Mesure et évaluation en éducation

Un processus itératif pour réduire l'impact de réponses aberrantes sur l'identification de patrons de réponses inappropriés

David Magis, Sébastien Béland and Gilles Raîche

Volume 36, Number 2, 2013

URI: https://id.erudit.org/iderudit/1024416ar
DOI: https://doi.org/10.7202/1024416ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Magis, D., Béland, S. & Raîche, G. (2013). Un processus itératif pour réduire l'impact de réponses aberrantes sur l'identification de patrons de réponses inappropriés. *Mesure et évaluation en éducation*, *36*(2), 87–110. https://doi.org/10.7202/1024416ar

Article abstract

The presence of response disturbances or aberrant response patterns is often assessed by the computation of person fit indexes. These indicate whether a pattern, as a whole, can be considered as abnormal with respect to the test characteristics. However, most often they require item parameters to be pre-calibrated and this calibration is performed upon the same data set. The presence of response disturbances might therefore impact the item calibration process and subsequently, the identification of person misfit. This paper presents a straightforward iterative process to reduce the risk of unfair item calibration due to the presence of response disturbances. The idea consists in iteratively removing the patterns flagged as aberrant from the item calibration process, and re-computing the person fit indexes with newly calibrated parameters. The process is illustrated by analyzing the data from an English skill assessment questionnaire in Quebec. applying the process to these data reveals: an increase in the number of response patterns flagged as aberrant, a somewhat weak impact on estimated item parameters, with a limited number of iterations required to reach convergence of the iterative process.

Un processus itératif pour réduire l'impact de réponses aberrantes sur l'identification de patrons de réponses inappropriés

David Magis
Université de Liège, Belgique
et Université du Québec à Montréal, Canada

Sébastien Béland
Université de Sherbrooke, Canada

Gilles Raîche
Université du Québec à Montréal, Canada

Mots clés : théorie de la réponse aux items, calibration des items, réponses aberrantes, patrons inappropriés, indice l_z, purification

La présence de réponses aberrantes est habituellement détectée par l'utilisation d'indices d'ajustement permettant de déterminer si le patron de réponses est inapproprié par rapport aux caractéristiques du test. Cette approche nécessite cependant une préestimation des paramètres d'items qui est souvent réalisée sur le même ensemble de données. La présence de réponses aberrantes pourrait donc influencer le processus de calibration et la détection de patrons inappropriés. Cet article présente un processus itératif pour réduire le risque d'une calibration biaisée causée par la présence de réponses aberrantes. La démarche consiste à retirer successivement les patrons identifiés comme inappropriés du processus de calibration des items. Ce processus est illustré en analysant les données d'un test de classement en anglais langue seconde (TCALS-II) au Québec. L'application du processus itératif aux données met en évidence une augmentation du nombre de patrons de réponses détectés comme inappropriés, présentant un impact relativement faible sur les paramètres d'items estimés et un nombre restreint d'itérations nécessaires pour obtenir une convergence du processus itératif.

Key words: item response theory, item calibration, aberrant responses, person fit; l_z index, purification

The presence of response disturbances or aberrant response patterns is often assessed by the computation of person fit indexes. These indicate whether a pattern, as a whole, can be considered as abnormal with respect to the test charac-

teristics. However, most often they require item parameters to be pre-calibrated and this calibration is performed upon the same data set. The presence of response disturbances might therefore impact the item calibration process and subsequently, the identification of person misfit. This paper presents a straightforward iterative process to reduce the risk of unfair item calibration due to the presence of response disturbances. The idea consists in iteratively removing the patterns flagged as aberrant from the item calibration process, and re-computing the person fit indexes with newly calibrated parameters. The process is illustrated by analyzing the data from an English skill assessment questionnaire in Quebec. Applying the process to these data reveals: an increase in the number of response patterns flagged as aberrant, a somewhat weak impact on estimated item parameters, with a limited number of iterations required to reach convergence of the iterative process.

PALAVRAS-CHAVE: teoria de resposta ao item, calibração do item, respostas aberrantes, padrões inapropriados, índice l_z, purificação

A presença de respostas aberrantes é habitualmente detetada pela utilização de índices de ajustamento que permitem determinar se o padrão de respostas é inapropriado em relação às características do teste. No entanto, esta abordagem necessita de uma pré-estimativa dos parâmetros de itens que é realizada sobre o mesmo conjunto de dados. A presença de respostas aberrantes poderia, assim, influenciar o processo de calibração e a deteção de padrões inapropriados. Este artigo apresenta um processo iterativo para reduzir o risco de uma calibração enviesada devido à presença de respostas aberrantes. O procedimento consiste em retirar sucessivamente os padrões identificados como inapropriados do processo de calibração dos itens. Este processo é ilustrado pela análise dos dados de um teste de avaliação de competências de Inglês como segunda língua (TCALS-II) no Quebeque. A aplicação do processo iterativo aos dados coloca em evidência um aumento do número de padrões de respostas detetadas como inapropriadas, apresentando um impacto relativamente fraco sobre os parâmetros dos itens estimados e um número restrito de iterações necessárias para obter uma convergência do processo iterativo.

Note des auteurs – Les auteurs tiennent à remercier Eric Frenette, rédacteur canadien, ainsi que deux arbitres anonymes pour leurs conseils judicieux. Cette étude a été subventionnée par une bourse de recherche post-doctorale « chargé de recherches » du Fonds national de la recherche scientifique (FNRS), Belgique, et le Pôle d'attraction interuniversitaire (PAI) P7/06 de l'État belge (Politique fédérale scientifique). La correspondance liée à cet article peut être adressée à David Magis, Département d'éducation et formation (B32), Université de Liège, boulevard du Rectorat 5, 4000 Liège, Belgique, téléphone : +32-4-366-3665, ou par courriel à l'adresse suivante : [david.magis@ulg.ac.be].

Introduction

Un des principaux objectifs de la théorie de la réponse aux items (TRI) est de fournir des modèles de mesure permettant d'estimer le niveau d'habileté des apprenants lorsqu'ils sont soumis à un test d'aptitude ou de connaissances. Ce niveau d'habileté représente un trait latent propre à chaque apprenant. Les modèles issus de la TRI sont des modèles pour traits latents reposant sur un minimum de conditions d'utilisation. L'approche classique de la TRI consiste à admettre deux hypothèses (ou conditions d'application) fondamentales, l'unidimensionnalité du trait latent et l'indépendance locale des items (Hambleton & Swaminathan, 1985). La première hypothèse présuppose que le test est conçu de façon telle à ne mesurer qu'un seul trait latent principal, une habileté en mathématiques par exemple. La seconde hypothèse stipule qu'à un niveau d'habileté fixé, les réponses aux items sont fournies de manière indépendante les unes par rapport aux autres. En d'autres termes, l'hypothèse d'indépendance locale suppose qu'une personne répondra à un item indépendamment des réponses fournies aux autres items. Sous ces deux conditions, des modélisations de réponses à l'item, tels que les modèles logistiques pour réponses dichotomiques, sont alors disponibles.

Cependant, même dans des conditions idéales d'application validant ces deux hypothèses, il se peut que des facteurs additionnels affectent certaines réponses des répondants, produisant de la sorte des réponses dites aberrantes. Ces facteurs peuvent être de nature différente, comme par exemple : stress lors de l'administration du questionnaire, tentative de fraude, réponse au hasard, fatigue à la fin du test (Brassard, Béland, & Raîche, 2011). Les patrons de réponses pour lesquels une ou plusieurs réponses sont aberrantes sont alors qualifiés de patrons inappropriés et l'estimation du niveau d'habileté devient alors problématique, car fortement dépendante du type et de la proportion de réponses aberrantes (Meijer & Sijtsma, 2001). Une approche récente consiste à estimer le niveau d'habileté selon une approche dite « robuste », en diminuant la contribution des réponses aberrantes lors du processus d'estimation des paramètres d'items et de personnes (Mislevy & Bock, 1982 ; Schuster & Yuan, 2011). Toutefois, dans le présent article, l'accent est mis sur une approche plus classique et moins technique, à savoir l'utilisation d'indices d'ajustement (*person fit indexes*).

L'identification de patrons de réponses inappropriés est réalisée en trois étapes :

a) la calibration des paramètres d'items selon le modèle de la TRI retenu (comme étant le plus approprié aux données étudiées),
b) l'estimation du niveau d'habileté des répondants en utilisant les items calibrés,
c) le calcul des indices d'ajustement et l'identification des patrons de réponses inappropriés par l'interprétation de l'indice d'ajustement.

Or, il est possible de supposer raisonnablement que la présence de réponses aberrantes influence les trois étapes décrites ci-dessus. Une raison possible est que cette calibration prend en compte l'entièreté des patrons de réponses fournis par les apprenants et que, la présence de réponses aberrantes étant négligeable par rapport à l'ensemble de données disponibles, la présence de patrons inappropriés n'affecterait que très peu cette calibration. À l'inverse, il a déjà été observé que des patrons de réponses inappropriés peuvent être très mal identifiés par les indices d'ajustement classiques (comme les indices *Infit* et *Outfit*), soit à cause de leur impact sur l'estimation du niveau d'habileté (Mislevy & Bock, 1982), soit sur leur effet sur la distribution de probabilité des indices d'ajustement (Karabatsos, 2003 ; Molenaar & Hoijtink, 1990 ; Nering, 1995 ; Reise, 1995). Toutefois, la calibration des items conditionnant le reste du processus, il semble inévitable de devoir s'y attarder et d'étudier de tels impacts. Par ailleurs, si des solutions ont été proposées pour corriger les indices d'ajustement en cas de problèmes d'estimation du niveau d'habileté (Snijders, 2001), une correction similaire du processus de calibration n'a pas encore été proposée.

Le présent article propose une méthode itérative permettant de diminuer progressivement l'impact de réponses aberrantes sur la calibration des items et, de fait, sur l'ensemble du processus d'identification des patrons inappropriés. Cette approche repose sur une méthode similaire à celle utilisée pour la détection du fonctionnement différentiel des items (Magis, Béland, Tuerlinckx, & De Boeck, 2010 ; Osterlind & Everson, 2009 ; Penfield & Camilli, 2007) appelée la purification des items. Dans ce contexte, la purification est appliquée uniquement lors de l'étape de calibration des items.

Repères théoriques

Dans cette section, les grandes étapes de l'identification de patrons de réponses inappropriés (calibration des items, estimation du niveau d'habileté des répondants et calcul des indices d'ajustement) sont décrites brièvement. Les items considérés sont restreints au type dichotomique, pour lesquels il existe plusieurs modèles (Baker & Kim, 2004 ; Hambleton & Swaminathan, 1985). Ce choix est en relation directe avec l'analyse ultérieure des données utilisées pour illustrer la méthode proposée. Le modèle qui sera utilisé dans cet article est le modèle logistique à trois paramètres (Birnbaum, 1968) qui peut être représenté sous la forme suivante :

$$P_j(\theta) = \Pr(X_j = 1 \mid \theta, a_j, b_j, c_j) = c_j + (1 - c_j)\ \frac{\exp[Da_j(\theta - b_j)]}{1+\exp[Da_j(\theta - b_j)]} \qquad (1)$$

Dans l'équation 1, θ représente le trait latent (ou niveau d'habileté) du répondant et X_j ($j = 1, \ldots, n$) est la réponse de l'apprenant au j-ème item d'un test constitué de n items au total. La constante D est utilisée afin de permettre une éventuelle compatibilité entre les modèles logistiques et les modèles de type probit (Bertrand & Blais, 2004 ; Lord & Novick, 1968 ; Hambleton & Swaminathan, 1985). Cette constante a été fixée à 1 dans le cadre de cette étude. Les paramètres d'items sont a_j , b_j et c_j et correspondent respectivement au niveau de discrimination, de difficulté et de pseudo-chance de l'item. Ce modèle a été retenu pour l'analyse originale des données (Raîche, 2002) et une nouvelle analyse est proposée dans la suite de cet article.

Calibration des items

La calibration des items consiste à fournir une estimation des paramètres d'items a_j , b_j et c_j sur la base de l'ensemble des réponses fournies par les apprenants. Les principales approches pour effectuer une telle calibration sont les méthodes du maximum de vraisemblance conjoint (Lord, 1980), du maximum de vraisemblance conditionnel (Andersen, 1970), du maximum de vraisemblance marginal (Bock & Aitkin, 1981) et l'estimation Bayésienne modale (Swaminathan & Gifford, 1985, 1986). Chaque méthode a ses avantages et ses inconvénients propres (voir Baker & Kim, 2004 ; Lord, 1986).

Dans cette étude, la méthode du maximum de vraisemblance marginal a été sélectionnée pour deux raisons. Tout d'abord, elle est connue pour donner des estimations stables et pratiquement dépourvues de biais (Baker & Kim,

2004). Ensuite, elle est implémentée de façon courante dans plusieurs logiciels, dont BILOG-MG3 (Zimowski, Muraki, Mislevy, & Bock, 2003) qui est, avec R (*R Development Core Team*, 2012), un des logiciels sélectionnés dans cette étude pour l'analyse des données.

Estimation de l'habileté

Une fois les paramètres d'items calibrés sur l'ensemble des patrons de réponses, le niveau d'habileté des répondants peut être estimé. Il existe un grand nombre de techniques, les plus connues étant le maximum de vraisemblance (Lord, 1980, 1986), le maximum *a posteriori* (Lord, 1986), la moyenne *a posteriori* (Mislevy & Bock, 1982) et le maximum de vraisemblance pondéré (Warm, 1989). La méthode du maximum de vraisemblance est la plus intuitive et la plus directe : l'estimateur obtenu jouit de belles propriétés statistiques asymptotiques (Baker & Kim, 2004 ; Hambleton & Swaminathan, 1985). Cet estimateur présente cependant plusieurs inconvénients dans des situations particulières qui peuvent toutefois être corrigés en adoptant un paradigme Bayésien en utilisant le maximum *a posteriori* ou la moyenne *a posteriori*. Enfin, le maximum de vraisemblance pondéré a été proposé pour corriger le biais de l'estimateur du maximum de vraisemblance, essentiellement pour des tests courts ou de longueur moyenne. Cette dernière technique d'estimation du niveau d'habileté a été sélectionnée dans cet article afin de disposer d'un estimateur obtenu dans un cadre traditionnel (non Bayésien) et dont le biais est quasi nul (Warm, 1989).

Indices d'ajustement

Finalement, avec les paramètres d'items et le niveau d'habileté des répondants estimés, il est possible de procéder à l'identification de patrons de réponses inappropriés. L'approche la plus courante consiste à calculer la valeur d'un indice d'ajustement du patron de réponses aux caractéristiques du test. Cet indice résume l'ajustement global du patron de réponses au modèle de réponse à l'item choisi. Une déviation importante par rapport à ce modèle se traduit habituellement par une valeur élevée de l'indice d'ajustement et ainsi la détection de la présence de réponses aberrantes dans ce patron. Il existe une quantité impressionnante d'indices d'ajustements (voir Karabatsos, 2003 ; Meijer & Sijtsma, 2001). Pour certains de ces indices, leur distribution de probabilité n'est pas connue et ceux-ci sont donc appliqués de manière plutôt subjective. Pour cette raison, cette étude s'intéresse plus spécifiquement à l'utilisation de l'indice de log-vraisemblance standardisé l_z (Drasgow, Levine, &

Williams, 1985) et plus précisément sa version modifiée l_z^* proposée par Snijders (2001). Étant donné l'aspect central de cet indice dans cet article ainsi que la complexité liée à l'écriture formelle de l'indice l_z^* (voir Magis, Béland, & Raîche, 2012), les auteurs présentent ci-dessous les étapes de calcul de cet indice dans le cadre de cette étude (c'est-à-dire sous le modèle logistique à trois paramètres et avec la méthode du maximum de vraisemblance pondéré).

Tout d'abord, l'indice l_z se présente comme suit :

$$l_z = \frac{\sum_{j=1}^{n}[X_j - P_j(\theta)]\,\omega_j(\theta)}{[\sum_{j=1}^{n} P_j(\theta)[1-P_j(\theta)]\,\omega_j(\theta)^2]^{1/2}} \tag{2}$$

avec

$$\omega_j(\theta) = \log \frac{P_j(\theta)}{1 - P_j(\theta)} \tag{3}$$

et $P_j(\theta)$ fourni par l'équation 1. Lorsqu'il est obtenu en utilisant le vrai niveau d'habileté, cet indice a une distribution asymptotique normale standard qui peut être utilisée pour déterminer un seuil de détection des patrons de réponse inappropriés. Toutefois, le vrai niveau d'habileté étant inconnu, l'on ne dispose que d'une estimation $\hat{\theta}$ (ici par maximum de vraisemblance pondéré) qui est utilisée dans l'équation 2 à la place du vrai niveau d'habileté θ. La modification de l'indice l_z proposée par Snijders (2001) consiste à corriger l'introduction de l'estimateur du niveau d'habileté à la place du vrai niveau d'habileté. Cette modification dépend à la fois du modèle de réponse à l'item et de l'estimateur d'habileté choisis (Magis et al., 2012). Dans notre contexte (modèle logistique à trois paramètres et estimateur par maximum de vraisemblance pondéré), l'indice corrigé l_z^* est obtenu comme suit. Pour tout item j, les notations suivantes sont posées :

$$r_j(\theta) = \frac{1}{P_j(\theta)[1-P_j(\theta)]}\,\frac{dP_j(\theta)}{d\theta}, \quad r_0(\theta) = \frac{J(\theta)}{2\,I(\theta)}, \quad c_n(\theta) = \sum_{j=1}^{n} \frac{\omega_j(\theta)}{I(\theta)}\,\frac{dP_j(\theta)}{d\theta} \tag{4}$$

et

$$\tilde{\omega}_j(\theta) = \omega_j(\theta) - c_n(\theta)\,r_j(\theta) \tag{5}$$

Dans ces conditions, l'indice l_z^* est obtenu ainsi :

$$l_z^* = \frac{\sum_{j=1}^{n}[X_j - P_j(\theta)]\,\omega_j(\theta) - c_n(\theta)\,r_0(\theta)}{[\sum_{j=1}^{n} P_j(\theta)[1 - P_j(\theta)]\,\tilde{\omega}_j(\theta)^2]^{1/2}} \tag{6}$$

en remplaçant θ par son estimateur $\hat{\theta}$ par maximum de vraisemblance pondéré.

Snijders (2001) a établi qu'avec cette modification l'indice l_z^* affiche une distribution de probabilité asymptotique normale standard. En fixant un seuil de signification α, il est dès lors possible de déterminer une valeur seuil pour l'identification de patrons de réponses inappropriés. Par exemple, en fixant ce seuil de signification à 1 %, le seuil est déterminé par le quantile à 0,01 de la loi normale standard, soit -2,326. Tout patron de réponses dont l'indice l_z^* est inférieur à -2,326 est donc considéré comme inapproprié au seuil de signification de 1 %.

Cette approche en trois étapes (calibration des items, estimation du niveau d'habileté et calcul des indices d'ajustement) est la plus classique et la plus valide de par l'introduction de l'indice modifié l_z^*. Toutefois, ni la phase de calibration des items, ni celle de l'estimation du niveau d'habileté n'est à l'abri d'une influence directe de réponses aberrantes. En ce qui concerne l'estimation du niveau d'habileté, l'impact est doublement important, car il dépend des items calibrés et du patron de réponses lui-même, tous deux possiblement affectés par les réponses aberrantes. Une approche plus robuste semble être la meilleure solution à ce jour pour diminuer l'impact des réponses aberrantes (Mislevy & Bock, 1982 ; Schuster & Yuan, 2011), mais celle-ci n'a pas encore été étudiée en détail dans le contexte de la détection de patrons de réponses inappropriés.

L'objectif principal de cette recherche est de se concentrer sur la première étape du processus, soit celle de la calibration des paramètres d'items. Cette calibration repose sur l'utilisation de la totalité des patrons de réponses, et certains d'entre eux sont probablement inappropriés et peuvent être détectés par l'utilisation d'indices d'ajustement. Il est donc proposé de procéder à une calibration répétée de façon itérative en retirant à chaque étape, les patrons de réponses identifiés comme inappropriés à l'étape précédente. Ce processus itératif est décrit à la section suivante. L'objectif secondaire est d'illustrer cette approche aux données obtenues lors de l'administration d'un test de classement en anglais, langue seconde.

Méthodologie

Le processus itératif

Le processus itératif de calibration des items peut se résumer comme suit. Supposons qu'une première analyse (calibration des items, estimation du niveau d'habileté, calcul des indices d'ajustement) a été effectuée au préalable et qu'un ensemble de patrons de réponses (identifiés comme inappropriés) a été identifié. Ceci constitue l'étape 1 du processus. Les étapes 2 et suivantes sont alors obtenues de la façon schématique suivante :

a) procéder à la calibration des items en retirant, de l'ensemble de données disponibles, les patrons de réponses qui sont actuellement identifiés comme inappropriés,

b) réestimer le niveau d'habileté des répondants en utilisant la nouvelle calibration des items pour tous les patrons de réponses,

c) calculer à nouveau les indices d'ajustement de tous les patrons de réponses avec les nouveaux estimateurs du niveau d'habileté et les nouvelles valeurs des paramètres d'items,

d) identifier les patrons de réponses inappropriés à l'aide des nouveaux indices d'ajustement,

e) si le sous-ensemble des patrons de réponses identifiés comme inappropriés correspond exactement au sous-ensemble obtenu à l'étape précédente, le processus itératif s'arrête. Sinon, retour à l'étape a) et les itérations recommencent.

L'idée de ce processus est claire : en retirant les patrons identifiés comme inappropriés, la calibration des items tend à être moins affectée par les réponses aberrantes et les paramètres d'items seront ainsi moins biaisés. Par ailleurs, l'estimation du niveau d'habileté se fera sur la base de ces nouveaux paramètres d'items et n'en sera ainsi qu'améliorée. Enfin, l'utilisation de l'indice d'ajustement l_z^* devrait permettre une meilleure identification des patrons réellement inappropriés (Snijders, 2001).

Il est légitime de se demander si le retrait successif de patrons de réponses risque d'affecter le degré de précision des estimations des paramètres d'items. Toutefois, la proportion de patrons de réponses inappropriés présents dans un ensemble de données est en général faible et une calibration adéquate des paramètres d'items requiert un échantillon suffisamment grand. Le retrait d'une fraction de patrons de réponses ne devrait donc pas mener à des jeux de

données trop petits pour obtenir une estimation fiable des paramètres d'items. En d'autres termes, le retrait de patrons de réponses identifiés comme inappropriés devrait améliorer la qualité de cette calibration des items sans pour autant trop affecter la précision des estimateurs des paramètres d'items.

Mise à l'échelle des paramètres d'items

Il convient également de mentionner un autre point technique de cette méthode itérative pour laquelle une solution peut être envisagée de manière simple. Lorsque la calibration des items se fait sur des ensembles de données différents, il est fort probable que les paramètres d'items ne se trouvent pas sur la même échelle pour les deux calibrations (Dorans, Pommerich, & Holland, 2007 ; Kolen & Brennan, 2004). Il est alors nécessaire, avant de procéder à l'identification des patrons de réponses inappropriés, d'effectuer une mise à l'échelle commune, paramètre par paramètre, en choisissant une échelle de référence (ou en d'autres mots, une calibration de référence qui fixe l'échelle des paramètres). Dans cette étude, l'échelle commune a été choisie comme celle obtenue lors de la calibration des items sur tous les patrons de réponses disponibles (c'est-à-dire à l'étape 1 du processus itératif) et les items calibrés lors des étapes suivantes sont remis à l'échelle fixée par la première calibration.

Il existe plusieurs méthodes pour effectuer cette mise à l'échelle commune (Dorans et al., 2007 ; Kolen & Brennan, 2004). Dans cette étude, le choix s'est porté sur la méthode dite «*mean/mean*». Elle consiste en une remise à l'échelle des paramètres de telle sorte que les paramètres de difficulté, d'une part, et de discrimination, d'autre part, soient centrés sur les mêmes valeurs moyennes que celles observées sur l'échelle de référence. En pratique, notons (a_{j1}, b_{j1}, c_{j1}) les paramètres calibrés du *j*-ème item lors de l'étape 1 (qui fournissent donc l'échelle de référence) et notons (a_{jk}, b_{jk}, c_{jk}) les mêmes paramètres mais calibrés lors de la *k*-ème étape du processus itératif. Il est possible d'obtenir alors les paramètres d'items $(a^*_{jk}, b^*_{jk}, c^*_{jk})$ calibrés lors de la k-ème étape, mais étalonnés sur l'échelle de référence, comme suit :

$$a^*_{jk} = \frac{a_{jk}}{A}, \qquad b^*_{jk} = A\, b_{jk} + B \qquad \text{et} \qquad c^*_{jk} = c_{jk} \tag{7}$$

avec

$$A = \frac{\sum_{j=1}^{n} a_{jk}}{\sum_{j=1}^{n} a_{j1}} \quad \text{et} \quad B = \frac{1}{n} \sum_{j=1}^{n} b_{j1} - A \frac{1}{n} \sum_{j=1}^{n} b_{jk} \tag{8}$$

Il s'agit ainsi d'une remise à l'échelle linéaire qui est réalisée facilement, une fois les paramètres d'items calibrés, par la détermination des constantes A et B de l'équation 8 et la transformation linéaire des paramètres dans l'équation 7.

Illustration

Le processus itératif décrit dans la section précédente est à présent illustré par l'analyse d'un ensemble de données issu du test de classement en anglais, langue seconde, au collégial (ou TCALS-II). Une description complète de l'ensemble de données est fournie en appendice. L'ensemble de données analysé est constitué de 85 items et de 1 373 patrons de réponses.

Les auteurs ont suivi la méthode d'analyse proposée par Raîche (2002). Le modèle logistique à trois paramètres a été retenu après vérification de son ajustement aux données et les items ont été calibrés selon cette modélisation. Par ailleurs, la méthode du maximum de vraisemblance pondéré (Warm, 1989) a été utilisée pour estimer le niveau d'habileté des répondants. L'indice d'ajustement l_z^* a été considéré pour détecter les patrons de réponses inappropriés.

La calibration des items du TCALS-II a été réalisée à l'aide du logiciel BILOG-MG3 (Zimowski et al., 2003) en utilisant les distributions *a priori* des paramètres d'items par défaut. La mise à l'échelle des paramètres d'items, l'estimation du niveau d'habileté et le calcul des indices d'ajustement ont été réalisés avec le logiciel statistique R (*R Development Core Team*, 2012). La librairie catR (Magis & Raîche, 2012) a été employée pour l'estimation du niveau d'habileté, tandis qu'un code spécifique a été rédigé pour la mise à l'échelle des paramètres d'items et le calcul des indices d'ajustement. Le code source est disponible sur simple demande aux auteurs.

Notons que deux seuils de signification α ont été sélectionnés : $\alpha = 1\,\%$ et $\alpha = 5\,\%$. Le premier est souvent utilisé pour l'identification de patrons inappropriés (Drasgow et al., 1985 ; Karabatsos, 2003 ; Snijders, 2001), tandis que le second est la valeur usuelle de ce seuil. De plus, un seuil plus élevé mène à une détection plus importante de patrons et cela peut impacter le processus itératif en termes de nombres d'itérations. Pour cette raison, il est intéressant de comparer la performance du processus itératif avec deux valeurs différentes de ce seuil.

Résultats

Pour les deux seuils de signification (1 % et 5 %), le processus itératif a fourni une réponse en un nombre relativement limité d'étapes, soit cinq étapes pour le seuil 1 % et sept étapes pour le seuil 5 %. Avant de se concentrer sur l'identification des patrons de réponses inappropriés, examinons les variations des distributions des paramètres d'items et du niveau d'habileté au cours des différentes étapes du processus.

Les tableaux 1 et 2 reprennent, respectivement pour les seuils de signification 1 % et 5 %, les valeurs moyennes, écarts-types, minimum et maximum des trois paramètres d'items (discrimination, difficulté et pseudo-chance) et du niveau d'habileté à chaque étape du processus itératif. La première étape représente les résultats obtenus lorsque la calibration s'effectue sur l'entièreté des patrons de réponses. Il est à noter que pour les paramètres d'items, ces statistiques ont été calculées avant la mise à l'échelle des paramètres par la méthode «*mean/mean*», ce qui permet de quantifier l'effet du retrait des patrons inappropriés sur les calibrations successives des items.

Tableau 1
Statistiques résumées des estimations successives des paramètres d'items et des niveaux d'habileté, avec un seuil de signification de 1 %

		Étapes				
Paramètre	*Statistique*	*1*	*2*	*3*	*4*	*5*
Discrimination	Moyenne	2,039	2,095	2,108	2,113	2,114
	Écart-type	0,808	0,854	0,862	0,863	0,861
	Minimum	0,288	0,291	0,294	0,293	0,295
	Maximum	4,472	4,371	4,338	4,335	4,316
Difficulté	Moyenne	-0,965	-0,984	-0,984	-0,983	-0,983
	Écart-type	0,780	0,792	0,792	0,792	0,790
	Minimum	-3,334	-3,378	-3,378	-3,376	-3,337
	Maximum	0,886	0,874	0,865	0,865	0,862
Pseudo-chance	Moyenne	0,195	0,192	0,191	0,192	0,192
	Écart-type	0,080	0,079	0,078	0,079	0,079
	Minimum	0,035	0,023	0,023	0,023	0,023
	Maximum	0,450	0,439	0,440	0,438	0,437

	Moyenne	-0,020	0,010	0,015	0,016	0,015
Habileté	Écart-type	1,016	1,035	1,036	1,037	1,038
	Minimum	-3,666	-3,737	-3,745	-3,746	-3,752
	Maximum	2,270	2,284	2,280	2,278	2,275

Tableau 2

Statistiques résumées des estimations successives des paramètres d'items et des niveaux d'habileté, avec un seuil de signification de 5 %

		Étapes						
Paramètre	*Statistique*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
	Moyenne	2,039	2,124	2,170	2,188	2,200	2,206	2,211
Discrimination	Écart-type	0,808	0,848	0,875	0,879	0,886	0,893	0,902
	Minimum	0,288	0,305	0,307	0,315	0,313	0,318	0,318
	Maximum	4,472	4,330	4,526	4,543	4,545	4,545	4,804
	Moyenne	-0,965	-0,981	-0,978	-0,978	-0,974	-0,974	-0,974
Difficulté	Écart-type	0,780	0,781	0,780	0,777	0,777	0,778	0,777
	Minimum	-3,334	-3,057	-3,031	-2,971	-2,966	-2,955	-2,954
	Maximum	0,886	0,889	0,898	0,892	0,895	0,897	0,895
	Moyenne	0,195	0,189	0,188	0,187	0,187	0,187	0,187
Pseudo-chance	Écart-type	0,080	0,078	0,076	0,076	0,076	0,076	0,076
	Minimum	0,035	0,024	0,023	0,024	0,024	0,024	0,024
	Maximum	0,450	0,436	0,433	0,430	0,434	0,437	0,437
	Moyenne	-0,020	0,024	0,039	0,047	0,050	0,054	0,056
Habileté	Écart-type	1,016	1,041	1,052	1,056	1,060	1,062	1,064
	Minimum	-3,666	-3,732	-3,742	-3,729	-3,728	-3,734	-3,738
	Maximum	2,270	2,278	2,280	2,289	2,300	2,313	2,313

Des conclusions semblables peuvent être tirées des résultats pour les deux seuils de signification. Tout d'abord, le niveau moyen de discrimination des items tend à augmenter légèrement, de même que sa dispersion. Ensuite, les niveaux de difficulté diminuent légèrement en moyenne et leur dispersion est

assez stable. Enfin, les paramètres de pseudo-chance restent presque inchangés au cours des différentes étapes. Finalement, en ce qui concerne le niveau d'habileté estimé, il est possible de constater, au fil des étapes, une légère augmentation des indices moyens d'habileté ainsi que de leur dispersion. En résumé, bien que le retrait successif des patrons identifiés comme inappropriés affecte logiquement la calibration des items et l'estimation du niveau d'habileté, cet impact reste toutefois faible.

Considérons à présent les indices l_z^* calculés à chaque étape du processus et la classification conséquente des patrons de réponses en tant qu'inappropriés ou non. Le tableau 3 reprend, pour chaque étape et chaque seuil de signification, le nombre de patrons de réponses identifiés comme inappropriés et ceux classés comme « normaux ». Il est à noter d'emblée que les patrons de réponses identifiés comme inappropriés à une étape le sont restés lors des étapes suivantes ; seuls de nouveaux patrons de réponses, jusqu'ici non identifiés comme inappropriés, ont été détectés en plus au cours du processus itératif.

Tableau 3
Évolution du nombre de patrons de réponses identifiés comme inappropriés ($l_z^* < Q$) et non identifiés ($l_z^* \geq Q$) après chaque étape du processus itératif

	$\alpha = 1\,\%$		$\alpha = 5\,\%$	
	$l_z^* \geq Q$	$l_z^* < Q$	$l_z^* \geq Q$	$l_z^* < Q$
Étape 1	1323	50	1268	105
Étape 2	1313	60	1224	149
Étape 3	1310	63	1202	171
Étape 4	1308	65	1194	179
Étape 5	1308	65	1188	185
Étape 6			1186	187
Étape 7			1186	187

Comme prévu, le nombre de patrons de réponses identifiés comme inappropriés à la première étape du processus est plus élevé pour le seuil de signification de 5 % (105 patrons) que pour le seuil de signification de 1 % (50 patrons). Cela est simplement dû au fait que le seuil de détection est plus élevé

dans le premier cas (-1,645) que dans le second (-2,326). Par ailleurs, les auteurs constatent que le nombre de patrons identifiés ultérieurement diminue graduellement au cours des étapes suivantes. L'absence de détection de nouveaux patrons de réponses inappropriés à la dernière étape du processus force ainsi ce processus itératif à s'interrompre.

Au final, 187 patrons de réponses sont identifiés en utilisant le seuil de signification de 5 % pour « seulement » 65 patrons avec le seuil de signification de 1 %. Toutefois, en comparant ces chiffres aux nombres de patrons identifiés après la première étape (et qui constituent les patrons habituellement détectés dans les analyses traditionnelles), soit respectivement 105 et 50, cela constitue une augmentation respective du nombre de patrons identifiés de 78 % et 30 %, ce qui n'est pas négligeable. En termes de proportions par rapport à l'échantillon total d'apprenants ayant répondu au TCALS-II, les patrons de réponses identifiés comme inappropriés à la fin du processus itératif représentent 13,6 % et 4,7 % (pour des seuils respectifs de 5 % et 1 %), alors qu'après la première étape ils représentaient 7,6 % et 3,6 % des patrons inclus dans les données. En somme, le processus itératif mène à un taux de détection plus élevé, l'accroissement du nombre de patrons identifiés comme inappropriés étant plus important pour des seuils de signification plus élevés.

Les couples de valeurs des indices l_z^* calculés à la première et à la dernière étape sont représentés dans la figure 1, pour le seuil de signification de 1 % (panneau de gauche) et de 5 % (panneau de droite). Il est à noter que seuls les couples de valeurs se trouvant dans une étendue de valeurs proche du seuil de détection sont représentés. Les lignes pointillées représentent les seuils de détection pour les niveaux de signification fixés et découpent les graphiques en quatre quadrants. Les patrons de réponses localisés dans les quadrants inférieur gauche et supérieur droit ne subissent pas de modification de classification au cours du processus ; les patrons du quadrant inférieur gauche sont identifiés comme inappropriés et ceux du quadrant supérieur droit comme « normaux ». Les quadrants inférieurs droits reprennent les patrons de réponses identifiés comme « normaux » à la première étape et comme inappropriés à la dernière étape. Finalement, les quadrants supérieurs gauches ne contiennent aucun patron de réponses, confirmant le fait que les patrons identifiés comme inappropriés à la première étape le sont restés à la dernière étape.

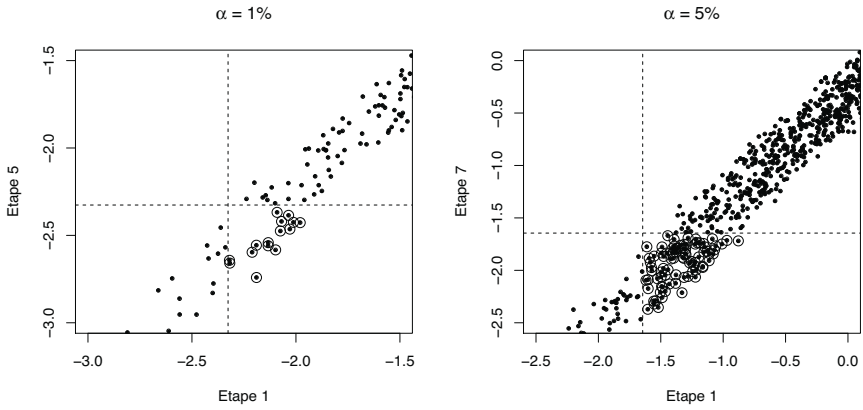

Figure 1 : ***Graphiques bivariés des indices l_z^* obtenus à la première et à la dernière étape du processus itératif, pour les seuils de signification 1 % (gauche) et 5 % (droite)***

Discussion

L'application du processus itératif aux données du TCALS-II a permis d'ouvrir des perspectives intéressantes pour des travaux ultérieurs.

Tout d'abord, le processus s'est terminé après un nombre relativement restreint d'itérations, ce qui ne représente donc pas un effort de calcul important. Il est à noter toutefois que ceci ne constitue en aucun cas un gage de garantie quant à la convergence systématique de cette méthode. Autrement dit, il n'est pas permis d'affirmer que le processus s'interrompra toujours de cette manière, ni de déterminer le nombre d'itérations nécessaire à cette convergence. Sur la base de l'exemple traité, il semble que le nombre de modifications successives dans l'identification des patrons de réponses inappropriés tend à décroître au fil des étapes. Il est donc possible d'envisager que plus le processus est long (en termes d'étapes), plus les modifications dans la calibration des items et l'identification des patrons inappropriés vont s'atténuer. Un critère d'arrêt du processus pourrait alors être établi en forçant l'arrêt des étapes itératives lorsqu'une transition d'une étape à la suivante n'apporte pas plus de modifications qu'un certain pourcentage minimum fixé à l'avance. Par exemple, lorsque l'écart entre les paramètres d'items issus de deux calibrations successives est systématiquement inférieur à un certain seuil fixé à l'avance, à l'image des seuils de tolérance pour la convergence d'algorithmes de maximisation.

L'exemple du TCALS-II a aussi mis en avant un phénomène intéressant : les patrons de réponses détectés comme inappropriés dès la première étape le restent jusqu'à la fin du processus. En d'autres termes, la méthode itérative tend à augmenter le nombre de patrons de réponses détectés comme inappropriés. Rien ne permet de présupposer que cette tendance se vérifie systématiquement, surtout si certains patrons de réponses inappropriés ont un indice d'ajustement très proche du seuil de décision et qui pourraient très bien être classés comme « normaux » à l'étape suivante du processus. La tendance observée dans cette étude peut être due aux données particulières traitées et donc n'être qu'un cas particulier, mais le fait est intéressant si l'on souhaite en tirer des conclusions pratiques pour cette étude. Il signifie que l'application du processus itératif aurait permis d'identifier 15 (avec le seuil de signification de 1 %) ou 82 (avec le seuil 5 %) apprenants supplémentaires comme ayant fourni un patron de réponses inappropriés pour le TCALS-II. Même si ce nombre est relativement petit par rapport à l'échantillon total d'apprenants, il ne faut pas négliger cet accroissement du nombre de patrons détectés, qui pourrait être une indication que la calibration des items sur l'ensemble des apprenants a un effet masquant pour certains patrons inappropriés.

Remarquons enfin que le choix du seuil de signification peut s'avérer central dans cette problématique. En effet, un seuil de signification plus élevé mène logiquement à une identification plus importante de patrons à la première étape, mais l'exemple illustre aussi que cela affecte la longueur du processus itératif et le nombre final de patrons identifiés comme aberrants. Ainsi, en termes d'impact pratique sur l'identification des apprenants ayant répondu de manière potentiellement inappropriée, cela engendrerait des efforts et un coût supplémentaire pour déterminer la cause de cette détection, et éventuellement le rappel des apprenants pour le passage d'une seconde épreuve. Le choix d'un seuil de signification moins élevé réduirait sensiblement cet effort, mais mènerait aussi probablement à une diminution de l'identification de patrons réellement inappropriés. Un compromis doit donc être trouvé entre ces deux facteurs, l'identification adéquate des patrons inappropriés et le coût (en termes de temps et d'argent) de la validation de ces identifications.

Conclusion

Les résultats obtenus sur la base de cette analyse empirique sont encourageants. Cependant, une étude plus complète et systématique doit être réalisée afin de confirmer ou d'infirmer les tendances dégagées dans cet article. Un aspect à la fois important et complexe de cette tâche consiste à simuler des données selon un modèle de réponse à l'item choisi, et d'insérer artificiellement des patrons de réponses inappropriés. La difficulté principale de cette approche est de générer ces patrons inappropriés de façon réaliste et systématique. Certains types de comportements atypiques, comme la tentative de fraude, les réponses au hasard ou l'inattention, peuvent être créés en suivant les propositions de Schuster et Yuan (2011) et St-Onge, Valois, Abdous, et Germain (2011) par exemple. Plus récemment, Raîche, Magis, Blais, et Brochu (2012) ont proposé une modélisation issue de la TRI qui permet de produire des patrons de réponses inappropriés selon des comportements réalistes. Il est à noter toutefois qu'il n'existe pas (encore) de méthode universelle pour générer des patrons de réponses inappropriés, car la source de l'aberrance est en général multiple et complexe.

Il convient toutefois de mentionner que ce processus itératif, à l'instar du processus de purification des items dans le contexte du fonctionnement différentiel des items (Magis et al., 2012), pourrait ne pas converger vers une solution stable. Il se pourrait en effet qu'un phénomène de boucle apparaisse entre plusieurs itérations du processus, certains patrons de réponses étant successivement identifiés comme inappropriés ou non. Cette situation, bien que probablement rare en pratique (l'exemple étudié dans cet article n'a pas mené à une boucle de ce type), peut néanmoins être considérée comme un inconvénient méthodologique de cette approche.

Toutefois, l'intérêt d'une étude par simulations de données est de pouvoir maîtriser le processus de génération des données et donc de connaître avec précision quels patrons de réponses sont réellement inappropriés et lesquels ne le sont pas. Ceci permettrait alors de déterminer l'efficacité réelle du processus itératif décrit dans cet article, tant sur le plan de l'identification des patrons de réponses inappropriés que sur leur impact sur la calibration des items.

Par ailleurs, le processus de génération des données permet de contrôler l'unidimensionnalité et l'indépendance locale des items, deux conditions nécessaires au bon déroulement de la procédure itérative. Dans l'exemple du TCALS-II, l'unidimensionnalité a été établie (voir en appendice) tandis que l'indépendance locale est en général acceptée dans les analyses s'y rapportant (Raîche, 2002). Il est toutefois utile de vérifier dans quelle mesure la violation d'une de ces hypothèses (ou des deux) a un impact immédiat sur la performance de la procédure.

RÉFÉRENCES

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society (Series B)*, *32*, 283-301.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. (2nd ed.). New York, NY: Marcel Dekker. doi: 10.2307/2532822

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

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. An application of the EM algorithm. *Psychometrika*, *37*, 29-51. doi: 10.1007/BF02293801

Brassard, P. D., Béland, S., & Raîche, G. (2011). Identification des comportements qui déterminent les patrons de réponses des étudiants qui tentent de se sous classer intentionnellement à un test. In G. Raîche, K. Paquette-Côté, & D. Magis (Eds.), *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation* (pp. 85-104). Québec, Canada : Presses de l'Université du Québec.

DeMars, C. E. (2010). *Item response theory*. Oxford, UK: Oxford University Press. doi: 10.1093/acprof:oso/9780195377033.001.0001

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales.* New York, NY: Springer. doi: 10.1007/978-0-387-49771-6

Drasgow, F., Levine, M.V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86. doi: 10.1111/j.2044-8317.1985.tb00817.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty six person-fit statistics. *Applied Measurement in Education*, *16*, 277–298. doi: 10.1207/S15324818AME1604_2

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practice* (2nd ed.). New York, NY: Springer. doi: 10.1007/978-1-4757-4310-4

Laurier, M., Froio, L., Pearo, C., & Fournier, M. (1998). *L'élaboration d'un test provincial pour le classement des étudiants en anglais langue seconde au collégial*. Québec, Canada : Direction générale de l'enseignement collégial, ministère de l'Éducation du Québec.

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

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157-162. doi: 10.1111/j.1745-3984.1986.tb00241.x

Lord, F. M., & M. R. Novick (Eds.). (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847-862. doi: 10.3758/BRM.42.3.847

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*, 1-31.

Magis, D., Béland, S., & Raîche, G. (2012). A didactic presentation of Snijders' l_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *12*, 37-57. doi: 10.3102/1076998610396894

Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135. doi: 10.1177/01466210122031957

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, *42*, 725-737. doi: 10.1177/001316448204200302

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*, 75-106. doi: 10.1007/BF02294745

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, *19*, 121-129. doi: 10.1177/014662169501900201

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125-167). Amsterdam, Netherlands: Elsevier.

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Raîche, G., Magis, D., Blais, J.-G., & Brochu, P. (2012). Taking atypical response patterns into account: a multidimensional measurement model from item response theory. In M. Simon, K. Ercikan, & M. Rousseau (Eds), *Improving large-scale assessment in education. Theory, issues, and practice* (pp. 238-259). New York, NY: Routledge.

Reise, S. R. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19*, 213-229. doi: 10.1177/014662169501900301

Schuster, C., & Yuan, K.-H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, *36*, 720-735. doi: 10.3102/1076998610396890

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331-342. doi: 10.1007/BF02294437

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement*, *35*, 419-432. doi: 10.1177/0146621610391777

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*, 349-364. doi: 10.1007/BF02294110

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589-601. doi: 10.1007/BF02295598

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, *54*, 427-450. doi: 10.1007/BF02294627

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262. doi: 10.1177/014662168100500212

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125-145. doi: 10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213. doi: 10.1111/j.1745-3984.1993.tb00423.x

Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. D. (2003). BILOG-MG 3 [Computer software]. Lincolnwood, IL: Scientific Software International.

Appendice

Le TCALS-II

Cette section fournit quelques renseignements sur l'ensemble de données qui a été analysé au cours de cette étude, le test de classement en anglais, langue seconde, au collégial (TCALS II).

Le TCALS-II est constitué d'un questionnaire de 85 items à choix multiples administré aux étudiants canadiens francophones au début de leurs études collégiales dans la province de Québec (Laurier, Froio, Pearo, & Fournier, 1998). Il fut administré pour la première fois en 1998 et est toujours en application actuellement. Le test a pour but d'évaluer leur niveau d'habileté en anglais, langue seconde, afin de les répartir en groupes de niveaux plus ou moins homogènes (Raîche, 2002).

Il a été établi (Laurier et al., 1998 ; Raîche, 2002) que le TCALS II possède un haut niveau de fidélité avec un coefficient α de Cronbach de l'ordre de 0,96. Par ailleurs, la dimensionnalité du TCALS II a été vérifiée par Raîche (2002) à l'aide d'une analyse en composantes principales appliquée aux coefficients de corrélation tétrachoriques. Cette analyse a mis en évidence que 25 % de la variance est expliquée par la première composante principale, tandis que la seconde n'explique que 2 % de la variance. Ce résultat a été confirmé par le test de l'éboulis de Cattell, appliqué aux mêmes coefficients de corrélation. L'unidimensionnalité du test est donc considérée comme établie (Laurier et al., 1998 ; Raîche, 2002).

En pratique, les données récoltées lors de l'administration du TCALS II en 1998 ont été utilisées pour cette analyse. La base de données contient un total de 1 373 patrons de réponses, dont 749 (54,6 %) fournis par des étudiantes et 624 (45,4 %) par des étudiants entrant au Collège de l'Outaouais en 1998. Les réponses polytomiques nominales ont été recodées en réponses dichotomiques et les réponses manquantes ont été codées comme de mauvaises réponses, en accord avec les analyses réalisées préalablement par Raîche (2002). L'ajustement des modèles logistiques à un, deux et trois paramètres a été effectué séparément et un test de rapport de vraisemblance a été utilisé pour comparer l'ajustement des modèles à un et deux paramètres, puis des modèles à deux et trois paramètres. Il apparaît que le modèle logistique à trois paramètres a un ajustement significativement meilleur que le modèle logistique à deux paramètres ($p < 0{,}001$), ce dernier ayant lui aussi un ajustement

meilleur que le modèle de Rasch ($p < 0{,}001$). Par ailleurs, la statistique Q1 de Yen (1981) a été considérée pour déterminer l'ajustement individuel des items au modèle à trois paramètres. Il a été observé que le modèle s'ajuste pour une grande majorité des items : 92,9 % des statistiques Q1 ont une p-valeur associée supérieure à 0,01 et 81,2 % ont une p-valeur associée supérieure à 0,05. Le modèle logistique à trois paramètres a donc été retenu pour ces raisons.

Finalement, l'investigation de la dépendance locale des items a été effectuée au moyen de la statistique Q3 de Yen (1984), calculée à partir du niveau d'habileté estimé au moyen du maximum de vraisemblance pondéré et avec les paramètres d'items calibrés selon le modèle logistique à trois paramètres. Les statistiques Q3 ont été calculées sur les corrélations linéaires entre les résidus (entre valeurs prédites par le modèle et réponses observées) pour toutes les paires d'items du TCALS II. Toutes ces statistiques sont inférieures à 0,20 sauf pour deux paires d'items, pour lesquels Q3 vaut 0,221 et 0,242. Selon le critère proposé par Yen (1993), l'indépendance locale entre paires d'items est donc considérée comme acceptable. Cette conclusion est en accord avec un argument proposé par notamment DeMars (2010), selon lequel une dépendance locale des items peut être détectée comme une seconde dimension importante dans une analyse de dimensionnalité. Étant donné l'absence d'une seconde dimension importante dans le TCALS II, la conclusion d'absence globale de dépendance locale des items est ainsi confirmée.

Date de réception : 6 novembre 2012

Date de réception de la version finale : 21 novembre 2013

Date d'acceptation : 22 novembre 2013