



Assurer la comparabilité : progression vers un score unique en litt ratie pour le TPCL

Rapport pr par  pour
l'Office de la qualit  et de la responsabilit  en  ducation (OQRE) par

Nizam Radwan
Psychom tricien, OQRE

Mark D. Reckase
Chercheur invit , OQRE; professeur, Michigan State University

W. Todd Rogers
Chercheur invit , OQRE; professeur, Universit  de l'Alberta

AVRIL 2010

À propos de l'Office de la qualité et de la responsabilité en éducation

L'Office de la qualité et de la responsabilité en éducation (OQRE) est un organisme provincial indépendant financé par le gouvernement de l'Ontario. Le mandat de l'OQRE est d'effectuer des tests à l'échelle de la province à des étapes clés de l'éducation de chaque élève aux cycles primaire et moyen et au secondaire et d'en rapporter les résultats aux éducatrices et éducateurs, aux parents et au public.

L'OQRE sert de catalyseur pour augmenter la réussite des élèves de l'Ontario en mesurant leur rendement en lecture, en écriture et en mathématiques par rapport aux attentes et contenus d'apprentissage du curriculum de l'Ontario. Les données obtenues fournissent une mesure de la qualité et de la responsabilité du système éducatif de l'Ontario.

Les résultats objectifs et fidèles des tests sont des données qui complètent les connaissances actuelles au sujet de l'apprentissage des élèves et sont un outil d'amélioration important à tous les niveaux : les élèves, les écoles, les conseils scolaires et la province.

À propos du programme de recherche de l'OQRE

L'OQRE entreprend des recherches pour deux raisons principales :

- maintenir des pratiques d'une qualité irréprochable et s'assurer que l'organisme demeure à l'avant-garde des évaluations à grande échelle;
- promouvoir l'utilisation des données de l'OQRE pour améliorer le rendement des élèves en explorant des moyens pour éclairer les politiques et les décisions prises par les éducatrices et éducateurs, les parents et le gouvernement.

Les projets de recherche de l'OQRE étudient en profondeur les facteurs qui influent sur le rendement des élèves et sur la qualité de l'éducation, et ils examinent les processus statistiques et psychométriques en produisant des données de test de haute qualité.



Office de la qualité et de la responsabilité en éducation

2, rue Carlton, bureau 1200, Toronto (Ontario) M5B 2M9, | 888 327-7377, www.oqre.on.ca

© Imprimeur de la Reine pour l'Ontario, 2010

Résumé

Le Test provincial de compétences linguistiques (TPCL) a été modifié en mars 2006 : il a été raccourci en réponse aux recommandations émises dans le cadre d'un examen externe du programme d'évaluations de l'OQRE. De plus, à ce moment-là, la décision a été prise de communiquer les résultats sous forme de score unique en littératie plutôt que de deux scores distincts en lecture et en écriture. En raison de ces changements, le TPCL administré en mars 2006 n'était pas équivalent à celui administré en octobre 2004.

Cependant, on souhaitait toujours appairer le TPCL de 2004 et celui de 2006 afin de générer des résultats permettant de mesurer le changement dans le rendement des élèves entre 2004 et 2006. Deux « tests » de 2004 ont été pris en compte pour établir l'appariement : le test intégral de 2004, noté de manière à générer un score unique en littératie, et un pseudotest construit à partir d'items compris dans le test de 2004 et destiné à imiter le test de 2006. L'appariement du TPCL de 2006 avec ces deux tests a donné des pourcentages semblables d'élèves ayant réussi le test. Le pseudotest a été choisi parce qu'il correspondait mieux au TPCL de 2006 et les scores appariés ont été traités comme des données comparables plutôt que mises en équivalence.

Introduction

Le Test provincial de compétences linguistiques (TPCL) mesure les compétences en lecture et en écriture des élèves de 10^e année fréquentant les écoles publiques et privées de la province de l'Ontario. Réussir le TPCL est l'une des exigences provinciales pour obtenir le diplôme d'études secondaires de l'Ontario. Les élèves qui font le TPCL une fois et qui ne le réussissent pas peuvent satisfaire à l'exigence en matière de littératie en suivant le Cours de compétences linguistiques des écoles secondaires de l'Ontario (CCLESO).

Le TPCL a été administré pour la première fois en 2002. Jusqu'en 2006, il fallait deux demi-journées pour le faire. Les items en lecture et en écriture étaient évalués durant les deux jours. Un modèle de décision conjonctif a été utilisé par le TPCL : un score pour la lecture et un score pour l'écriture ont été communiqués, et les élèves devaient réussir à la fois la composante lecture et la composante écriture pour satisfaire à l'exigence en matière de littératie.

Le TPCL a été raccourci en 2006 à la suite des recommandations émises dans le cadre d'un examen externe exhaustif du programme d'évaluations de l'OQRE (Wolfe, Childs et Elgie, 2004). Le nouveau test, qui fait environ la moitié du test initial, requiert une demi-journée. En outre, le modèle de décision conjonctif a été remplacé par un modèle de décision compensatoire. Un seul score en littératie est communiqué au lieu de scores distincts en lecture et en écriture. Les composantes lecture et écriture sont pondérées de façon presque égale, de manière à ce que chacune compte pour environ la moitié du score intégral. Cela a été fait pour reproduire le modèle de décision précédent

qui pondérerait explicitement de manière égale la composante lecture et la composante écriture.

Malgré les changements apportés entre 2004 et 2006, on souhaitait toujours appairer le TPCL de 2006 à celui de 2004 afin de pouvoir évaluer les changements constatés dans le rendement des élèves entre 2004 et 2006¹. Toutefois, il faut satisfaire à plusieurs conditions préalables avant de pouvoir mettre en équivalence les deux tests (Kolen et Brennan, 2004).

Premièrement, les tests doivent avoir une *conception équivalente*, c'est-à-dire qu'il devrait être possible d'interpréter de manière valide les scores générés par les tests en matière de concept ou de domaine. En outre, la moyenne, l'écart-type et la fiabilité devraient être les mêmes et la corrélation dont l'atténuation est corrigée en raison d'une non-fiabilité devrait présenter une valeur de 1,00 entre les deux tests.

Deuxièmement, d'après la *condition d'équité*, pour chaque groupe d'élèves faisant le test qui présentent des habiletés semblables ou un rendement égal, la distribution conditionnelle des scores du test X (après la mise en équivalence) devrait être identique à celle des scores du test Y. Pour chaque groupe d'élèves ayant le même score vrai, τ , la distribution statistique conditionnelle des scores du test X (après la mise en équivalence) devrait être la même que celle des scores du test Y :

$$E_{j/\tau} [eq_Y(X): \tau] = E_{j/\tau}(Y/\tau),$$

où $E_{j/\tau}$ représente la prédiction reprise, $j = 1, 2, \dots, N_\tau$, $N_\tau \rightarrow \infty$ les élèves présentant le score vrai τ ,

¹ Le TPCL pour l'année scolaire 2004-2005 a été administré en octobre 2004, tandis que celui de 2005-2006 a été administré en mars 2006. Aucun test n'a été administré en 2005.

eq_Y représente une fonction de mise en équivalence qui transforme les scores du test X en fonction de la grille de notation du test Y,

$(X) : \tau$ représente la distribution des scores transformés du test X pour les élèves affichant le score vrai τ et

(Y / τ) représente la distribution des scores du test Y pour les élèves affichant le score vrai τ .

Troisièmement, on devrait retrouver l'*invariance de la population*. La transformation de la mise en équivalence pour appliquer les scores du test X à la grille de notation du test Y devrait être identique pour tous les échantillons provenant de la même population.

Quatrièmement, la *condition de symétrie* exige que la transformation du test X au test Y soit la même que la transformation du test Y au test X. Cela diffère de la régression, où la régression de Y sur X n'est pas nécessairement la même que celle de X sur Y².

La première des quatre conditions, soit la conception équivalente, est particulièrement intéressante dans la présente étude étant donné le changement apporté à la longueur du TPCL et l'adoption d'un modèle de décision compensatoire. Par conséquent, cette étude visait principalement à déterminer si le TPCL de 2004 et celui de 2006 présentaient une conception équivalente. Si c'est le cas, les scores pourraient faire l'objet d'une mise en équivalence. Sinon, selon le niveau de similarité, les deux étendues

² Si la moyenne et l'écart-type des deux tests sont respectivement égaux, la régression de Y sur X et de X sur Y sera par conséquent la même.

des scores pourraient être échelonnées à des fins de comparabilité (American Educational Research Association, American Psychological Association et National Council on Measurement in Education, 1999).

L'étude visait ensuite à évaluer le processus d'appariement pour la mise en équivalence ou l'échelonnage. Comme il a été mentionné précédemment, en 2004 le concept hypothétique de littératie a été estimé au moyen de deux sous-tests, en lecture et en écriture. On a supposé que les élèves étaient classés selon une échelle continue unique, les tests ont produit deux estimations distinctes en littératie et celles-ci ont servi à estimer les emplacements des élèves en déterminant si leur score se trouvait au-dessus des seuils de rendement respectifs des deux tests. En 2006, un score unique a été utilisé pour décider si les élèves avaient réussi ou non le test. Ce score, qui se fonde sur une notation combinée des items de test en lecture et en écriture, devrait être une approximation plus proche du concept hypothétique de littératie que le score du TPCL de 2004. Dans le cas où cette hypothèse est démontrée, une complication survient dans le processus d'appariement puisque le TPCL de 2004 ne comprend pas la même pondération en lecture et en écriture que le TPCL de 2006. Dans ce cas, il faut tenir compte de la différence constatée dans la pondération durant le processus d'appariement. Donc, un autre objectif s'est ajouté à cette étude, soit celui d'enquêter sur la façon de rendre compte d'une telle différence durant le processus d'appariement lorsque le test de 2006 a été jugé comme définition opérationnelle du construit.

Examen du concept d'équivalence pour le TPCL de 2004 et celui de 2006

Définition de la littératie

La définition de la littératie adoptée pour le TPCL est la suivante :

« Dans le cadre du TPCL, la littératie comprend l'acquisition d'habiletés en lecture et en écriture nécessaires pour comprendre différents genres de textes et pour communiquer par écrit au moyen d'une variété de genres de textes, tel que prévu dans *Le curriculum de l'Ontario*, dans toutes les matières, jusqu'à la fin de la 9^e année. » (*Document-cadre*, 2007, p. 9)

Cette définition est toujours la même depuis l'introduction du TPCL en 2002.

Cette définition est élargie par la description de ce que les élèves doivent utiliser et faire lors du TPCL. Pour la lecture, les élèves doivent utiliser des stratégies pour interagir avec divers textes narratifs, informatifs et avec éléments graphiques pour conceptualiser et comprendre la signification de différents genres de textes, démontrer leur compréhension d'une signification explicite et implicite et faire un lien entre leur compréhension de ce qu'ils ont lu et leur expérience et leurs connaissances personnelles. Pour l'écriture, on demande aux élèves de rédiger deux réponses courtes, une série de paragraphes qui expriment et appuient l'opinion qu'ils ont émise à partir de la tâche d'écriture et une nouvelle journalistique qu'ils rédigent à partir de la tâche d'écriture. Dans leurs réponses, les élèves démontrent leur capacité à communiquer des idées et de l'information de manière claire et cohérente (adapté du *Document-cadre*, 2007, p. 9).

Comparabilité du TPCL de 2004 et de celui de 2006

Le TPCL de 2004 et celui de 2006 ont été conçus pour évaluer les compétences décrites dans la définition de la littératie et de son élargissement. Toutefois, le TPCL de 2006 était

pratiquement deux fois plus court que celui de 2004 et présentait un score unique en littératie, plutôt que deux scores distincts en lecture et en écriture. D'après les critères de sélection, les items de 2004 et de 2006 (les textes de lecture et les items à choix multiple et à réponse construite ainsi que les tâches d'écriture s'y rattachant pour les tests de langues française et anglaise) devaient présenter les similarités ci-dessous :

- a. niveaux de traitement cognitif et emploi des stratégies de lecture et d'écriture;
- b. niveaux de difficulté des textes de lecture;
- c. difficulté des items de lecture et d'écriture.

(Bien que le TPCL et le Ontario Secondary School Literacy Test [OSSLT], le test de compétences linguistiques en anglais, soient conçus séparément, le nombre d'items et de points est le même.)

La première et la deuxième rangées des deux parties du tableau 1 contiennent respectivement une description du nombre d'items et de points pour chacune des trois compétences en lecture et pour l'écriture, qui, ensemble, constituent la littératie, pour le TPCL de 2004 et celui de 2006. Si les textes de lecture et les tâches d'écriture longue, utilisés en 2004 et en 2006, mesuraient les mêmes habiletés générales, la distribution des items de lecture à choix multiple, des items de lecture à réponse courte et des items d'écriture à réponse courte différait quelque peu dans les deux tests. Le test provincial de 2006 comprenait huit items à choix multiple qui mesuraient les conventions d'écriture et la grammaire, alors que ce n'était pas le cas pour le test de 2004.

En outre, tandis que les items à choix multiple dans les tests de 2004 et de 2006 présentaient une conception équivalente, l'équivalence de la conception des tâches pour

les items de lecture à réponse courte était plus ténue. Par exemple, pour une paire d'items, les élèves de 2004 devaient donner une explication alors que les élèves de 2006 devaient rédiger un résumé. Les deux items d'écriture à réponse courte utilisés en 2006 ont été jumelés à un item d'écriture à réponse longue en 2004 dans lequel les élèves devaient résumer un long texte de lecture.

En fin de compte, le degré de concordance entre les tests de 2004 et de 2006 n'était pas suffisamment élevé pour les considérer comme étant de conception équivalente. Les différences constatées dans la distribution des items et le manque de concordance entre les items à réponse construite ont exclu une telle affirmation. Par conséquent, l'échelonnage aux fins de comparabilité a été adopté (American Educational Research Association, et autres, 1999).

Tableau 1

Distribution des items et des points en lecture et en écriture par année

Année		Habilités en lecture						Habilités en écriture			Total
		L1		L2		L3		CM	TÉC	TÉL	
		CM	RC	CM	RC	CM	RC				
Nombre d'items											
2004	<i>k</i>	14	16	21	24	5	20	0	0	4	104
	%	13,5	15,1	20,2	23,1	4,8	19,2	0	0	3,8	
2006	<i>k</i>	17	0	17	3	6	3	8	2	2	58
	%	29,3	0,0	29,3	5,2	10,3	5,2	13,8	3,4	3,4	
2004 PT	<i>k</i>	14	0	21	3	5	3	0	0	3	49
	%	30,6	0,0	42,9	6,1	10,2	6,1	0,0	0,0	6,1	
Nombre de points											
2004	<i>n</i>	14	21	21	31	5	33	0	0	16	141
	%	9,9	14,9	14,9	22,0	3,5	23,4	0	0	11,3	
2006	<i>n</i>	17	0	17	9	6	9	8	10	40	116
	%	14,6	0,0	14,6	7,8	5,2	7,8	6,9	8,6	34,5	
2004 PT	<i>n</i>	14	0	21	6	5	6	0	0	48	100
	%	14,0	0,0	21,0	6,0	5,0	6,0	0,0	0,0	48,0	

Remarques : *k* et *n* correspondent respectivement au nombre d'items et de points; % représente le pourcentage correspondant.

L1 – Compréhension des idées et des renseignements explicites (directement énoncés)

L2 – Compréhension des idées et des renseignements implicites (faire des inférences)

L3 – Établissement de liens entre les connaissances et expériences personnelles et les idées et les renseignements contenus dans les textes lus

CM – Items à choix multiple

RC – Items de lecture à réponse construite

TÉC – Tâches d'écriture courte

TÉL – Tâches d'écriture longue

PT – Pseudotest

La différence entre le nombre d'items et le nombre de points L1, L2 et L3 est attribuable à l'utilisation d'items à choix multiple, notés de manière dichotomique (0, 1), et d'items à réponse construite, notés de manière polytomique (0, 1, 2).

Les points sont accordés pour le développement du sujet et les conventions linguistiques.

Échelonnage aux fins de comparabilité

Bien que l'on pense généralement que le concept hypothétique existe avant de concevoir un test, le concept est souvent défini sur le plan opérationnel par le premier modèle de test produit. Lorsqu'un test mesure le rendement, le concept est très complexe et constitue généralement une combinaison pondérée d'habiletés et de connaissances précises. Par exemple, la littératie est une combinaison pondérée de connaissance du vocabulaire et de la syntaxe, de la compréhension de la lecture, de connaissances préalables, de connaissance de la grammaire et de l'usage, du sens de l'organisation, de connaissance de l'orthographe et de bien d'autres composantes. La pondération de ces différentes composantes peut se faire de manière explicite (par exemple, la lecture et l'écriture peuvent être énoncées de manière à avoir des coefficients de pondération égaux) ou elle peut être déterminée par un certain nombre d'items de chaque type compris dans le test ainsi que par la qualité de l'information fournie par les items. La façon dont les réponses aux items sont combinées pour obtenir un score a une influence sur la pondération des composantes.

Si l'on peut supposer qu'un modèle particulier de test est bien conçu et bien construit, alors le score du test qui est communiqué doit refléter la pondération souhaitée des composantes et être hautement associé au construit. Cela veut dire que le classement des élèves selon le score communiqué devrait être étroitement lié au classement qui serait observé si les emplacements des élèves sur le construit même étaient disponibles. Cette observation sous-entend que le but de l'échelonnage aux fins de comparabilité est d'estimer l'emplacement du seuil de réussite ou de non réussite sur l'échelle de

communication des résultats du test (p. ex., en 2006) qui correspond à l'emplacement sur le construit qui découle de décisions précédentes (p. ex., en 2004).

Le processus d'échelonnage aux fins de comparabilité peut présenter une complication. En effet, des versions antérieures d'un test peuvent afficher une pondération différente pour des composantes de la version actuelle qui est utilisée comme définition du concept. Cela signifie qu'il faudra tenir compte des différences constatées dans les pondérations au moment de procéder à l'échelonnage aux fins de comparabilité. Dans le cas envisagé ici, on a présumé que le test provincial de 2006 reflétait avec exactitude la pondération des composantes lecture et écriture dans le construit linguistique de l'OQRE. Le test de 2004, lui, comportant deux parties notées séparément, la pondération des composantes a donc été traitée différemment. Par conséquent, il a été nécessaire de tenir compte de cette différence pour l'échelonnage des tests de 2004 et de 2006 aux fins de comparabilité.

Le groupe d'experts en psychométrie pour l'OQRE a formulé des recommandations concernant l'analyse des données et les procédures psychométriques pour la transition de 2004 à 2006 qui permettraient d'apparier les scores du TPCL de 2004 et de celui de 2006. Le groupe a recommandé l'utilisation de deux modèles de tests reflétant les différentes pondérations des composantes lecture et écriture aux fins d'échelonnage du TPCL de 2004 et de celui de 2006. Le premier était le modèle de test intégral, dans lequel le test de 2004 avec ses deux composantes distinctes a été traité comme un test générant un score. Le deuxième était le modèle du pseudotest, où il fallait créer un test « parallèle » au TPCL de 2006 au moyen des items compris dans le TPCL de

2004. Le deuxième volet de cette étude visait à comparer ces deux modèles de tests afin de déterminer lequel produirait les résultats d'échelonnage supérieurs.

Modèles de tests pour l'échelonnage

Modèle de test intégral

Le modèle de test intégral se fondait sur l'hypothèse que l'ensemble complet des items administrés dans le test provincial de 2004 fournissait la pondération adéquate pour les compétences de lecture et d'écriture du construit linguistique, selon la définition, lorsque les deux composantes ont été analysées ensemble comme test unique³. Cela dit, la pondération fournie par le nombre de points bruts dans l'ensemble complet des items correspondait à la pondération découlant du test de 2006 lorsque les items de lecture et d'écriture ont été étalonnés ensemble.

Modèle du pseudotest

En 2006, un score unique a été utilisé pour déterminer la réussite ou la non réussite. Ce score était fondé sur la notation combinée des items de lecture et d'écriture. Par conséquent, le score unique devrait être une approximation plus proche du construit linguistique que le test de 2004 avec ses parties distinctes de lecture et d'écriture.

En raison de ces changements, un pseudotest de 2004 a été construit à partir des items de la composante test de 2004 qui « imitaient » le mieux le test provincial de 2006. Le type et le nombre d'items sélectionnés devaient correspondre aussi près que possible au type et au nombre d'items et au nombre total de points pour chacune des trois

³ Les items de la mise à l'essai sont intégrés à chaque TPCL suivant un plan d'échantillonnage de matrice. Par conséquent, il existe plus d'un test chaque année. Chaque test contient les mêmes items de la composante test ainsi qu'un ensemble d'items de mise à l'essai. Les items pour le test de l'année suivante sont choisis à partir des items de mise à l'essai intégrés, et ce sont ces items qui sont utilisés pour appairer les deux années.

compétences en lecture et pour la composante écriture du TPCL de 2006. On a tenté de s'assurer que les items sélectionnés du test provincial de 2004 satisfaisaient mieux ces critères que le test intégral de 2004. Les trois tâches d'écriture longue comprises dans le pseudotest ont été notées à l'aide d'une grille de notation à quatre points en 2004, tandis que les deux tâches d'écriture longue comprises dans le test provincial de 2006 ont été notées avec une grille de notation à six points. Par conséquent, pour obtenir un coefficient de pondération de score égal au coefficient de pondération des items d'écriture de 2006, les scores des trois tâches d'écriture longue de 2004 ont été multipliés par quatre.

Le nombre d'items et de points bruts pour le pseudotest de 2004 est indiqué dans la dernière rangée de chaque partie du tableau 1. La comparaison des valeurs de ces deux rangées avec les rangées correspondantes pour le TPCL de 2004 et celui de 2006 indique que le pseudotest a donné une concordance plus proche du TPCL de 2006 que le test intégral. Par conséquent, la conception d'un pseudotest de 2004 à partir d'items de la composante test contenus dans le TPCL de 2004 visant à imiter le TPCL de 2006 a été jugée réussie.

Échantillon d'étalonnage et d'échelonnage

Un ensemble de règles d'exclusion a été mis en œuvre pour la sélection de l'échantillon d'étalonnage et d'échelonnage afin de s'assurer que l'échantillon était représentatif de la population provinciale. D'abord, les élèves admissibles antérieurement ont été retirés. Ensuite, les catégories suivantes des élèves admissibles pour la première fois ont été exclues des échantillons d'étalonnage de 2004 et de 2006 :

1. les élèves n'ayant fait aucun travail ou un travail incomplet dans une importante partie du test;

2. les élèves bénéficiant d'adaptations, sauf dans le cas de temps supplémentaire;
3. les élèves ayant bénéficié d'une exemption, d'un report ou ayant suivi le Cours de compétences linguistiques des écoles secondaires de l'Ontario (CCLESO);
4. les élèves ayant reçu un enseignement à domicile.

Après l'application des critères d'exclusion, le nombre d'élèves admissibles pour la première fois dans les échantillons d'échelonnage du TPCL était respectivement de 4 645 et de 137 496 pour le TPCL et le OSSLT de 2004, et de 5 009 et de 146 280 pour ceux de 2006. Le nombre d'élèves francophones était beaucoup plus petit que celui des élèves anglophones.

Étalonnage

Un modèle de Rasch modifié (Rasch, 1960, 1980) a été utilisé pour étalonner les items à choix multiple et le modèle modifié à crédit partiel (MCP) (Masters, 1982) a été utilisé pour étalonner les items à réponse construite. Ces modèles ont été sélectionnés pour le TPCL en raison du petit nombre d'élèves de langue francophone. En raison du nombre de versions des tests présentant différents items de mise à l'essai intégrés à l'ensemble commun d'items de la composante test, le nombre d'élèves francophones qui ont répondu à chaque item de mise à l'essai variait de 400 à 600. En revanche, le nombre d'élèves anglophones qui ont répondu à chaque item de mise à l'essai s'élevait à 7 000, et environ 1 500 réponses d'élèves aux items à réponse construite ont été notées pour l'échelonnage. Cependant, la politique de l'OQRE était de maintenir la procédure d'analyse identique pour les élèves francophones et anglophones. Conséquemment, en raison d'une taille d'échantillon insuffisante pour les items de mise à l'essai en français, des modèles plus

complexes comme les modèles à deux ou trois paramètres n'ont pas été utilisés pour les tests.

Les étalonnages ont été réalisés à l'aide du logiciel PARSCALE 4.1 (Muraki et Bock, 2003). Le paramètre a pour les items à choix multiple et à réponse construite a été défini à 0,588. Le paramètre de pseudochance pour les items à choix multiple a été défini à 0,20 ($\frac{1}{k+1}$; où k représente le nombre d'options) pour rendre compte de la possibilité que des élèves ayant de très faibles compétences puissent répondre correctement à un item. Par conséquent, le modèle de Rasch modifié est donné par la formule suivante :

$$P_i(\theta) = 0,20 + (1 - 0,20) \frac{\exp^{D \cdot 0,588(\theta - b_i)}}{1 + \exp^{D \cdot 0,588(\theta - b_i)}}$$

où $P_i(\theta)$ représente la probabilité qu'une ou un élève (ayant la compétence θ) réponde à l'item i correctement;

D est un facteur d'échelonnage = 1,7;

et

b_i représente le paramètre de difficulté pour l'item i .

Le modèle à crédit partiel a été utilisé pour estimer un item et les paramètres d'habiletés pour les items à réponse construite. Le modèle à crédit partiel est donné par la formule suivante :

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^x 0,588(\theta - \delta_{ij})\right]}{\sum_{r=0}^{m_i} \exp\left[\sum_{j=0}^r 0,588(\theta - \delta_{ij})\right]}$$

où $P_{ix}(\theta)$ représente la probabilité qu'une ou un élève (ayant la compétence θ)
reçoive un score particulier x pour l'item i ,

et

δ_{ij} représente le paramètre de difficulté d'étape pour l'item i qui est associé à un
score de catégorie j .

La première analyse réalisée visait à vérifier la qualité psychométrique des items
de mise à l'essai de 2004 avant leur inclusion dans l'étalonnage des items de la
composante test de 2006 et l'appariement des scores des tests de 2006 et de 2004. Cette
vérification s'est faite en trois étapes, à savoir :

1. étalonner les items de la composante test de 2004 et les items de mise à l'essai
sélectionnés pour le test provincial de 2006;
2. étalonner les items de la composante test de 2006;
3. comparer les paramètres des items qui sont communs pour les années 2004 et
2006. Les points correspondant aux paires de paramètres d'items devraient
apparaître le long d'une droite diagonale. Les items correspondant aux points
se trouvant à l'extérieur de la bande de confiance de 95 % qui entoure cette
droite sont jugés comme des observations aberrantes (Hu, Rogers et
Vukmirovic, 2008) et sont éliminés de l'échelonnage.

D'après cette analyse, tous les items de mise à l'essai ont été retenus.

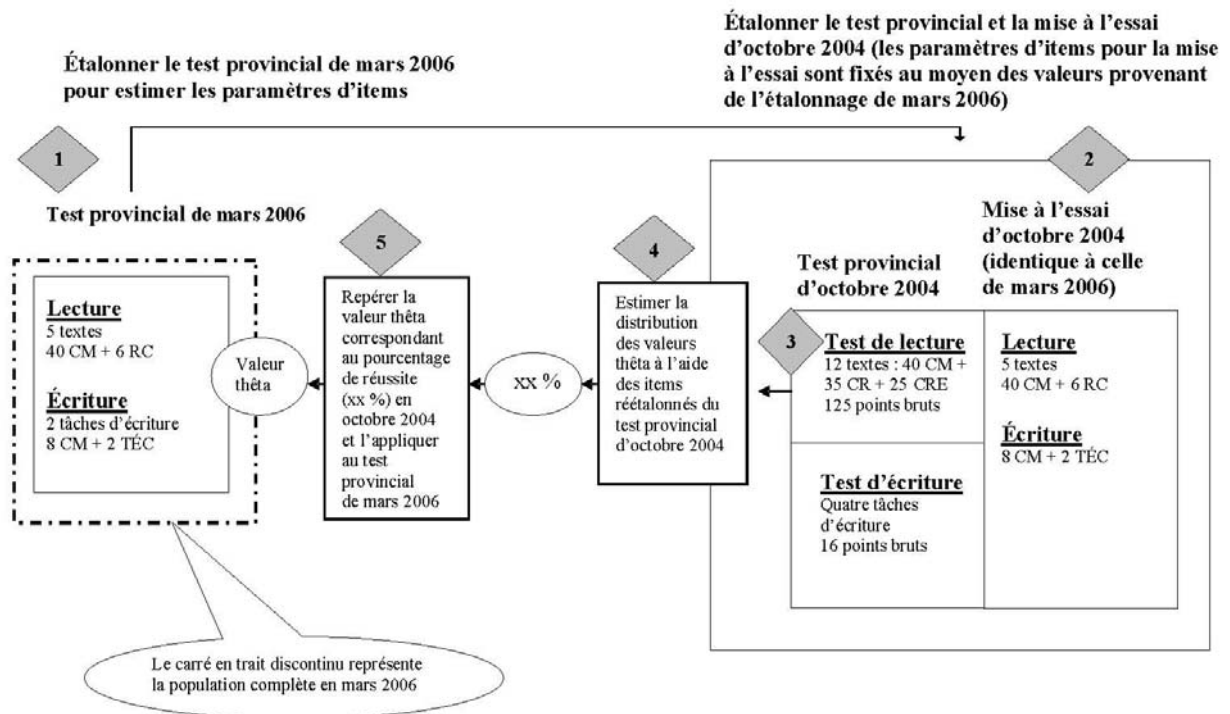
Procédure d'échelonnage

La conception du groupe sans équivalent à items communs et à paramètre à date fixe a été utilisée pour échelonner les scores du TPCL de 2004 et de celui de 2006. Des items communs ont été sélectionnés parmi les items de mise à l'essai de 2004, puis utilisés comme items de la composante test en 2006. Pour mettre les scores de 2006 et de 2004 sur la même échelle, les paramètres d'items de l'étalonnage de 2006 ont été fixés lorsque les paramètres de l'évaluation de 2004 ont été réétalonnés.

Le processus employé pour mettre le TPCL de 2004 et celui de 2006 sur une échelle commune est schématisé à la figure 1 pour le modèle de test intégral et à la figure 2 pour le modèle du pseudotest. Les étapes comprises dans chaque procédure sont décrites sous le schéma correspondant.

Figure 1

Procédure d'étalonnage pour le modèle de test intégral



Échelonnage pour le test intégral

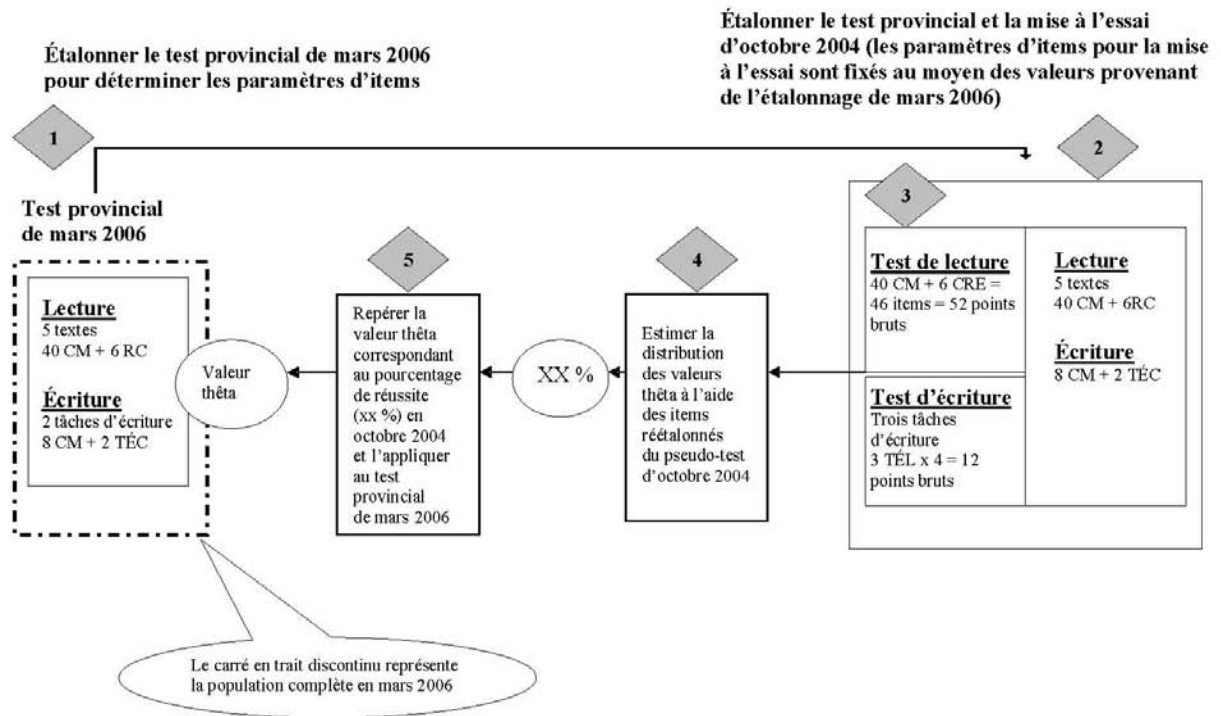
1. Étalonner tous les items de la composante test de 2006 au moyen de toutes les données du test provincial.
 - a. Créer un fichier de données renfermant les données pour les items de 2006 qui ont été notés, comme les items de la composante test pour l'échantillon d'étalonnage.
 - b. Exécuter le programme d'étalonnage TRI et sauvegarder les paramètres d'items pour chaque item et les estimations de compétence pour chaque élève. Utiliser les paramètres d'items pour échelonner les scores du test de 2004 sur l'échelle de 2006. Utiliser la distribution des compétences pour déterminer la proportion des élèves qui ont réussi le test en 2006.
2. Déterminer les items provenant de l'ensemble d'items de la composante test de 2006 qui étaient identiques dans la mise à l'essai de 2004. Utiliser uniquement les items identiques pour apparier les résultats de 2004 à l'échelle de 2006.
3. Réétalonner les composantes lecture et écriture de 2004 comme un test unique de compétences linguistiques, puis appliquer les scores à l'échelle de 2006.
 - a. Créer un fichier de données renfermant :
 - i. les réponses des élèves aux items de la composante test de 2004;
 - ii. les réponses des élèves aux items de mise à l'essai de 2004 qui ont été utilisés tels quels dans le test de 2006;
 - iii. une décision de réussite ou de non réussite pour chaque élève.
 - b. Appliquer les règles d'exclusion au fichier de données de 2004 pour obtenir les données prévues pour l'étalonnage.

- c. Exécuter le programme d'étalonnage TRI et fixer les paramètres d'items, pour les items de mise à l'essai utilisés dans les tests provinciaux de 2006, aux valeurs figurant à l'étape d'étalonnage 1b. Tous les items de la composante test de l'étalonnage de 2004 présentent des paramètres d'items sur l'échelle de 2006.
4. Utiliser les estimations des paramètres d'items pour les items de la composante test de 2004 figurant à l'étape 3c dans le calcul des estimations de compétence du modèle de Rasch pour l'ensemble complet des données du test provincial de 2004.
 - a. Fixer les paramètres d'items aux valeurs estimées à l'étape 3c.
 - b. Estimer le niveau de compétence pour chaque élève dans l'échantillon d'étalonnage.
5. Déterminer la valeur de compétence selon le modèle de Rasch de manière à ce que la proportion d'élèves se situant au-dessus de cette valeur corresponde à la proportion d'élèves ayant réussi le test en 2004.
 - a. Compiler le taux de réussite des élèves ciblés à l'étape 2.
 - b. Déterminer le point dans la distribution de l'estimation des compétences figurant à l'étape 4b avec la même proportion au-dessus de la valeur que le taux de réussite dont il est question à l'étape 5a.
 - c. Utiliser la valeur de l'étape 5b comme seuil de rendement (réussite/non réussite) dans la distribution des compétences de 2006 (se reporter à l'étape 1b).

L'ensemble complet des items de la composante test de 2004 a été utilisé pour définir l'échelle des compétences du modèle de Rasch. Par conséquent, il est probable que le seuil de réussite ou de non réussite dont il est question à l'étape 5c était très fiable.

Figure 2

Procédure d'étalonnage pour le modèle du pseudotest



Échelonnage pour le pseudotest

1. Étalonner tous les items de la composante test de 2006 à l'aide des données du test provincial.
 - a. Créer un fichier de données renfermant les données provenant des items qui ont été notés, comme les items de la composante test pour l'échantillon d'étalonnage de 2006.
 - b. Exécuter le programme d'étalonnage TRI et sauvegarder les paramètres d'items pour chaque item et l'estimation des compétences pour chaque

élève. Utiliser les paramètres d'items pour mettre les scores de 2004 sur l'échelle de 2006. Utiliser la distribution des compétences pour déterminer la proportion des élèves qui ont réussi le test de 2006.

2. Cibler les items dans l'ensemble d'items de la composante test de 2006 qui étaient identiques à ceux du pseudotest de 2004.
3. Étalonner le pseudotest de 2004 comme test unique sur l'échelle de 2006.
 - a. Créer un fichier de données renfermant :
 - i. les résultats provenant des items du pseudotest de 2004;
 - ii. les items de la mise à l'essai de 2004 qui ont été utilisés tels quels dans le test de 2006.
 - b. Appliquer les règles d'exclusion aux fichiers de données du pseudotest de 2004 pour obtenir les données aux fins d'étalonnage.
 - c. Exécuter le programme d'étalonnage TRI en fixant les paramètres d'items, pour les items de la mise à l'essai utilisés dans les tests provinciaux de 2006, aux valeurs provenant de l'étalonnage indiqué à l'étape 1b précédente. Tous les items du pseudotest provenant de l'étalonnage présentaient alors des paramètres d'items sur l'échelle de 2006.
4. Utiliser l'estimation des paramètres d'items pour les items du pseudotest dont il est question à l'étape 3c afin de calculer l'estimation des compétences selon le modèle de Rasch pour l'ensemble des données du test provincial de 2004.
 - a. Fixer les paramètres d'items aux valeurs estimées à l'étape 3c.
 - b. Estimer le niveau de compétence de chaque élève provenant de l'échantillon.

5. Déterminer la valeur de compétence du modèle de Rasch en utilisant la même proportion qui se trouve au-dessus de cette valeur comme taux de réussite de 2004.
 - a. Compiler le taux de réussite pour les élèves faisant partie de l'étape 4b.
 - b. Déterminer le point dans l'estimation de la distribution des compétences figurant à l'étape 4b, qui présente la même proportion au-dessus de la valeur comme le taux de réussite mentionné à l'étape 5a.
 - c. Utiliser la valeur de l'étape 5b comme seuil de rendement (réussite/non réussite) dans la distribution des compétences de 2006.

Résultats et discussion

Les résultats de l'échelonnage qui se fondent sur les échantillons d'appariement sont résumés au tableau 2 (les élèves francophones se trouvent dans la partie inférieure et les élèves anglophones dans la partie supérieure). Les pourcentages des élèves francophones et anglophones ayant réussi le test et faisant partie des échantillons d'échelonnage de 2004 étaient respectivement de 86,3 % et de 88,6 %. Le pourcentage des élèves francophones ayant réussi le test et faisant partie de l'échantillon d'échelonnage de 2006 était de 85,8 % pour le test intégral et de 86,4 % pour le pseudotest. Le pourcentage des élèves anglophones ayant réussi le test et faisant partie de l'échantillon d'échelonnage de 2006 était de 89,2 % pour le test intégral et de 87,8 % pour le pseudotest.

Le modèle du test intégral a entraîné une baisse de 0,5 point de pourcentage dans l'échantillon d'échelonnage de 2006 pour les élèves francophones et une hausse de 0,6 point de pourcentage dans l'échantillon d'échelonnage de 2006 pour les élèves anglophones, en comparaison avec le pourcentage des élèves ayant réussi le test et faisant

partie de l'échantillon de l'échelonnage de 2004. Le modèle du pseudotest a causé une hausse de 0,1 point de pourcentage dans l'échantillon d'échelonnage de 2006 pour les élèves francophones et une baisse de 0,8 point de pourcentage dans l'échantillon d'échelonnage de 2006 pour les élèves anglophones. Malgré la différence dans la taille de l'échantillon, les résultats d'échelonnage en français et en anglais sont comparables.

Tableau 2

Résultats de l'échelonnage aux fins de comparabilité pour le test intégral et le pseudotest : langues française et anglaise

Modèle de test pour l'échelonnage	Pourcentage des élèves ayant réussi le test	
	2004	2006
Élèves francophones		
Test intégral	86,3	85,8
Pseudotest	86,3	86,4
Élèves anglophones		
Test intégral	88,6	89,2
Pseudotest	88,6	87,8

Les modèles du test intégral et du pseudotest ont produit des résultats semblables pour les échantillons d'échelonnage du TPCL de 2006. Cependant, la mise en œuvre de chaque modèle présente des avantages et des inconvénients relatifs. Dans le cas du test intégral, le fait de ne pas avoir à construire un pseudotest à partir d'un ensemble fixe d'items administrés précédemment est un avantage, mais la pondération des habiletés et des connaissances du test de l'année 1 (découlant des points bruts) risque de ne pas correspondre avec celle du test de l'année 2, ce qui est un inconvénient possible. Une pondération incohérente entre les deux années peut causer un décalage involontaire dans

la proportion de réussite dans l'année 2 (se reporter à Schaeffer, Henderson-Montero, Julian et Bene, 2002; Sykes et Hou, 2003; Rogers et Nowicki, 2009).

Le pseudotest devrait mieux correspondre au construit mesuré durant la deuxième année. Cependant, le pseudotest a besoin d'être construit, et le nombre d'items permettant d'établir une correspondance avec les spécifications actuelles portant sur le construit à mesurer peut être insuffisant. De plus, le pseudotest comprendra moins d'items que le test intégral, ce qui peut réduire la fiabilité des scores du test et nuire aux résultats de l'échelonnage. Dans le cas présent, les valeurs du coefficient alpha de Cronbach étaient respectivement de 0,81 et de 0,82 pour les pseudotests de langues française et anglaise et de 0,86 pour les tests provinciaux de langues française et anglaise.

Conclusion

Le modèle du pseudotest a été choisi pour échelonner le TPCL de 2004 et celui de 2006 aux fins de comparabilité, après que le nombre d'items a été réduit d'environ de moitié et que le modèle de décision conjonctif a été remplacé par un modèle de décision compensatoire. Le pseudotest ressemblait plus aux spécifications du test provincial de 2006 et à la proportion des items par type et par habileté que le test intégral. Même s'il a semblé que le test intégral de 2004 a fourni une pondération adéquate des composantes lecture et écriture relativement au construit linguistique, le modèle du pseudotest fournit de manière explicite une pondération plus appropriée.

Un des principaux défis à relever dans les évaluations à grande échelle est de tenir compte des changements qui doivent être apportés dans le test en réponse aux changements apportés au curriculum, à la durée de l'administration du test ou au modèle de décision. Malgré ces changements, le souhait de savoir si le rendement des élèves

varie d'une année à l'autre demeure. Le modèle du pseudotest énoncé dans le présent document offre une méthode viable d'échelonnage aux fins de comparabilité lorsqu'un changement est apporté à la structure du test, à la durée de son administration ou à la communication des résultats.

Références

- American Educational Research Association, American Psychological Association et National Council on Measurement in Education. 1999, *Standards for Educational and Psychological Testing*, Washington, DC, American Educational Research Association.
- Hu, H., W. T. Rogers et Z. Vukmirovic. 2008, Investigation of IRT-based equating methods in the presence of outlier common items, *Applied Psychological Measurement*, vol. 32, n° 4, p. 311-333.
- Kolen, M. J., et R. L. Brennan. 2004, Test equating, scaling, and linking: Methods and practices (2^e édition), New York, NY, Springer.
- Linn, R. L. 1993, Linking results of distinct assessments, *Applied Measurement in Education*, vol. 6, n° 1, p. 83-102.
- Masters, G. N. 1982, A Rasch model for partial credit scoring, *Psychometrika*, vol. 47, p. 149-174.
- Mislevy, R. J. 1992, *Linking educational assessments: Concepts, issues, methods, and prospects*, Princeton, NJ, ETS Policy Information Center.
- Muraki, E., et R.D. Bock. 2003, *PARSCALE: IRT item analysis and test scoring for rating-scale data, (Version 4.1)* [logiciel] Chicago, IL, Scientific Software International.
- Office de la qualité et de la responsabilité en éducation. 2007, *Document-cadre, Test provincial de compétences linguistiques*, Toronto, ON, auteur.
- Rasch, G. 1960/1980, *Probabilistic models for some intelligence and attainment tests*. (Copenhague, Institut danois pour la recherche pédagogique), édition augmentée,

- 1980, avec avant-propos et postface de B. D. Wright. Chicago, IL, The University of Chicago Press.
- Rogers, W. T., et D. Nowicki. April 2009, *A Comparison of Three Weighting Procedures for High- and Low-Stakes Examinations with Mixed Item Formats in Different Subject Areas*, document présenté lors de la réunion annuelle du National Council on Measurement and Evaluation, San Diego, CA.
- Schaeffer, G. A., D. Henderson-Montero, M. Julian et N. H. Bene. 2002, A Comparison of Three Scoring Methods for Tests With Selected-Response and Constructed-Response Items, *Educational Assessment*, vol. 8, p. 317-340.
- Sykes, R. C., et L. Hou. 2003, Weighting constructed-response items in IRT-based exams, *Applied Measurement in Education*, vol. 16, p. 257-275.
- Wolfe, R., R. Childs et S. Elgie. Mai 2004, *Final report of the external evaluation of EQAO's assessment process*, Toronto, ON, Institut d'études pédagogiques de l'Ontario.



Office de la qualité et de la responsabilité en éducation

2, rue Carlton, bureau 1200, Toronto (Ontario) M5B 2M9, | 888 327-7377, www.oqre.on.ca

© Imprimeur de la Reine pour l'Ontario, 2010