



Office de
la qualité et
de la responsabilité
en éducation

OQRE Série d'études de recherche

• NUMÉRO 3 • AOÛT 1999

MISSION DE L'OQRE

L'OQRE assurera une plus grande responsabilité et contribuera à améliorer la qualité de l'éducation en Ontario. Cela s'effectuera au moyen d'évaluations basées sur des données objectives, fidèles et pertinentes et de la publication en temps voulu de ces données et de la formulation de recommandations aux fins d'amélioration du système.

La présente étude a été écrite sous contrat pour l'Office de la qualité et de la responsabilité en éducation (OQRE). Les opinions exprimées sont celles des auteurs et ne sont pas nécessairement celles de l'OQRE.

© Imprimeur de la Reine pour l'Ontario 1999

Tous droits réservés. Aucune partie du présent document ne peut être reproduite, emmagasinée dans un système de recherche documentaire ou diffusée par moyen électronique, mécanique, sous forme de photocopie, d'enregistrement ou autre, sans l'accord préalable de l'Office de la qualité et de la responsabilité en éducation par l'intermédiaire de la chef des communications et du marketing.

Office de qualité et de la responsabilité en éducation
2, rue Carlton, bureau 1200
Toronto (Ontario) M5B 2M9

Téléphone : 1-888-327-7377

Site Web : www.eqao.com

ISBN 0-7778-9136-0

Avant-propos

L'OQRE croit que les évaluations à grande échelle peuvent contribuer à apporter des changements constructifs au système d'éducation si elles amènent les éducatrices et les éducateurs, les parents et les élèves à réfléchir à ce qui se passe dans la classe et à en discuter.

Le mandat de l'OQRE est fondé sur deux engagements clés : la responsabilité et l'amélioration. Nous rendons compte aux parents, aux éducatrices et aux éducateurs et au public sur le rendement des élèves et la qualité de l'éducation dans le système scolaire de l'Ontario financé par les deniers publics. Nous veillons aussi à ce que ces renseignements servent de catalyseur pour améliorer l'apprentissage et l'enseignement.

L'une des façons de favoriser l'amélioration est de soutenir la recherche sur les facteurs qui ont une influence sur le rendement des élèves et la qualité de l'éducation. C'est ce que fait l'OQRE avec sa *Série d'études de recherche* qui fournit des analyses et des renseignements sur un vaste éventail de questions en rapport avec l'éducation.

Cette série a deux objectifs importants. Le premier est d'augmenter nos connaissances sur ce qui donne de bons résultats dans les classes, dans les écoles et dans les conseils scolaires. Le second est de combler les lacunes entre ce que nous savons et ce que nous faisons. En donnant aux éducatrices et aux éducateurs les connaissances, les occasions et la motivation nécessaires pour adapter et perfectionner leurs stratégies et leurs méthodes sur la base de données exactes et fidèles, la recherche peut avoir un effet positif et durable sur l'apprentissage des élèves. L'OQRE s'intéresse à toutes les formes de recherches systématiques qui favorisent le changement constructif et l'amélioration.

J'ai le plaisir de présenter la présente étude de la *Série d'études de recherche de l'OQRE*.

La directrice générale,

Joan M. Green
Août 1999

TABLE DES MATIÈRES

Résumé	1
Cadre méthodologique	2
Théorie de la généralisabilité.....	2
Classement et erreur de classement	5
Application de l'analyse du classement et des erreurs de classement aux évaluations de l'OQRE	7
Collecte et sommaire des données	7
Le modèle statistique	8
Résultats	10
Discussion et conclusion.....	12
Références.....	13
Les auteurs	14

Perspectives psychométriques pour l'OQRE : Théorie et applications de la généralisabilité

Richard Wolfe

*Institut d'études pédagogiques de l'Ontario
Université de Toronto*

David Wiley

Université Northwestern

Ross Traub

*Institut d'études pédagogiques de l'Ontario
Université de Toronto*

La présente communication examine les perspectives psychométriques à partir desquelles définir et évaluer la généralisabilité des mesures et des rapports produits par l'Office de la qualité et de la responsabilité en éducation (OQRE). L'exactitude des mesures est exprimée en termes des composantes de la variabilité de la mesure : tâches, individus, correctrices et correcteurs et interactions connexes. Deux perspectives psychométriques sont utilisées pour analyser la généralisabilité des mesures des évaluations de l'OQRE. La première est l'analyse de la généralisabilité des sources de variation dans la notation des tests (c.-à-d. les composantes de la variance vraie et de la variance erronée); la seconde fait appel à des modèles de catégorisation vraie du rendement des élèves et à un cadre de probabilité pour définir le classement correct et le classement incorrect. Les données issues des évaluations de 3^e année et de 9^e année administrées en 1997-1998 par l'OQRE en français et en anglais sont examinées par une méthode d'analyse de structure latente qui estime les taux de classement correct et de classement incorrect des correctrices et des correcteurs. La probabilité de classement correct pour ces évaluations est estimée s'établir entre 70 pour 100 et 80 pour 100. Ce résultat est raisonnablement élevé et comparable à ce qui est observé dans d'autres évaluations à grande échelle du rendement.

Cadre méthodologique

Théorie de la généralisabilité

La méthode la plus complète utilisée par les spécialistes des mesures éducatives pour analyser les erreurs dans la notation des exercices de rendement effectués par les élèves est basée sur la théorie de la généralisabilité. Cette méthode représente une extension et une expansion des méthodes utilisées en psychométrie dans le cadre de la théorie traditionnelle de fidélité, comme les coefficients de fidélité fournis par la répétition des tests et les échanges de correctrices et de correcteurs. Elle s'appuie fortement sur la notion d'échantillonnage des tâches d'évaluation et des conditions de mesure à partir de populations plus importantes soumises à ces tests et dans ces conditions.

La théorie de la généralisabilité considère la fidélité de la mesure du point de vue de la généralisation de ces échantillons aux populations pertinentes. Les populations sont caractérisées par des paramètres qui représentent la variabilité dans la performance au test mesurée. Ces paramètres caractérisent deux types de variabilité :

1. la variabilité entre les individus dans les caractéristiques de la performance que l'on retrouve couramment dans les différentes tâches et dans les différentes conditions, c'est-à-dire les variations générales de la performance individuelle qui transcendent les circonstances particulières de la mesure;
2. la variation particulière de la performance individuelle qui ne reflète que des tâches et des conditions particulières.

Le premier type représente les stabilités de mesure désirées tandis que le deuxième représente les erreurs de mesure. Les études et les analyses de généralisabilité évaluent la magnitude de ces aspects de la variation de la performance en procédant à des analyses de la variance. Ces analyses décomposent la variation observée en composantes de variance représentant les personnes, les tâches et d'autres conditions de la mesure. Les magnitudes relatives et

absolues des différentes composantes de la variance permettent de prédire l'exactitude qui serait obtenue avec des systèmes de mesure différents utilisant des conditions et des nombres différents des tâches étudiées. La théorie de la généralisabilité a été formulée pour la première fois par Cronbach, Gleser, Nanda et Rajaratnam (1972).

Dans les évaluations de performance, les erreurs de mesure sont le plus souvent associées aux tâches et à la notation. C'est-à-dire qu'elles apparaissent dans les variations de la notation d'un individu pour des tâches similaires et dans la variation des notations pour une même tâche. Plus précisément, ces erreurs sont évaluées relativement à la variation :

- de la notation de différents exercices considérés comme interchangeables aux fins de l'évaluation;
- des notations assignées à une seule tâche également considérées comme interchangeables aux fins de l'évaluation.

Plus ces variations sont importantes, plus grande est l'erreur de mesure et plus faible l'exactitude. Dans la plupart des évaluations de performance, les exercices notés dans un contexte de notation particulier pour une matière particulière et une classe particulière sont habituellement considérés comme interchangeables, particulièrement lorsque les correctrices et correcteurs sont formés de la même façon de manière à atteindre le même niveau de compétence.

Les données pour une analyse de généralisabilité de ces types d'erreurs sont souvent obtenues à partir d'une situation où chaque membre d'un groupe de personnes soumises à l'évaluation exécute chaque tâche en une suite constante et où chaque réponse est notée par le même ensemble de correctrices et de correcteurs. Une analyse des données produites dans ce type de structure, lorsqu'il y a plus d'une tâche dans la série, produit les estimations des composantes de variabilité illustrées au Tableau 1.

Cadre méthodologique

Théorie de la généralisabilité

Tous les types d'erreurs énumérés dans le Tableau 1 contribuent à l'erreur de mesure. Ceci est vrai quel que soit le nombre de tâches effectuées par la répondante ou le répondant dans une série de tâches interchangeableables ou le nombre de notations interchangeableables assignées à une tâche. Cependant, suivant la conception de l'évaluation, p. ex., le nombre exact de tâches effectuées par une répondante ou un répondant ou le nombre de correctrices et de correcteurs, les effets des différentes composantes varieront. Si seulement une tâche d'un type donné est effectuée par les répondantes et répondants ou si seulement une notation est obtenue pour une tâche, certaines de ces composantes ne seront pas estimables.

On peut faire remonter la plupart de ces types d'erreurs à deux sources générales : les différences dans la mesurabilité des tâches et les différences entre les correctrices et correcteurs dans l'évaluation de performance. Comme les tâches sont sélectionnées, et donc fixées, avant que la notation ait lieu, les différences entre les tâches peuvent être conçues comme antérieures aux effets de la notation et indépendantes de ces effets. Par ailleurs, les

effets de la notation ne peuvent se produire que dans le contexte des tâches assignées. De ce point de vue, il est logique de définir les composantes d'erreurs associées aux différences entre les tâches (effets des tâches) comme reliées aux tâches mais non à la notation et de définir les effets de la notation comme reliées à la notation, qu'elles soient ou non reliées aux tâches. C'est-à-dire que les résultats d'une tâche particulière différeront habituellement de façon aléatoire de ceux qui seraient obtenus après une évaluation de performance à un large éventail de tâches interchangeableables. La notation, en revanche, dépend de chaque tâche : les correctrices et les correcteurs ne notent que des tâches précises. De ce point de vue, les composantes de la variance d'erreur du Tableau 1 seraient classées comme indiqué dans le Tableau 2.

L'erreur due aux différences entre les tâches à accomplir, à l'exclusion de la notation, s'exprime en fidélité comme :

$$r_t = \sigma_p^2 / (\sigma_p^2 + \sigma_t^2 + \sigma_{pt}^2).$$

L'erreur due à la notation, une fois les tâches fixées (de sorte que la notation vraie d'une tâche particulière est la notation

TABLEAU 1 Types de variation

Variance de la composante	Type de variation
Variation vraie	
σ_p^2	Variation générale « vraie » entre les personnes dont on évalue le rendement
Variation aléatoire	
σ_s^2	Variation entre les correctrices et correcteurs en indulgence moyenne
σ_t^2	Variation entre les tâches en difficulté moyenne
σ_{ps}^2	Variation entre les correctrices et les correcteurs et les personnes en indulgence envers les individus
σ_{pt}^2	Variation entre les tâches et les personnes en difficulté de la tâche pour les individus
σ_{st}^2	Variation entre les correctrices et les correcteurs et les tâches en difficulté de la tâche notée
σ_{pst}^2	Variation entre les correctrices et les correcteurs, les tâches et les personnes en difficulté de la tâche et en indulgence envers les individus

Cadre méthodologique

Théorie de la généralisabilité

observée une fois l'erreur ajoutée) s'exprime en fidélité comme :

$$r_s = (\sigma_p^2 + \sigma_t^2 + \sigma_{pt}^2) / (\sigma_p^2 + \sigma_t^2 + \sigma_{pt}^2 + \sigma_s^2 + \sigma_{ps}^2 + \sigma_{st}^2 + \sigma_{pst}^2).$$

Comme la fidélité totale, reflétant les deux sources d'erreurs, est :

$$r_{tot} = \sigma_p^2 / (\sigma_p^2 + \sigma_t^2 + \sigma_{pt}^2 + \sigma_s^2 + \sigma_{ps}^2 + \sigma_{st}^2 + \sigma_{pst}^2),$$

on a donc :

$$r_{tot} = r_t r_s.$$

C'est-à-dire que la fidélité totale est le produit de la fidélité liée à l'erreur due aux tâches et de la fidélité liée à l'erreur due à la notation.

Bien que toutes les composantes ci-dessus soient présentes dans le processus d'évaluation de l'OQRE, la conception actuelle de l'évaluation impose des limites quant à celles qui peuvent être évaluées. Des études de recherche pourraient être faites pour dégager les composantes qui peuvent effectivement être intégrées à la principale étape de recueil des données et incluses dans l'analyse subséquente.

TABLEAU 2 Composantes d'erreurs pour les tâches et la notation

Source	Composantes d'erreurs
Tâche	$\sigma_t^2, \sigma_{pt}^2$
Notation	$\sigma_s^2, \sigma_{ps}^2, \sigma_{st}^2, \sigma_{pst}^2$

Cadre méthodologique

Classement et erreur de classement

L'analyse des erreurs de mesure suivant la méthode de généralisabilité traite les réponses analysées comme s'il était possible de leur assigner des valeurs sur une échelle continue. Une autre méthode assigne les réponses à une ou plusieurs catégories discrètes et ordonnées. De ce point de vue, l'erreur de mesure vient de l'assignation de notations à des catégories incorrectes. La façon naturelle de quantifier ces erreurs de classement est en termes de probabilité de classement correct ou d'erreur de classement, c'est-à-dire de probabilités comme : $P(i|j)$ = la probabilité d'une réponse étant assignée à la catégorie i alors que j est en fait la catégorie correcte.

L'erreur de classement peut être due à chacun des types d'erreurs énumérés dans les Tableaux 1 et 2. Plus particulièrement, si un individu était vraiment dans la catégorie j , une tâche t pourrait avoir la probabilité $P(i|j;t)$ de classer cette personne dans la catégorie i . Cette probabilité d'erreur de classement liée à une tâche émanerait des mêmes sources que celles qui sont résumées par σ_t^2 et σ_{pt}^2 . Une moyenne de ces probabilités d'erreurs de classification de toutes les tâches, à savoir la moyenne $\{P(i|j;t)\}$, résume la relation générale parmi les tâches entre les catégories vraies et observées.

Les relations d'erreurs de classement peuvent être présentées sous forme de matrice d'erreur de classement, comme illustré dans le Tableau 3.

Ces matrices convertissent les valeurs vraies en valeurs observées. C'est-à-dire que si y_p est un vecteur représentant le classement d'un individu p dans une catégorie observée (c.-à-d. que le vecteur = 1 pour la catégorie observée et 0 pour chaque autre catégorie), et si x_p est le vecteur vrai correspondant, alors $E_m(y_p) = Mx_p$, où E_m est l'attente pour tout le processus d'erreur de classement (performance de la tâche ou notation). De même, $E_p[E_m(y_p)] = E_p[Mx_p] = ME_p[x_p]$, où E_p est l'attente pour l'ensemble de la population. On remarque que $E_p[x_p] = \pi$, les proportions vraies des catégories dans la population, et $E_p[E_m(y_p)] = p$, les proportions observées des catégories dans la population. C'est ainsi que $p = M\pi$. C'est-à-dire que les matrices d'erreurs de classement transforment les distributions vraies pour l'ensemble des catégories de groupes de répondantes et de répondants en distributions observées pour l'ensemble des catégories de tâches ou de notations. Si nous indexons ces valeurs pour une tâche particulière, soit t , alors $p_t = M_t\pi$.

Si l'erreur de classement due à la notation de la tâche t est M_{ts} , alors $p_{ts} = M_{ts}p_t = M_{ts}M_t\pi$, si bien que la matrice d'erreur de classement de la performance vraie après sélection et la notation des tâches est le produit de l'erreur de classement liée à la matrice de sélection de la tâche et de la matrice d'erreur liée au classement de la notation. Il faut également noter que $\Sigma_t(p_{ts}) = \Sigma_t(M_{ts}M_t\pi) = [\Sigma_t(M_{ts}M_t)]\pi$, si bien que la

TABLEAU 3 Formulation de la matrice d'erreurs de classement **M**

Catégorie observée	Catégorie vraie			
	1	2	3	4
1	P(1 1)	P(1 2)	P(1 3)	P(1 4)
2	P(2 1)	P(2 2)	P(2 3)	P(2 4)
3	P(3 1)	P(3 2)	P(3 3)	P(3 4)
4	P(4 1)	P(4 2)	P(4 3)	P(4 4)
Total	1,00	1,00	1,00	1,00

Cadre méthodologique

Classement et erreur de classement

somme et la moyenne des erreurs de classement pour l'ensemble des tâches représentent les erreurs de classement pour la somme et la moyenne des proportions observées.

Chaque entrée diagonale dans la matrice d'erreur de classement représente la probabilité de classement correct, c'est-à-dire que la tâche d'un individu dans une catégorie vraie particulière sera notée dans cette catégorie. Ces entrées diagonales représentent l'exactitude ou la fidélité de la mesure. (Il faut noter que les indices de fidélité classiques, tels que décrits ci-dessus pour la théorie de la généralisabilité, sont des rapports de corrélation et non des probabilités. Nous trouvons les probabilités d'analyse de structure latente plus interprétables, surtout quand des échelles catégoriques sont utilisées comme dans les évaluations de l'OQRE. Bien que l'étendue des probabilités et des rapports de corrélation soit la même, zéro à un, l'échelle ne l'est pas, si bien que les deux ne sont pas directement comparables. Nous présentons une façon de transformer les indices classiques de fidélité en probabilités plus loin dans cet article.)

Application de l'analyse des classements et des erreurs de classement aux évaluations de l'OQRE

Dans les évaluations de 3^e année et de 9^e année de l'OQRE, nous devons considérer chaque élève comme ayant effectué une seule tâche, à savoir la procédure d'évaluation tout entière. Compte tenu des priorités et des objectifs généraux de l'OQRE, il n'était pas faisable de répéter les tâches dans aucun des tests, c'est-à-dire de séparer des réponses d'élèves pouvant être notées séparément et être considérées comme interchangeables. Cela signifie que les composantes de variance qui dépendent de la tâche (variance de la tâche, tâche par interaction des élèves, tâche par interaction des correctrices et correcteurs et tâche par élève par interaction des correctrices et correcteurs) ne pouvaient pas être évaluées. Nous devons donc nous concentrer sur les variances et les coefficients de fidélité dus aux correctrices et correcteurs.¹

COLLECTE ET SOMMAIRE DES DONNÉES

Les données de la présente analyse dérivent d'un système de répétition de la correction utilisé par l'OQRE. Une fois qu'entre 500 à 1 000 cahiers ont été corrigés, on procède à une analyse rapide de la distribution provisoire des notes. Un échantillon stratifié de cahiers est choisi avec une distribution relativement uniforme sur l'étendue des notes. La note générale ou moyenne est utilisée dans cette sélection. Pour l'évaluation provinciale en anglais, qui est administrée à plus de 100 000 élèves, environ 200 cahiers sont choisis aux fins de calibrage. Pour l'évaluation en français et lors des mises à l'essai (c.-à.-d. l'évaluation de 9^e année) quelque 100 cahiers sont choisis aux fins de calibrage. Une fois sélectionnés, les cahiers sont réinsérés aussi

souvent que possible dans le flot des cahiers à corriger sans indication visible qu'il s'agit d'une répétition. Comme il y a un grand nombre de cahiers à corriger et un grand nombre de correctrices et de correcteurs, il est très rare qu'un cahier de calibrage soit noté deux fois par la même personne. Suivant la rapidité de la correction et le nombre de cahiers et de correctrices et correcteurs, chaque cahier de calibrage est corrigé à nouveau de trois à six fois.

Les données des corrections répétées sont résumées, séparément pour chaque évaluation et pour chaque note qui doit être analysée, en tabulation en croix. On trouvera un exemple de tabulation en croix au Tableau 4. Dans la forme A de l'exemple, un cahier choisi au hasard corrigé par deux correctrices ou correcteurs sélectionnés au hasard a 21,15 pour 100 de chances d'être noté deux fois au niveau 2 et 12,61 pour 100 de chances d'être noté une fois au niveau 2 et une fois au niveau 3. Les figures correspondantes pour la forme B sont 18,33 pour 100 et 13,61 pour 100.

La distribution générale des notes peut être déterminée à partir de la tabulation en croix en ajoutant les chiffres en diagonale à la moitié de la somme des chiffres hors diagonale (puisque la tabulation en croix enregistre les notes en groupes de deux). Les distributions sont 14,98 pour 100, 34,95 pour 100 et 16,43 pour 100 pour les niveaux 1, 2, 3 et 4 pour la forme A; et 19,45 pour 100, 34,58 pour 100, 32,08 pour 100 et 13,90 pour 100 pour la forme B. Bien que le même nombre de cahiers aient été choisis aux fins de réinsertion dans chacun des quatre niveaux, après la première correction des cahiers, on assiste à une certaine régression

¹Dans toutes les évaluations de l'OQRE, il y a des notes séparées pour les principales compétences et domaines, et pour les sujets ou matières. Elles fournissent des évaluations différenciées des élèves mais ne sont pas remplaçables ou interchangeables aux fins de détermination de la fidélité. Dans plusieurs évaluations et dans les mises à l'essai, il y a eu des notations « analytiques » où des notes partielles ont été appliquées aux différentes questions et aux différents segments du travail des élèves. Celles-ci corrêlaient bien avec les notes globales. Les questions à réponse choisie et à réponse courte soumises à des échantillons d'élèves corrêlaient également bien avec les notes globales. Quoique ceci augure bien de la fidélité des notes globales (les corrêlations ne pourraient pas exister sans fidélité au moins à ce niveau), il n'est pas possible de quantifier la fidélité à partir de ces connexions parce que les méthodes pour ce type d'analyse n'ont pas encore été développées.

Application de l'analyse des classements et des erreurs de classement aux évaluations de l'OQRE

vers le milieu dans les notations moyennes, c'est-à-dire que les pourcentages de niveaux 1 et 4 sont diminués alors que ceux des niveaux 2 et 3 sont augmentés.

LE MODÈLE STATISTIQUE

Les sommaires des données dépendent de la conjonction entre les données de distribution des notes et les taux d'accord entre les correctrices et correcteurs. Pour désagréger ces effets, nous appliquons une analyse de structure latente comme illustré à la figure 1. (Pour l'explication initiale de l'analyse de structure latente, voir Lazarsfeld et Henry, 1968.)

Pour comprendre l'analyse de structure latente, nous devons considérer qu'aux fins d'application réelle, nous ne savons pas quelle est la notation « vraie » d'un cahier. Nous pourrions essayer de la dériver d'un consensus d'experts mais, même dans ce cas, nous avons peu de chances d'obtenir un accord à 100 pour 100. En l'absence d'un classement « vrai », nous ne pouvons pas calculer directement la proportion de classement correct et incorrect. En revanche, les proportions de l'accord, comme indiqué au Tableau 4, dépend à la

fois de la distribution des notes vraies hypothétiques (les catégories latentes de cahiers) et des probabilités de classement des correctrices et correcteurs (voir le Tableau 3).

L'analyse de la catégorie latente propose un modèle simple pour le processus de notation. D'abord, les cahiers ont un classement vrai, bien qu'inconnu, latent (ou vrai) et commun à toutes les correctrices et les correcteurs. Ensuite, quand une correctrice ou un correcteur assigne une note à un cahier, il s'agit d'un classement correct avec une certaine probabilité. Avec la probabilité complémentaire, la correctrice ou le correcteur relève ou rabaisse la note d'un niveau. Au niveau 1, (le niveau le plus bas), on considère qu'il ne peut y avoir qu'une note correcte ou relevée, notes rabaisée et correcte étant semblables et au niveau 4 (le niveau le plus élevé), il ne peut y avoir qu'une note correcte ou rabaisée, notes relevée et correcte étant semblables. Aux autres niveaux, il y a des probabilités constantes mais différenciées de relèvement et de rabaissement.

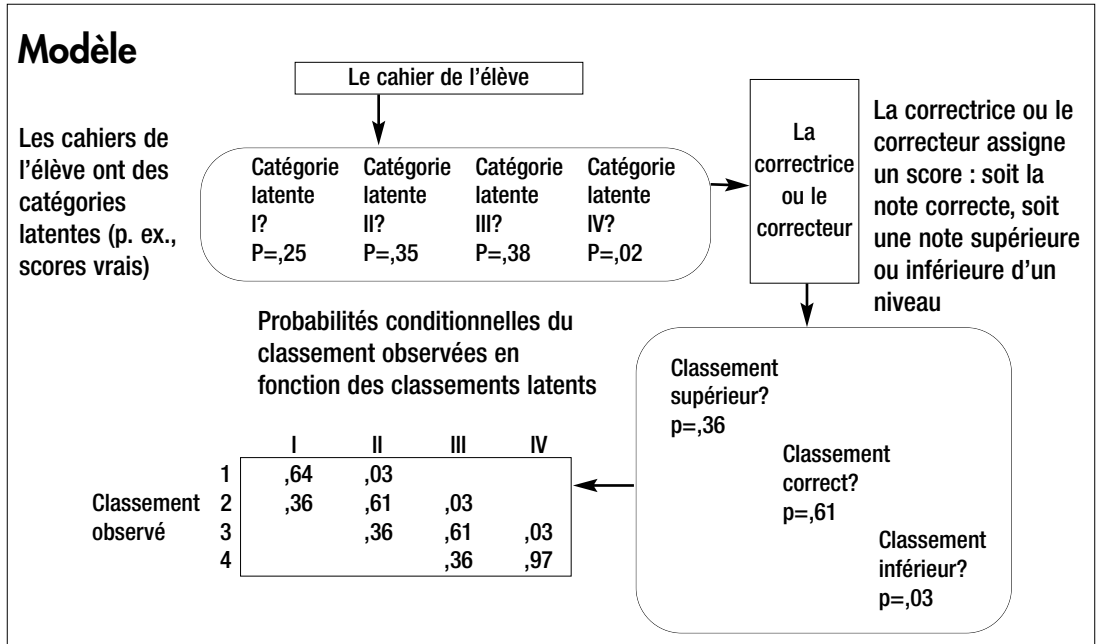
TABLEAU 4 Exemple de résumés de données de classement et d'erreurs de classement

Forme A Niveau	Niveau 1	Niveau 2	Niveau 3	Niveau 4
1	7,70			
2	14,01	21,15		
3	0,56	12,61	18,21	
4	0	0,98	17,60	7,14
Forme B Niveau	Niveau 1	Niveau 2	Niveau 3	Niveau 4
1	10,28			
2	15,83	18,33		
3	2,50	13,61	17,08	
4	0	3,06	13,89	5,42

Remarque : Dans chaque tabulation en croix, la somme des entrées est 100 et elles donnent une estimation de la distribution des notes comme si un cahier choisi au hasard était noté par deux correctrices ou correcteurs choisis au hasard.

Application de l'analyse des classements et des erreurs de classement aux évaluations de l'OQRE

FIGURE 1 Modèle de catégorie latente pour estimer la fidélité de la notation comme erreurs de classement



Données attendues

		I	II	III	IV
1		7,7			
2		14,0	21,2		
3		,6	12,6	18,2	
4			1,0	17,6	7,1

Tabulation en croix des moyennes théoriques définies par les paramètres (la somme est 100 pour 100)

Données observées

		I	II	III	IV
1		10,1			
2		12,4	16,4		
3		,6	16,7	19,0	
4			,7	17,0	7,1

Tabulation en croix des moyennes observées, calculées à partir des données calibrées (la somme est 100 pour 100)

Remarque : Dans l'analyse statistique, nous essayons de trouver les paramètres d'un modèle avec lequel les données observées ont le plus de chance de coïncider, c'est-à-dire qui soit le plus proche des données attendues, compte tenu des paramètres.

Application de l'analyse des classements et des erreurs de classement aux évaluations de l'OQRE

Les proportions des cahiers dans les catégories vraies sont des paramètres du modèle à estimer. Les probabilités de notation correcte, de relèvement et de rabaissement de la note représentent trois autres paramètres. Cet ensemble de paramètres est estimé à partir du sommaire des données par la méthode de la probabilité maximale. Cela signifie qu'à l'issue d'une itération numérique, des valeurs de paramètres sont trouvées qui rendent les proportions sommaires observées aussi probables que possible.

De retour à la Figure 1, nous voyons qu'un diagramme du modèle est présenté dans la rangée du haut avec les paramètres pour les catégories latentes et le classement ou l'erreur de classement tel qu'estimé à partir des données dans la rangée du bas. La rangée du milieu montre la valeur attendue des données compte tenu des paramètres et l'on peut voir qu'il s'agit d'une proche approximation des données.

Ce modèle particulier de catégorie latente est un cas spécial du modèle présenté dans la section précédente. La simplification est nécessaire à cause des tailles relativement

faibles des échantillons. Le modèle semble être adéquat pour correspondre aux données, en partie parce qu'il y a très peu de cas où les correctrices et correcteurs diffèrent de plus d'un niveau dans leur jugement.

RÉSULTATS

Les résultats de l'application du modèle de catégorie latente aux évaluations de 1997–1998 sont présentés dans les Tableaux 5 et 6. Comme on l'a fait remarquer plus tôt, on peut considérer les probabilités de classement vrai comme une bonne mesure sommaire de la fidélité, c'est-à-dire de la probabilité qu'une ou un élève sera noté « correctement » par une correctrice ou un correcteur. Ces probabilités s'établissent entre 70 et 80 pour 100. Ceci est raisonnablement élevé et comparable aux niveaux que l'on trouve habituellement dans d'autres évaluations à grande échelle (voir, par exemple, Dunbar, Koretz et Hoover, 1991). De plus, ces taux de classement sont comparables à ce qui serait obtenu avec les tests traditionnels à question à réponse choisie (voir la section suivante).

TABLEAU 5 Analyse de catégorie latente pour les niveaux assignés en 1997–1998 en français

Niveau	LC0	LC1	LC2	LC3	LC4	bas	haut	vrai
Français 3 ^e année Lecture	0,10	0,05	0,30	0,55	0,00	0,13	0,12	0,74
Français 3 ^e année Écriture	0,10	0,06	0,33	0,53	0,00	0,09	0,16	0,74
Français 3 ^e année Math	0,12	0,05	0,29	0,46	0,08	0,12	0,08	0,79
Français 9 ^e année Math	0,07	0,13	0,40	0,38	0,02	0,25	0,05	0,69

TABLEAU 6 Analyse de catégorie latente pour les niveaux assignés en 1997–1998 en anglais

Niveau	LC0	LC1	LC2	LC3	LC4	bas	haut	vrai
Anglais 3 ^e année Lecture	0,10	0,08	0,35	0,42	0,05	0,11	0,07	0,81
Anglais 3 ^e année Écriture	0,09	0,09	0,46	0,36	0,00	0,06	0,20	0,74
Anglais 3 ^e année Math	0,09	0,11	0,41	0,38	0,00	0,08	0,18	0,73
Anglais 9 ^e année Math	0,12	0,15	0,44	0,28	0,01	0,02	0,26	0,72

Application de l'analyse des classements et des erreurs de classement aux évaluations de l'OQRE

Taux de classement dans les tests conventionnels à réponse choisie

Pour des raisons de contenu et de validité corrélative, les évaluations de l'OQRE utilisent essentiellement des tâches à exécuter qui peuvent être mesurées de façon globale. Il est intéressant de comparer les coefficients de fidélité qui pourraient être obtenus si l'OQRE utilisait à la place des évaluations de performance des tests traditionnels. Les élèves pourraient être classés en quatre niveaux (pour correspondre aux niveaux de rendement de l'OQRE) en divisant la distribution des notes à un test gradué conventionnel comme un test à réponse choisie, suivant des seuils de note appropriés. En utilisant la théorie classique de la fidélité des tests, on peut estimer les probabilités de classement correct qui seraient obtenues par cette méthode.

On propose à titre d'exemple qu'un test gradué ait donné lieu à une distribution de niveaux finals similaires aux niveaux de l'évaluation de 3^e année administrée par l'OQRE en 1997–1998, avec 10 pour 100 pour le niveau 1, 40 pour 100 pour le niveau 2, 45 pour 100 pour le niveau 3 et 5 pour 100 pour le niveau 4. (Ces chiffres représentent une distribution approximative moyenne à travers les matières et les langues de l'OQRE).

Considérons des tests gradués avec des coefficients de fidélité classiques (ρ) de 0,70, 0,80 et 0,90, correspondant à une exactitude moyenne, élevée et très élevée pour les tests éducationnels commerciaux en mathématiques et en français. Pour chaque test, nous pourrions calculer les probabilités d'assignation de note au niveau correct

pour les élèves qui sont au centre des niveaux 1 à 4, comme indiqué au Tableau 7.

Par exemple, au Tableau 7, le chiffre 73 pour le niveau 2, avec une fidélité de 0,70 indique qu'il y a 73 pour 100 de chance qu'une ou un élève au vrai centre du niveau 2 de ce test (ce qui signifie, dans ce cas, au 30^e percentile de la distribution vraie des notes) sera classé au niveau 2. (Une analyse plus détaillée indique que les élèves ont 6 pour 100 de chance d'être classés au niveau 1, 21 pour 100 de chance d'être classés au niveau 3 et pratiquement aucune chance d'être classés au niveau 4.)

Nous voyons que les probabilités de classement correct sont plus élevées pour les niveaux 2 et 3 parce qu'il s'agit de larges catégories et que nous considérons seulement les élèves qui se trouvent au centre. Pour les élèves plus proches du seuil des niveaux, les probabilités de classement correct seraient plus faibles. Les chances de classement correct sont aussi substantiellement plus faibles dans les niveaux extérieurs, 1 et 4, parce que ces niveaux sont plus étroits et parce qu'il y a une régression générale vers le centre.

La leçon à tirer est que les tests conventionnels à réponse choisie, même s'ils ont une haute fidélité, produiront des erreurs de classement substantielles lorsqu'ils seront utilisés pour assigner des niveaux aux élèves. En outre, les tests à réponse choisie sont jugés présenter des défauts considérables en matière de validité et de partialité. Par ailleurs, les taux de classement correct par une seule correctrice ou un seul correcteur tels que produits dans les évaluations de performance de l'OQRE sont d'une magnitude comparable.

TABLEAU 7 Probabilité de classement correct des élèves au centre du niveau pour les tests avec des coefficients de fidélité différents

Fidélité	Niveau 1	Niveau 2	Niveau 3	Niveau 4
0,70	57	73	80	50
0,80	66	82	88	60
0,90	81	94	96	75

Discussion et conclusion

On fait parfois appel aux coefficients de fidélités inter-tâches et inter-correcteurs pour ce type de test et de correction. Cependant, une analyse de la fidélité inter-tâches n'est pas faisable ici parce qu'il n'y a qu'une tâche par élève. Aux fins de la présente analyse, nous nous sommes donc attachés à la fidélité inter-correcteurs. Un échantillon de cahiers avec une distribution uniforme sur l'échelle de notation a été réinséré entre trois et six fois dans l'ensemble des cahiers à corriger. Un cahier n'a pratiquement jamais été corrigé par la même correctrice ou le même correcteur. Ce qui était calculé dans le modèle présenté ici est l'exactitude de la note qui était assignée à chaque cahier comme indiqué par la probabilité de classement correct. (Ce calcul est clairement différent de la probabilité que deux correctrices ou correcteurs coïncident dans leur notation. Bien que l'on puisse calculer la probabilité que deux correctrices ou correcteurs coïncident, dans ce modèle cette probabilité inclurait des cas où les correctrices et les correcteurs coïncident dans leurs erreurs.)

Il est important de souligner que les distributions de structure latente ne sont pas les mêmes que la distribution originale des notes. On peut présumer qu'elles donnent une meilleure image de la distribution des classements « vrais » de ces élèves parce que les distributions observées sont déformées par les erreurs de classement. Il faut remarquer que l'échantillon de cahiers utilisé dans l'exercice de calibration a été choisi pour représenter une distribution plus ou moins uniforme sur l'échelle de notation, si bien que les proportions de structure latente peuvent ne pas refléter la population tout entière. En tous cas, on peut voir que le nombre de cas dans la catégorie latente 4 est très faible.

Ces analyses nous permettent de commencer une évaluation de la suffisance de la fidélité des mesures et des rapports de l'OQRE. Généralement parlant, les probabilités de classement correct sont

raisonnablement élevées. Ces taux sont comparables à ce que l'on trouve dans d'autres évaluations à grande échelle et ils sont équivalents aux taux de classement qui seraient obtenus dans des tests plus traditionnels à réponse choisie, qui sont généralement considérés comme présentant un degré très élevé de fidélité.

Références

Cronbach, L. J., Gleser, G. C., Nanda, H. et Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York : Wiley, 1972.

Dunbar, S.B., Koretz, D.M. et Hoover, H.D. Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4(4), 1991, p. 289-303

Lazarsfeld, P. F. et Henry, N. W. *Latent structure analysis*. Boston : Houghton Mifflin, 1968.

Les auteurs

Richard Wolfe est professeur agrégé dans le programme de mesure et d'évaluation du Département du curriculum, de l'enseignement et de l'apprentissage de l'Institut d'études pédagogiques de l'Ontario de l'Université de Toronto. Son champ de recherches comprend la conception et l'analyse des évaluations à grande échelle.

David Wiley est professeur, en congé, dans le programme des sciences de l'apprentissage, et ancien doyen de l'École d'éducation et de politique sociale de l'Université Northwestern. Ses recherches portent sur les statistiques et sur la théorie du curriculum, de l'apprentissage et de l'évaluation.

Ross Traub est professeur dans le programme de mesure et d'évaluation du Département du curriculum, de l'enseignement et de l'apprentissage de l'Institut d'études pédagogiques de l'Ontario de l'Université de Toronto. Ses recherches portent sur la psychométrie théorique et appliquée.