classés d'au moins un niveau qui a été effectivement détecté par l'indice l_z, ce dernier semble très efficace pour permettre l'identification d'une étudiante ou d'un étudiant qui utilise la stratégie de réponses inversées. Le tableau 5.12 présente ces taux de détection en fonction du niveau où aurait dû être classé l'étudiante ou l'étudiant. Pour chaque niveau de classement (niveau réel), le pourcentage d'étudiantes et d'étudiants sous-classés d'au moins un niveau et détectés par l'indice l_z est indiqué.

Quand 10 % des premières réponses sont données au hasard, le taux de détection des cas réels de sous-classement est égal à 100 % pour les étudiantes et les étudiants des niveaux 3 et 4. Au niveau 2, 66,20 % des cas sont détectés, tandis que seulement 30,28 % des cas sont détectés au niveau 1. Lorsque 20 % des réponses sont données au hasard, même au niveau 1, 87,53 % des cas de sous-classement sont détectés, tandis qu'aux niveaux 2, 3 et 4 le taux de détection est supérieur à 98,00 %. Quel que soit le niveau de classement, l'indice l_z constitue un indicateur extrêmement puissant pour détecter l'utilisation d'une stratégie de réponses inversées.

De plus, rappelons-nous les remarques que nous avons formulées lors de l'analyse des tableaux 5.8, 5.9, 5.10 et 5.11, suggérant que les étudiantes et les étudiants faibles réussissent moins bien à se sous-classer avec la stratégie de réponses inversées qu'avec la stratégie de réponses au hasard. Puisque l'indice l_z permet aussi de mieux dépister les étudiantes et les étudiants faibles qui utilisent la stratégie de réponses inversées que ceux et celles qui utilisent la stratégie de réponses au hasard, il est encore plus évident que la stratégie de réponses inversées est totalement inappropriée pour permettre aux étudiantes et aux étudiants faibles de se sous-classer.

Tableau 5.12
Efficacité du dépistage du sous-classement par la stratégie de réponses inversées

	Efficacité de la détection de l'indice l_z	
Niveau réel	Non détecté	Détecté
Niveau 1 (0 %)	94,73 %	5,27 %
(10 %)	69,72 %	30,28 %
(20 %)	12,47 %	87,53 %
Niveau 2 (0 %)	95,06 %	4,94 %
(10 %)	33,80 %	66,20 %
(20 %)	1,16 %	98,84 %
Niveau 3 (0 %)	94,33 %	5,67 %
(10 %)	0,00 %	100,00 %
(20 %)	0,00 %	100,00 %
Niveau 4 (0 %)	95,73 %	4,27 %
(10 %)	0,00 %	100,00 %
(20 %)	0,00 %	100,00 %

Les figures 5.4 et 5.5 illustrent graphiquement l'efficacité les résultats présentés au tableau 5.12. La figure 5.4 traite des résultats relatifs à l'application des stratégies à 10 % des premières questions du test, tandis que la figure 5.5 illustre l'impact des stratégies à 20 % des premières questions.

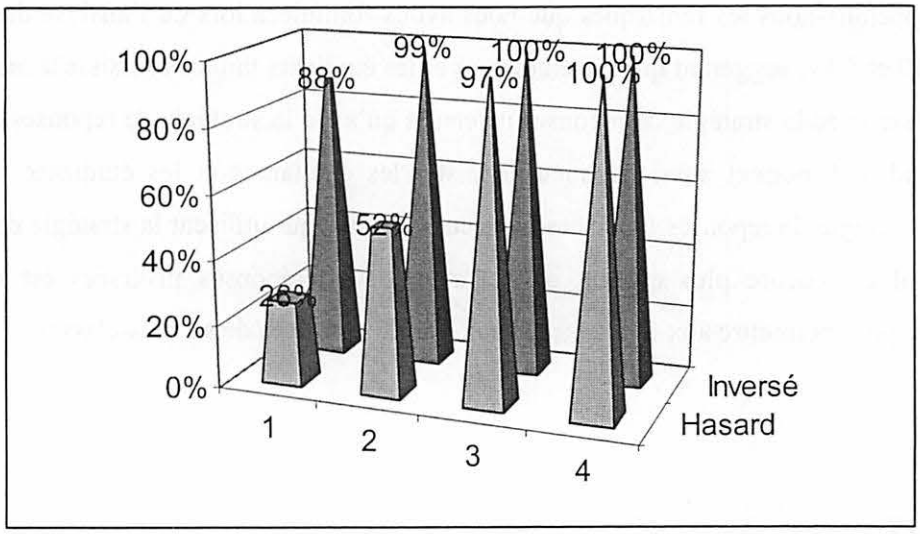

Figure 5.4 Taux de détection des patrons de réponses aberrants lorsque 10 % des questions sont affectées en fonction du niveau de classement réel et de la stratégie adoptée

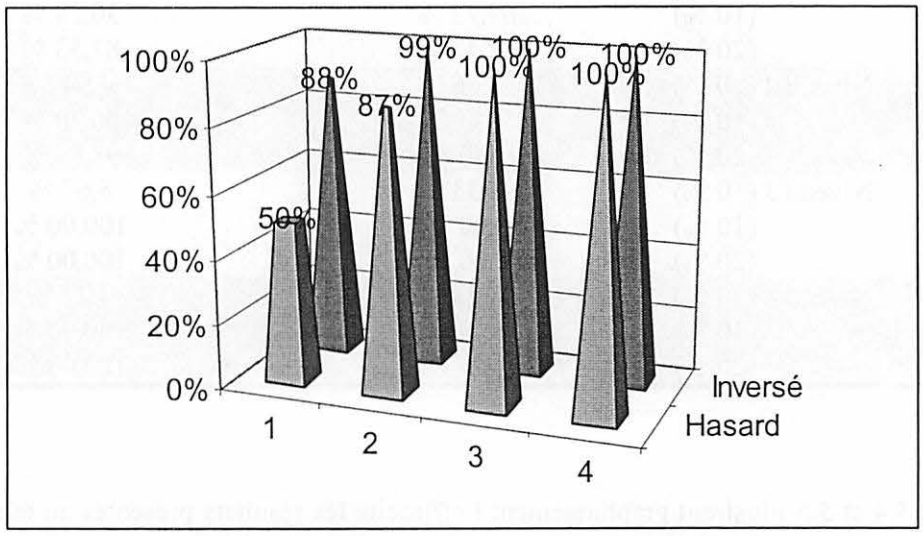

Figure 5.5 Taux de détection des patrons de réponses aberrants lorsque 20 % des questions sont affectées en fonction du niveau de classement réel et de la stratégie adoptée

5.3 Une application aux 1er et 2e tours du SRAM de la cohorte 2002-2003 au collège de l'Outaouais

Nous terminons ce chapitre par l'application de l'utilisation l'indice l_z dans la détection des cas potentiels de sous-classement chez les étudiantes et les étudiants de la cohorte 2002-2003 admis et inscrits au Collège de l'Outaouais au premier et au second tours du Service d'admission du Montréal métropolitain (SRAM).

On se rappellera que nous avions prévu, au chapitre quatre, conjointement avec le personnel enseignant du département des langues secondes du collège de l'Outaouais, d'utiliser l'indice χ_B^2 intégré au logiciel Bilog pour détecter les étudiantes et les étudiants qui auraient éventuellement tenté de se sous-classer au TCALS II. Ceux-ci seraient ensuite convoqués pour rédiger une composition écrite. 108 étudiantes et étudiants ont été détectés par Bilog et ont été convoqués. Les résultats de cette démarche ne nous sont pas encore totalement disponibles. Ils le seront uniquement suite à la rentrée scolaire de l'automne 2002. Toutefois, pour le moment, nous avons jugé pertinent de vérifier si les étudiantes et les étudiants détectés à partir de l'indice χ_B^2 sont les mêmes que nous pouvons maintenant détecter à partir de l'indice l_z. À cette fin, nous avons appliqué l'indice l_z aux résultats des étudiantes et des étudiants admis et inscrits au 1er tour du SRAM au TCALS II et avons noté quels étaient les étudiantes et les étudiants qui avaient aussi été détectés par l'indice χ_B^2.

Nous avons remarqué que plusieurs des étudiantes et des étudiants détectés par l'indice l_z étaient les mêmes que ceux identifiés par χ_B^2, mais toutefois, pas tous. Au tableau 5.13, seulement les 25 premiers étudiants et étudiantes identifiés par l'indice l_z sont présentés. Les résultats de tous les étudiants et étudiantes sont présentés à l'annexe 2. À la colonne χ_B^2, un *Oui* indique que le logiciel Bilog nous avait aussi permis d'identifier l'étudiante ou l'étudiant. Un *Non* indique que Bilog n'avait pas permis l'identification. Des 69 étudiantes et étudiants identifiés par l_z au 1er tour du SRAM, seulement 42, soit 60,90 %, ont aussi été identifiés par χ_B^2. De plus, l'analyse du

tableau 5.13 nous permet de constater que lorsque l'indice l_z prend une valeur supérieure à –2,43, tous les étudiantes et étudiants sont aussi détectés par l'indice χ_B^2.

Tableau 5.13
25 premiers patrons de réponses détectés par l'indice l_z ainsi que par l'indice χ_B^2 intégré au logiciel BILOG (1er tour du SRAM de la cohorte 2002-2003 au collège de l'Outaouais)

Habileté	Score (sur 85)	l_z	χ_B^2
-1,30	37	-5,72	Oui
-1,20	39	-4,03	Oui
-0,70	46	-3,93	Oui
-1,50	34	-3,81	Oui
-1,40	35	-3,76	Oui
-1,60	32	-3,66	Oui
-1,20	39	-3,52	Oui
-1,06	41	-3,22	Oui
-1,20	39	-3,06	Oui
-1,30	37	-3,00	Oui
-1,50	34	-2,95	Oui
-1,67	31	-2,93	Oui
-1,60	32	-2,87	Oui
-1,70	31	-2,83	Oui
-0,90	43	-2,82	Oui
-1,00	42	-2,79	Oui
-1,60	32	-2,67	Oui
-1,10	40	-2,67	Oui
-1,30	37	-2,66	Oui
-1,50	34	-2,62	Oui
-1,40	35	-2,52	Oui
-1,50	34	-2,44	Oui
0,10	59	-2,43	Non
-1,90	28	-2,39	Non
-1,40	35	-2,36	Oui

* Un *Oui* indique que le patron de réponses a aussi été détecté par la procédure proposée par le logiciel Bilog

La date d'obtention des résultats des étudiantes et des étudiants admis au 2e tour du SRAM ne nous a pas permis de les inclure à l'exercice de rappel pour rédiger une composition écrite et nous n'avons donc pas appliqué l'indice χ_B^2 à ceux-ci. Toutefois, nous avons tout de même appliqué l'indice l_z aux résultats des ces étudiantes et étudiantes au TCALS II. Une fois regroupés, les étudiantes et les étudiants des premier et second tours du SRAM sont au nombre 1 361. De ces 1 361, l'indice l_z a permis d'identifier 141 étudiantes et étudiants dont le patron de réponses est douteux, soit 10,36 % de la cohorte. Cela nous donne une idée du pourcentage d'étudiantes et d'étudiants que cette démarche nous permettra de détecter à l'avenir. Tous les scores des 141 étudiants et étudiantes identifiés par l'indice l_z sont présentés à l'annexe 2.

Chapitre 6

Pour conclure, quelques recommandations et des suites à donner

Ce projet de recherche avait pour objectif de développer un indice, ou des indices, d'ajustement inadéquat spécifiques au dépistage du sous-classement intentionnel par les étudiantes et les étudiants du réseau collégial au test de classement en anglais, langue seconde, le TCALS II. Depuis longtemps, le personnel enseignant des départements des langues secondes des établissements d'enseignement collégial du Québec doit composer avec ce problème de sous-classement de la part des étudiantes et des étudiants. Nous désirions donc y apporter une solution.

Dans un premier temps, nous avons étudié les qualités métrologiques du test et son adéquation pour classer des étudiantes et des étudiants dont le niveau d'habileté en anglais, langue seconde, est plutôt élevé, tel que celui des étudiantes et des étudiants du Collège de l'Outaouais. Nous avons remarqué que le TCALS II est constitué d'items dont le niveau de difficulté est relativement faible et qui ne permettent donc pas d'estimer avec la précision souhaitée le niveau d'habileté des étudiantes et des étudiants dont le niveau d'habileté est élevé en anglais, langue seconde. Le postulat d'unidimensionnalité du test affirmé par Laurier, Froio, Pearo et Fournier (1998), est aussi soutenu par nos analyses. Enfin, les corrélations de Pearson entre les valeurs obtenues de 1998-1999 à 2002-2003 quant à l'estimation des paramètres d'items sont toujours supérieures à 0,96, signifiant leur stabilité au cours des ans. Ce dernier résultat pourrait nous laisser croire que la sécurité du test n'a pas été affectée au cours des ans : l'hypothèse serait toutefois à confirmer plus sérieusement. Tout compte fait, malgré quelques faiblesses, le TCALS II semble constituer un test adéquat pour estimer le niveau d'habileté d'une étudiante ou d'un étudiant en anglais, langue seconde.

La recherche s'est poursuivie par la vérification directement auprès des étudiantes et des étudiants des stratégies éventuellement utilisées par eux pour se sous-classer au TCALS II. Suite à une rencontre de groupe, les étudiantes et les étudiants ont fait ressortir quatre grandes familles de

stratégies, comportant certaines variantes, qui leurs sembleraient efficaces. La stratégie qui est proposée le plus fréquemment est de donner des réponses au hasard. La seconde stratégie proposée, consiste à donner une mauvaise réponse aux questions du test. Selon nous, ces deux premières stratégies nécessitent l'élaboration d'un indice de détection plutôt sophistiqué. Les étudiantes et les étudiants ont aussi suggéré d'omettre de répondre à des questions ou encore de répéter consécutivement la même valeur de la clé de réponse. Ces deux dernières stratégies, pour leur part, ne nécessitent aucun mécanisme sophistiqué de détection. Une simple observation visuelle ou une opération informatique peuvent nous permettre de détecter ces étudiantes et étudiants facilement.

Ces informations en main, nous avons, par la suite, élaboré des indices qui pourraient permettre de détecter les étudiantes et les étudiants qui utilisent, soit la stratégie de réponses au hasard, soit la stratégie de mauvaises réponses, que nous nommeront dorénavant la stratégie de réponses inversées. Sept indices ont été considérés. Quatre de ces indices sont dits non optimaux car ils visent la détection de patrons de réponses aberrants non spécifiques au sous-classement. Il s'agit de l_z, *W*, *Zeta* et χ_B^2. Les trois autres, élaborés par nous, sont dits optimaux car ils sont prévus spécifiquement pour détecter, soit un patron de réponses au hasard, soit un patron de réponses inversées. Il s'agit des indices I_{hasard}, $I_{inversé}$ et I_{sous}.

Puisque nous ne pouvions connaître à l'avance quels étaient les étudiantes et les étudiants qui avaient réussi à se sous-classer au TCALS II, nous avons du mettre à l'épreuve ces indices à partir d'une simulation par ordinateur. Nos résultats démontrent la supériorité de l'indice l_z pour détecter le comportement de sous-classement au test. Son efficacité est d'ailleurs remarquable, principalement aux niveaux de classement 3 et 4. On arrive alors à détecter au moins 96 % des étudiantes et des étudiants qui ont réussi à se sous-classer au test, qu'ils aient donné des réponses au hasard ou des réponses inversées. Même pour les étudiantes et les étudiants des niveaux 1 et 2, l'efficacité du dépistage par l'indice l_z peut être assez élevée. Nous avouons que ces résultats vont au-delà de ce que nous avions espéré.

6.1 Quelques recommandations

Suite à ces analyses et à ces résultats, nous sommes maintenant en mesure de formuler quelques recommandations quant au développement futur du TCALS II et quant à l'application des mécanismes de dépistage du sous-classement volontaire de la part des étudiantes et des étudiants au TCALS II dans les collèges.

En premier lieu, il ressort clairement que l'ajout de questions plus difficiles au TCALS II est nécessaire. Le personnel enseignant des départements des langues secondes du réseau collégial devra se préoccuper de cette faiblesse du test et mettre en place une opération d'élaboration de nouvelles questions et de calibration de celles-ci en fonction des paramètres d'items actuels du TCALS II. À notre avis, ce sera une excellente opportunité pour élaborer un plus grand nombre d'items avec comme objectif de créer une banque d'items suffisamment grande pour planifier la mise en œuvre de versions adaptatives du TCALS II. C'est d'ailleurs une des propositions qui était faite à l'intérieur d'un des premiers rapports rédigé sur les mécanismes de classement des étudiantes et des étudiants dans le cours d'anglais, langue seconde, dans le réseau collégial (Fournier, 1992). On y proposait déjà la création d'un test adaptatif dont un des avantages, par ailleurs, serait de diminuer l'importance des comportements de sous-classement volontaire.

En second lieu, nous recommandons l'utilisation de l'indice l_z pour détecter les cas de sous-classement intentionnels de la part des étudiantes et des étudiants. Il ne faut toutefois pas perdre de vue que l'indice l_z permet de détecter les patrons de réponses aberrants en général. Il ne faut donc pas prendre pour acquis que l'identification d'une étudiante ou d'un étudiant par cet indice signifie nécessairement que celle-ci ou celui-ci a tenté de se sous-classer. Le patron de réponses peut provenir d'une étudiante ou d'un étudiant qui n'était tout simplement pas disposé à répondre au test ce jour-là. Il se peut que l'administration du test se soit déroulée dans des conditions peu usuelles et perturbatrices : cela pourrait se produire pour une étudiante ou un étudiant mal situé par rapport à la source sonore lors de l'administration de la sous-section de compréhension auditive. Il pourrait aussi s'agir d'une étudiante ou d'un étudiant dont la culture et les

connaissances générales sont suffisamment différentes pour interférer avec les réponses au test. À ce moment, le TCALS II ne mesurerait pas seulement le niveau d'habileté en anglais, langue seconde : il y aurait alors une indication de multidimensionnalité du test. Dans tous ces cas, il est totalement justifié de mettre en doute le résultat au test et d'investiguer plus à fond quant aux compétences langagières de l'étudiante ou de l'étudiant. Des procédures appropriées, adaptées aux réalités de l'institution d'enseignement, doivent être mises en place. Selon nos observations au Collège de l'Outaouais, les mesures spéciales adaptées à ces étudiantes et à ces étudiants ne toucheraient d'ailleurs qu'environ 10 % des nouvelles et des nouveaux étudiants admis annuellement.

En troisième lieu, nous recommandons de mettre en place un processus de détection des omissions et des répétitions consécutives du même choix de réponse aux questions du TCALS II. Selon nous, une étudiante ou un étudiant qui omet de répondre ou répète consécutivement un même choix de réponse à plus de 10 % des questions du test, devrait retenir notre attention. Cette opération peut déjà être réalisée facilement dans tous les collèges. Une version améliorée de notre outil informatique pourrait, bien sûr, éventuellement intégrer cette détection.

En dernier lieu, il semble nécessaire d'appliquer un mécanisme de suivi à l'identification d'un patron de réponses aberrant. On pourrait, par exemple, convoquer les étudiantes et les étudiants détectés et leur administrer une composition écrite, une entrevue orale en personne ou, à la limite, uniquement par téléphone. Nous tenons à noter que le seul fait de convoquer des étudiantes et des étudiants pour une seconde procédure de classement envoie un message clair à ceux-ci : il est possible de détecter les tentatives de sous-classement volontaire. Selon les échos que nous avons reçus en Outaouais, l'application de cette procédure a fait jaser beaucoup dans la région. À elle seule, cette pratique pourrait éventuellement être dissuasive et ainsi diminuer la fréquence des tentatives de sous-classement.

6.2 Les suites à donner au projet

Les informations contenues à l'intérieur de ce rapport permettent de réaliser l'informatisation du calcul de l'estimateur du niveau d'habileté ainsi que de l'indice l_z. Pour assurer la réalisation de nos travaux, nous avons dû nous-même procéder à cette tâche. Toutefois, le programme réalisé en langage SAS 6.0 est actuellement un peu lourd à utiliser et nécessitera quelques améliorations pour permettre sa diffusion dans les collèges. Nous espérons, d'ici peu, développer une version adaptée du logiciel pour les intervenants des services informatiques du réseau collégial. Entre temps, nous pourrions accepter de traiter les données du TCALS II en provenance des collèges.

Nous désirons aussi continuer les travaux relatifs à l'administration au collège de l'Outaouais d'une composition écrite et vérifier quels ont été les cas de sous-classement détectés par l'administration de cette composition écrite. Il nous sera alors possible d'identifier formellement des étudiantes et des étudiants sous-performants, mais surtout d'étudier les patrons de réponses de celles-ci et de ceux-ci. Nous croyons pouvoir ainsi percer plus à fond les stratégies de sous-classement utilisées. Par exemple, les étudiantes et les étudiants appliquent-ils vraiment leur stratégie de sous-classement uniquement au début du test? Ou seulement à la fin? Ou encore, plus ou moins de façon aléatoire? De plus, certains auteurs (Nickerson, 2002) soutiennent que la simulation d'un patron de réponses au hasard par un ordinateur ne correspond pas du tout à la notion du hasard qu'a un humain et, de suite, au comportement adopté par ce dernier. À titre d'exemple, même si la répétition d'un même choix de réponses peut théoriquement se produire aléatoirement, un humain évitera généralement de reproduire la même réponse plusieurs fois consécutives s'imaginant que cela est contraire aux lois du hasard. Il serait ainsi utile de découvrir à quoi correspond réellement une séquence de réponses au hasard au TCALS II.

Il sera important de continuer l'expérimentation, ainsi que les simulations, avec divers indices de détection de patrons de réponses aberrants et d'appliquer ces indices à diverses sections du test. Outre leur efficacité à détecter les tentatives de sous-classement, nous avons découvert d'intéressantes propriétés à ces indices, principalement quant à leur distribution de probabilité, et

nous tenons à poursuivre leur étude. Nous espérons ainsi en arriver à une simplification éventuelle de leur application.

Conjugué à nos recherches antérieures sur le testing adaptatif par ordinateur, ce projet de recherche prépare aussi le terrain à des travaux d'élaboration d'une procédure adaptative qui permettrait d'administrer un test dont le niveau de difficulté des items varierait non seulement en fonction de la valeur de l'estimateur du niveau d'habileté de l'étudiante ou de l'étudiant, mais aussi selon le niveau d'aberrance du patron de réponses. Actuellement, à notre connaissance, aucun travail en ce sens n'a été entrepris.

Enfin, nous espérons appliquer ces résultats au développement d'un environnement informatique destiné à soutenir une version adaptative du TCALS II. Cette version adaptative pourrait alors intégrer des mécanismes de dépistage autant du comportement spécifique de sous-classement volontaire que des patrons de réponses non spécifiques plutôt aberrants.

ANNEXES

6.1 Quelques recommandations

Suite à ces analyses et à ces résultats, nous sommes maintenant en mesure de formuler quelques recommandations quant au développement futur du TCALS II et quant à l'application des mécanismes de dépistage du sous-classement volontaire de la part des étudiantes et des étudiants au TCALS II dans les collèges.

En premier lieu, il ressort clairement que l'ajout de questions plus difficiles au TCALS II est nécessaire. Le personnel enseignant des départements des langues secondes du réseau collégial devra se préoccuper de cette faiblesse du test et mettre en place une opération d'élaboration de nouvelles questions et de calibration de celles-ci en fonction des paramètres d'items actuels du TCALS II. À notre avis, ce sera une excellente opportunité pour élaborer un plus grand nombre d'items avec comme objectif de créer une banque d'items suffisamment grande pour planifier la mise en œuvre de versions adaptatives du TCALS II. C'est d'ailleurs une des propositions qui était faite à l'intérieur d'un des premiers rapports rédigé sur les mécanismes de classement des étudiantes et des étudiants dans le cours d'anglais, langue seconde, dans le réseau collégial (Fournier, 1992). On y proposait déjà la création d'un test adaptatif dont un des avantages, par ailleurs, serait de diminuer l'importance des comportements de sous-classement volontaire.

En second lieu, nous recommandons l'utilisation de l'indice l_z pour détecter les cas de sous-classement intentionnels de la part des étudiantes et des étudiants. Il ne faut toutefois pas perdre de vue que l'indice l_z permet de détecter les patrons de réponses aberrants en général. Il ne faut donc pas prendre pour acquis que l'identification d'une étudiante ou d'un étudiant par cet indice signifie nécessairement que celle-ci ou celui-ci a tenté de se sous-classer. Le patron de réponses peut provenir d'une étudiante ou d'un étudiant qui n'était tout simplement pas disposé à répondre au test ce jour-là. Il se peut que l'administration du test se soit déroulée dans des conditions peu usuelles et perturbatrices : cela pourrait se produire pour une étudiante ou un étudiant mal situé par rapport à la source sonore lors de l'administration de la sous-section de compréhension auditive. Il pourrait aussi s'agir d'une étudiante ou d'un étudiant dont la culture et les

connaissances générales sont suffisamment différentes pour interférer avec les réponses au test. À ce moment, le TCALS II ne mesurerait pas seulement le niveau d'habileté en anglais, langue seconde : il y aurait alors une indication de multidimensionnalité du test. Dans tous ces cas, il est totalement justifié de mettre en doute le résultat au test et d'investiguer plus à fond quant aux compétences langagières de l'étudiante ou de l'étudiant. Des procédures appropriées, adaptées aux réalités de l'institution d'enseignement, doivent être mises en place. Selon nos observations au Collège de l'Outaouais, les mesures spéciales adaptées à ces étudiantes et à ces étudiants ne toucheraient d'ailleurs qu'environ 10 % des nouvelles et des nouveaux étudiants admis annuellement.

En troisième lieu, nous recommandons de mettre en place un processus de détection des omissions et des répétitions consécutives du même choix de réponse aux questions du TCALS II. Selon nous, une étudiante ou un étudiant qui omet de répondre ou répète consécutivement un même choix de réponse à plus de 10 % des questions du test, devrait retenir notre attention. Cette opération peut déjà être réalisée facilement dans tous les collèges. Une version améliorée de notre outil informatique pourrait, bien sûr, éventuellement intégrer cette détection.

En dernier lieu, il semble nécessaire d'appliquer un mécanisme de suivi à l'identification d'un patron de réponses aberrant. On pourrait, par exemple, convoquer les étudiantes et les étudiants détectés et leur administrer une composition écrite, une entrevue orale en personne ou, à la limite, uniquement par téléphone. Nous tenons à noter que le seul fait de convoquer des étudiantes et des étudiants pour une seconde procédure de classement envoie un message clair à ceux-ci : il est possible de détecter les tentatives de sous-classement volontaire. Selon les échos que nous avons reçus en Outaouais, l'application de cette procédure a fait jaser beaucoup dans la région. À elle seule, cette pratique pourrait éventuellement être dissuasive et ainsi diminuer la fréquence des tentatives de sous-classement.

6.2 Les suites à donner au projet

Les informations contenues à l'intérieur de ce rapport permettent de réaliser l'informatisation du calcul de l'estimateur du niveau d'habileté ainsi que de l'indice l_z. Pour assurer la réalisation de nos travaux, nous avons dû nous-même procéder à cette tâche. Toutefois, le programme réalisé en langage SAS 6.0 est actuellement un peu lourd à utiliser et nécessitera quelques améliorations pour permettre sa diffusion dans les collèges. Nous espérons, d'ici peu, développer une version adaptée du logiciel pour les intervenants des services informatiques du réseau collégial. Entre temps, nous pourrions accepter de traiter les données du TCALS II en provenance des collèges.

Nous désirons aussi continuer les travaux relatifs à l'administration au collège de l'Outaouais d'une composition écrite et vérifier quels ont été les cas de sous-classement détectés par l'administration de cette composition écrite. Il nous sera alors possible d'identifier formellement des étudiantes et des étudiants sous-performants, mais surtout d'étudier les patrons de réponses de celles-ci et de ceux-ci. Nous croyons pouvoir ainsi percer plus à fond les stratégies de sous-classement utilisées. Par exemple, les étudiantes et les étudiants appliquent-ils vraiment leur stratégie de sous-classement uniquement au début du test? Ou seulement à la fin? Ou encore, plus ou moins de façon aléatoire? De plus, certains auteurs (Nickerson, 2002) soutiennent que la simulation d'un patron de réponses au hasard par un ordinateur ne correspond pas du tout à la notion du hasard qu'a un humain et, de suite, au comportement adopté par ce dernier. À titre d'exemple, même si la répétition d'un même choix de réponses peut théoriquement se produire aléatoirement, un humain évitera généralement de reproduire la même réponse plusieurs fois consécutives s'imaginant que cela est contraire aux lois du hasard. Il serait ainsi utile de découvrir à quoi correspond réellement une séquence de réponses au hasard au TCALS II.

Il sera important de continuer l'expérimentation, ainsi que les simulations, avec divers indices de détection de patrons de réponses aberrants et d'appliquer ces indices à diverses sections du test. Outre leur efficacité à détecter les tentatives de sous-classement, nous avons découvert d'intéressantes propriétés à ces indices, principalement quant à leur distribution de probabilité, et

nous tenons à poursuivre leur étude. Nous espérons ainsi en arriver à une simplification éventuelle de leur application.

Conjugué à nos recherches antérieures sur le testing adaptatif par ordinateur, ce projet de recherche prépare aussi le terrain à des travaux d'élaboration d'une procédure adaptative qui permettrait d'administrer un test dont le niveau de difficulté des items varierait non seulement en fonction de la valeur de l'estimateur du niveau d'habileté de l'étudiante ou de l'étudiant, mais aussi selon le niveau d'aberrance du patron de réponses. Actuellement, à notre connaissance, aucun travail en ce sens n'a été entrepris.

Enfin, nous espérons appliquer ces résultats au développement d'un environnement informatique destiné à soutenir une version adaptative du TCALS II. Cette version adaptative pourrait alors intégrer des mécanismes de dépistage autant du comportement spécifique de sous-classement volontaire que des patrons de réponses non spécifiques plutôt aberrants.

ANNEXES

ANNEXE 1

Composition de classement d'anglais, langue seconde, au collégial

Tel qu'il avait été convenu avec le personnel enseignant du département des langues secondes du collège de l'Outaouais, une composition écrite a été administrée aux étudiantes et aux étudiants rappelés suite à la détection par l'indice χ_B^2 d'un patron de réponses douteux. Les consignes de réalisation de cette composition sont présentées ici.

Composition de classement d'anglais, langue seconde, au collégial

Cégep is a time of transition. Everyone writing this essay has attended secondary school, most of you in the very recent past. Take a few minutes to reflect on your secondary school experience, comparing it with what you expect your cégep experience to be ...

Write a composition where you include reflections both on your past school life and your goals next autumn at cégep.

Remember that this is a **composition**. Therefore, it must have a minimum of three paragraphs, including your introduction and conclusion.

You have 30 minutes to write this composition. There is no set number of words required but you **must use the entire 30 minutes.**

ANNEXE 2

PATRONS DE RÉPONSES DÉTECTÉS PAR L'INDICE l_z ET PAR L'INDICE χ^2_B PROPRE À BILOG

(Cohorte 2002-2003 - 1er et 2e tours du SRAM au collège de l'Outaouais)

Le tableau A2.1 présente les 141 cas détectés par l'indice l_z parmi les 1 361 étudiantes et étudiants admis et inscrits au collège de l'Outaouais à partir des premier et second tours du SRAM pour l'année scolaire 2002-2003. Sont indiquées les valeurs de l'indice l_z, de l'estimateur du niveau d'habileté et de son erreur-type, ainsi que, pour le 1er tour du SRAM uniquement, si oui ou non l'indice χ^2_B a aussi permis de détecter l'étudiante ou l'étudiant. Lorsque rien n'est indiqué à la colonne χ^2_B, c'est qu'il s'agit d'une étudiante ou d'un étudiant du second tour du SRAM, clientèle pour laquelle l'indice χ^2_B n'a pas été calculé.

Tableau A2.1

Résultats de la détection des étudiantes et des étudiants des deux premiers tours du SRAM au collège de l'Outaouais (2002-2003)

OBS	l_z	HABILETE	TOTAL	S	χ^2_B
1	-5.72	-1.30	37	0.14	OUI
2	-4.69	-0.80	45	0.14	
3	-4.24	-0.80	45	0.14	
4	-4.03	-1.20	39	0.14	OUI
5	-3.98	-1.20	39	0.14	
6	-3.97	-1.50	34	0.15	
7	-3.93	-0.70	46	0.15	OUI
8	-3.81	-1.50	34	0.15	OUI
9	-3.76	-1.40	35	0.14	OUI
10	-3.71	-1.20	39	0.14	
11	-3.66	-1.60	32	0.15	OUI
12	-3.52	-1.20	39	0.14	OUI
13	-3.22	-1.06	41	0.14	OUI
14	-3.11	-1.50	34	0.15	
15	-3.06	-1.30	37	0.14	
16	-3.06	-1.20	39	0.14	OUI
17	-3.00	-1.30	37	0.14	OUI
18	-3.00	-0.90	43	0.14	
19	-2.95	-1.50	34	0.15	OUI
20	-2.93	-1.67	31	0.15	OUI
21	-2.87	-1.60	32	0.15	OUI
22	-2.83	-1.70	31	0.15	OUI
23	-2.82	-0.90	43	0.14	OUI
24	-2.79	-1.00	42	0.14	OUI
25	-2.76	-1.80	29	0.15	
26	-2.75	-1.70	31	0.15	
27	-2.67	-1.60	32	0.15	OUI
28	-2.67	-1.10	40	0.14	OUI
29	-2.66	-1.30	37	0.14	OUI
30	-2.62	-1.50	34	0.15	OUI
31	-2.55	-1.50	34	0.15	
32	-2.52	-1.40	35	0.14	OUI
33	-2.49	-0.40	51	0.16	
34	-2.47	-0.60	48	0.15	
35	-2.44	-1.50	34	0.15	OUI
36	-2.43	0.10	59	0.20	NON
37	-2.42	0.30	62	0.21	
38	-2.39	-1.90	28	0.16	NON
39	-2.36	-1.40	35	0.14	OUI
40	-2.36	-0.60	48	0.15	
41	-2.35	-1.40	35	0.14	OUI
42	-2.34	-1.40	35	0.14	
43	-2.34	-1.30	37	0.14	OUI
44	-2.33	-1.40	35	0.14	
45	-2.30	-1.00	42	0.14	OUI
46	-2.30	-0.90	43	0.14	
47	-2.29	-0.90	43	0.14	
48	-2.28	-1.90	28	0.16	NON

Tableau A2.1

Résultats de la détection des étudiantes et des étudiants des deux premiers tours du SRAM au collège de l'Outaouais (2002-2003) (suite)

OBS	l_z	HABILETE	TOTAL	S	χ^2_B
49	-2.18	-1.70	31	0.15	NON
50	-2.16	-0.30	53	0.16	
51	-2.15	-1.10	40	0.14	
52	-2.10	-1.90	28	0.16	
53	-2.07	-1.70	31	0.15	
54	-2.05	-1.90	28	0.16	NON
55	-2.05	-1.60	32	0.15	OUI
56	-2.03	-0.40	51	0.16	
57	-2.02	-1.60	32	0.15	OUI
58	-1.99	-1.50	34	0.15	
59	-1.92	-0.60	48	0.15	
60	-1.92	0.20	60	0.20	NON
61	-1.91	-1.60	32	0.15	
62	-1.91	-0.40	51	0.16	
63	-1.91	-0.30	53	0.16	NON
64	-1.88	-1.90	28	0.16	NON
65	-1.87	-1.90	28	0.16	
66	-1.87	-1.20	39	0.14	
67	-1.87	-0.80	45	0.15	OUI
68	-1.85	-1.90	28	0.16	NON
69	-1.85	-1.10	40	0.14	OUI
70	-1.80	-0.30	53	0.16	NON
71	-1.80	0.30	62	0.21	OUI
72	-1.77	-1.10	40	0.14	
73	-1.77	-0.80	45	0.14	OUI
74	-1.76	-1.40	35	0.14	
75	-1.75	0.10	59	0.19	
76	-1.74	-1.20	39	0.14	
77	-1.74	0.40	63	0.22	
78	-1.72	-1.20	39	0.14	
79	-1.72	-0.90	43	0.14	OUI
80	-1.72	-0.90	43	0.14	
81	-1.70	-1.50	34	0.15	OUI
82	-1.70	-1.10	40	0.14	
83	-1.70	0.00	57	0.19	
84	-1.69	-1.60	32	0.15	NON
85	-1.69	-0.10	56	0.17	NON
86	-1.68	-0.70	46	0.15	OUI
87	-1.66	-0.40	51	0.16	NON
88	-1.65	-1.20	39	0.14	OUI
89	-1.63	-1.80	29	0.16	
90	-1.61	0.20	60	0.20	NON
91	-1.60	-0.70	46	0.15	OUI
92	-1.60	0.40	63	0.22	
93	-1.59	-1.50	34	0.15	
94	-1.58	-0.10	56	0.17	
95	-1.55	-1.40	35	0.14	
96	-1.55	-1.20	39	0.14	OUI
97	-1.54	-1.60	32	0.15	

Tableau A2.1

Résultats de la détection des étudiantes et des étudiants des deux premiers tours du SRAM au collège de l'Outaouais (2002-2003) (suite)

OBS	l_z	HABILETE	TOTAL	S	χ^2_B
98	-1.54	-1.20	39	0.14	
99	-1.51	0.20	60	0.20	NON
100	-1.50	-1.50	34	0.15	
101	-1.48	-1.20	39	0.14	OUI
102	-1.48	-1.20	39	0.14	
103	-1.46	-1.20	39	0.14	OUI
104	-1.44	-1.70	31	0.15	
105	-1.44	-1.60	32	0.15	OUI
106	-1.43	-1.80	29	0.15	
107	-1.42	-1.50	34	0.15	
108	-1.41	0.70	68	0.26	NON
109	-1.38	1.00	73	0.31	
110	-1.36	-0.10	56	0.17	NON
111	-1.35	-2.00	26	0.16	
112	-1.35	0.00	57	0.18	
113	-1.34	0.00	57	0.18	
114	-1.32	1.00	73	0.31	
115	-1.30	-2.00	26	0.17	OUI
116	-1.29	-1.70	31	0.15	
117	-1.28	1.00	73	0.31	
118	-1.27	0.10	59	0.20	
119	-1.16	0.70	68	0.26	
120	-1.13	0.40	63	0.22	NON
121	-1.13	0.40	63	0.23	
122	-1.13	0.50	65	0.24	
123	-1.08	0.80	70	0.28	
124	-1.08	1.00	73	0.31	NON
125	-1.07	0.40	63	0.22	
126	-1.07	0.70	68	0.26	
127	-1.04	-2.10	25	0.17	
128	-1.04	0.50	65	0.24	NON
129	-1.02	0.40	63	0.22	
130	-1.02	1.00	73	0.31	
131	-1.01	0.10	59	0.20	
132	-1.01	1.00	73	0.31	
133	-0.99	0.50	65	0.24	NON
134	-0.98	0.70	68	0.26	NON
135	-0.97	0.40	63	0.22	
136	-0.97	1.20	76	0.35	NON
137	-0.96	0.50	65	0.24	NON
138	-0.95	0.70	68	0.26	
139	-0.94	1.50	80	0.41	NON
140	-0.93	1.20	76	0.35	NON
141	-0.92	0.40	63	0.22	NON

Références

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied measurement in education, 7*(4), 255-278.

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

Bay, L. et Nering, M.L. (1998). *A demonstration of using person-fit statistics in standard setting.* Texte présenté lors du congrès annuel de l'American Educational Research Association tenu à San Diego. (Document ERIC no ED 421 533)

Bedrick, E.J. (1997). Approximating the conditional distribution of person fit indexes for checking the Rasch model. *Psychometrika, 62*(2), 191-199.

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

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

Birenbaum, M., Kelly, A.E., Tatsuoka, K.K. (1992a). *Diagnosing knowledge states in algebra using the rule space model.* Research Report RR-92-57-ONR, Princeton : Educational Testing Service.

Birenbaum, M., Kelly, A.E., Tatsuoka, K.K. (1992b). *Toward a stable diagnostic representation of students' errors in algebra.* Research Report RR-92-58-ONR, Princeton : Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring examinee's ability. Dans F.M. Lord et M.R. Novick (Éds) : *Statistical theories of mental test scores*. Reading : Addison-Wesley.

Blais, J.G. (1987). *Effets de la violation du postulat d'unidimensionnalité dans la théorie des réponses aux items*. Thèse de doctorat inédite. Montréal : Université de Montréal.

Blais, J.-G. et Raîche, G. (2002a, soumis). *Features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules*. Dans G. Engelhard (Éd.) : Objective measurement – Theory into practice. Greenwich, Connecticut : Ablex.

Blais, J.-G. et Raîche, G. (2002b). *Some features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules*. Texte présenté au 11e Biennal International Objective Measurement Workshop tenu à New Orleans. [À paraître sur ERIC]

Bock, R.D. (1997). The nominal categories model. Dans W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory*. New York : Springer-Verlag.

Bock, R.D. et Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431-444.

Bond, T.G. et Fox, C.M. (2001). *Applying the Rasch model : Fundamental measurement in the human sciences*. Mawah, NJ : Laurence Erlbaum Associates.

Bracey, G. et Rudner, L.M. (1992). *Person-fit statistics : High potential and many unanswered questions* (rapport de recherche no EDO-TM-92-5). Washington, DC : Office of Educational Research and Improvement. (Document ERIC no ED 355 249)

Ciha, T.E. et collab. (1974). Parents as identifiers of giftedness, ignored but accurate. *Gifted Child Quaterly, 18*(3), 191-195.

Coleman, L.J. et Cross, T.L. (2001). *Being gifted in school : An introduction to development, guidance, and teaching*. Waco, Texas : Prufrock Press.

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

Drasgow, F. et Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology, 72*(1), 10-18.

Drasgow, F. et Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*(1), 59-67.

Drasgow, F., Levine, M.V. et McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*(1), 59-79.

Drasgow, F., Levine, M. et McLaughlin, M.E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*(2), 171-191.

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.

Drasgow, F., Levine, M.V. et Zickar, M.J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9*(1), 47-64.

Feldhusen, J.F. et Jarwan, F.A. (2000). Identification of gifted and talented youth for educational program. Dans K.A. Heller, F.J. Mönks, R.J. Sternberg et R.F. Subotnik (Éds) : *International hanbook of giftedness and talent.* Amsterdam : Elsevier.

Fischer, G.H. (1995). Derivations of the Rasch model. Dans G.H. Fischer et I.W. Molenaar (Éds) : *Rasch models – Foundations, recent developments, and applications.* New York : Springer-Verlag.

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

Gagné, F. (1994). Are teachers really poor talent detectors? Comments on Pegnato and Birch's (1959) study of the effectiveness and efficiency of various identification techniques. *Gifted Child Quaterly, 38*(3), p. 124-126.

Gitomer, D.H. et Rock, D. (1993). Addressing process variables in test analysis. Dans N. Frederiksen, R.J. Mislevy et I.I. Bejar (Éds) : *Test theory for a new generation of tests.* Hillsdale : Lawrence Erlbaum Associates.

Goldstein, H. (1994a). Mathematical and ideological assumptions in the modelling of test item responses. Dans D. Laveault, B.D. Zumbo, M.E. Gessaroli et M.W. Boss (Éds) : *Modern theories of measurement - Problems and issues.* Ottawa : Université d'Ottawa.

Goldstein, H. (1994b). Présupposés mathématiques et idéologiques des modèles de réponses aux items. *Mesure et évaluation en éducation, 17*(2), 107-114.

Goldstein, H., Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

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

Harwell, M., Stone, C.A., Hsu, T.C.,et Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

Junker, B.W., Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*(1), 65-81.

Klauer, K.C. (1995). The assessment of person fit. Dans G.H. Fischer et I.W. Molenaar (Éds) : *Rasch models – Foundations, recent developments, and applications.* New York : Springer-Verlag.

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, QC : Direction générale de l'enseignement collégial, ministère de l'Éducation du Québec.

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

Levine, M.V. et Drasgow, F. (1982). Appropriateness measurement : Review, critique and validating studies. . *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

Levine, M.V. et Drasgow, F. (1983). Appropriateness measurement : Validating studies and variable ability models. Dans D.J. Weiss (Éd.) : *New horizons in testing – Latent trait test theory and computerized adaptive testing.* New York : Academic Press.

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

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

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

Liou, M. (1993). Exact person tests for assessing model-data fit in the Rasch model. *Applied Psychological Measurement, 17*(2), 187-195.

Lord, F.M. (1952). A theory of test scores. *Psychmetric Monographs*, no 7.

Luecht, R.M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*(4), 389-404.

McDonald, R.P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*(4), 379-396.

McDonald, R.P. (1997). Normal-ogive multidimensional model. Dans W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory.* New York : Springer-Verlag.

McDonald, R.P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*(2), 99-114.

Meijer, R.R. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*(4), 321-336.

Meijer, R.R., Molenaar, I.W. et Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*(2), 111-120.

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), 77-90.

Meijer, R.R. et Nering, M.L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*(4), 321-336.

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.

Meijer, R.R. et Sijtsma, K. (2001). Methodology review : Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107-135.

Meijer, R.R. et van Krimpen-Stoop, E.M.L.A. (1998). *Simulating the null distribution of person-fit statistics for conventional and adaptive tests* (Rapport de recherche no 98-02). Enschede, Pays-Bas : University de Twende. (Document ERIC no ED 421 548).

Meijer, R.R. et van Krimpen-Stoop, E.M.L.A. (2000). Person fit across subgroups : An achievement testing example. Dans A. Boomsma, M.A.J. van Duijn et T.A.B. Snijders (Éds) : *Essays on item response theory.* New York : Springer-Verlag.

Michell, J. (2002). *Measurement: A beginner's guide.* Texte présenté au congrès annuel de l'American Educational Research Association tenu à New Orleans.

Ministère de l'Éducation du Québec (1993). *Régime d'enseignement collégial.* Québec : Direction générale de l'enseignement collégial, ministère de l'Éducation du Québec.

Mislevy, R.J., Bock, R.D. (1986). *PC-BILOG : Item analysis and test scoring with binary logistic models.* Mooresville, Indiana : Scientific Software Inc.

Mislevy, R.J., Stocking, M.L. (1987). *A consumer's guide to LOGIST and BILOG.* Research report RR-87-4. Princeton, NJ : Educational testing service.

Mokken, R.J. (1997). Nonparametric models for dichotomous responses. Dans W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory.* New York : Springer-Verlag.

Mokken, R.J., Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*(4), 417-430.

Molenaar, I.W. et Hoijtink, H. (1996). Person-fit and the Rasch model with an application to knowledge of logical quantors. *Applied Measurement in Education, 9*(1), 27-46.

Molenaar, I.W. (1995). Some background for item response theory and the Rasch model. Dans G.H. Fischer et I.W. Molenaar (Éds) : *Rasch models - Foundations, recent developments, and applications*. New York : Springer-Verlag.

Molenaar, I.W. (1997). Nonparametric models for polytomous responses. Dans W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory*. New York : Springer-Verlag.

Molenaar, I.W. et Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*(1), 73-106.

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.

Nickerson, R.S. (2002). The production and perception of randomness. *Psychological Review, 109*(2), 330-357.

Pegnato, C.V. et Birch, J.W. (1959). Locating gifted children in junior high schools : A comparison of methods. *Exceptional Children, 25*, 300-304.

Raîche, G. (2001a). *La distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt : selon l'erreur type et selon le nombre d'items administrés*. Thèse de doctorat inédite. Montréal : Université de Montréal.

Raîche, G. (2001b). *Principes et enjeux du testing adaptatif par ordinateur : de la loi des petits nombres à la loi des grands nombres*. Texte présenté lors du 69^{e} congrès de l'Association canadienne française pour l'avancement de la science tenu à Sherbrooke.

Raîche, G. (2002a). Comment dépister les étudiants qui cherchent à se sous-classer au test de classement en anglais, langue seconde dans le réseau collégial québécois. *Bulletin de l'Association pour la recherche au collégial, 15*(3), 8-9.

Raîche, G. (2002b, soumis). La simulation d'un test adaptatif basé sur le modèle de Rasch. *Mesure et évaluation en éducation*.

Raîche, G., Arbach, G. et Blais, J.G. (2002a). *Des élèves qui ratent intentionnellement un test de classement !* Texte présenté lors du colloque annuel de l'Association québécoise de pédagogie collégiale tenu à Québec.

Raîche, G., Arbach, G. et Blais, J.G. (2002b). *Le dépistage du sous-classement volontaire au test de classement en anglais, langue seconde, au collégial.* Texte présenté lors du congrès annuel de l'Association francophone pour le savoir tenu à Sainte-Foy.

Raîche, G. et Blais, J.-G. (2002a, à paraître). Étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt dans le contexte de l'application du modèle de Rasch. *Mesure et évaluation en éducation.*

Raîche, G. et Blais, J.-G. (2002b, à paraître). *Considerations about expected a posteriori estimation in adaptive testing : Adaptive a priori, adaptive correction for bias, and adaptive integration interval.* Dans G. Engelhard (Éd.) : Objective measurement – Theory into practice. Greenwich, Connecticut : Ablex.

Raîche, G. et Blais, J.-G. (2002c). *Practical considerations about expected a posteriori estimation in adaptive testing : Adaptive a priori, adaptive correction for bias, and adaptive integration interval.* Texte présenté lors du 11e Biennal International Objective Measurement Workshop tenu à New Orleans. [À paraître sur ERIC]

Ramsey, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630.

Ramsey, J.O. (1993a). *TESTGRAF : a program for the graphical analysis of multiple choice test and questionnaire data.* Montréal : Département de psychologie, Université McGill.

Ramsey, J.O. (1993b). *TESTGRAF : Some graphics tools for the analysis of examination data.* Texte présenté lors de la 16e session d'étude de l'ADMÉÉ à Laval. Montréal : Association pour le développement de la mesure et de l'évaluation en éducation.

Ramsey, J.O. (1997). A functional approach to modelling test data. Dans W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory.* New York : Springer-Verlag.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago : MESA Press, 1980.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412.

Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

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

Roberts, J.S., Donoghue, J.R., Laughlin, J.E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied psychological measurement, 24*(1), 3-32.

Rost, J. (1995). The growing family of Rasch models. Dans A. Boomsma, M.A.J. van Duijn et T.A.B. Snijders (Éds) : *Essays on item response theory*. New York : Springer-Verlag.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*(2), 203-219.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39*(1), 111-121.

Samejima, F. (1997). Graded response model. Dans W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory*. New York : Springer-Verlag.

SAS (1996). *SAS language*. Carey, NC : SAS Institute, 1990..

Smith, R.M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51*(3), 541-565.

Smith, R.M. (2002). *The family approach to assessing fit in Rasch measurement*. Texte présenté au 11e International Biennual Objective Measurement Workshop tenu à New Orleans.

Smith, R. et Suh, K.K. (2002). *Rasch fit statistics as a direct test of the invariance parameter estimation*. Texte présenté au 11e International Biennual Objective Measurement Workshop tenu à New Orleans.

Snijders, T.O.A. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. Psychometrika, 66(3), 331-342.

Stout, W. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*(2), 293-325.

Tatsuoka, K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied Measurement in Education, 9*(1), 65-76.

Thissen, D. (1988). *MULTILOG : multiple, categorical item analysis and test scoring using item response theory*. Mooresville, IN : Scientific Software.

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

Wainer, H. et Mislevy, R.J. (1990). Item response theory, item calibration and proficiency estimation. Dans H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (Éds) : *Computerized adaptive testing - A primer*. Hillsdale : Lawrence Erlbaum Associates.

Whitmore, J.R. (1980). *Giftedness, conflict, and underachievement.* Boston : Allyn and Bacon.

Wilson, M. (1992). The ordered partition model : An extension of the partial credit model. *Applied psychological measurement, 16*(4), 309-325.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement : Issues and Practice, 16*(4), 33-45 et 52.

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

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

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (Rapport de recherche no RR-95-2). Princeton, NJ : Educational Testing Service. (Document ERIC no ED 395 035)

Yamamoto, K. et Everson, H.T. (1995). *Modeling the mixture of IRT and pattern responses by a modified hybrid model* (Rapport de recherche no RR-95-16). Princeton, NJ : Educational Testing Service. (Document ERIC no ED 395 036)

Yamamoto, K., Gitomer, D.H. (1993). Application of a HYBRID model to a test of cognitive skill representation. Dans N. Frederiksen, R.J. Mislevy et I.I. Bejar (Éds) : *Test theory for a new generation of tests.* Hillsdale : Lawrence Erlbaum Associates.