



Étude comparative de quatre méthodes de calibrage de tests

Rapport préparé pour
l'Office de la qualité et de la responsabilité en éducation (OQRE) par

Xiao Pang
Psychométricienne, OQRE

Ebby Madera
Psychométricienne, OQRE

Nizam Radwan
Psychométricien, OQRE

Su Zhang
Psychométricienne, OQRE

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

Remerciements

Cette étude a été dirigée par Michael Kozlow et les chercheurs invités de l'OQRE Todd Rogers et Mark Reckase, qui ont donné des conseils sur la rédaction de la proposition et sur l'étude elle-même. Ils ont prodigué des recommandations précieuses et détaillées en matière de procédures de recherche, ont fait part de leurs commentaires à différentes étapes de l'analyse, et ont lu et commenté le rapport final. Le soutien intellectuel et technique offert par Qi Chen a permis d'accélérer la procédure d'analyse. Yunmei Xu a apporté un soutien opportun pour compléter l'analyse. Les auteures et auteurs de ce rapport leur sont reconnaissants de leurs contributions importantes qui ont permis d'améliorer la qualité de cette étude.

Résumé

Cette étude évaluait l'efficacité de quatre méthodes de calibrage courantes à mesurer les gains réels des élèves, soit le calibrage concomitant (CC), le calibrage selon un paramètre d'item commun fixe (PICF), le calibrage par courbes caractéristiques du test (CCT) de Stocking et Lord et le calibrage à l'aide de la moyenne/sigma (M/S). Nous avons évalué l'efficacité des quatre procédures à l'aide de données simulées pour un test de format à items multiples. Cinq conditions de gain (-0,3, -1,0, 0,0, 0,1 et 0,3 sur l'échelle θ) ont été intégrées à la simulation pour simuler le Test provincial de compétences linguistiques (TPCL), l'Ontario Secondary School Literacy Test (OSSLT), les Tests en lecture, écriture et mathématiques, cycles primaire et moyen, de langue française et de langue anglaise, et le Test de mathématiques, 9^e année, cours appliqué, de langue anglaise. Vingt répétitions ont été effectuées. Les pourcentages estimés des niveaux de rendement multiples et dans les catégories réussite et non réussite ont été comparés aux pourcentages réels respectifs obtenus des distributions θ connues. Les résultats des sept tests ont montré que les procédures de calibrage PICF, CCT et M/S basées sur des calibrations distinctes fonctionnaient aussi bien et beaucoup mieux que la procédure CC.

Introduction

L'un des objectifs de l'Office de la qualité et de la responsabilité en éducation (OQRE) est de fournir des preuves concrètes des variations dans le rendement des élèves d'une année à l'autre en Ontario.¹ L'Office administre chaque année des tests, en français et en anglais, aux cycles primaire (3^e année) et moyen (6^e année) en lecture, en écriture et en mathématiques, et en 9^e année, cours théorique et appliqué de mathématiques. Les résultats de ces tests sont diffusés en termes du pourcentage d'élèves dans cinq niveaux de rendement (manque d'indices pour attribuer le niveau 1 [MI du niveau 1] ou inférieur au niveau 1, et les niveaux 1, 2, 3 et 4). La norme de rendement acceptable établie par la province est le niveau 3. En plus de ces tests, l'OQRE est aussi responsable de l'administration de deux tests de compétences linguistiques : le Test provincial de compétences linguistiques (TPLC), de langue française, et l'Ontario Secondary School Literacy Test (OSSLT), de langue anglaise, tous deux étant un crédit obligatoire pour l'obtention du diplôme d'études secondaires.²

Lorsque l'on signale une variation sur le plan du rendement entre deux années, il est important d'établir la distinction entre les différences dues à la difficulté des tests utilisés pour évaluer les élèves et les gains ou pertes de rendement véritables entre les deux années. L'objectif du calibrage est d'ajuster les différences de difficulté inhérentes aux tests pour que seules les véritables différences de rendement figurent.

Il existe toutefois différentes procédures pour effectuer le calibrage de tests, dont certaines sont basées sur la théorie classique des scores (TCS) et d'autres sur la théorie de la réponse à l'item (TRI). Des études ont démontré que le calibrage basé sur la TCS et sur la TRI produit des résultats similaires à ceux obtenus par le calibrage horizontal. Hills, Subhiyah et Hirsch (1988), par exemple, ont trouvé des résultats similaires avec le calibrage concomitant (CC) vertical à l'aide du modèle de Rasch et le modèle TRI à trois paramètres, et un calibrage distinct à l'aide du modèle TRI à trois paramètres avec le calibrage selon un paramètre d'item commun fixe (PICF) et le calibrage à l'aide de la moyenne/sigma (M/S) (Marco, 1977). Kolen et Brennan (1995) ont toutefois fait remarquer que puisque plusieurs grands programmes

¹ L'OQRE est un organisme indépendant du ministère de l'Éducation de l'Ontario qui administre des tests à l'échelle de la province.

² Les élèves qui ne réussissent pas le TPCL peuvent le faire de nouveau l'année suivante ou suivre le Cours de compétences linguistiques des écoles secondaires de l'Ontario.

d'évaluation utilisent des modèles TRI pour concevoir et calibrer leurs tests, il est souvent logique de recourir aux méthodes de calibrage basées sur la TRI. Par conséquent, puisque l'OQRE utilise des procédures basées sur la TRI pour calibrer et mettre en équivalence chaque item de ses tests, nous avons restreint dans le cadre de la présente étude les méthodes de calibrage aux méthodes basées sur cette théorie.

Les procédures de calibrage TRI les plus souvent utilisées sont la procédure CC (Wingersky et Lord, 1984), qui est basée sur le calibrage concomitant d'un échantillon constitué des élèves évalués dans chacune des deux années à mettre en équivalence; la procédure PDCF; la procédure des courbes caractéristiques du test (CCT) (Loyd et Hoover, 1980) et la procédure M/S. Les procédures PDCF, CCT et M/S sont basées sur des calibrages distincts des deux échantillons. Malheureusement, ces procédures ne produisent pas toujours les mêmes résultats. Il est, par conséquent, essentiel de bien comprendre le comportement de chacune des méthodes de calibrage pour assurer que l'interprétation de la variance estimée est valide.

L'OQRE utilise actuellement différentes méthodes de calibration TRI et la procédure de calibrage PDCF. Mais aucune étude ne s'est penchée sur l'efficacité de cette démarche pour déceler les gains ou les différences entre les résultats des tests de l'OQRE de deux années différentes, ou si l'une ou l'autre des méthodes de calibrage TRI est mieux en mesure de déceler ces variations.

Objectif de l'étude

L'objectif de cette étude est d'évaluer l'efficacité des quatre procédures de calibrage indiquées ci-dessus (CC, PDCF, CCT et M/S) à déceler les variations réelles du rendement des élèves au fil des ans. On a spécifiquement comparé avec quel degré de précision les résultats produits par les quatre procédures représentaient des variations connues dans les pourcentages d'élèves de chaque niveau de rendement pour les tests du cycle primaire (3^e année), du cycle moyen (6^e année) et de 9^e année et dans les deux catégories de rendement du TPCL et de l'OSSLT (réussite et non réussite).

Examen des méthodes de calibrage

Deux façons de faire peuvent être utilisées avec les méthodes de calibrage par groupe non équivalent d'items communs et basé sur la TRI : le calibrage concomitant ou le calibrage

distinct. Avec le calibrage concomitant et l'approche de calibrage (Lord et Winkersky, 1984), les réponses des élèves aux deux tests à mettre en équivalence sont combinées dans un seul fichier de données en alignant les items communs. Les tests sont ensuite mis en équivalence simultanément. Par conséquent, les paramètres estimés des items des tests sont placés sur une échelle commune. Les scores d'habileté des élèves aux deux tests sont estimés séparément à l'aide des paramètres d'items échelonnés correspondants, et les moyennes des deux tests sont ensuite comparées pour déterminer la direction et l'ampleur de la variation. En théorie, la méthode CC devrait donner des résultats plus stables que les méthodes de calibrage distinctes qui recourent aux transformations, et on s'attend aussi à ce que la méthode CC réduise l'impact des fluctuations des échantillons dans l'estimation du paramètre de pseudo-chance à cause de l'augmentation du nombre de participants aux résultats faibles.

Avec deux calibrations distinctes, les calibrations sont effectuées séparément pour les deux tests et les items communs sont utilisés pour placer les deux tests sur une échelle commune. Le test utilisé pour définir l'échelle commune est le « test de référence », et le deuxième test est le « test calibré ». On peut utiliser une transformation linéaire pour placer les paramètres des items du test calibré sur l'échelle du test de référence en fonction des items communs aux deux tests. Les procédures de calibrage qui utilisent une transformation linéaire incluent l'approche moyenne/moyenne (M/M) (Loyd et Hoover, 1980), la méthode M/S (Marco, 1977) et l'approche CCT (Li, Lissitz et Yang, 1999; Stocking et Lord, 1983). Bien qu'il soit théoriquement approprié d'utiliser la procédure M/M ou M/S, les paramètres sont utilisés séparément pour estimer les coefficients de calibrage. La méthode CCT, au contraire, est une procédure d'estimation simultanée qui tient mieux compte de l'information fournie (Li et autres, 1999).

La méthode PICF est une méthode alternative de calibrage et de calibration en deux étapes. Avec cette méthode, on commence par calibrer le test de référence. Une fois que le test de calibrage est calibré, les paramètres de ses items communs sont fixés pour les valeurs estimées obtenues par le biais de la calibration du test de référence. Ainsi, la distribution du score du test calibré est placée sur l'échelle du test de référence (pour une description technique détaillée de la méthode PICF, voir Kim, 2006). La procédure PICF est censée produire des résultats supérieurs à ceux produits par les procédures M/M, M/S et CCT du fait de l'évitement des fonctions de transformation inappropriées.

Bien que certaines études aient été effectuées pour évaluer différentes approches de calibrage TRI (Hanson et Belguin, 2002; Hills, Subhiyah et Hirsch, 1988; Kim et Cohen, 1998; Kolen et Brennan, 2004; Petersen, Cook et Stocking, 1983; Prowker et Camilli, 2006; Bishop, Shari, Lei et Domaleski, 2006; Hu, Rogers et Vulkmirovic, 2008; Kim, 2006; Lee et Ban, 2010; Wingersky, Cook et Eignor, 1987), un nombre limité d'études a été réalisé pour comparer le comportement des approches concomitante et séparée.

Petersen et autres (1983) ont comparé la procédure linéaire TCS, les méthodes concomitantes TRI 1PL et 3PL et la méthode de transformation linéaire 3PL. Ils ont découvert que les différentes méthodes donnaient lieu à des résultats similaires lorsque les tests à calibrer étaient parallèles et que les groupes des deux années étaient équivalents.

Wingersky et autres (1987) ont étudié les effets des caractéristiques des items liants sur les résultats de calibrage de la note réelle par TRI et ont conclu que la procédure CCT était influencée par la présence d'items liants qui fonctionnent différemment pour les deux groupes utilisés pour fournir les données. Hills et autres (1988) ont comparé le calibrage linéaire TCS, les méthodes concomitantes TRI 1PL et 3PL, la méthode PICF 3PL et la méthode de transformation linéaire 3PL lorsque les tests étaient parallèles et que les groupes étaient équivalents. Ils ont découvert que les différentes méthodes donnaient lieu à des résultats similaires. Lors de leur calibration des six tests à l'aide d'un seul ensemble d'items de copies types, Li, Griffith et Tam (1997) ont découvert que les approches PICF et CCT donnaient des estimations des habiletés calibrées comparables, sauf avec la méthode de calibrage CCT dans le cas des élèves aux aptitudes extrêmes.

Lee et Ban (2010) ont comparé quatre procédures de calibrage TRI différentes (CC, CCT, Haebara et transformation des compétences) et ont découvert que les procédures de calibration séparées donnaient de meilleurs résultats que les procédures CC et de transformation des compétences. Kim et Cohen (1998) ont comparé les procédures CCT et concomitante à deux procédures d'estimation différentes – le maximum de vraisemblance marginale et le maximum marginal « a posteriori » – en recourant à des items à choix multiple. Ils ont découvert que les deux procédures donnaient lieu à des résultats similaires, sauf lorsque le nombre d'items communs était petit, auquel cas l'approche séparée donnaient des résultats plus précis. Linn et autres (1980) ont comparé les procédures de calibrage CCT, M/M, M/S et M/S pondéré. Leurs résultats

ont indiqué que les différences dans les coefficients de calibrage entre ces méthodes étaient minimales entre des échantillons de taille variée et un nombre d'items communs variés.

Hanson et Belguin (2002) ont comparé les procédures CC, M/M, M/S, CCT et CCT d'Haerbrara à l'aide de simulations par ordinateur. Ils ont découvert qu'en général, la procédure CC produisait moins de biais et d'erreur aléatoires entre les répétitions et que les procédures de calibration distincte et de transformation. Keller et autres (2004) ont évalué la capacité des quatre méthodes (CC, M/S, CCT et PICF) de détecter les variations dans la distribution des aptitudes des candidats en recourant à des données simulées basées sur une distribution d'aptitudes normales standard. Ils ont découvert que la méthode M/S donnait les meilleurs résultats et que la méthode PICF donnait les pires résultats.

Hu et autres (2008) ont effectué une étude par simulation pour étudier dix variations de quatre méthodes de calibrage (CC, M/S, CCT et PICF) en l'absence et en la présence d'observations aberrantes dans l'ensemble d'items communs. Leur conclusion était que « les transformations CCT et par M/S donnaient les meilleurs résultats... Les calibrations CC et PICF avaient une interaction complexe avec l'équivalence de groupe ainsi qu'avec le nombre et les scores bruts des observations aberrantes » (p. 311). Lorsqu'il n'y avait aucune observation aberrante dans l'ensemble d'items communs, leur conclusion était que les quatre méthodes étaient sensibles, mais pas toutes dans la même mesure, à la présence des groupes non équivalents; lorsqu'il n'y avait aucune différence entre les habiletés moyennes de deux groupes, les quatre procédures étaient équivalentes (biais et erreur aléatoire), mais lorsque les habiletés moyennes des deux groupes différaient par un écart-type, les méthodes M/S et CCT donnaient lieu à des biais et des erreurs aléatoires moindres, la méthode PICF donnait lieu à des biais et des erreurs aléatoires modérés, et la méthode CC produisait le plus grand nombre de biais et d'erreurs aléatoires.

Les résultats des études examinées étaient incohérents. De plus, aucune étude approfondie n'a comparé les quatre méthodes de calibrage à l'aide a) de données constituées d'items à notation dichotomique et polytomique; b) de données dont la distribution n'est pas nécessairement normale; c) de groupes non équivalents; et d) de tests de matières variées.

Comme indiqué ci-dessus, les tests de l'OQRE sont constitués d'un mélange d'items à notation dichotomique et polytomique. Les scores ne suivent pas une courbe de distribution

normale. Les populations ne sont pas équivalentes d'une année à l'autre. Actuellement, l'OQRE utilise une calibration distincte suivie de la méthode PICF pour lier deux années consécutives de chaque test.

Les procédures CC et PICF ont été comparées à l'aide des tests de l'OQRE du cycle primaire, du cycle moyen et de 9^e année des années 2005-2006 et 2006-2007. Certaines différences dans la variation en pourcentages des élèves placés dans les quatre niveaux de rendement (niveaux 1, 2, 3 et 4) dans les années adjacentes ont été observées pour les deux procédures de calibrage. Par conséquent, l'objectif de cette étude est d'examiner en profondeur dans quelle mesure ces deux procédures de calibrage pourraient déceler les variations réelles reflétées dans le pourcentage des élèves dans chaque catégorie de rendement entre deux années consécutives. Simultanément, les approches CCT et M/S ont été étudiées afin de déterminer laquelle des quatre méthodes décelait le mieux la variation réelle.

Méthode

Cette étude s'efforce de vérifier le degré de précision avec lequel les estimations de compétences TRI calibrées produites par les méthodes de calibrage CC, M/S, CCT et PICF décelaient les variations réelles dans le rendement des élèves entre les années 2005-2006 et 2006-2007 (respectivement l'an 1 et l'an 2). Dans les situations de test réelles, il est impossible de connaître les vraies variations, puisque les capacités réelles des élèves sont inconnues. Si l'information requise ne peut autrement raisonnablement être obtenue analytiquement, des études de simulation devraient être effectuées (Psychometric Society, 1979; Lehman et Bailey, 1968). Des procédures de simulation informatiques ont par conséquent été utilisées dans le cadre de la présente étude, lors desquelles les variations estimées dans le rendement des élèves ont été comparées aux variations réelles connues.

Simulations de données

Modèle de calibrage. Afin de simuler des cas réalistes, les données simulées ont été générées pour reproduire le modèle de matrice de groupes d'items communs non équivalents de l'OQRE. Dans ce modèle, des ensembles différents d'items mis à l'essai sont intégrés dans des versions multiples du test provincial pour l'an 1 à l'aide d'une matrice. Les items du test ayant de bonnes propriétés psychométriques sont ensuite utilisés dans le test provincial de l'an 2 et servent de lien entre les tests de l'an 1 et de l'an 2. La figure 1 représente visuellement ce modèle

de calibrage. Le bloc supérieur gauche contient le test provincial pour l’an 1 et le bloc supérieur droit contient un ensemble différent d’items mis à l’essai intégrés au test (M_1, M_2 et M_3). Les items mis à l’essai qui correspondent au devis du test, qui possèdent de bonnes caractéristiques psychométriques et qui produisent la fonction d’information du test désirée sont utilisés pour concevoir le test provincial pour l’an 2, contenu dans le bloc inférieur droit. Par conséquent, tous les items du test provincial pour l’an 2 ont préalablement été mis à l’essai, ce qui donne un lien beaucoup plus solide entre les deux tests que ceux établis par d’autres modèles de calibrage dans lesquels un nombre limité d’items communs sont utilisés pour calibrer les tests (généralement 20 items ou au moins 20 % du nombre total d’items) (Angoff, 1984; Kolen et Brennan, 2004).³

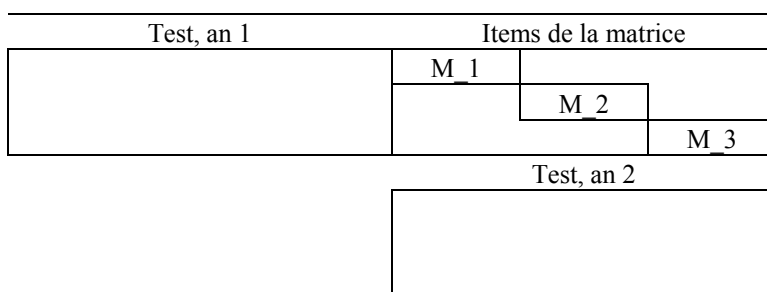


Figure 1 : Modèle de calibrage pour les tests de l’OQRE

Ce modèle comporte d’autres avantages. Plus d’items peuvent être mis à l’essai durant la période normale d’administration du test. De plus, puisque les items mis à l’essai sont intégrés parmi les items du test provincial de l’an 1 et que les élèves ne savent pas quels items sont des items mis à l’essai et quels items font parti du test provincial, ils sont motivés à bien répondre à tous les items. On permet d’éliminer l’effet de la fatigue qui influencerait les réponses aux items mis à l’essai s’ils étaient placés à la fin du test. Finalement, le risque lié à la sécurité du test est grandement réduit. Ces avantages rendent intéressant le modèle de matrice de groupe d’items communs non équivalents, qui a été utilisé dans plusieurs programmes d’évaluation à grande échelle (Hu et autres, 2008).

Taille de l’échantillon. Puisque la population d’élèves pour chaque test et que le nombre de cahiers de test de la matrice mis à l’essai varie, la taille de l’échantillon pour chaque cahier de la matrice varie aussi. En général, les tests des cycles primaire et moyen de langue anglaise et

³ À l’exception des tâches d’écriture longue. Ces tâches demandent trop de temps pour qu’elles soient incluses comme items mis à l’essai dans le test provincial.

l'OSSLT sont ceux pour lesquels la taille des échantillons est la plus grande (p. ex., $n \approx 5\ 000$ dans le cas de l'OSSLT), alors que les tests des cycles primaire et moyen de langue française et le TPCL sont ceux pour lesquels la taille des échantillons est la plus petite (p. ex., $n \approx 526$ dans le cas du TPCL). La taille des échantillons pour le Test de mathématiques, 9^e année, cours appliqué, de langue anglaise, est moyenne ($n \approx 1\ 500$). Pour déterminer l'effet possible des matières sur les taux de récupération, différents tests et sous-tests ont été sélectionnés pour la simulation : écriture, cycle primaire de langue française; écriture, cycle primaire de langue anglaise; lecture, cycle moyen de langue anglaise; mathématiques, cycle moyen de langue française; mathématiques, cours appliqué de 9^e année de langue anglaise; et le TPCL et l'OSSLT.

Caractéristiques des tests. Les caractéristiques psychométriques des tests et des sous-tests utilisés dans cette étude sont présentées dans le tableau 1. Comme indiqué, les tests comportent différents nombres d'items à choix multiple et à réponse construite et différents nombres de scores bruts pour les items à réponse construite. Par exemple, dans le cas du TPCL et de l'OSSLT, les catégories de scores varient de trois à dix pour les items de lecture à réponse construite et les tâches d'écriture courte et longue. Pour l'an 2, le nombre d'items a été réduit en lecture, cycle primaire de langue anglaise, et en mathématiques, cours appliqué de 9^e année de langue anglaise; cependant, la structure de base et le contenu du test à mesurer étaient similaires entre les deux années (consultez les documents-cadres affichés sur le site Web de l'OQRE : www.oqre.on.ca). Pour tous les tests, le nombre total de points pour les items à choix multiple est inférieur au nombre de points pour les items à réponse construite.

La moyenne des distributions θ est négative, mais près de zéro. Cela est probablement dû à l'asymétrie négative des distributions (Lord, 1980, p. 49-50). Les déviations standards des distributions θ sont légèrement inférieures à un. Certaines distributions étaient légèrement leptocurtiques (p. ex., lecture, cycle moyen de langue anglaise), d'autres étaient essentiellement mésocurtiques (p. ex., OSSLT et écriture, cycle primaire de langue anglaise), alors que d'autres étaient légèrement platycurtiques (p. ex., écriture, cycle primaire de langue française).

Tableau 1

Caractéristiques psychométriques des tests et des sous-tests de l'étude

Test	Année	N	Nombre d'items		Moyenne	DS	Asym.	Apla.
			CM	RC				
OSSLT	1	150 000	39	12 (50) ^a	-0,01 ^c	0,81	-0,25	3,03
	2	150 000	39	12 (50)	-0,01	0,85	-0,26	3,03
TPCL	1	5 260	39	12 (50)	-0,02	0,88	-0,11	2,80
	2	5 260	39	12 (50)	-0,02	0,79	-0,20	3,10
Écriture, primaire, angl.	1	110 000	8	6 (36)	-0,01	0,92	-0,14	3,03
	2	110 000	8	6 (36)	-0,01	0,91	-0,13	3,03
Écriture, primaire, fr.	1	6 000	8	6 (36)	0,01	0,91	-0,14	2,90
	2	6 000	8	6 (36)	0,01	0,90	-0,05	2,88
Lecture, moyen, angl.	1	132 000	32	12 (48)	-0,01	0,93	-0,13	3,29
	2	132 000	25	10 (40)	-0,01	0,92	-0,24	3,34
Math., moyen, fr.	1	5 600	28	8 (32)	0,00	0,93	-0,05	2,64
	2	5 600	28	8 (32)	0,00	0,94	-0,07	2,74
Math., cours appliqué, 9 ^e année, angl.	1	37 500 ^b	24	12 (30)	-0,03	0,92	-0,04	3,20
	2	34 441	24	7 (28)	-0,03	0,90	-0,17	3,18

^a $x(y)$: nombre d'items à réponse construite et nombre total de points possibles pour ces items.

^b Échantillons de l'hiver et du printemps combinés.

^c Les statistiques descriptives étaient basées sur l'échelle θ des calibrations opérationnelles.

Modèles TRI. Le modèle TRI utilisé pour générer les réponses aux items pour le TPCL et l'OSSLT était un modèle de Rasch aux estimations fixées à 0,20 pour les items à choix multiple et au paramètre a fixé à 0,588. Cette valeur du paramètre a définit la discrimination à 1,0 parce que le paramètre a est multiplié par 1,7 dans le modèle. Dans le cas des sous-tests du cycle

primaire, du cycle moyen et de 9^e année, le modèle à deux paramètres couplé à un paramètre de pseudo-chance fixé a été utilisé pour les items à choix multiple. Le modèle à crédit partiel généralisé a été utilisé pour les items à réponse construite de tous les tests et sous-tests. Ces modèles TRI semblent être les plus appropriés dans le cas des tests de l'OQRE (Xie, 2006).

Étapes pour la simulation des données. Les deux questions qui suivent ont guidé la conception de la simulation informatique de chaque test :

- a. Quelles sont les véritables variations (en pourcentage) pour chaque niveau de rendement?
- b. Quels seraient les gains pour chaque niveau de rendement dans une situation de test réelle une fois les quatre processus de calibrage appliqués, et à quel point ces variations estimées sont-elles près des variations réelles?

Les étapes de simulation des données qui suivent ont été effectuées pour aider à répondre à ces questions.

1. Pourcentages réels

Pour déterminer les pourcentages réels de chaque catégorie de rendement, les distributions θ connues pour l'an 1 et l'an 2 ont été simulées à l'aide de la famille Pearson de type quatre (moyenne, écart-type, asymétrie et aplatissement) des distributions θ tirées des tests provinciaux de l'an 1 et de l'an 2, respectivement (voir le tableau 1). Puisque les variations réelles entre les deux années ne sont pas connues, cinq variations « réelles » possibles (-0,3, -0,1, 0,0, 0,1 et 0,3 unité sur l'échelle θ) ont été modélisées dans les simulations de données pour refléter différentes variations du rendement. Ces valeurs couvrent l'intervalle de variations du rendement qui pourrait être observé dans des conditions éducatives réalistes, bien que les conditions $\pm 0,3$ représentent des variations plus importantes que celles qui sont généralement observées dans les tests de l'OQRE. Pour créer cinq conditions de gain pour l'an 2, les cinq gains ont été ajoutés à la moyenne de la distribution θ de l'an 1. La distribution θ connue pour l'an 2 a ensuite été simulée pour chacune des cinq conditions de gain pour chaque test ou sous-test sélectionné. La taille des échantillons utilisés dans les simulations a été déterminée pour être proche des échantillons de calibration utilisés en pratique pour chaque test. Dans les échantillons de calibration, les élèves auxquels on a permis d'utiliser des versions spéciales et les élèves qui

n'ont pas répondu ont été exclus. Dans le cas de l'OSSLT, les élèves qui étaient admissibles à faire le test antérieurement ont aussi été exclus de l'échantillon de calibration.

On a déterminé les scores seuils sur la distribution θ connue de l'an 1 en utilisant les pourcentages publiés par l'OQRE pour chaque niveau de rendement. Ces scores seuils ont ensuite été appliqués aux cinq distributions θ connues de l'an 2 pour déterminer le pourcentage réel pour chaque niveau de rendement.

2. Pourcentages empiriques

Pour obtenir les pourcentages empiriques, la matrice de données qui simule les tests de l'OQRE doit être simulée. La simulation des données comporte deux étapes : a) simuler l'ensemble de données complet pour les élèves de l'an 1 et de l'an 2; b) utiliser l'ensemble de données complet pour générer la matrice de données pour la calibration. Pour simuler l'ensemble de données complet, les paramètres des items des tests provinciaux de l'an 1 et de l'an 2 ont été combinés en un fichier. La distribution θ connue de l'an 1 a aussi été combinée avec chacune des cinq distributions θ connues de l'an 2. Les vecteurs d'items-réponses pour les élèves ont ensuite été générés pour les tests de l'an 1 et de l'an 2 pour chaque condition de gain en fonction du fichier de paramètres et des distributions θ combinées (voir figure 2).

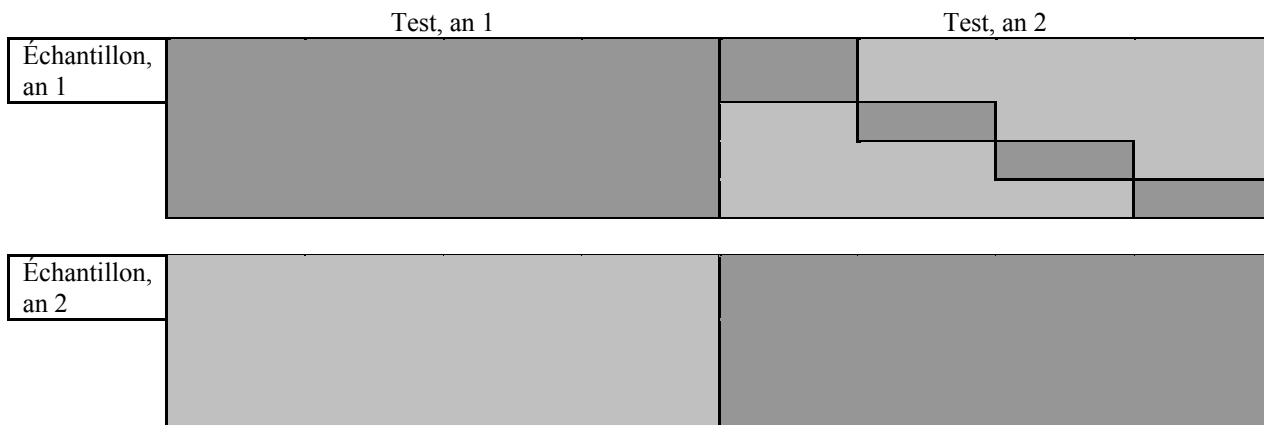


Figure 2. Structure complète des données

L'axe vertical du diagramme représente les élèves. Les élèves au-dessus du point milieu de l'axe vertical sont les élèves de l'an 1 et ceux au-dessous du point milieu sont ceux de l'an 2. L'axe horizontal représente les items. Les items du côté gauche du point milieu de l'axe horizontal sont inclus dans le test administré la première année et les items du côté droit sont inclus dans le test administré la deuxième année. Pour créer la matrice d'échantillon de calibrage

de l'an 1 et l'échantillon de calibrage du test provincial de l'an 2, les parties en gris pâle du diagramme ont été supprimées de l'ensemble de données complet. On estime que la meilleure manière d'obtenir des renseignements valables sur les variations du rendement des élèves consisterait à faire faire les deux tests à chaque cohorte d'élèves. Par conséquent, créer les échantillons de calibrage depuis l'ensemble de données complet idéal semblait raisonnable.

Une fois les ensembles de calibrage usuels créés, le calibrage a été effectué par les méthodes de calibrage CC, PICAL, CCT et M/S pour obtenir les pourcentages empiriques. Dans le cas de la procédure CC, les ensembles de données de l'an 1 et de l'an 2 ont été combinés et calibrés ensemble. Dans le cas des procédures CCT et M/S, les ensembles de données de l'an 1 et de l'an 2 ont d'abord été calibrés séparément. Les procédures CCT et M/S ont ensuite été appliquées pour obtenir les coefficients de transformation linéaire pour réduire le test de l'an 1 (calibré) au test de l'an 2 (référence). Dans le cas des procédures PICAL, les deux tests ont été calibrés séparément, les paramètres de la matrice des items de l'an 1 fixé aux valeurs des paramètres des items du test provincial de l'an 2 pour réduire le test de l'an 1 à l'échelle de l'an 2. Des procédures similaires à celles utilisées lors de l'étape 1 ont été appliquées pour identifier un score seuil et obtenir un pourcentage empirique pour chaque niveau de rendement et chaque condition de gain.

Programmes informatiques. Les réponses des participants aux items ont été simulées à l'aide des logiciels Datagenpcmv et Datagen3plt de Matlab. Datagenpcmv a été utilisé pour simuler les réponses aux items à réponse construite et datagen3plt a été utilisé pour simuler les réponses aux items à choix multiple. Les distributions des réponses aux items simulées ont été comparées aux distributions des réponses aux items réels et elles montrent des structures très similaires pour chacun des tests sélectionnés. On a utilisé PARSCALE pour effectuer les calibrations parce que l'OQRE l'utilise pour la calibration TRI opérationnelle et la notation. MULTILOG et PARSCALE génèrent des estimations de paramètres similaires (Childs et Chen, 1999; Hanson et Beguin, 2002), mais PARSCALE produit un paramètre d'emplacement global de l'item et reproduit les paramètres de catégorie en les centrant à zéro (Childs et Chen, 1999). De plus, PARSCALE peut gérer un nombre de participants beaucoup plus élevé.

Évaluation des méthodes de calibrage PICAL, CC, CCT et M/S. Le rendement des méthodes de calibrage PICAL, CC, CCT et M/S a été évalué en comparant le pourcentage estimé au pourcentage réel correspondant pour chacun des quatre niveaux de rendement pour les tests

du cycle primaire, du cycle moyen et de 9^e année et pour les catégories réussite et non réussite du TPCL et de l'OSSLT. On a effectué 20 répétitions de chaque simulation. On estime aussi que l'inclusion d'une grande variété de tests était très importante pour cette étude.

Les statistiques descriptives des pourcentages empiriques pour les 20 répétitions ont été traitées pour chaque niveau de rendement, méthode de calibrage et condition variable. Chaque pourcentage estimé moyen a été comparé au pourcentage réel correspondant pour déterminer le biais dans l'estimation empirique :

$$Biais = \frac{1}{n} \sum_{i=1}^n (\hat{\Delta}_{il} - \Delta_l) = \bar{\hat{\Delta}}_{il} - \Delta_l, \quad (1)$$

où $\bar{\hat{\Delta}}_{il} = \frac{1}{n} \sum_{i=1}^n \hat{\Delta}_{il}$,

Δ_l est la valeur réelle pour le niveau de rendement l ,

$\hat{\Delta}_{il}$ est la valeur estimée pour la répétition i du niveau de rendement l , et

$n = 20$ est le nombre de répétitions (Sinharay et Holland, 2007).

Si le biais est négatif, le pourcentage réel est sous-estimé; si le biais est positif, le pourcentage réel est surestimé.

La stabilité des pourcentages empiriques sur l'ensemble des répétitions a été évaluée à l'aide de l'erreur moyenne quadratique (EMQ) :

$$EMQ = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\Delta}_{il} - \Delta_l)^2} \quad (2)$$

Plus la valeur EMQ est petite, plus les valeurs estimées se rapprochent des valeurs réelles.

Dans le cadre de cette étude, les valeurs de biais et d'EMQ égales ou inférieures à 1 % étaient considérées comme étant négligeables. Les différences de biais produites par les deux méthodes devaient excéder 0,50 pour être considérées comme des différences significatives. Plusieurs programmes d'évaluation à grande échelle considèrent une variation de 1 % d'une année à l'autre comme étant négligeable.

Résultats

Les résultats du TPCL et de l'OSSLT sont présentés d'abord, suivis des résultats des sous-tests du cycle primaire, du cycle moyen et de 9^e année sélectionnés pour cette étude.

OSSLT et TPCL

OSSLT. Les résultats de l'OSSLT sont indiqués dans la partie supérieure du tableau 2. La structure des biais des quatre méthodes de calibrage est complexe. Par exemple, alors que la méthode CC a décelé la variation réelle de zéro, cette procédure n'a pas permis de déceler aussi bien les autres variations. Alors que les procédures PDCF, CCT et M/S ont décelé les variations des pourcentages d'élèves n'ayant pas réussi le test aussi bien sur l'ensemble des conditions de variabilité, les procédures CCT et M/S ont décelé les variations -0,1 et 0,3 beaucoup mieux que ne l'ont fait les procédures PDCF et CC. Des surestimations ont été observées dans les variations positives (c.-à-d. des augmentations du pourcentage d'élèves n'ayant pas réussi le test), le biais des procédures CC étant plus prononcé que celui des procédures PDCF, CCT et M/S (p. ex., 3,64 % contrairement à 0,72 %, 0,38 % et 0,37 % pour un gain réel de 0,3 sur l'échelle θ). Des sous-estimations ont été observées dans les variations négatives, le biais de la procédure de calibrage CC étant supérieur à celui des trois autres procédures, pour une variation réelle de -0,3 (-5,76 % comparativement à -0,99 %, -1,06 % et -1,05 %). Globalement, les méthodes CCT et M/S ont donné des résultats légèrement supérieurs à la méthode PDCF, et ces trois méthodes ont donné des résultats très supérieurs à la méthode CC. Exprimé en termes d'EMQ moyen sur l'ensemble des cinq conditions de gain, les quatre méthodes de calibrage se classent comme suit : CCT (0,45 %), M/S (0,49 %), PDCF (0,59 %) et CC (2,52 %).

TPCL. Comme le montre la partie inférieure du tableau 2, la méthode CC était encore une fois la pire pour déceler les variations dans les deux directions, surestimant largement les variations positives (2,48 % pour une variation de 0,1 et 4,29 % pour une variation de 0,3) et sous-estimant les variations négatives (-0,89 % pour une variation de -0,1 et -4,57 % pour un gain de -0,3). Les procédures PDCF, CCT et M/S ont aussi surestimé le gain de zéro. Fait intéressant, à l'exception du gain de -0,3, les procédures PDCF, CCT et M/S surestimaient le reste des gains, -0,1 et 0,1 étant les plus prononcés. Encore une fois, les méthodes PDCF, CCT et M/S se sont classées premières, avec une EMQ moyenne d'environ 1 % pour les cinq conditions de

gain. La procédure CC a présenté une EMQ moyenne substantiellement plus élevée de 2,66 %. Finalement, la magnitude de l'EMQ pour le TPCL était généralement plus élevée que le biais pour l'OSSLT.

Tableau 2

Résultats de calibrage pour l'OSSLT et le TPCL. Pourcentage d'élèves n'ayant pas réussi le test pour la distribution thêta de l'an 2

Méthode		Variation				
		-0,3	-0,1	0	0,1	0,3
OSSLT $N = 150\ 000$						
	Valeur réelle	23,76 %	17,72 %	15,04 %	12,70 %	8,65 %
CC	Biais	- 5,76 %	-1,69 %	-0,03 %	1,39 %	3,64 %
	EMQ	5,76 %	1,70 %	0,11 %	1,39 %	3,64 %
PICF	Biais	- 0,99 %	-0,58 %	-0,11 %	0,12 %	0,72 %
	EMQ	1,00 %	0,78 %	0,18 %	0,14 %	0,83 %
CCT	Biais	- 1,06 %	-0,37 %	-0,15 %	0,06 %	0,38 %
	EMQ	1,07 %	0,41 %	0,22 %	0,11 %	0,44 %
M/S	Biais	- 1,05 %	-0,34 %	-0,20 %	0,02 %	0,37 %
	EMQ	1,09 %	0,41 %	0,31 %	0,22 %	0,41 %
TPCL $N = 5\ 260$						
	Valeur réelle	24,43 %	17,62 %	14,33 %	11,43 %	7,64 %
CC	Biais	-0,17 %	0,64 %	0,90 %	1,34 %	0,95 %
	EMQ	4,59 %	0,97 %	0,76 %	2,52 %	4,48 %
PICF	Biais	-0,20 %	0,45 %	0,66 %	1,35 %	0,99 %
	EMQ	0,71 %	0,76 %	0,75 %	1,44 %	1,09 %
CCT	Biais	-0,17 %	0,64 %	0,90 %	1,34 %	0,95 %
	EMQ	0,72 %	1,12 %	1,05 %	1,46 %	1,05 %
M/S	Biais	0,00 %	0,00 %	0,01 %	0,01 %	0,01 %
	EMQ	0,76 %	1,09 %	0,99 %	1,12 %	0,84 %

Tests des cycles primaire et moyen et de 9^e année

Il y a quatre catégories de rendement pour le programme de tests des cycles primaire et moyen et de 9^e année. L'OQRE publie aussi une catégorie de rendement inférieur au niveau 1, mais très peu d'élèves figurent dans cette catégorie; elle a donc été combinée avec le niveau 1 dans le cadre de cette étude. Les pourcentages estimés et réels d'élèves dans chaque catégorie de rendement ont été comparés pour les procédures CC, PICAL, CCT et M/S. Il y a donc 20 estimés de biais (4 niveaux \times 5 conditions de gain) pour chaque test ou sous-test et procédure de calibrage. Les résultats pour les cinq sous-tests (écriture, cycle primaire de langue française et de langue anglaise; lecture, cycle moyen de langue anglaise; mathématiques, cycle moyen de langue française; et mathématiques, cours appliqué de 9^e année de langue anglaise) sont présentés ci-dessous.

Écriture, cycle primaire de langue anglaise. Les résultats en écriture, cycle primaire de langue anglaise, sont présentés dans le tableau 3. D'abord, le biais pour la condition de changement zéro était négligeable pour chaque niveau de rendement et pour chacune des méthodes de calibrage CC, PICAL, CCT et M/S. Le biais était aussi négligeable pour toutes les autres conditions de variation pour le niveau 1 de rendement pour les quatre méthodes de calibrage. Des différences ont commencé à figurer entre les méthodes de calibrage pour les autres conditions et niveaux de rendement. Le biais de la méthode CC, par exemple, était plus élevé que celui des autres procédures pour les niveaux 2, 3 et 4 de rendement pour des variations de -0,3, -0,1, 0,1 et 0,3. Par exemple, pour une variation de -0,3 au niveau 2 de rendement, le biais était de -8,50 % pour la procédure CC, de -2,47 % pour la procédure PICAL, de -2,37 % pour la procédure CCT et de -2,34 % pour la procédure M/S. Les résultats pour les méthodes PICAL, CCT et M/S étaient comparables : chaque méthode a essentiellement décelé le gain de 0,1 à tous les niveaux de rendement, les variations de -0,3 et de -0,1 au niveau 4 de rendement et la variation de 0,3 au niveau 3 de rendement. Il est intéressant de noter que ces trois procédures étaient biaisées positivement au niveau 3 de rendement et biaisées négativement au niveau 2 de rendement pour des gains de -0,3 et de -0,1, mais biaisées négativement au niveau 4 pour un gain de 0,3. Les résultats en écriture, cycle primaire de langue anglaise, révélaient le rendement inférieur de la méthode de calibrage CC, avec le biais et l'EMQ les plus élevés, et le rendement similaire des procédures PICAL, CCT et M/S pour les cinq conditions de variation.

Écriture, cycle primaire de langue française. Les résultats de calibrage en écriture, cycle primaire de langue française, sont présentés dans le tableau 4. Une comparaison entre ces résultats et ceux en écriture, cycle primaire de langue anglaise, révèle des différences entre les deux sous-tests. Par exemple, alors que le biais était encore minime pour toutes les procédures de calibrage au niveau 1 de rendement pour toutes les conditions de variation, le biais était relativement important pour chaque procédure de calibrage pour les niveaux 2, 3 et 4 de rendement. Bien que le sens du biais fût visible pour toutes les variations hormis 0,0, il ne suivait pas une tendance claire.

Tableau 3

Résultats de calibrage en écriture, cycle primaire de langue anglaise : pourcentage à chaque niveau de rendement pour les distributions théta de l'an 2 (N = 110 000)

Variation	Niveau	CC				PICF				CCT				M/S				
		4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	
-0,3	Val. réelle	2,87	54,13	42,15	0,85													
	Biais	2,22	6,50	-8,50	-0,23	0,33	2,07	-2,47	0,08	0,35	1,96	-2,37	0,06	0,36	1,93	-2,34	0,06	
	EMQ	2,23	6,51	8,51	0,23	0,34	2,10	2,50	0,09	0,38	2,01	2,42	0,08	0,40	1,99	2,42	0,08	
-0,1	Val. réelle	4,87	60,52	34,05	0,55													
	Biais	0,86	2,10	-2,94	-0,02	-0,01	0,94	-0,99	0,05	0,01	0,97	-1,02	0,05	0,07	1,00	-1,12	0,05	
	EMQ	0,87	2,10	2,95	0,03	0,11	0,98	1,03	0,06	0,10	1,00	1,07	0,06	0,16	1,05	1,18	0,06	
0	Val. réelle	6,19	63,25	30,14	0,42													
	Biais	-0,41	0,74	-0,42	0,09	-0,63	1,30	-0,76	0,09	-0,62	1,33	-0,80	0,08	-0,62	1,31	-0,78	0,09	
	EMQ	0,47	0,81	0,45	0,10	0,73	1,43	0,82	0,10	0,71	1,47	0,87	0,09	0,70	1,47	0,93	0,10	
0,1	Val. réelle	7,70	65,50	26,51	0,29													
	Biais	-1,40	-0,80	2,04	0,16	-0,52	0,34	0,11	0,07	-0,55	0,38	0,10	0,08	-0,54	0,35	0,11	0,08	
	EMQ	1,42	0,85	2,05	0,16	0,67	0,52	0,23	0,08	0,62	0,46	0,24	0,08	0,65	0,44	0,30	0,09	
0,3	Val. réelle	11,64	67,90	20,31	0,15													
	Biais	-4,66	-1,49	5,96	0,20	-1,31	0,48	0,76	0,07	-1,31	0,49	0,76	0,06	-1,24	0,35	0,82	0,07	
	EMQ	4,67	1,50	5,96	0,20	1,36	0,54	0,79	0,07	1,36	0,54	0,79	0,07	1,34	0,49	0,89	0,07	

Tableau 4

Résultats de calibrage en écriture, cycle primaire de langue française : pourcentage à chaque niveau de rendement pour les distributions thêta de l'an 2 (N = 6 000)

Variation	Niveau	CC				PICF				CCT				M/S				
		4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	
-0,3	Val. réelle	9,27	57,28	32,20	1,23													
	Biais	3,42	2,88	-5,83	-0,47	-0,03	0,01	-0,01	0,03	-0,08	-0,13	0,16	0,06	-0,18	-0,47	0,53	0,12	
	EMQ	3,45	2,97	5,87	0,48	0,54	1,24	1,12	0,28	0,72	1,30	1,25	0,30	1,36	1,64	2,03	0,37	
-0,1	Val. réelle	12,81	60,05	26,35	0,78													
	Biais	1,27	1,06	-2,19	-0,14	-0,56	0,22	0,31	0,03	-0,59	0,14	0,43	0,03	-0,46	-0,53	0,88	0,12	
	EMQ	1,34	1,25	2,26	0,16	0,92	1,10	0,87	0,14	1,00	1,12	1,07	0,15	1,50	1,25	1,89	0,23	
0	Val. réelle	16,49	60,45	22,63	0,42													
	Biais	-0,80	0,49	0,24	0,08	-0,65	0,35	0,23	0,08	-0,52	0,24	0,20	0,09	0,18	-0,23	-0,04	0,11	
	EMQ	0,89	0,78	0,59	0,12	1,09	1,09	0,84	0,14	1,19	1,15	0,99	0,14	2,07	1,40	1,71	0,16	
0,1	Val. réelle	18,71	60,57	20,38	0,33													
	Biais	-2,01	0,32	1,56	0,14	-0,10	-0,45	0,43	0,12	0,23	-0,57	0,23	0,12	1,44	-1,26	-0,31	0,13	
	EMQ	2,07	0,78	1,65	0,17	1,08	1,35	0,86	0,18	1,32	1,58	0,97	0,18	3,15	2,49	2,03	0,22	
0,3	Val. réelle	26,06	59,75	14,02	0,17													
	Biais	-7,38	2,13	5,05	0,21	-2,01	0,73	1,22	0,07	-1,58	0,42	1,10	0,08	-0,31	-0,56	0,80	0,08	
	EMQ	7,39	2,18	5,05	0,22	2,40	1,45	1,33	0,09	2,18	1,56	1,22	0,10	2,93	2,38	1,40	0,11	

Avec la procédure CC, des estimations négatives ont été observées pour des variations de -0,1 et -0,3 aux niveaux 1 et 2, alors que des estimations positives ont été décelées pour les niveaux 3 et 4. Dans les cas des variations positives, des estimations positives ont été observées pour les niveaux 1, 2 et 3, et des estimations négatives ont été décelées pour le niveau 4. Dans le cas des procédures PICF et CCT, à l'exception d'une variation de 0,0 au niveau 4, le biais était essentiellement positif. Dans le cas de la procédure M/S, des biais positifs et négatifs ont été observés pour les niveaux 1 et 2 et les niveaux 3 et 4, respectivement. La procédure CC présentait un biais et une EMQ légèrement plus importants dans toutes les conditions à l'exception d'un gain de 0,0. Avec toutes les procédures, plus la condition de gain était élevée par rapport à 0,0, plus grands étaient le biais et l'EMQ.

Pour l'ensemble des cinq conditions de variation, l'importance du biais et de l'EMQ avaient tendance à être supérieurs que ceux en écriture, cycle primaire de langue anglaise. Cela pourrait être attribué à la différence dans la taille des échantillons, mais le même résultat n'a pas été observé dans le cas du TPCL et de l'OSSLT (où la différence entre le nombre d'élèves de langue française et de langue anglaise était à peu près identique).

Mathématiques, cycle moyen de langue française. Les résultats de mathématiques, cycle moyen de langue française, sont présentés dans le tableau 5. Encore une fois la procédure CC a donné des résultats dont le biais et l'EMQ étaient plus importants aux niveaux 2 et 4 pour une variation de -0,1 et de -0,3 et aux niveaux 2 à 4 pour une variation de 0,3. Par comparaison, les méthodes PICF et CCT ont donné sensiblement les mêmes résultats, avec un biais négligeable et une EMQ bien moindre pour tous les niveaux de rendement et pour toutes les conditions de gain. Par comparaison avec les procédures PICF et CCT, la procédure M/S n'a pas donné d'aussi bons résultats. Dans plusieurs cas le biais était négligeable, mais l'EMQ était assez importante, ce qui est une indication que le comportement de la procédure M/S n'était pas aussi stable que celui des méthodes PICF et CCT. Par exemple, pour une variation de 0,3 au niveau 3, le biais était de -0,25, mais l'EMQ était de 3,24. Les estimations positives et négatives ont été décelées et aucun motif clair ne ressortait. L'ampleur du biais et de l'EMQ tendent à être moindres que ceux observés en écriture, cycle primaire de langue française et de langue

anglaise, dans la plupart des cas. Les trois tests ont montré que la procédure CC affichait le rendement le moins bon. Le rendement des méthodes PICF, CCT et M/S était très similaire. Mais la méthode M/S donnait une EMQ plus importante dans certains cas, indiquant un rendement moins stable que celui des deux autres procédures.

Tableau 5

Résultats de calibrage en mathématiques, cycle moyen de langue française : pourcentage à chaque niveau de rendement pour les distributions θ de l'an 2 ($N = 5\ 600$)

Variation	Niveau	CC				PICF				CCT				M/S				
		4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	
-0,3	Val. réelle	19,44	52,51	26,64	1,41													
	Biais	3,99	0,20	-3,52	-0,67	0,39	0,14	-0,18	-0,34	0,73	-0,03	-0,34	-0,35	0,98	-0,49	-0,22	-0,27	
	EMQ	4,01	0,66	3,57	0,69	0,70	0,79	0,93	0,40	0,90	0,90	0,91	0,42	1,29	1,98	1,22	0,49	
-0,1	Val. réelle	26,69	50,63	21,85	0,82													
	Biais	1,55	0,41	-1,63	-0,33	0,46	0,32	-0,55	-0,22	0,68	0,05	-0,52	-0,21	1,47	-0,93	-0,41	-0,13	
	EMQ	1,62	0,75	1,68	0,35	0,75	0,92	0,87	0,28	0,89	0,82	0,80	0,27	1,87	2,44	1,69	0,31	
0	Val. réelle	29,3	52,22	17,93	0,55													
	Biais	0,62	-0,68	0,08	-0,02	0,36	-0,50	0,15	0,00	0,72	-0,77	0,06	0,00	1,48	-1,78	0,26	0,04	
	EMQ	0,95	1,03	0,42	0,15	1,09	1,13	0,62	0,17	1,26	1,32	0,43	0,17	2,13	2,80	1,08	0,23	
0,1	Val. réelle	32,28	50,33	17,05	0,34													
	Biais	-0,64	0,37	0,15	0,12	0,06	0,30	-0,32	0,08	0,56	-0,13	-0,37	0,08	1,36	-0,96	-0,29	0,11	
	EMQ	0,81	0,83	0,63	0,16	0,77	1,22	0,67	0,14	1,01	1,27	0,77	0,14	1,91	2,35	1,22	0,18	
0,3	Val. réelle	42,01	45,44	12,35	0,2													
	Biais	-4,92	2,73	2,10	0,10	-1,00	0,75	0,22	0,03	-0,54	0,24	0,25	0,05	-0,14	-0,25	0,32	0,07	
	EMQ	4,93	2,88	2,23	0,13	1,12	0,94	0,64	0,09	0,64	0,61	0,48	0,14	1,51	3,24	1,62	0,22	

Lecture, cycle moyen de langue anglaise. Bien que la structure des biais ne fût pas identique, la taille des biais en lecture, cycle moyen de langue anglaise (tableau 6), avait tendance à être plus similaire aux valeurs observées pour l'OSSLT que pour les autres tests et sous-tests. Cette découverte est possiblement attribuable à la taille plus importante des échantillons (132 000).

Bien que la méthode CC ait décelé la variation réelle de 0,0, elle n'a pas aussi bien décelé les autres variations. Dans le cas des variations négatives, des sous-estimations ont été observées pour les niveaux 1 et 2 de rendement, et des surestimations pour les niveaux 3 et 4, le biais étant plus prononcé pour la condition de variation -0,3. Le schéma était inversé dans le cas des variations positives. Bien que la procédure CC ait produit de meilleurs résultats en lecture, cycle moyen de langue anglaise, que dans le cas des autres tests et sous-tests, par comparaison avec les trois autres procédures, un biais et une EMQ beaucoup plus importants ont été observés pour la variation de -0,3 à tous les niveaux de rendement, ainsi que pour la variation de 0,3 des niveaux 2 et 4. En revanche, les trois autres procédures ont connu un rendement aussi bon, la taille du biais et de l'EMQ s'établissant à moins de 1 %.

Mathématiques, cours appliqué de 9^e année de langue anglaise. Les résultats de ce sous-test étaient très similaires à ceux en lecture, cycle moyen de langue anglaise. Globalement, la méthode CC a donné les pires résultats (voir le tableau 7). Des surestimations et des sous-estimations substantielles ont été observées pour les deux conditions de variation extrêmes. Dans le cas d'une variation de -0,3, des surestimations ont été produites pour les niveaux 1 et 2, et des sous-estimations pour les niveaux 3 et 4. Le schéma était inversé dans le cas d'une variation de 0,3. Le biais le plus important était supérieur à 5 %. Les trois autres procédures ont donné de bien meilleurs résultats que la procédure CC, et aussi bien, les méthodes PICF et CCT donnant des biais et une EMQ légèrement plus importants que la méthode M/S au niveau 2 pour une variation de 0,3. Autrement, l'ampleur du biais était inférieure à 1 % pour toutes les conditions et niveaux de rendement.

Tableau 6

Résultats de calibrage en lecture, cycle moyen de langue anglaise : pourcentage à chaque niveau de rendement pour les distributions θ de l'an 2 (N = 132 000)

Variation	Niveau	CC				PICF				CCT				M/S				
		4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	
-0,3	Val. réelle	4,31	53,82	32,45	9,42													
	Biais	1,52	2,93	-2,19	-2,16	0,34	0,05	0,24	-0,53	0,36	-0,01	0,23	-0,47	0,37	-0,07	0,22	-0,43	
	EMQ	1,52	2,94	2,19	2,17	0,34	0,19	0,29	0,54	0,36	0,20	0,28	0,49	0,38	0,25	0,28	0,47	
-0,1	Val. réelle	6,76	59,39	27,08	6,77													
	Biais	0,55	0,40	-0,17	-0,78	0,12	-0,34	0,49	-0,27	0,13	-0,36	0,50	-0,26	0,16	-0,38	0,48	-0,26	
	EMQ	0,56	0,42	0,22	0,78	0,14	0,38	0,52	0,28	0,15	0,41	0,53	0,28	0,20	0,44	0,52	0,28	
0	Val. réelle	8,20	61,21	24,96	5,63													
	Biais	-0,12	-0,40	0,76	-0,30	-0,03	-0,55	0,73	-0,15	-0,01	-0,52	0,66	-0,13	0,00	-0,57	0,68	-0,10	
	EMQ	0,13	0,41	0,80	0,31	0,10	0,60	0,76	0,17	0,10	0,54	0,68	0,15	0,12	0,63	0,70	0,16	
0,1	Val. réelle	9,79	63,19	22,33	4,68													
	Biais	-0,97	-0,82	1,79	0,00	-0,11	-0,57	0,82	-0,14	-0,21	-0,46	0,81	-0,14	-0,19	-0,45	0,79	-0,14	
	EMQ	0,98	0,83	1,80	0,05	0,33	0,67	0,82	0,15	0,24	0,51	0,82	0,16	0,23	0,54	0,81	0,18	
0,3	Val. réelle	14,25	65,20	17,47	3,08													
	Biais	-3,86	0,18	3,36	0,32	-0,90	0,08	0,94	-0,12	-0,91	0,10	0,93	-0,12	-0,89	0,05	0,94	-0,11	
	EMQ	3,86	0,22	3,36	0,33	0,91	0,20	0,94	0,12	0,93	0,21	0,94	0,13	0,91	0,23	0,95	0,12	

Tableau 7

Résultats de calibrage en mathématiques, cours appliqué de 9^e année de langue anglaise : pourcentage à chaque niveau de rendement pour les distributions thêta de l'an 2 (N = 34 441)

Variation	Niveau	CC				PICF				CCT				M/S				
		4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	
-0,3	Val. réelle	2,64	27,09	43,10	27,17													
	Biais	0,96	5,01	-0,43	-5,53	0,13	0,55	0,22	-0,90	0,08	0,35	0,27	-0,70	0,18	0,13	-0,27	-0,04	
	EMQ	1,00	5,26	0,58	5,77	0,15	0,60	0,47	0,99	0,14	0,53	0,61	0,90	0,27	0,64	0,86	1,01	
-0,1	Val. réelle	4,17	33,86	41,82	20,15													
	Biais	0,39	1,07	-0,32	-1,14	0,01	-0,07	0,39	-0,33	0,02	0,00	0,30	-0,33	0,19	0,08	-0,12	-0,14	
	EMQ	0,42	1,16	0,41	1,23	0,17	0,29	0,52	0,46	0,24	0,38	0,59	0,46	0,37	0,69	0,74	0,77	
0	Val. réelle	5,23	37,07	40,43	17,27													
	Biais	-0,22	-0,52	0,52	0,22	-0,14	-0,25	0,38	0,00	-0,12	-0,25	0,32	0,05	0,08	-0,19	-0,05	0,17	
	EMQ	0,41	0,62	0,83	0,29	0,19	0,35	0,46	0,23	0,28	0,50	0,51	0,49	0,42	0,73	0,68	0,79	
0,1	Val. réelle	6,44	40,47	38,74	14,35													
	Biais	-0,65	-2,02	1,09	1,58	-0,35	-0,55	0,58	0,33	-0,44	-0,41	0,71	0,14	-0,37	-0,39	0,61	0,16	
	EMQ	0,77	2,12	1,32	1,62	0,37	0,66	0,69	0,41	0,48	0,55	0,81	0,32	0,49	0,73	0,86	0,54	
0,3	Val. réelle	9,54	46,30	34,23	9,93													
	Biais	-2,15	-4,39	3,38	3,16	-0,73	-0,90	1,14	0,49	-0,62	-0,91	1,01	0,52	-0,40	-0,74	0,72	0,41	
	EMQ	2,23	4,49	3,49	3,23	0,78	0,99	1,16	0,54	0,72	1,01	1,07	0,58	0,63	0,99	0,94	0,60	

Discussion

La présente étude examinait le rendement des procédures de calibrage CC, PICF, CCT et M/S pour déceler les variations dans le rendement des élèves entre deux années successives à l'aide du biais et de l'EMQ. Cinq conditions de variation (-0,3, -0,1, 0, 0,1 et 0,3) sur l'échelle θ ont été considérées. Sept des tests ou sous-tests de l'OQRE ont été choisis, à la complexité et à la taille des échantillons variable : le TPCL et l'OSSLT; écriture, cycle primaire de langue française et de langue anglaise; mathématiques, cycle moyen de langue française; lecture, cycle moyen de langue anglaise; mathématiques, cours appliqué de 9^e année de langue anglaise.

Les résultats ont révélé que le biais et l'EMQ suivaient une structure complexe sur l'ensemble des cinq conditions de variation et des sept tests. La procédure CC produisait un biais substantiellement plus important que les trois autres procédures de calibrage, sauf dans les cas de la condition de gain zéro. L'ampleur du biais était un important 8,50 % en valeur absolue. Avec toutes les procédures de calibrage, un biais plus important a été observé pour les deux conditions de variation extrêmes ($\pm 0,3$), mais pas toujours pour un test donné. Cet effet était plus prononcé avec la méthode CC. Les méthodes PICF, CCT et M/S produisaient pour leur part des biais beaucoup moins importants. Le biais maximal était d'environ 2 % en valeur absolue. Des biais moindres ont été observés pour les tests de langue anglaise que pour les tests de langue française, ce qui est sans doute imputable à la différence dans la taille des échantillons. Dans les cas du TPCL et de l'OSSLT, des sous-estimations ont été observées du côté des gains négatifs, et des surestimations du côté des gains positifs. Les sous-estimations et les surestimations variaient pour les autres tests comportant des catégories de rendement multiples.

Le biais et l'EMQ substantiels révélés par la procédure CC par rapport au changement du rendement des groupes indiquaient l'échec général de cette procédure pour le traitement des groupes non équivalents. Cette constatation va dans le sens des constatations observées par des études antérieures impliquant des groupes équivalents et non équivalents (Petersen et autres, 1983; Hills et autres, 1988; Hu et autres, 2008).

Parmi les trois procédures de calibration et de calibrage distinct, la procédure M/S semblait être la méthode la plus sensible à la taille des échantillons. De plus grandes

variations des biais entre les répétitions ont été observées avec la procédure M/S lorsque les tests et sous-tests de langue française étaient utilisés. La procédure M/S utilisait la moyenne et l'écart-type des paramètres pour déterminer les coefficients de transformation linéaire. Selon Yen et Fitzpatrick (2006), cette méthode comporte deux limitations majeures : a) les paramètres des items sont traités comme des entités indépendantes et les observations aberrantes peuvent considérablement modifier les résultats; et b) si un ensemble d'estimations des paramètres d'items est basé sur un échantillon plus petit de participants qu'un autre, les variances de l'erreur des deux ensembles de paramètres des items seront différentes. Dans ce cas, si les variances des deux ensembles d'estimations sont définies comme étant égales (comme avec la procédure M/S), un biais systématique des résultats aura lieu par un « rétrécissement » de l'ensemble des résultats avec une variance de l'erreur plus élevée pour correspondre à l'ensemble avec une variance de l'erreur moindre. C'est-à-dire que les variances des estimations seront équivalentes, mais les variances des valeurs réelles ne le seront pas (Yen et Fitzpatrick, 2006). Dans le cas des tests de langue française, les échantillons étaient petits du fait de la structure de la matrice utilisée pour faire la mise à l'essai des items durant l'an 1. Par exemple, en mathématiques, cycle moyen de langue française, la matrice comportait 10 blocs, pour un échantillon constitué d'environ 560 élèves. Cela pourrait avoir donné lieu à des estimations des paramètres des items, ce qui aurait un effet défavorable sur la moyenne et l'écart-type.

Contrairement à la procédure M/S, la procédure de calibrage CCT a obtenu une liaison des coefficients en minimisant la différence entre les courbes caractéristiques du test pour les deux administrations. Selon Yen et Fitzpatrick (2006), l'un des avantages les plus importants de cette méthode est qu'elle minimise les différences dans les scores attendus plutôt que dans les scores ou paramètres observés. Ce faisant, les paramètres ou thêtas sont alignés, réduisant par conséquent la variance entre les répétitions.

Dans le cas des tests aux niveaux de rendement multiples, le biais le plus important a été observé en écriture, cycle primaire de langue française et de langue anglaise. Ces deux sous-tests comportaient seulement 14 items et 38 scores maximaux possibles. Ces résultats semblent suggérer qu'un test court peut influencer, à des degrés variés, le rendement des méthodes de calibrage, la méthode CC étant substantiellement

influencée, suivie par la procédure M/S. Le sous-test en écriture, cycle primaire de langue française, avait de bonnes chances de produire l'effet confondant d'un test court et d'un petit échantillon sur les méthodes de calibrage (voir le tableau 4).

Bien que la procédure PICF ait donné de meilleurs résultats que les autres méthodes de calibrage dans cette étude, il faut être prudent dans son application. Yen et Fitzpatrick (2006) ont souligné que les caractéristiques importantes de la procédure PICF est qu'en maintenant fixes les paramètres des items communs, cela oblige le programme d'estimation des paramètres à accommoder les valeurs des paramètres des items communs (c.-à-d. à les traiter comme s'il s'agissait de valeurs réelles). Par conséquent, il est essentiel, dans l'application de cette méthode, que les paramètres des items communs correspondent au nouvel échantillon de données. Autrement, la méthode PICF peut avoir un impact adverse sur l'estimation des paramètres des items non communs. Lorsque la méthode PICF est utilisée pour établir des relations, le déplacement des paramètres des items mis à l'essai au test provincial devrait être examiné avec soin. Un item avec un déplacement de vaste amplitude devrait être rejeté du processus d'établissement de relations.

Étant donné la complexité des structures de test, matières et niveaux de rendement multiples des tests utilisés dans cette étude, un résumé de plus haut niveau des résultats était nécessaire afin d'obtenir une impression générale du rendement des quatre procédures de calibrage. Le tableau 8 contient la valeur moyenne absolue des biais et de l'EMQ pour les cinq conditions de changement pour le TPCL et l'OSSLT pour chaque méthode de calibrage.

Tableau 8

Biais et EMQ moyens pour les tests avec des catégories réussite et non réussite

Test	Gain	CC		PICF		CCT		M/S	
		Biais	EMQ	Biais	EMQ	Biais	EMQ	Biais	EMQ
OSSLT <i>N</i> = 150 000 <i>n</i> = 51 Moyenne au-dessus des conditions	-0,3	-5,76	5,76	-0,99	1,00	-1,06	1,07	-1,05	1,09
	-0,1	-1,69	1,70	-0,58	0,78	-0,37	0,41	-0,34	0,41
	0	-0,03	0,11	-0,11	0,18	-0,15	0,22	-0,20	0,31
	0,1	1,39	1,39	0,12	0,14	0,06	0,11	0,02	0,22
	0,3	3,64	3,64	0,72	0,83	0,38	0,44	0,37	0,41
		2,50	2,52	0,50	0,59	0,40	0,45	0,39	0,49
TPCL <i>N</i> = 5 260 <i>n</i> = 51 Moyenne au-dessus des conditions	-0,3	-4,57	4,59	-0,20	0,71	-0,17	0,72	-0,14	0,76
	-0,1	-0,89	0,97	0,45	0,76	0,64	1,12	0,63	1,09
	0	0,73	0,76	0,66	0,75	0,90	1,05	0,73	0,99
	0,1	2,48	2,52	1,35	1,44	1,34	1,46	0,95	1,12
	0,3	4,29	4,48	0,99	1,09	0,95	1,05	0,71	0,84
		2,59	2,66	0,73	0,95	0,80	1,08	0,63	0,96

Remarque : *N* = taille de l'échantillon; *n* = nombre d'items

Le tableau 9 contient la valeur moyenne absolue des biais et de l'EMQ pour les cinq conditions de changement et quatre niveaux de rendement pour chaque sous-test du cycle primaire ou moyen et pour chaque méthode de calibrage.

Tableau 9

Biais et EMQ moyens des tests avec des niveaux de rendement multiples

Test	Gain	CC		PICF		CCT		M/S	
		Biais	EMQ	Biais	EMQ	Biais	EMQ	Biais	EMQ
Math., moyen, français $N = 5\ 600$ $n = 35$	-0,3	2,09	2,23	0,26	0,71	0,36	0,78	0,49	1,24
	-0,1	0,98	1,10	0,39	0,70	0,37	0,69	0,73	1,58
	0	0,35	0,64	0,25	0,75	0,39	0,80	0,89	1,56
	0,1	0,32	0,61	0,19	0,70	0,29	0,80	0,68	1,41
	0,3	2,46	2,54	0,50	0,70	0,27	0,47	0,20	1,65
Moyenne au-dessus des conditions		1,24	1,42	0,32	0,71	0,34	0,71	0,60	1,49
Lecture, moyen, anglais $N = 132\ 000$ $n = 36$	-0,3	2,20	2,20	0,29	0,34	0,27	0,33	0,27	0,35
	-0,1	0,48	0,49	0,31	0,33	0,31	0,34	0,32	0,36
	0	0,39	0,41	0,37	0,41	0,33	0,37	0,34	0,40
	0,1	0,90	0,92	0,41	0,49	0,40	0,43	0,39	0,44
	0,3	1,93	1,94	0,51	0,55	0,52	0,55	0,50	0,55
Moyenne au-dessus des conditions		1,18	1,19	0,38	0,42	0,37	0,41	0,36	0,42
Écriture, moyen, anglais $N = 110\ 000$ $n = 14$	-0,3	4,36	4,37	1,23	1,26	1,18	1,22	1,17	1,22
	-0,1	1,48	1,49	0,50	0,55	0,51	0,56	0,56	0,61
	0	0,42	0,46	0,69	0,77	0,71	0,79	0,70	0,80
	0,1	1,10	1,12	0,26	0,38	0,28	0,35	0,27	0,37
	0,3	3,08	3,08	0,65	0,69	0,66	0,69	0,62	0,70
Moyenne au-dessus des conditions		2,09	2,10	0,67	0,73	0,67	0,72	0,66	0,74
Écriture, moyen, français $N = 6\ 000$ $n = 14$	-0,3	3,15	3,19	0,02	0,79	0,11	0,89	0,31	1,28
	-0,1	1,17	1,25	0,28	0,76	0,30	0,84	0,49	1,18
	0	0,40	0,60	0,33	0,79	0,26	0,87	0,13	1,31
	0,1	1,01	1,17	0,27	0,87	0,29	1,01	0,77	1,94
	0,3	3,69	3,71	1,01	1,32	0,79	1,26	0,43	1,70
Moyenne au-dessus des conditions		1,88	1,98	0,38	0,91	0,35	0,97	0,43	1,48
Math., 9 ^e année, anglais $N = 34\ 441$ $n = 31$	-0,3	2,98	3,15	0,45	0,55	0,35	0,55	0,15	0,70
	-0,1	0,73	0,81	0,20	0,36	0,16	0,42	0,13	0,64
	0	0,37	0,54	0,19	0,31	0,19	0,45	0,12	0,65
	0,1	1,34	1,46	0,45	0,53	0,42	0,54	0,38	0,65
	0,3	3,27	3,36	0,81	0,87	0,76	0,85	0,57	0,79
Moyenne au-dessus des conditions		1,74	1,86	0,42	0,52	0,38	0,56	0,27	0,69

Remarque : N = taille de l'échantillon; n = nombre d'items

Tel qu'indiqué dans les tableaux 8 et 9, l'EMQ moyenne est substantiellement plus importante dans le cas de la procédure CC que pour les trois autres procédures de calibrage, pour l'ensemble des tests. La méthode PICF a produit les estimations les plus exactes et précises, avec une EMQ moyenne de moins de 1 % supérieure pour les niveaux de rendement et les conditions de gain, indépendamment du test, de la durée du test et de la taille de l'échantillon. La procédure CCT a donné d'aussi bons résultats que la PICF dans tous les cas sauf pour le TPCL et l'OSSLT, pour lesquels l'EMQ moyenne était légèrement supérieure (1,08 %). La méthode M/S donnait de moins bons résultats avec les tests impliquant des petits échantillons (écriture, cycle primaire de langue française, et mathématiques, cycle moyen de langue française), avec une EMQ moyenne substantiellement plus élevée. Toutes les procédures de calibrage semblaient dans une certaine mesure être plus influencées par la taille de l'échantillon que par la durée du test; mais la procédure M/S semblait être beaucoup plus sensible à la taille de l'échantillon. Une vue d'ensemble de l'étude a montré que les méthodes PICF et CCT ont donné d'aussi bons résultats, suivies de la méthode M/S. La procédure CC s'est classée dernière.

Les constatations de cette étude diffèrent des constatations d'autres chercheurs. Par exemple, Keller et autres (2004) ont trouvé que, entre les procédures de calibrage CC, CCT, PICF et M/S, la procédure M/S était celle qui décelait le mieux les variations. La procédure CC avait tendance à sous-estimer les variations négatives du rendement des élèves et à surestimer les variations positives. Skorupski et autres (2003) ont découvert que les deux procédures CC et PICF sous-estimaient les gains, alors que Keller (2002) a trouvé que ces deux procédures avaient tendance à surestimer les gains. Dans la présente étude, aucune structure claire n'a été observée sur l'ensemble des catégories de rendement entre et sur l'ensemble des sous-tests et des tests. Bien que des EMQ moyennes aient été signalées, elles ont été observées sur l'ensemble des catégories et ne devraient pas masquer la présentation des résultats par catégorie de rendement. Comme mentionné dans la section des résultats, les schémas du rendement des quatre procédures de calibrage sur l'ensemble des conditions de gain et catégories de rendement étaient quelque peu complexes.

Plusieurs facteurs ont sans doute contribué à ce que les constatations de cette étude diffèrent de celles des études antérieures. D'abord, alors que dans les autres

recherches le θ réel et la variation étaient inconnus ou encore qu'elles étudiaient uniquement la récupération du θ réel et des paramètres, cette étude a examiné le pourcentage pour chaque niveau de rendement, classifié pour les autres procédures de calibrage. Deuxièmement, la plupart des chercheurs précédents utilisaient des données réelles ou des données simulées basées sur une distribution normale des aptitudes, alors que les données dans le cadre de cette étude ont été simulées à partir des distributions de capacité réelles des élèves, dont la forme ne suivait pas une courbe normale, et des paramètres des items du test provincial reproduisant sept tests réels différents. Troisièmement, les données des études antérieures étaient habituellement simulées pour imiter des tests constitués seulement d'items à notation dichotomique, alors que les données simulées dans cette étude reproduisaient des items à notation dichotomique et polytomique. Quatrièmement, étant donné le concept de la matrice utilisée par l'OQRE dans la conception des tests, le nombre d'items utilisés pour lier les tests des deux années était beaucoup plus grand, et les paramètres de liaison estimés dans l'échantillon du test provincial (plutôt que de plus petits échantillons d'items mis à l'essai) étaient plus stables. Tous ces éléments ont conduit à un effet de relation beaucoup plus fort dans cette étude. Finalement, cette étude couvrait un large éventail de matières qui utilisaient des échantillons et des populations de taille variable (de langue française et de langue anglaise), alors que les études antérieures se concentraient souvent sur une seule matière.

En bref, la contribution principale de la présente étude est qu'elle a fourni des renseignements beaucoup plus complets en ce qui a trait au rendement des quatre méthodes de calibrage que les études antérieures.

Conclusion

Les résultats obtenus dans la présente étude sur cinq conditions de variation et sept tests suggèrent qu'il est raisonnable que l'OQRE continue d'utiliser ses procédures actuelles (PICF) pour calibrer les tests du cycle primaire, du cycle moyen et de 9^e année ainsi que le TPCL et l'OSSLT. Aucune des trois autres procédures étudiées dans cette étude n'a invariablement donné de meilleurs résultats que ceux produits par la méthode PICF.

Dans une étude future, lors du calibrage à l'aide de PARSCALE, des distributions de pondération de mise à jour et des cycles EM multiples (MWU-MEM) utilisés dans la méthode PICF (Kim, 2006) devraient être examinés. Selon Kim, la méthode d'estimation MWU-MEM permettait le mieux de déceler les paramètres entre les autres méthodes d'estimation. De plus, plusieurs facteurs pourraient aussi être inclus dans des études futures. On pourrait s'intéresser à l'impact de données manquantes, une situation fréquente dans les évaluations à grande échelle. On pourrait examiner l'impact des observations aberrantes dans l'ensemble des items communs (Hu et autres, 2008) sur les méthodes de calibrage. On pourrait étudier plus explicitement l'effet de différentes proportions d'items polytomiques et leur interaction avec des échantillons de tailles multiples. Une approche en la matière pourrait examiner les populations de langue française et de langue anglaise pour tous les tests. Finalement, on devrait augmenter le nombre de répétition afin de déterminer si l'uniformité se maintient entre les répétitions.

Références

- Angoff, W. H. 1984, *Scales, norms, and equivalent scores*, Princeton, NJ, Educational Testing Service.
- Bishop, N. S., S. Sharairi, D. Swift, P. Lei et D. Domaleski. 2006, *Comparing achievement growth over multiple years using different IRT scale transformation procedures*, article présenté à l'assemblée annuelle du National Council on Measurement in Education, San Francisco, avril 2006.
- Childs, R. A. et W. H. Chen. 1999, Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for the two polytomous IRT models, *Applied Psychological Measurement*, vol. 23, n° 4, p. 371-379.
- Cohen, A. S. et Seock-Ho Chen. 1998, An investigation of linking methods under the graded response model, *Applied Psychological Measurement*, vol. 22, n° 2, p. 116-130.
- Hanson, B. A. et A. A. Beguin. 2002, Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design, *Applied Psychological Measurement*, vol. 26, n° 1, p. 3-24.

- Hills, J. R., R. G. Subhiyah et T. M. Hirsch. 1988, Equating minimum-competency tests: Comparison of methods, *Journal of Educational Measurement*, vol. 25, n° 3, p. 221-231.
- Hu, H., W. T. Roger 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.
- Keller, L. A., W. P. Skorupski, H. Swaminathan et M. G. Jodoin. 2004, An evaluation of Item Response Theory equating procedures for capturing changes in examinee distributions with mixed-format tests, article présenté à l'assemblée annuelle du National Council on Measurement in Education, avril 2004.
- Kim, S. 2006, A comparative study of IRT fixed parameter calibration methods. *Journal of Education Measurement*, vol. 43, p. 355-381.
- Kim, S. et A. S. Cohen. 1998, A comparison of linking and concurrent calibration under item response theory, *Applied Psychological Measurement*, vol. 22, p. 131-143.
- Kolen, M. J. et R. L. Brennan. 2004, *Test equating, scaling, and linking: Methods and Practices*, New York, Springer.
- Lee, W. et J. Ban. 2010, A comparison of IRT linking procedures, *Applied Measurement in Education*, vol. 23, p. 23-48.
- Lehman, R. S. et D. E. Bailey. 1968, *Digital computing: Fortran IV and its applications in behavioral science*, New York, John Wiley.
- Li, Y. H., W. D. Griffith et H. P. Tam. 1997, *Equating multiple tests via an IRT linking design utilizing a single set of anchor items with fixed common item parameters during calibration process*, article présenté à l'assemblée annuelle de la Psychometric Society, Knoxville, TN.
- Li, Y. H., R. W. Lizzitz et Y-N. Yang. 1999, *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items*, article présenté à l'assemblée annuelle du National Council on Measurement in Education, Montréal, QC.
- Linn, R. L., M. V. Levine, C. N. Hastings et J. L. Wardrop. 1980, *An investigation of item bias in a test of reading comprehension* (Technical Report n° 163), Urbana, IL, Center for the Study of Reading, University of Illinois.

- Lord, F. M. 1980, *Applications of item response theory to practical testing problems*, New Jersey, Hillsdale.
- Loyd, B. H. et H. D. Hoover. 1980, Vertical equating using the Rasch model, *Journal of Educational Measurement*, vol. 17, p. 179-193.
- Marco, G. L. 1977, Item characteristic curve solutions to three intractable testing problems, *Journal of Educational Measurement*, vol. 14, p. 139-160.
- Petersen, N. S., L. L. Cook et M. L. Stocking. 1983, IRT versus conventional equating methods: A comparative study of scale stability, *Journal of Educational Statistics*, vol. 8, p. 137-156.
- Prowker A. et G. Camilli. 2006, *Long-term stability of two common equating methods: what no child left behind could mean for equating practices*, article présenté à l'assemblée annuelle du National Council on Measurement in Education, San Francisco, avril 2006.
- Psychometric Society. 1979, Publication policy regarding Monte Carlo studies, *Psychometrika*, vol. 44, p. 133-134.
- Sinharay, S. et P. W. Holland. 2007, Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, vol. 44, n° 3, p. 249-275.
- Skorupski, W. P., M. G. Jodoin, L. A. Keller et H. Swaminathan. 2003, *An evaluation of equating procedures for capturing growth*, article présenté à l'assemblée du National Council on Measurement in Education, Chicago, IL.
- Stocking, M. L. et F. M. Lord. 1983, Developing a common metric in item response theory, *Applied Psychological Measurement*, vol. 7, p. 201-210.
- Wingersky, M. S., L. L. Cook et D. R. Eignor. 1987, *Specifying the characteristics of linking items used for item response theory item calibration*, (ETS research Report 87-24), Princeton, NJ, Educational Testing Service.
- Xie, Y. 2006, *Study of psychometric issues for EQAO assessments*, article non publié, Office de la qualité et de la responsabilité en éducation.
- Yen, M. W. et A. R. Fitzpatrick. 2006, Item Response Theory. *Educational Measurement* (4^e éd.).



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