Mesure et évaluation en éducation

Étude de nouveaux indices de détection de la réponse au hasard et de l'inattention selon différentes valeurs de l'habileté dans le contexte de la modélisation de Rasch

Sébastien Béland, Gilles Raîche, David Magis and Martin Riopel

Volume 39, Number 1, 2016

URI: https://id.erudit.org/iderudit/1036707ar
DOI: https://doi.org/10.7202/1036707ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

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

Explore this journal

Cite this article

Béland, S., Raîche, G., Magis, D. & Riopel, M. (2016). Étude de nouveaux indices de détection de la réponse au hasard et de l'inattention selon différentes valeurs de l'habileté dans le contexte de la modélisation de Rasch. *Mesure et évaluation en éducation*, *39*(1), 95–118. https://doi.org/10.7202/1036707ar

Article abstract

Some students may guess at random or be inattentive in a testing situation. Several approaches have been developed to detect these types of behavior. The use of person-fit index is the most studied approach and seems very promising. In this study, we focus on three popular indices which have many features to facilitate their interpretation: l_z, ZU and ZW. Nevertheless, previous studies have shown that these three indices are strongly affected by the fact that the ability of a student is estimated rather than known. Snijders (2001) proposed a corrected version of the l_z index (named l_z*) to take account of this problem. Magis, Béland, and Raîche (2014) have already used the Snijders correction to create two person-fit indexes: U* and W*. It is now time to extend our understanding of the corrected indexes l_z*, U*, and W*, and standardized indices l_z, ZU, and ZW. To do this, we conduct two studies using different values of the student's ability: an analysis of type I errors (probability of being wrong in identifying inappropriate response patterns), and an analysis of the power of detection of theses indexes. Our results show that the corrected indices l_z* and W* are most interesting because their scores are approximately normally distributed and allow to adequately detect guessing at random and inattention response patterns.

Étude de nouveaux indices de détection de la réponse au hasard et de l'inattention selon différentes valeurs de l'habileté dans le contexte de la modélisation de Rasch

Sébastien Béland
Université de Montréal

Gilles Raîche
Université du Québec à Montréal

David Magis
Université de Liège

Martin Riopel
Université du Québec à Montréal

Mots clés : théorie de la réponse à l'item, indice de détection de patrons de réponses inappropriés, réponse au hasard, inattention

Certains étudiants peuvent répondre au hasard ou être inattentifs dans une situation de testing. Plusieurs approches ont déjà été développées pour détecter ce type de réponse. Parmi celles-ci, l'utilisation d'indices de détection (person-fit indexes) de patrons de réponses inappropriés est l'approche qui est la plus étudiée et qui semble la plus prometteuse. Dans le cadre de cette étude, nous nous concentrons sur trois indices de détection populaires qui présentent des caractéristiques permettant d'en faciliter l'interprétation : l_z, ZU et ZW. Des études antérieures ont montré que ces trois indices sont fortement affectés par le fait que l'habileté d'un étudiant est estimée plutôt que réelle. Snijders (2001) a proposé une version corrigée de l'indice l_z (nommée l_z) afin de tenir compte de cette difficulté. Magis, Béland et Raîche (2014) ont déjà corrigé deux autres indices selon l'approche de Snijders : U* et W*. Il reste cependant à analyser plus en détail le comportement des indices corrigés l_z*, U* et W* et des indices standardisés l_z, ZU et ZW. Pour ce faire, nous effectuons deux études selon différentes valeurs de l'habileté, soit une analyse des erreurs de type I des indices (probabilité de se tromper en identifiant un patron de réponses inapproprié) et une analyse de leur puissance de détection. Ces analyses permettront de démontrer que ce sont généralement les indices corrigés l_z* et W* qui sont les plus intéressants à utiliser puisque leurs scores suivent approximativement la loi normale et qu'ils permettent de bien détecter la réponse au hasard et l'inattention.*

Keywords: item response theory, person-fit index of response patterns, guessing at random, inattention

Some students may guess at random or be inattentive in a testing situation. Several approaches have been developed to detect these types of behavior. The use of

person-fit index is the most studied approach and seems very promising. In this study, we focus on three popular indices which have many features to facilitate their interpretation: l_z, ZU and ZW. Nevertheless, previous studies have shown that these three indices are strongly affected by the fact that the ability of a student is estimated rather than known. Snijders (2001) proposed a corrected version of the l_z index (named l_z) to take account of this problem. Magis, Béland, and Raîche (2014) have already used the Snijders correction to create two person-fit indexes: U* and W*. It is now time to extend our understanding of the corrected indexes l_z*, U*, and W*, and standardized indices l_z, ZU, and ZW. To do this, we conduct two studies using different values of the student's ability: an analysis of type I errors (probability of being wrong in identifying inappropriate response patterns), and an analysis of the power of detection of theses indexes. Our results show that the corrected indices l_z* and W* are most interesting because their scores are approximately normally distributed and allow to adequately detect guessing at random and inattention response patterns.*

Palavras-chave: Teoria da resposta ao item, índice de deteção de padrões de respostas inadequadas, respostas ao acaso, desatenção

Alguns alunos podem responder de forma aleatória ou desatenta numa situação de teste. Várias abordagens têm sido utilizadas para detetar este tipo de resposta. A utilização de índices de deteção (person-fit indexes) de padrões de respostas inapropriadas é a abordagem mais estudada e a que parece ser mais promissora. Neste estudo, concentramo-nos em três índices de deteção populares que apresentam características que permitem facilitar a interpretação: I_z, ZU e ZW. Estudos anteriores demonstraram que estes três índices são fortemente afetados pelo facto de que a habilidade de um aluno é mais estimada do que real. Snijders (2001) propôs uma versão corrigida do índice I_z (denominado I_z) para ter em conta esta dificuldade. Magis, Béland e Raîche (2014) já corrigiram dois outros índices segundo a abordagem de Snijders: U* e W*. Resta, porém, analisar mais em detalhe o comportamento dos índices corrigidos I_z*, U* e W* e os índices padronizados l_z, ZU e ZW. Para fazer isso, realizámos dois estudos usando diferentes valores de habilidade, seja uma análise dos erros do tipo I dos índices (probabilidade de estar errado na identificação de respostas inapropriadas) e uma análise do seu poder de deteção. Estas análises permitirão demonstrar que são geralmente os índices corrigidos l_z* e W* que são os mais interessantes, uma vez que as suas pontuações seguem aproximadamente a lei normal e permitem detetar adequadamente a resposta ao acaso ou desatenta.*

Note des auteurs : Cet article est issu de la thèse de doctorat de Sébastien Béland. Gilles Raîche et Martin Riopel en ont été respectivement le directeur et le codirecteur. David Magis, pour sa part, outre son soutien et ses encouragements, a effectué les dérivations mathématiques des indices corrigés et a développé le code R qui a permis de calculer les indices corrigés.

La correspondance liée à cet article peut être adressée à : [sebastien.beland@umontreal.ca].

Introduction

Il est connu que certains étudiants répondent de façon inappropriée aux épreuves d'évaluation. Par exemple, ils peuvent répondre au hasard à une série d'items (Hendrawan, Glas & Meijer, 2005 ; Karabatsos, 2003 ; Kogut, 1987 ; Meijer, Muijtjens & van der Vleuten, 1996). Dans cette situation, des étudiants dont l'habileté est plutôt faible tenteraient la chance en répondant, sans réfléchir, aux items du test. Des étudiants peuvent aussi être inattentifs (Emons, Glas, Meijer & Sijtsma, 2003 ; Karabatsos, 2003 ; Raîche, Magis, Blais & Brochu, 2012). Ce comportement touche alors les sujets qui présentent une habileté élevée et qui n'auraient pas obtenu de bonnes réponses à des items qu'ils devraient normalement maîtriser.

Plusieurs approches ont déjà été développées pour détecter les individus qui répondent au hasard ou qui sont inattentifs dans une situation de testing (Zickar & Drasgow, 1996). Parmi celles-ci, l'utilisation d'indices de détection (*person-fit indexes*) de patrons de réponses inappropriés est certainement celle qui est la plus étudiée et qui semble la plus prometteuse (Karabatsos, 2003 ; Sijtsma & Meijer, 2001).

La majorité des auteurs s'entendent pour déclarer que les indices de détection de patrons de réponses inappropriés peuvent être classés en deux grandes catégories : les indices qui ne reposent pas sur les paramètres de la théorie de la réponse à l'item (appelés indices non paramétriques) et les indices qui s'y réfèrent (appelés indices paramétriques). Notons que la seconde catégorie d'indices présente plusieurs avantages, dont celui d'être plus faciles à interpréter puisque leurs scores suivent généralement les quantiles d'une distribution connue, par exemple la loi normale.

Dans le cadre de cet article, nous nous centrerons sur trois indices de détection paramétriques qui présentent des caractéristiques permettant d'en faciliter l'interprétation : l_z (Drasgow, Levine & Williams, 1985), ZU (selon Karabatsos, 2003) et ZW (selon Karabatsos, 2003). Toutefois, il s'est avéré que ces indices sont tous fortement modifiés par le fait que, l'habileté réelle d'un étudiant étant généralement inconnue, elle doit être estimée, ce qui a un impact non négligeable et biaisant sur le calcul des indices

de détection (Li & Olejnik, 1997 ; Molenaar & Hoijtink, 1990). Pour contrer cet effet, Snijders (2001) a proposé une version corrigée de l'indice l_z, utilisée dans les écrits scientifiques sous la notation l_z*, et qui règlerait ce problème.

Magis, Béland et Raîche (2014) ont déjà appliqué la correction de Snijders (2001) pour créer les indices modifiés U* et W*. Par contre, il reste à faire des analyses afin de mieux comprendre leur comportement. Dans ce dessein, nous examinerons le comportement des indices corrigés l_z*, U* et W* ainsi que des indices standardisés l_z, ZU et ZW à partir de deux études, soit : une analyse de leurs erreurs de type I empiriques et théoriques ainsi qu'une analyse de leur puissance de détection.

La suite de cet article est divisée en quatre sections. Dans la prochaine section, le cadre théorique est explicité. Ensuite, une deuxième section décrit le devis méthodologique du projet. La troisième section présente les résultats des deux études effectuées, soit l'analyse des erreurs de type I empiriques et théoriques et l'analyse de la puissance des indices. La discussion, dans la quatrième section, sera suivie par la conclusion.

Cadre théorique

Le calcul des indices de détection paramétriques nécessite l'utilisation d'un modèle probabiliste. Dans le cadre de cet article, nous sélectionnerons le modèle de Rasch (1960). Ce choix est basé sur le fait que cette modélisation est très utilisée dans l'industrie des tests psychométriques et dans les milieux universitaires. Mentionnons aussi que c'est une généralisation de cette modélisation, appelée en anglais le *mixed-coefficients multinomial logit model* (Adams & Wu, 2007), qui a été utilisée pour analyser les données du test PISA 2012.

En contexte scolaire, le modèle de Rasch à réponse dichotomique permet de calculer la probabilité qu'un répondant obtienne une bonne réponse à un item en se basant uniquement sur la difficulté de l'item. Supposons que x_i (x_i = 0 ou 1) représente la réponse dichotomique du répondant à l'item i. Mathématiquement, nous écrivons ce modèle comme suit :

$$P_i(\theta) = P(x_i = 1 \,|\, \theta, b_i) = \frac{\exp[\theta - b_i]}{1 + \exp[\theta - b_i]} \quad , \tag{1}$$

où θ est le paramètre d'habileté du sujet et b_i le paramètre de difficulté de l'item. Ce modèle présente l'avantage de mettre sur une même unité de mesure l'habileté du répondant et la difficulté de l'item. Cela facilite l'interprétation de ces paramètres puisqu'ils peuvent être construits sur l'algèbre des scores standardisés, ou scores z, à moyenne égale à 0 et à variance égale à 1. Enfin, les paramètres contenus dans l'équation 1 peuvent être estimés à l'aide de plusieurs méthodes qui ont déjà été présentées en détail par Baker et Kim (2004), par Hambleton et Swaminathan (1985) et par Lord (1980).

Les indices de détection paramétriques

À la base, le score produit par un indice de détection ne donne pas beaucoup d'information permettant de juger de la qualité d'un patron de réponses. Pour que ce score soit intelligible, le chercheur doit établir une valeur-seuil (*cut score*) qui servira à discriminer les patrons de réponses appropriés et les patrons de réponses qui ne le sont pas. Les écrits scientifiques ont démontré que les indices paramétriques sont plus faciles à interpréter, car ils sont beaucoup plus nombreux à présenter des valeurs-seuils connues (Sijtsma & Meijer, 2001).

Dans une situation où le score d'un indice suit les quantiles d'une loi de probabilité connue, il est possible d'interpréter ce score à l'aide d'un test d'hypothèses classique. Les sections suivantes présentent deux catégories d'indices qui peuvent être interprétés à partir d'un tel test d'hypothèses.

Les indices de type vraisemblance

Initialement présenté par Levine et Rubin (1979), l'indice l_0 calcule le maximum du logarithme népérien de la vraisemblance d'un patron de réponses :

$$l_0(\theta) = \log L(\theta) = \sum_{i=1}^{I} \left\{ x_i \log P_i(\theta) + (1 - x_i) \log Q_i(\theta) \right\}, \qquad (2)$$

où $Q_i(\theta) = 1 - P_i(\theta)$. Cette approche permet de déterminer jusqu'à quel point le vecteur de réponses examiné se conforme aux probabilités prévues par le modèle (par exemple, le modèle de Rasch). Dans la situation où l'étudiant répond conformément à l'un de ces modèles, la fonction de vraisemblance tendra à atteindre sa valeur maximale. À l'opposé, plus le

résultat de l_0 serait faible, plus le vecteur de réponses s'écarterait du modèle établi sur la base de l'ensemble des vecteurs normatifs, les résultats du répondant pouvant alors être considérés comme étant inappropriés.

Bien qu'intéressant, cet indice présente un problème de taille : son interprétation est difficile, car il n'est pas indépendant de l'habileté estimée d'un étudiant et il n'existe pas de valeur butoir à partir de laquelle un patron de réponses est considéré comme étant inapproprié. Pour ces raisons, Drasgow, Levine et Williams (1985) ont développé une version standardisée de l_0 : l'indice l_z. Mathématiquement, l_z est une mise en scores z des résultats obtenus par l'approche de Levine et Rubin (1979) :

$$l_z = \frac{l_0 - E(l_0)}{V(l_0)^{1/2}}. \tag{3}$$

Selon Drasgow et ses collaborateurs, cette transformation permettrait à l_z de se distribuer approximativement selon les quantiles d'une loi normale $N(0,1)$ lorsque les épreuves d'évaluation présentent suffisamment d'items. Par exemple, si le seuil de signification de la déviance du patron de réponses attendu α est fixé à 0,01, l'obtention d'une valeur inférieure au point de coupure -2,33 permettra de considérer que les réponses qu'a fournies un étudiant à une épreuve d'évaluation sont inappropriées.

Bien que cet indice soit le plus cité dans les recherches en mesure et en évaluation en éducation, son interprétation ne serait pas sans problèmes. En effet, Molenaar et Hoijtink (1990, 1996) ainsi que Nering (1995, 1997) ont déjà démontré que la distribution de cet indice ne suivrait pas exactement les quantiles d'une loi normale lorsque l'habileté θ d'un étudiant est estimée à l'aide de méthodes telles que la méthode du maximum de vraisemblance, la méthode d'estimation du maximum a posteriori ou la méthode du maximum de vraisemblance pondérée. De plus, cet indice est moins efficace dans le cadre de tests contenant un nombre limité d'items (Li & Olejnik, 1997).

D'un autre côté, l_z semble présenter une bonne puissance de détection. Par exemple, Drasgow, Levine et McLaughlin (1987) ainsi que Raîche (2002) et Raîche et Blais (2003) ont procédé à la comparaison de plusieurs indices et ils ont démontré que l_z avait une excellente puissance de détection. Karabatsos (2003) fait le même constat en comparant 36 indices : l_z est l'un des indices paramétriques qui présentent le plus haut pourcentage de détection.

Les indices de type carré moyen

Wright et Stone (1979) ont développé l'indice *U*, aussi appelé *outfit mean square* (littéralement « carré moyen déviant »), qui peut être noté comme suit :

$$U = \frac{1}{I-1}\sum_{i=1}^{I}\frac{[x_i - P_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}, \tag{4}$$

où $P_i(\theta)$ est une moyenne et $P_i(\theta)\,Q_i(\theta)$ une variance. Il est important de comprendre que *U* n'est pas une statistique pondérée, ce qui lui confère une grande sensibilité aux scores extrêmes par rapport à la difficulté de chacun des items ou à l'habileté des personnes.

Selon Karabatsos (2003), cet indice peut être standardisé en s'inspirant de la transformation Wilson-Hilferty :

$$ZU = \left(U^{1/3} - 1\right)\frac{3}{S_U} + \frac{S_U}{3}, \tag{5}$$

où

$$S_U = \sqrt{\frac{\sum_{i=1}^{I}\frac{1}{P_i(\theta)\,Q_i(\theta)} - 4I}{I^2}}. \tag{6}$$

Avec cette transformation, *ZU* devrait asymptotiquement suivre les quantiles d'une loi normale.

L'indice *W* (Wright & Masters, 1982), aussi appelé *infit mean square*, est un indice pondéré qui donne plus de poids aux protocoles de réponses dont le score est près de la difficulté de chacun des items ou de l'habileté. Pour cette raison, plusieurs le préfèrent à *U*. Mathématiquement, *W* s'écrit de la façon suivante :

$$W = \frac{\sum_{i=1}^{I}[x_i - P_i(\theta)]^2}{\sum_{i=1}^{I}P_i(\theta)Q_i(\theta)}. \tag{7}$$

Encore ici, il est possible d'appliquer la même standardisation que pour *U* et ainsi d'obtenir *ZW* (Karabatsos, 2003) :

$$ZW = \left(W^{1/3} - 1\right)\frac{3}{S_W} + \frac{S_W}{3}, \tag{8}$$

où

$$S_w = \frac{\sum_{i=1}^{I} P_i(\theta)Q_i(\theta) - 4\sum_{i=1}^{I}[P_i(\theta)Q_i(\theta)]}{\sum_{i=1}^{I} P_i(\theta)Q_i(\theta)}. \quad (9)$$

Les indices ZU et ZW ont été étudiés dans quelques études. Ainsi, Drasgow, Levine et McLaughlin (1987), Li et Olejnik (1997) ainsi que Noonan, Boss et Gessaroli (1992) ont démontré que ZU et ZW ne suivent pas vraiment les quantiles d'une loi normale. Enfin, Al-Mahrazi (2003) a démontré que la puissance de détection des indices ZU et ZW était limitée. Il recommande même d'éviter d'utiliser ces indices puisqu'ils sont très fortement affectés par certaines caractéristiques du test (par exemple, le nombre d'items ou la valeur des paramètres d'items estimés).

Les solutions déjà proposées pour tenir compte du fait que l'habileté doit être estimée

Dans les écrits scientifiques portant sur les indices de détection de patrons de réponses inappropriés, Sijtsma et Meijer (2001) ont soulevé l'importance de chercher de nouvelles solutions afin de pallier le fait que l'habileté réelle d'un répondant étant inconnue, elle doive être estimée, et que l'utilisation d'un niveau estimé est un problème si la validité de l'indice de détection dépend des aléas de l'estimation. À ce jour, seuls quelques auteurs se sont penchés sur ce problème (Glas & Meijer, 2003 ; Raîche & Blais, 2005 ; Raîche, Magis, Blais & Brochu, 2012). À cause de sa flexibilité et parce qu'elle a été étudiée par quelques auteurs, nous nous concentrerons uniquement sur la correction proposée par Snijders (2001).

La correction de Snijders

Snijders (2001) a développé un indice, nommé l_z*, qui permettrait de corriger la moyenne et la variance de la distribution de l_z (voir l'équation 3). D'abord, cet auteur démontre que l_z peut être réécrit sous la forme simplifiée suivante :

$$\frac{W_l(\theta)}{V[W_l(\theta)]^{1/2}}, \quad (10)$$

où le numérateur est égal à :

$$W_I(\theta) = \sum_{i=1}^{I} [x_i - P_i(\theta)]\, w_i(\theta)\,. \tag{11}$$

Dans l'équation 11, $w_i(\theta)$ est un facteur permettant de pondérer les écarts entre la réponse à un item et la probabilité d'obtenir une bonne réponse à cet item : $x_i - P_i(\theta)$ Sachant que x_i est une variable de type Bernoulli, $W_I(\theta)$ présente une moyenne et une variance qui sont respectivement égales à :

$$E[W_I(\theta)] = 0 \quad \text{et} \quad V[W_I(\theta)] = I\, \sigma_I{}^2(\theta)\,, \tag{12}$$

où

$$\sigma_I{}^2(\theta) = \frac{1}{I} \sum_{i=1}^{I} w_i(\theta)^2\; P_i(\theta)\, Q_i(\theta) \tag{13}$$

(Snijders, 2001, p. 332-335). Ainsi, pour établir la forme que prend l_z à l'équation 10, nous devons fixer la pondération $w_i(\theta)$, qui est présentée à l'équation 11, à :

$$\sigma_I{}^2(\theta) = \frac{1}{I} \sum_{i=1}^{I} w_i(\theta)^2\; P_i(\theta)\, Q_i(\theta) \tag{14}$$

en déduisant que

$$l_0 - E(l_0) = W_I(\theta) \quad \text{et que} \quad V(l_0) = I\, \sigma_I{}^2(\theta) = V[W_I(\theta)]. \tag{15}$$

Enfin, la correction de Snijders (2001) de l'indice l_z peut s'écrire :

$$l_z{}^* = \frac{l_0(\theta) - E[l_0(\theta)] + c_I(\theta)\, r_0(\theta)}{\tilde{V}[l_0(\theta)]^{1/2}}\,, \tag{16}$$

où

$$c_I(\hat{\theta}) = \frac{\sum_{i=1}^{I} P_i{}'(\theta)\, w_i(\theta)}{\sum_{i=1}^{I} P_i{}'(\theta)\, r_i(\theta)}\,, \tag{17}$$

sachant que $r_0(\theta)$ dépend de la méthode d'estimation de l'habileté (par exemple, maximum de vraisemblance ou vraisemblance maximale pondérée) et que $r_i(\theta)$ dépend du modèle de réponse à l'item utilisé (par exemple, le modèle de Rasch), et où la variance présente au dénominateur peut s'écrire comme suit :

$$\tilde{V}[l_0(\theta)] = \sum_{i=1}^{I} \tilde{w}_i(\theta)^2 \; P_i(\theta)\, Q_i(\theta)\,, \tag{18}$$

sachant que la pondération modifiée est égale à :

$$\tilde{w}_i(\theta) = w_i(\theta) - c_l(\theta)\, r_i(\theta). \tag{19}$$

Malgré le fait que l'article de Snijders ait été publié il y a plus de 12 ans, très peu d'études ont été conduites de façon approfondie sur cette correction. Par exemple, Van Krimpen-Stoop et Meijer (1999) ont démontré que l_z et l_z* obtiennent des résultats comparables en contexte de testing adaptatif. De leur côté, Sijtsma et Meijer (2001) ainsi que De la Torre et Deng (2008) ont discuté de la pertinence de cette approche pour améliorer la détection de patrons de réponses inappropriés. Enfin, Magis, Raîche et Béland (2011) ont présenté un article didactique permettant de faciliter la compréhension de l'article initial de Snijders (2001), en plus d'en faire une analyse sur un ensemble de données en langues. Leurs résultats ont démontré que l_z* présente des qualités métriques supérieures à sa version non corrigée, l_z.

Les indices U et W* (Magis, Béland & Raîche, 2014)*

Comme nous l'avons déjà soulevé un peu plus haut, l'indice l_z peut être réécrit sous la forme suivante :

$$\frac{W_I(\theta)}{V[W_I(\theta)]^{1/2}}.$$

À l'aide de quelques manipulations algébriques, il est possible de transposer le format présenté à l'équation 10 aux indices U et W en fixant uniquement le poids $w_i(\theta)$ approprié (équation 11). Dans le cas de U, ce facteur de poids de l'équation devient :

$$w_i(\theta) = \frac{Q_i(\theta) - P_i(\theta)}{I\; P_i(\theta) Q_i(\theta)}, \tag{20}$$

alors qu'il prend la forme suivante pour l'indice de détection W :

$$w_i(\theta) = \frac{Q_i(\theta) - P_i(\theta)}{\sum_{i=1}^{I} P_i(\theta) Q_i(\theta)}. \tag{21}$$

Ensuite, l'indice U^* peut se réécrire sous la même forme que l_z^* (équation 16) :

$$U^* = \frac{U(\theta) - E[U(\theta)] + c_1(\theta)\, r_0(\theta)}{\tilde{V}[U(\theta)]^{1/2}} \tag{22}$$

et cette transformation est aussi applicable pour l'indice W^* :

$$W^* = \frac{W(\theta) - E[W(\theta)] + c_1(\theta)\, r_0(\theta)}{\tilde{V}[W(\theta)]^{1/2}}. \tag{23}$$

Tous les indices corrigés par Snijders (2001) sont censés être interprétés à l'aide des quantiles de la loi normale. Le lecteur intéressé trouvera plus de détails dans l'article de Magis, Béland et Raîche (2014). Leurs résultats démontrent que les indices corrigés sont généralement plus efficaces pour détecter la réponse au hasard et l'inattention que leur version non corrigée, respectivement U et W.

Méthodologie

À l'instar de la majorité des études qui ont porté sur les indices de détection de patrons de réponses inappropriés, nous avons sélectionné la simulation assistée par ordinateur pour étudier le comportement des indices standardisés l_z, ZU, ZW et des indices corrigés l_z^*, U^* et W^*. Il est à noter que les considérations éthiques ne s'appliquent pas dans le cadre de cette étude puisque nous ne faisons pas intervenir de participants humains.

Les informations générales

Nous générerons deux longueurs de test (30 et 80 items). De plus, nous utiliserons le modèle de Rasch pour estimer la probabilité d'obtention d'une bonne réponse à un item. Voici la procédure appliquée pour générer les paramètres de cette simulation Monte-Carlo. Premièrement, nous nous inspirons de certains éléments de la méthodologie de Van Krimpen-Stoop et Meijer (1999) en générant cinq valeurs de θ allant de -2 à 2, soit :

$$\theta = -2;\ -1;\ 0;\ 1;\ 2.$$

Par la suite, 15 000 patrons de réponses seront simulés par valeurs de θ. Ensuite, les 30 ou 80 paramètres de difficulté b_i seront générés par une pige au hasard dans la loi normale. Les paramètres θ seront estimés en utilisant la méthode du maximum de vraisemblance pondérée (Warm, 1989).

Les deux études

Les résultats seront répartis en deux études. Dans la première étude, les erreurs de détection de type I font l'objet d'examen, les protocoles de réponses ne comportant pas de patrons inappropriés. L'erreur de type I théorique est calculée à partir des seuils α de la loi normale. De son côté, l'erreur de type I empirique est calculée à partir des seuils α de la fonction de densité des scores des indices obtenus par simulation informatique. Les résultats seront présentés pour trois valeurs de α : 0,01; 0,05; et 0,1.

Dans la seconde étude, nous utiliserons la modélisation développée par Raîche, Magis, Blais et Brochu (2012) et qui est disponible dans la librairie irtProb (Raîche, 2014) du logiciel R afin de générer des patrons de réponses inappropriés. Notons qu'il existe d'autres méthodes telles que celle de Levine et Drasgow (1982), mais elles n'ont pas été retenues à cause de leur caractère artificiel. Deux paramètres de cette modélisation peuvent être utilisés pour générer des réponses au hasard et des réponses inattentives : les paramètres de pseudo-chance et d'inattention personnelle. Dans cet article, nous les utiliserons pour générer une valeur de pseudo-chance personnelle ($C = 0{,}3$) ainsi qu'une valeur d'inattention personnelle ($D = 0{,}3$). Nous rapporterons le pourcentage de détection de ces réponses inappropriées selon un seuil d'erreur α égal à 0,05.

Résultats

Les résultats des deux études sont présentés successivement dans cette section.

Étude 1

Le tableau 1 présente les résultats de l'étude des erreurs de type I sur des données avec 30 items. Au seuil $\alpha = 0{,}01$, nous observons que ce sont les indices corrigés par la méthode de Snijders qui présentent des erreurs de type I empiriques les plus près de la valeur attendue. De plus, c'est l'in-

dice W^* qui s'en approche le plus. De leur côté, les indices standardisés ont tendance à sous-estimer l'erreur de type I théorique ; ils seraient plutôt conservateurs.

Au seuil $\alpha = 0{,}05$, ce sont les indices l_z^* et W^* qui présentent les erreurs de type I empiriques les plus proches de la valeur théorique. Les indices standardisés, de leur côté, présentent des erreurs empiriques systématiquement en dessous de 0,05. Notons tout de même que l_z est l'indice standardisé qui s'approche le plus de l'erreur de type I théorique.

Enfin, au seuil $\alpha = 0{,}10$, ce sont encore les indices l_z^* et W^* qui offrent la meilleure approximation de l'erreur de type I empirique avec la valeur prescrite. De leur côté, les indices standardisés présentent des erreurs empiriques systématiquement en dessous de 0,10.

Tableau 1
Erreurs de type I empiriques (30 items)

	l_z	l_z^*	ZU	U^*	ZW	W^*
			$\alpha = 0{,}01$			
$\theta = -2$	0,000	0,019	0,005	0,012	0,000	0,010
$\theta = -1$	0,007	0,019	0,012	0,014	0,003	0,014
$\theta = 0$	0,012	0,016	0,021	0,025	0,006	0,013
$\theta = 1$	0,006	0,016	0,007	0,022	0,002	0,015
$\theta = 2$	0,001	0,015	0,001	0,021	0,000	0,016
			$\alpha = 0{,}05$			
$\theta = -2$	0,003	0,058	0,013	0,030	0,000	0,050
$\theta = -1$	0,029	0,060	0,018	0,029	0,019	0,056
$\theta = 0$	0,045	0,052	0,032	0,038	0,006	0,052
$\theta = 1$	0,023	0,052	0,026	0,056	0,017	0,052
$\theta = 2$	0,008	0,052	0,014	0,056	0,004	0,056
			$\alpha = 0{,}10$			
$\theta = -2$	0,010	0,096	0,035	0,050	0,002	0,093
$\theta = -1$	0,058	0,010	0,037	0,048	0,048	0,100
$\theta = 0$	0,084	0,094	0,052	0,059	0,080	0,094
$\theta = 1$	0,051	0,093	0,057	0,091	0,045	0,095
$\theta = 2$	0,022	0,088	0,042	0,088	0,016	0,099

Le tableau 2 présente les résultats pour l'étude des erreurs de type I pour 80 items. Au seuil $\alpha = 0{,}01$, ce sont les indices l_z^* et W^* qui présentent les erreurs de type I empiriques les plus près de l'erreur de type I théorique. De leur côté, les indices standardisés ont généralement tendance à sous-estimer l'erreur de type I théorique. Notons que la similitude entre les erreurs de type I est à son plus haut niveau lorsque $\theta = 0$.

Au seuil $\alpha = 0{,}05$, ce sont encore une fois l_z* et W* qui présentent les erreurs de type I empiriques les plus comparables à la valeur de l'erreur de type I théorique. Les indices standardisés présentent des erreurs empiriques systématiquement en dessous du seuil $\alpha = 0{,}05$.

Enfin, le seuil $\alpha = 0{,}10$ explicite le fait que les indices l_z* et W* présentent les plus grandes similitudes entre les erreurs de type I empiriques et l'erreur de type I théorique. De leur côté, les indices standardisés présentent des erreurs empiriques systématiquement sous la valeur de 0,05. Encore une fois, la similitude entre les erreurs de type I est à son plus haut niveau lorsque $\theta = 0$.

Tableau 2
Erreurs de type I empiriques (80 items)

	l_z	l_z*	ZU	U*	ZW	W*
			α = 0,01			
θ = -2	0,000	0,0131	0,008	0,027	0,000	0,010
θ = -1	0,003	0,0133	0,009	0,024	0,002	0,012
θ = 0	0,012	0,0128	0,010	0,016	0,008	0,012
θ = 1	0,003	0,0137	0,005	0,019	0,001	0,011
θ = 2	0,000	0,0123	0,003	0,025	0,000	0,010
			α = 0,05			
θ = -2	0,002	0,052	0,027	0,061	0,000	0,048
θ = -1	0,023	0,053	0,037	0,060	0,018	0,051
θ = 0	0,048	0,051	0,047	0,055	0,043	0,049
θ = 1	0,023	0,053	0,031	0,060	0,020	0,052
θ = 2	0,004	0,053	0,024	0,066	0,001	0,052
			α = 0,10			
θ = -2	0,008	0,097	0,056	0,094	0,002	0,095
θ = -1	0,053	0,096	0,070	0,096	0,047	0,095
θ = 0	0,093	0,097	0,090	0,096	0,089	0,096
θ = 1	0,060	0,097	0,069	0,104	0,051	0,096
θ = 2	0,014	0,089	0,058	0,110	0,007	0,100

Étude 2

La figure 1 présente les pourcentages de détection de la réponse au hasard pour 30 items et lorsque le paramètre de pseudo-chance personnelle C est égal à 0,3. Lorsque θ augmente, nous remarquons que le pourcentage de patrons détectés tend à diminuer. Cela est logique puisque les individus ayant une valeur θ élevée n'ont pas besoin de deviner : ils connaissent la bonne réponse à un item et, ainsi, l'impact du paramètre de pseudo-chance personnelle C est moins important, car moins d'items difficiles

sont générés à cause du protocole choisi dans cette étude. Les indices l_z, l_z* et W* présentent les pourcentages de détection équivalents aux valeurs $-2 \leq \theta \leq 0$. Ces indices sont aussi ceux qui présentent les pourcentages de détection les plus élevés à ces mêmes valeurs. Mentionnons que c'est W* qui présente les pourcentages de détection les plus élevés lorsque $\theta = 1$. Pour la valeur $\theta = 2$, c'est plutôt l'indice standardisé ZU qui présente le plus haut pourcentage de détection. Ce résultat contraste grandement avec le score de cet indice pour les valeurs $-2 \leq \theta \leq 1$ qui présentait les pourcentages de détection les plus faibles avec ZW.

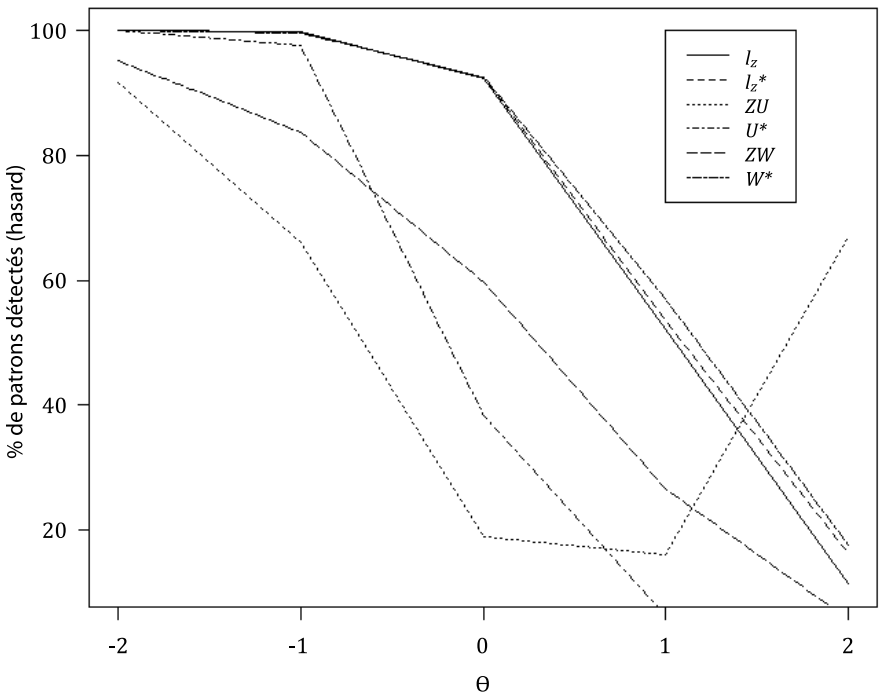

Figure 1. ***Puissance des indices à détecter la réponse au hasard à valeur C = 0,3, au seuil α = 0,05 et pour 30 items***

La figure 2 présente les pourcentages de détection de l'inattention pour 30 items lorsque le paramètre d'inattention personnelle D est égal à 0,3. La relation va dans le sens inverse de celle observée à la figure précédente : lorsque θ augmente, le pourcentage de patrons détectés tend aussi à augmenter. Cela est attendu puisque ce sont les individus ayant une valeur θ élevée qui sont touchés par l'inattention (ils échouent à une question, alors qu'ils devraient obtenir une bonne réponse).

L'indice ZU présente le plus haut taux de détection à la valeur θ = -2. Notons que ce sont les deux autres indices standardisés qui présentent les pourcentages de détection les moins élevés à cette valeur. À la valeur θ = -1, c'est l'indice corrigé W^* qui présente le pourcentage de détection le plus élevé. Aux valeurs $0 \leq \theta \leq 2$, les indices W^*, l_z et l_z^* présentent les pourcentages de détection les plus élevés. Notons de plus que leurs pourcentages de détection sont similaires pour ces valeurs θ.

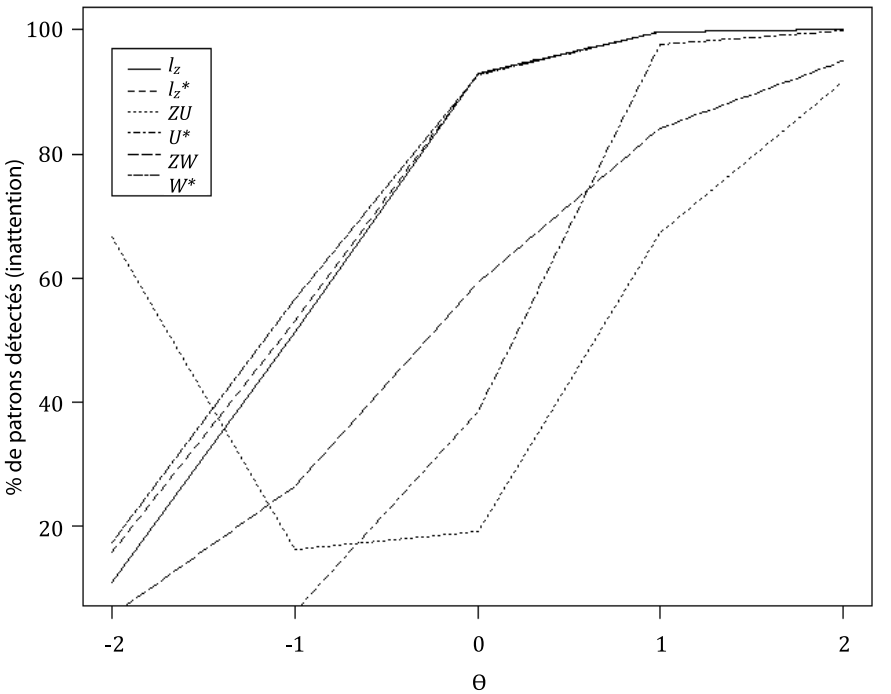

Figure 2. ***Puissance des indices à détecter l'inattention à valeur D = 0,3, au seuil α = 0,05 et pour 30 items***

La figure 3 présente les pourcentages de détection de la réponse au hasard pour 80 items lorsque C = 0,3. Nous observons que les pourcentages de détection sont légèrement plus élevés lorsque nous analysons des tests de 80 items plutôt que de 30 items. Comme dans la figure 1, nous observons que lorsque θ augmente, le pourcentage de patrons détectés tend à diminuer.

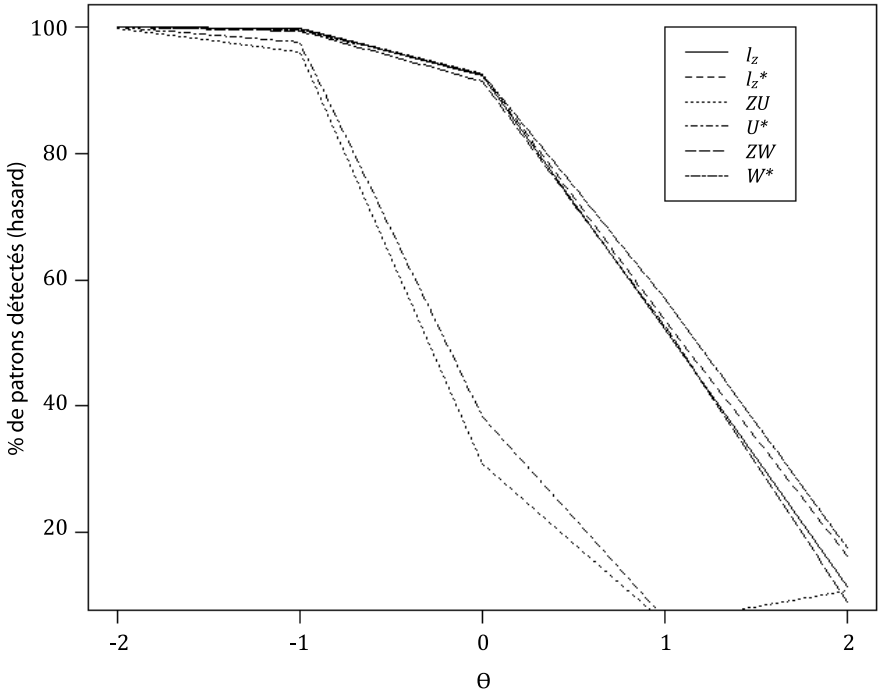

Figure 3. ***Puissance des indices à détecter la réponse au hasard à valeur C = 0,3, au seuil α = 0,05 et pour 80 items***

Les indices l_z, l_z* et *W** présentent des pourcentages de détection équivalents aux valeurs $-2 \leq \theta \leq 0$. Ces indices sont aussi ceux qui présentent les pourcentages de détection les plus élevés à ces valeurs de θ. C'est *W** qui présente les pourcentages de détection les plus élevés au seuil $\theta = 1$ et $\theta = 2$. De leur côté, les indices ZU et *U** présentent les pourcentages de détection les moins élevés.

La figure 4 présente les pourcentages de détection de l'inattention pour 80 items lorsque $D = 0{,}3$. Encore une fois, les pourcentages de détection sont généralement plus élevés lorsque le test comporte 80 items plutôt que 30 items.

L'indice *W** présente le plus haut taux de détection aux valeurs $\theta = -2$ et $\theta = -1$. Aux valeurs $0 \leq \theta \leq 2$, les indices W*, l_z et l_z* présentent les pourcentages de détection les plus élevés. Notons de plus que leurs pourcentages de détection sont similaires. Enfin, les indices ZU et *U** présentent des pourcentages de détection plus faibles que ceux des autres indices.

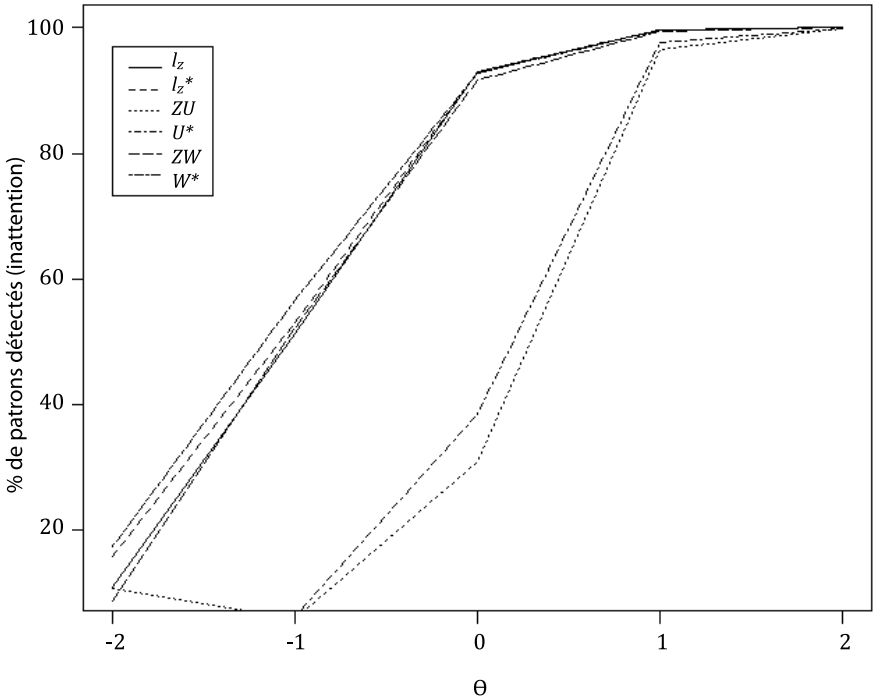

Figure 4. ***Puissance des indices à détecter l'inattention à valeur D = 0,3, au seuil α = 0,05 et pour 80 items***

Discussion

La discussion sera présentée pour chacune des deux études.

Étude 1

Nos résultats vont dans le même sens que ceux de van Krimpen-stoop et Meijer (1999). Ces auteurs ont obtenu des erreurs de type I bornées entre 0,04 et 0,07 pour l_z et des erreurs de type I bornées entre 0,07 et 0,09 pour l_z* (test de 20 items et $\alpha = 0{,}05$). Ces résultats, qui s'apparentent aux nôtres, montrent que les erreurs de type I s'approchent généralement du seuil α pour les indices de détection de type vraisemblance.

De la Torre et Deng (2008) ont produit une étude de comparaison de diverses approches au sein de laquelle se trouvait l_z*. Certains éléments de leur étude portant sur les erreurs de type I de cet indice sont rapportés au tableau 3.

Tableau 3
Certains résultats tirés de l'étude des erreurs de type I de lz*
(De la Torre & Deng, 2008, p. 167-168)

		$\alpha = 0{,}01$	$\alpha = 0{,}05$	$\alpha = 0{,}10$
30	$\theta = -2$	0,013	0,052	0,097
	$\theta = 0$	0,011	0,047	0,086
	$\theta = 2$	0,013	0,050	0,066
50	$\theta = -2$	0,012	0,051	0,093
	$\theta = 0$	0,012	0,048	0,094
	$\theta = 2$	0,014	0,054	0,024

Ces résultats montrent aussi de façon globale qu'il existe des similitudes entre les erreurs de type I empiriques et théoriques de l_z*. Nous avons observé des résultats comparables aux tableaux 1 et 2 de cet article.

Magis, Béland et Raîche (2014) ont montré que l'indice *W** présentait des erreurs de type I qui s'approchent du seuil α. Par exemple, pour des ensembles de données de 80 items analysés à l'aide du modèle de Rasch, ces auteurs ont calculé des erreurs de type I égales à 0,013 pour $\alpha = 0{,}01$, à 0,052 pour $\alpha = 0{,}05$ et à 0,097 pour $\alpha = 0{,}10$. Les résultats que nous avons obtenus dans des conditions comparables montrent que *W** a des erreurs de type I empiriques et théoriques similaires.

Dans le cas de *U**, Magis, Béland et Raîche (2014) ont obtenu des résultats comparables à ceux de *W** puisque les erreurs de type I de cet indice sont égales à 0,021 pour $\alpha = 0{,}01$, à 0,059 pour $\alpha = 0{,}05$ et à 0,097 pour $\alpha = 0{,}10$. Dans ce cas-ci, nos résultats n'ont pas permis d'obtenir des erreurs de type I s'accordant aussi bien avec les seuils α correspondants. Une piste d'explication permettant de comprendre ce résultat est liée au fait que la stratégie de génération de données adoptée par Magis, Béland et Raîche (2014) est différente de celle utilisée dans cet article.

Étude 2

Observons les résultats de deux études pertinentes. Karabatsos (2003) a étudié le taux de détection de la réponse au hasard des indices l_z, ZU et ZW. Ses résultats ont montré que ces trois indices présentent un taux de détection avoisinant 90%. Nos résultats ont démontré que l_z présente un résultat qui va dans ce sens pour $\theta = -2$ (30 items) et aux valeurs $\theta = -2$ (80 items) et $\theta = -1$ (80 items). Par contre, nous n'avons pas obtenu des

résultats aussi élevés que Karabatsos pour les indices ZU et ZW avec 30 items. Lorsque nous analysons 80 items, il est possible d'obtenir des résultats comparables à ceux de Karabatsos uniquement à la valeur $\theta = -2$ pour ZU et à la valeur $\theta = 1$ pour ZW. Dans toutes les autres situations, nos pourcentages de détection sont systématiquement plus faibles.

De leur côté, Magis, Béland et Raîche (2014) ont utilisé le modèle de Rasch pour analyser la détection de la réponse au hasard et de l'inattention. Dans toutes les situations d'analyse, nous observons que nos pourcentages de détection pour U^* et W^* sont plus élevés que ceux obtenus par ces auteurs.

Les limites

Au moins deux limites peuvent se dégager de nos analyses. Premièrement, la méthode de génération de la réponse au hasard et de l'inattention dans cet article diffère de celle des autres études que nous avons citées. Pour cette raison, nos résultats ne sont pas parfaitement comparables à ce qui a été fait ailleurs puisque c'est la première fois que cette modélisation est utilisée pour générer des patrons de réponses inappropriés. Cependant, cette approche est plus réaliste que celle adoptée dans d'autres études. En effet, la génération du hasard est directement intégrée dans le modèle utilisé, soit, celui de Raîche, Magis, Blais et Brochu (2012).

Deuxièmement, la posture adoptée ici est essentiellement descriptive. Cette stratégie nous semble justifiée, car nous avions d'abord l'obligation d'explorer le comportement des indices dans différentes situations de simulation. Il faudra donc raffiner plusieurs de nos résultats afin de mieux comprendre le comportement des indices à l'étude.

Conclusion

Nous avons utilisé des simulations Monte-Carlo pour étudier le comportement des indices de détection corrigés l_z*, U* et W* et des indices standardisés l_z, ZU et ZW. Les résultats des deux études réalisées peuvent être synthétisés comme suit. Dans l'étude 1, les résultats portant sur différentes valeurs de θ ont démontré que ce sont l_z* et W* dont les erreurs de type I empiriques respectent le plus la valeur théorique prescrite. Ensuite, l'étude 2 a démontré que l_z* et W* sont les indices qui présentent les pourcentages de détection les plus élevés dans la majorité des situations de simulation. Néanmoins, les résultats peuvent différer pour les valeurs θ extrêmes. Par exemple, ZU présente le plus haut pourcentage de détection pour la réponse au hasard lorsque $\theta = 2$ (30 items) et il présente aussi le plus haut pourcentage de l'inattention lorsque $\theta = -2$ (30 items). Rappelons que cet indice présentait l'un des plus faibles pourcentages de détection à toutes les autres valeurs de θ.

Ce projet a fait émerger quelques pistes de recherche intéressantes. Premièrement, nous avons uniquement concentré nos analyses sur trois indices standardisés et sur trois indices corrigés. Il serait pertinent d'étendre la méthode de Snijders (2001) à d'autres indices, tels que les indices de prudence (*extended caution indices*) de Tatsuoka et Linn (1983) et l'indice *Zeta* de Tatsuoka (1996).

Deuxièmement, nous avons généré des matrices de données qui devaient respecter les postulats de base des modèles de réponse à l'item. Comme nous savons que, dans la réalité, il est fréquent que ces postulats soient violés, il serait pertinent d'étudier la puissance des indices analysés en présence d'une multidimensionnalité de θ ou d'une dépendance locale entre les items.

Troisièmement, les résultats d'épreuves d'évaluation et de questionnaires de recherche en éducation comportent souvent des non-réponses de la part des étudiants. Ce problème peut survenir lorsqu'un étudiant a omis de répondre à une section d'un test ou lorsqu'il a volontairement décidé de ne pas répondre à certaines questions. À l'exception de Zhang et Walker (2008), très peu d'études ont tenté d'analyser la puissance des indices en présence de réponses manquantes. Ainsi, il serait pertinent de vérifier le comportement des indices corrigés dans un tel contexte.

Réception : 20 juin 2015
Version finale : 14 janvier 2016
Acceptation : 4 février 2016

RÉFÉRENCES

Adams, R. L., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalised form of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 57-75), New York, NY: Springer.

Al-Mahrazi, R. (2003). *Investigating a new modification of the residual-based person fit index and its relationship with other indices in dichotomous item response theory* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

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

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

De la Torre, J., & Deng, W. (2008). Improving person fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, *45*, 159-177. doi: 10.1111/j.1745-3984.2008.00058.x

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59-79. doi: 10.1177/014662168701100105

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polytomous 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

Emons, W. H. M., Glas, C. A. W., Meijer, R. R., & Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement*, *27*, 459-478. doi: 10.1177/0146621603259270

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person-fit analysis in item response theory models. *Applied Psychological Measurement*, *26*, 217-233. doi: 10.1177/0146621603027003003

Hambleton, R. K., & Swaminathan, H. (1985). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications.

Hendrawan, I., Glas, C. A. W., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement*, *29*, 26-44. doi: 10.1177/0146621604270902

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

Kogut, J. (1987). *Detecting aberrant item response patterns in the Rasch model.* Research report No. 87-3. Enschede, Netherlands: University of Twente.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, *35*, 42-56. doi: 10.1111/j.2044-8317.1982.tb00640.x

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269-290. doi: 10.3102/10769986004004269

Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, *21*, 215-231. doi: 10.1177/01466216970213002

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

Magis, D., Béland, S., & Raîche, G. (2014). Snijders's correction of Infit and Outfit indexes with estimated ability level: An analysis with the Rasch model. *Journal of Applied Measurement*, *15*, 82-93.

Magis, D., Raîche, G., & Béland, S. (2011). 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*, *37*, 57-81. doi: 10.3102/1076998610396894

Meijer, R. R., Muijtjens, A. M. M., & van der Vleuten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education*, *9*, 77-89.

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

Molenaar, I. W., & Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education*, *9*, 27-45.

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

Nering, M. L. (1997). The distribution of indexes of person-fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, *21*, 115-127. doi: 10.1177/01466216970212002

Noonan, B. W., Boss, M. W., & Gessaroli, M. E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indices. *Applied Psychological Measurement*, *16*, 345-352.

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

Raîche, G. (2014). *irtProb: Utilities and probability distributions related to multidimensional person item response models*. Retrieved from: https://cran.r-project.org/web/packages/irtProb/index.html

Raîche, G. & Blais, J.-G. (2003). Efficacité du dépistage des étudiantes et des étudiants qui cherchent à obtenir un résultat faible au test de classement en anglais, langue seconde, au collégial. Dans J.-G. Blais et G. Raîche (dir.), *Regards sur la modélisation de la mesure en éducation et en sciences sociales* (pp. 73-90). Saint-Nicolas, Québec : Presses de l'Université Laval.

Raîche, G., & Blais, J.-G. (July, 2005). *Characterization of the distribution of the l_z index of person fit according to the estimated proficiency level*. Paper presented at the 70th annual convention of the Psychometric Society, Tilburg, Netherlands.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Sijtsma, K., & Meijer, R. R. (2001). The person responses function as a tool in person-fit research. *Psychometrika*, *66*, 191-208. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/tocv66n2.pdf

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

Tatsuoka, K. (1996). Use of generalized person-fit indexes, Zetas for statistical pattern classification. *Applied Measurement in Education*, *9*, 65-76. doi: 10.1207/s15324818ame0901_6

Tatsuoka, K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, *7*, 81-96. doi: 10.1177/014662168300700111

Van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, *23*(4), 327-345. doi: 10.1177/01466219922031446

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

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Zhang, B., & Walker, C. M. (2008). Impact of missing data on person model fit and person trait estimation. *Applied Psychological Measurement, 32*, 466-480.

Zickar, M. J., & Dragow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, *20*, 71-87. doi: 10.1177/014662169602000107