Application de l'indice l_z pour l'élimination de données de recherche en langues

François Pichette
TÉLUQ – Université du Québec

Sébastien Béland
Université de Montréal

Gilles Raîche
Université du Québec à Montréal

Mots clés : indice l_z, élimination de données, technique de vérification de phrases, tests SVT

L'indice de détection de patrons de réponses inappropriés l_z (Drasgow, Levine & Williams, 1985) a été appliqué à un test d'habileté en lecture en langue seconde de 64 items soumis à 171 étudiants universitaires. L'objectif était de confronter un rejet intuitif de données de recherche à une élimination suggérée par l_z. En outre, l_z a été mis à l'épreuve pour détecter 12 participants additionnels ayant répondu par pseudo-hasard. Les résultats suggèrent que, bien que l_z détecte efficacement des patrons de réponses aberrants pour de grands groupes et qu'il soit préférable à l'élimination intuitive, cet indice présente des limites pour l'analyse de plus petites matrices de données.

Key words: l_z index, data elimination, sentence verification technique, SVT tests

With the intent to detect inappropriate response patterns, the l_z index (Drasgow, Levine & Williams, 1985) was applied to a 64-item test of second language reading ability administered to 171 university students. Our goal was to compare intuitive rejection of research data to data elimination suggested by l_z. In addition, l_z was challenged to detect 12 additional participants who had responded by pseudo-chance. Results suggest that although l_z detects efficiently aberrant response patterns for large groups and that it proves superior to intuitive rejection, that index has limitations when it comes to analyzing smaller data matrices.

Palavras-chave: índice l_z, eliminação de dados, técnica de verificação de frase, teste SVT

O índice de deteção de respostas inapropriadas l_z (Drasgow, Levine & Williams, 1985) foi aplicado a um teste de competência de leitura em segunda língua com 64 itens administrado a 171 estudantes universitários. O objetivo era confrontar uma rejeição intuitiva de dados de investigação com uma eliminação sugerida pelo l_z. Além disso, l_z foi testado para detetar 12 participantes adicionais que responderam de modo pseudo-aleatório. Os resultados sugerem que, apesar de l_z detetar eficazmente padrões de respostas aberrantes por grandes grupos e que é preferível à eliminação intuitiva, este índice tem limitações para a análise das matrizes mais pequenas de dados.

Note des auteurs : La correspondance liée à cet article peut être adressée à : [francois.pichette@teluq.ca], [sebastien.beland@umontreal.ca] et [raiche.gilles@uqam.ca].

Introduction

Les professeurs de langue doivent fréquemment tester leurs étudiants à des fins de placement au niveau de compétence approprié, de même qu'à des fins d'évaluation de l'apprentissage. En outre, ceux qui font de la recherche doivent également évaluer des habiletés langagières chez les participants à leur recherche et ils nourrissent les mêmes attentes que pour les situations d'évaluation des apprentissages. Ainsi, ils souhaitent non seulement que les participants possèdent l'habileté visée et fournissent l'effort attendu d'eux dans l'accomplissement de la tâche évaluée, mais aussi que leur performance mesurée reflète leur compétence. En bref, il est souhaité que la mesure de compétence soit valide.

Toutefois, lors de la compilation des données ainsi obtenues, il arrive fréquemment que certains des résultats soient considérés comme aberrants ; ils sont à la fois différents des attentes et des résultats obtenus par les autres personnes testées, et présentent un écart extrême par rapport à la valeur centrale. Lorsqu'un participant à une étude obtient des scores beaucoup plus élevés ou plus bas que les autres, le chercheur pourrait se limiter à éliminer, sur la base de considérations intuitives, ce type de données aberrantes.

Cette élimination de données jugées aberrantes est une opération très délicate. L'ampleur de l'élimination peut se situer n'importe où entre le rejet d'un participant sur 100 ayant quitté la salle sans terminer le test, jusqu'à la conservation d'une poignée seulement de participants après l'élimination « à l'œil » de données jugées trop éloignées de la moyenne. (Pour un survol récent des méthodes courantes d'élimination des données en langues, voir Pichette, Béland, Jolani et Leśniewska, 2015.) Quoi qu'il en soit, il semble difficile de défendre toute décision d'éliminer des données, quelles qu'elles soient, sans une méthode objective.

Le but de la présente étude est de confronter l'élimination intuitive de données de recherche à une élimination objective, sur la base d'un indice de détection de patrons de réponses inappropriés, dans ce cas-ci l'indice l_z (Drasgow, Levine & Williams, 1985).

Il importe en effet que les participants à une étude qui possèdent l'habileté nécessaire accomplissent de bonne foi et avec l'effort voulu les tâches attendues d'eux afin de mesurer ce qui est à mesurer. Il importe donc que les données recueillies soient fiables et valides. Or, malgré les précautions prises, il peut arriver que des participants prennent part à une étude dans de mauvaises dispositions, par exemple en croyant à tort que cela leur assurera des bénéfices futurs de la part du professeur qui les teste, pour ensuite répondre au hasard afin de terminer rapidement. Il peut aussi arriver qu'une personne mente quant aux prérequis exigés pour participer, sans quoi elle aurait pu être exclue de la recherche en raison de son profil qui aurait pu fausser les données recueillies. Ces situations, et bien d'autres semblables, font en sorte que certains participants qui n'auraient pas dû participer à l'étude ou qui n'auraient pas dû y participer de la façon dont ils l'ont fait peuvent venir fausser les résultats de l'étude. Ils doivent donc être exclus des analyses a posteriori. Une façon de remédier à ce problème serait d'identifier de telles personnes par l'intermédiaire de patrons de réponses aberrants.

Cet article sera divisé comme suit : le cadre théorique sera d'abord présenté, suivi des détails méthodologiques, des résultats obtenus et de la discussion. Enfin, nous terminerons l'article par une conclusion.

Cadre théorique

L'indice l_z appartient à une série d'indices de détection qui reposent sur l'application d'une modélisation probabiliste de réponse à l'item. Cette modélisation, s'inspirant de la modélisation par régression logistique, permet de calculer la probabilité $P_i(\theta)$ d'une bonne réponse d'un participant j à un item i. Lorsque possible, l'indice du participant sera omis. Par exemple, dans le cadre de la modélisation logistique à trois paramètres (Bertrand & Blais, 2004), nous pouvons représenter cette probabilité par l'équation suivante :

$$P_i(\theta) = \Pr(X_i = 1 \mid \theta, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \quad (1)$$

où θ est un paramètre d'habileté du participant, b_i un paramètre de difficulté de l'item, a_i un paramètre de discrimination de l'item et c_i un paramètre de pseudo-hasard de l'item. Il est à noter que la modélisation

associée à l'équation 1 se réduit au modèle à deux paramètres lorsque $c_i = 0$ et au modèle à un paramètre (aussi appelé *modèle de Rasch*) lorsque $c_i = 0$ et $a_i = 1$.

L'indice l_z découle de l'indice l_0 de Levine et Rubin (1979). En bref, l_0 calcule tout simplement le logarithme népérien de la probabilité d'un patron de réponses :

$$l_0 = \log L(\theta) = \sum_{i=1}^{n} \left\{ X_i \log P_i(\theta) + (1 - X_i) \log Q_i(\theta) \right\}. \quad (2)$$

Ici, $P_i(\theta)$ a déjà été donné en équation (1) et $Q_i(\theta) = 1\text{-}P_i(\theta)$. Ainsi, le logarithme népérien de la probabilité est calculé pour chacun des patrons de réponses. Malheureusement, cet indice présente un problème de taille : son interprétation est pratiquement impossible à établir, car aucune valeur-seuil ne lui est associée pour statuer sur le caractère approprié ou non des patrons de réponses. Pour cette raison, Drasgow, Levine et Williams (1985) ont élaboré une version standardisée de l_0: l_z. Mathématiquement, cet indice correspond à :

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

où $E(l_0)$ et $V(l_0)$ sont respectivement la moyenne et la variance de l_0. Puisque cet indice se distribuerait selon une loi normale centrée réduite (Drasgow, Levine & Williams, 1985), l'interprétation des résultats est aisée. Par exemple, si nous fixons le seuil de signification à 0,05, une valeur inférieure au point de coupure -1,64 permettra de détecter un patron de réponses comme étant inapproprié. À l'opposé, une valeur positive élevée démontre que le patron de réponses observé est tout simplement plus probable que le patron de réponses attendu : une situation qui n'est pas associée à un patron de réponses inapproprié.

Les raisons principales qui ont motivé le choix de l_z résident dans le fait qu'il est reconnu comme l'un des plus puissants et populaires indices statistiques pour détecter les patrons de réponses inappropriés (Drasgow & Levine, 1986 ; Li & Olejnik, 1997 ; Nering, 1997 ; Nering & Meijer, 1998 ; Raîche, Magis, Béland & Blais, 2011 ; Raîche, Magis, Blais & Brochu, 2013 ; Reise & Due, 1991). Par contre, il est important de signaler que cet indice permet seulement de détecter si un patron de réponses est approprié ou inapproprié. Ainsi, il est impossible de se prononcer sur la nature du comportement (p. ex., la fatigue ou le stress) ou de la stratégie adoptée

par le participant (p. ex., la tricherie volontaire ou le sous-classement intentionnel) ayant fourni un patron de réponses inapproprié. Le lecteur intéressé à en connaître davantage sur le sujet est invité à consulter Brassard (2011), Cronbach (1946), Meijer (1996), Johnson (1998) ou Ro (2001).

L'indice l_z a été utilisé dans plusieurs recherches en éducation (p. ex., Dodeen & Darabi, 2009 ; Meijer, 2003 ; Raîche, 2002 ; Reise & Flannerey, 1996). Dans certains cas, il a trouvé des applications dans la vie courante. Par exemple, Raîche (2002) l'a utilisé pour détecter les personnes qui obtiendraient volontairement de mauvais résultats dans un test de classement en anglais langue seconde pour obtenir ensuite facilement des notes élevées dans un cours qui serait alors très facile pour eux. De leur côté, Dodeen et Darabi (2009) ont appliqué l'indice l_z à une série de quatre tests de personnalité en mathématique. Ils soutiennent que cet indice permet de mieux comprendre le comportement des étudiants qui ont participé à la recherche. Enfin, Karabatsos (2003) a comparé 11 indices comparatifs (aussi appelés indices non paramétriques) et 25 indices dérivés des modèles de réponse à l'item (aussi appelés indices paramétriques). Ses analyses ont démontré que l_z figure parmi les indices dérivés des modèles de réponse à l'item les plus efficaces. Par contre, les indices comparatifs sont ceux qui ont présenté les plus hauts pourcentages de détection.

La présente étude vise à explorer l'application de l'indice l_z dans le contexte de l'élimination de participants à une recherche qui auraient répondu inadéquatement aux items d'une épreuve qu'ils ont passée. La présente étude poursuit deux objectifs principaux : (a) comparer l'élimination intuitive et l'élimination par l_z, et (b) examiner le potentiel que présente l_z pour détecter des participants ayant répondu par pseudo-hasard au test utilisé.

Méthodologie

Participants

Les participants à l'étude sont 183 étudiants universitaires bilingues anglais-français, soit 129 femmes et 54 hommes. De ce nombre, six seulement ont indiqué avoir l'anglais comme langue maternelle ; les autres ont tous indiqué le français comme langue maternelle. Après vérification, l'exclusion des six participants anglophones n'aurait aucun impact notable sur les statistiques présentées ici, c'est-à-dire que les alphas de Cronbach

et les indices l_z n'ont varié que légèrement à la seconde décimale, et que le pourcentage de données manquantes et les items identifiés par les indices n'en ont pas été affectés. Les données de ces six participants ont donc été conservées. Le niveau de compétence en anglais des francophones, comme indiqué par eux-mêmes, variait d'intermédiaire à avancé. Les débutants n'ont pu figurer parmi les participants puisque, en raison de leur compétence linguistique limitée, ils ne seraient pas en mesure d'effectuer les tâches de lecture demandées.

Instrument

Technique de vérification de phrases

Le test qui a servi à recueillir les données à analyser a été créé en 2008 pour mesurer l'habileté en compréhension en lecture de l'anglais (Pichette, Lafontaine & de Serres, 2009). Il est basé sur la technique de vérification de phrases (*sentence verification technique*, *SVT*; Royer, Hastings & Hook, 1979) utilisée traditionnellement par des enseignants pour vérifier la lisibilité de textes destinés à des groupes d'élèves. La structure habituelle du test, basé sur la technique de vérification de phrases développée par Royer et ses collaborateurs, consiste pour des étudiants à lire quatre passages de 12 phrases prélevés dans le document à utiliser, puis à indiquer de mémoire, après chacun de ces passages, si oui ou non des items sous la forme de phrases individuelles correspondaient au passage lu. Ce type de test a démontré au fil des années des alphas de Cronbach se situant entre 0,70 et 0,80 pour une version basée sur quatre passages totalisant 64 items, tout en faisant ressortir des corrélations de 0,50 à 0,73 avec des scores sur des tests normalisés de compréhension en lecture (Royer, 2004). Royer (2004) conclut que ces tests sont des outils efficaces qui mesurent ce qu'ils prétendent mesurer, car ils sont également sensibles à la variation en habileté de lecture et au degré de difficulté des textes lus, et car ils mesurent la compréhension du passage plutôt que de la phrase. L'historique de cette technique et les détails de nature psychométrique qui y sont associés sont présentés dans Royer (2004).

La technique de vérification de phrases devenant caduque en raison de l'essor des outils de mesure automatique de lisibilité, le principe a été recyclé dans la création d'un instrument pour mesurer l'habileté en compréhension en lecture de l'anglais. Le processus de mise sur pied suit des recommandations formulées par Royer pour créer un outil gratuit, plus simple et plus facile à faire passer que les tests normalisés actuels. À cette

fin, quatre textes en anglais d'intérêt général de 12 phrases chacun ont été choisis, modifiés et soumis à des échelles de lisibilité différentes pour s'assurer qu'ils présentent un intervalle suffisant de degrés de difficulté.

La méthode préconisée depuis les débuts de la technique de vérification de phrases consiste, pour son utilisateur, à créer 16 items pour chaque passage d'un test par technique de vérification de phrases. Ces items se répartissent en quatre catégories distinctes. D'abord, quatre des 12 phrases du texte sont paraphrasées, c'est-à-dire qu'elles transmettent le même sens avec des mots différents. Puis, quatre autres phrases font l'objet de changement de sens, c'est-à-dire que seuls un ou deux mots ont été changés, tout en affectant le sens entier de la phrase. Enfin, les quatre phrases restantes sont laissées intactes, puis quatre phrases plausibles sont ajoutées comme leurres. Dans le cadre de cette étude, au lieu d'opter pour une répartition égale des items dans les quatre catégories, les paraphrases et les changements de sens se retrouvent surpondérés, car il s'agit des plus discriminants lorsque la performance aux items associés à la technique de vérification de phrases est corrélée à celle obtenue à des tests de compréhension en lecture normalisés (Marchant, Royer & Greene, 1988). Le test utilisé comprend donc cinq paraphrases, cinq changements de sens, deux phrases intactes et quatre leurres. Dans les tests par technique de vérification de phrases, la difficulté peut se retrouver à deux niveaux différents : certains textes sont plus difficiles que d'autres, tandis que, à l'intérieur de chaque texte, certains éléments sont plus difficiles que d'autres. L'annexe présente un exemple de texte et de ses items accompagnateurs.

Technique de vérification de phrases modifiée

Comme l'indice l_z a déjà été suggéré pour identifier des étudiants pouvant ne pas avoir répondu correctement à des tests, nous avons décidé de mettre un peu plus à l'épreuve cet indice en nous demandant s'il serait capable d'identifier des personnes que le chercheur saurait avoir répondu au hasard. À cette fin, une version différente du test a été créée par l'enlèvement des quatre textes à lire, de sorte que le test ne renfermait que les titres des textes, suivis des items.

Nous aurions pu procéder autrement et demander simplement à des répondants supplémentaires de répondre *oui* ou *non* 64 fois de suite en l'absence à la fois des textes et des items, ce qui nous aurait fourni des patrons de réponses que nous aurions pu imputer au hasard. Toutefois, dans un test comme celui qui a été développé, les répondants arrivent certes

à choisir leurs réponses en vertu de leur compréhension du texte lu, habileté que le chercheur prétend mesurer et qui découle d'une série de processus cognitifs complexes (Bernhardt, 1991 ; Giasson, 2007 ; Grabe, 2009). Or, en plus de la compréhension du texte, d'autres habiletés, facteurs et processus peuvent guider la personne testée dans le choix des réponses fournies, entre autres les inférences et recoupements basés sur les items en présence, de même que des habiletés, facteurs et processus semblables qui sous-tendent l'habileté à faire des tests, quelle qu'en soit la nature (Cohen, 1992-1993 ; Scharnagl, 2005). Ainsi, comme il importe de limiter au maximum la différence entre les personnes testées avec cette méthode et l'ensemble des participants, seuls les textes ont été éliminés, ce qui mène à des patrons de réponses auxquels nous associons le terme de *pseudo-hasard* plutôt que celui de *hasard* puisqu'ils découlent des processus cognitifs évoqués précédemment. Le pseudo-hasard sous-tendu par cette modification du test nous a semblé plus approprié que le hasard pur, que nous aurions pu facilement obtenir par des moyens mécaniques ou autres.

Déroulement

Ce nouvel instrument a été soumis en premier lieu à 171 de nos participants, qui y ont répondu sur une copie papier dans une salle de classe. La durée moyenne de passation du test fut de 21 minutes, avec un éventail de durée de 16 à 24 minutes. Par la suite, la version tronquée du test en format électronique a été soumise à 12 participants supplémentaires au profil semblable au reste des participants, c'est-à-dire des francophones universitaires locuteurs d'anglais langue seconde de niveau de compétence non débutant.

Méthode d'analyse des résultats

Une fois le test effectué par les participants, nous avons dû composer avec des données manquantes, qui étaient de l'ordre de 5% des données recueillies. Ces données manquantes ont été traitées comme étant de mauvaises réponses et se sont vu assigner un score de zéro. Il est à noter que ce type de traitement a aussi été adopté par Raîche (2002).

Le test de vérification de phrases modifié a été calibré à l'aide du modèle de réponse à l'item à deux paramètres. Outre l'estimation du niveau de difficulté de chacun des items, ce modèle permet de donner un poids différent à ceux-ci à l'aide d'un paramètre de discrimination. Le modèle à trois paramètres n'a pas été retenu à cause de problèmes d'estimation du

paramètre de pseudo-hasard, surtout avec les petits échantillons, comme ici. L'estimation des paramètres d'item et du paramètre d'habileté des participants a été conduite à l'aide de la librairie irtoys, disponible dans le logiciel R (Partchev, 2011). C'est la méthode du maximum de vraisemblance qui a été retenue pour estimer θ et la méthode du maximum de vraisemblance conjoint pour les paramètres a_i et b_i. Enfin, nous avons généré nous-mêmes le code[1] pour calculer l'indice l_z.

Considérations éthiques

Le projet de recherche a été approuvé par les comités d'éthique de la recherche avec les êtres humains des deux universités où les participants ont été testés. Les participants ont été informés au préalable des détails de la recherche : objectifs, déroulement, durée prévue, participation volontaire et droit de retrait en tout temps, etc. Ils ont tous signé un formulaire de consentement à cet effet avant la collecte des données. Ils ont été assurés par écrit de la confidentialité dans la gestion des données, qui implique entre autres que les données sont anonymisées et qu'aucun nom ne paraîtra dans aucun rapport. Le fait pour les participants de fournir leur nom était d'ailleurs optionnel.

Résultats

Comparaison entre élimination intuitive et élimination par l_z

Un type d'élimination relevé dans les recherches en langues consiste à exclure des analyses les participants ayant obtenu des scores trop faibles, sans égard au profil de données qu'ils présentent, par exemple en excluant les données situées au-dessus d'un certain nombre ou pourcentage d'erreurs (p. ex., Borghi, Glenberg & Kaschak, 2004 ; Glenberg et al., 2008 ; Guasch, Sanchez-Casas, Ferre & García-Albea, 2011 ; Yanguas, 2009), ou les participants qui présentent un certain écart par rapport à la moyenne des autres participants (p. ex., Bolger, Balass, Landen & Perfetti, 2008 ; Pothos, Chater & Ziori, 2006). Ainsi, un collègue qui n'est pas impliqué dans la présente recherche a éliminé 10 des 171 premiers participants de façon intuitive en excluant les participants ayant des résultats anormalement bas ou ceux qui ont un score élevé pour les textes difficiles et bas pour les textes faciles. Après l'élimination des données de ces 10 personnes,

nous obtenons, pour les 161 participants restants, un coefficient α de Cronbach de 0,68, ce qui suggère un niveau de fidélité moyen pour notre test ; avant cette élimination, ce coefficient était égal à 0,83.

Une fois l_z appliqué à nos données pour l'ensemble des 171 participants, un premier exercice à faire est de scruter l'élimination intuitive sous la loupe de l_z. Parmi les paramètres, l'indice l_z considère l'habileté des participants, exprimée sous la forme de scores z. Cette habileté est estimée en utilisant l'approche de vraisemblance maximale. Pour nos données, nous obtenons une étendue de l'estimation des niveaux d'habileté se situant de -3 à +4. Le tableau 1 fait ressortir un fort chevauchement entre les participants à l'habileté la plus faible et l'identité des personnes éliminées à l'œil : parmi les huit participants les moins habiles, nous retrouvons sept des 10 personnes éliminées intuitivement.

Tableau 1
Huit participants retirés intuitivement par ordre croissant des notes

Note	θ	N° de participant	Élimination intuitive	Erreurs de Guttman	l_z
36	-2,91	P166	√	402	-1,60
48	-2,71	P167	√	532	-0,44
39	-2,61	P89	√	228	0,84
48	-2,24	P64	√	241	0,67
53	-2,08	P87	√	243	1,28
55	-1,80	P121		163	1,00
61	-1,60	P123	√	271	-0,27
58	-1,60	P107	√	347	-2,24

Par contre, parmi ces 171 participants, l'indice l_z ne suggère l'élimination que d'une des 10 personnes que l'intuition d'un collègue avait suggéré d'éliminer (participant 107) ainsi que de trois participants différents (participants 2, 157 et 162) pour lesquels le coefficient l_z de -1,64 (seuil de 0,05 dans un test unilatéral) et moins suggère un profil de réponses inapproprié à éliminer des analyses ultérieures.

Cette autre méthode d'identification des patrons de réponses aberrants a légèrement modifié les valeurs de l'alpha de Cronbach et de corrélation que nous avions obtenues pour nos tests. Alors que nous avions un α de Cronbach de 0,68 à la suite de l'élimination intuitive, celui-ci est plutôt de 0,83, donc considérablement plus élevé suivant l'élimination par l_z que l'élimination intuitive.

Réponses par pseudo-hasard à la lumière de l_z

Comme nous l'avions prévu, la moyenne au test a été de beaucoup inférieure en l'absence des textes à lire, se situant près du taux de chance avec une moyenne de 56,80% (écart-type de 6,20) et une étendue de 48,40 à 67,20.

Le second objectif de cette étude était de vérifier si l'indice l_z permet de détecter des participants ayant répondu par hasard ou par pseudo-hasard. Avec les données de ces participants intégrées aux matrices de données, comme l'indique le tableau 2, six des 10 personnes à exclure des analyses tel que le suggère l_z font partie des 12 personnes ayant fait le test sans avoir lu les textes. Les six autres patrons de réponses sont tout de même associés à des indices l_z négatifs ou à peu près nuls.

Au tableau 2, une analyse des erreurs de Guttman a aussi été menée a posteriori pour vérifier le rapport entre les bonnes réponses aux items difficiles et les mauvaises réponses aux items faciles, puis pour comparer celles-ci aux résultats obtenus à partir de l'indice l_z. Dans ce cas-ci, la différence entre les 12 participants additionnels et les 171 participants réguliers est encore assez claire, ces répondants en situation de réponse au hasard ayant obtenu des erreurs de Guttman trois fois plus fréquemment que la moyenne de ces erreurs pour les participants réguliers.

Tableau 2
Comparaison des coefficients de lz et du nombre d'erreurs de Guttman

N° de participant	l_z	**Erreurs de Guttman**
P1-P171 (moyenne)	0,02	125,30
H1	-1,03	329
H2	-3,05	492
H3	-1,63	459
H4	-4,71	495
H5	-0,93	371
H6	-0,20	345
H7	0,05	332
H8	-2,11	355
H9	-5,35	481
H10	-4,59	437
H11	-0,63	297
H12	-1,18	257
H1-H12 (moyenne)	-2,11	387,50

Note. P = participant régulier; H = participant au hasard.

Le tableau 3 ci-dessous montre que 11 des 12 personnes ayant répondu par pseudo-hasard se retrouvent parmi les erreurs de Guttman les plus élevées. Les neuf participants dont l'indice l_z avait suggéré l'élimination ainsi que neuf autres participants qui n'avaient pas été identifiés par l'indice l_z s'y trouvent aussi. Le nombre d'erreurs de Guttman confirme également le caractère inadéquat de l'élimination intuitive puisque seulement quatre des 10 éliminations par cette méthode semblent justifiées (P157, P107, P166 et P167). Dans ce cas-ci, la puissance de détection des erreurs de Guttman semble donc supérieure à celle de l'indice l_z.

Tableau 3
Erreurs de Guttman en tête de liste

N⁰ de participant	l_z	**Erreurs de Guttman**
…	…	…
P157	-1,68	294
H11	-0,63	297
P2	-2,40	299
H1	-1,03	329
H7	0,05	332
H6	-0,20	345
P107	-2,24	347
H8	-2,11	355
P94	-0,73	361
P162	-1,84	366
H5	-0,93	371
P166	-1,60	402
H10	-4,59	437
H3	-1,63	459
H9	-5,35	481
H2	-3,05	492
H4	-4,71	495
P167	-0,44	532

Note. P = participant régulier; H = participant au hasard.

Discussion

Comparaison entre élimination intuitive et par l'indice l_z

Tel qu'il est mis en lumière dans le tableau 1, l'élimination intuitive a permis la détection de sept des huit plus faibles participants, avec les trois autres éliminations qui les suivent de près dans l'échelle des habiletés en ordre croissant. Cette propension à éliminer les plus faibles suggère que l'élimination intuitive est grandement influencée par la performance globale des participants. Ainsi, pour éliminer un participant, il ne suffit pas qu'il ait obtenu un score faible dans un test facile combiné avec un résultat élevé dans un test difficile puisqu'un seul niveau de difficulté est considéré (celui du texte). En effet, même dans un tel cas, il est possible que, pour chaque texte, ce participant ait mieux réussi les questions faciles que les questions difficiles, ce qui veut dire que son patron de réponses n'est pas nécessairement si aberrant, malgré son score global faible ; du moins, pas assez aberrant pour justifier un rejet de ses données.

En second lieu, le tableau 2 fournit l'argument le plus convaincant pour utiliser une méthode dite objective (l_z ou autre) pour l'élimination de données de recherche. Un α de Cronbach de 0,68 découlant de l'élimination intuitive suggère que les estimations du niveau d'habileté à partir de notre instrument n'étaient pas suffisamment précises, ce qui aurait pu nous inciter à retravailler inutilement cet outil. Par contre, l'élimination statistique de données à partir de l'indice l_z a contredit cette impression en faisant ressortir un α de Cronbach respectable et rassurant de 0,83.

Réponses par pseudo-hasard à la lumière de l_z

Le fait qu'il existe une différence de détection entre l'approche selon le nombre d'erreurs de Guttman et l'indice l_z n'est pas vraiment surprenant : d'autres auteurs ont aussi établi que certains indices sont plus efficaces que d'autres afin de détecter des patrons de réponses spécifiques. Par exemple, Karabatsos (2003) a démontré que la réponse au hasard est généralement le comportement le plus facile à détecter. Par contre, ce dernier a obtenu des résultats opposés aux nôtres : dans son étude, l_z présentait un taux de détection plus élevé que le nombre d'erreurs de Guttman. Néanmoins, il est important de soulever que les données analysées par Karabatsos étaient créées artificiellement par ordinateur, alors que, dans le cadre de notre étude, nous analysions des données réelles. Il faut cependant bien comprendre que l'identification des patrons de réponses aber-

rants à partir des erreurs de Guttman et de l'indice l_z, telle qu'elle a été appliquée dans cette étude, n'est pas strictement comparable. En fait, la stratégie de détection à partir des erreurs de Guttman ne tient compte que du niveau de difficulté des items. L'indice l_z a été utilisé à la suite de l'estimation des paramètres d'items, selon une modélisation logistique qui introduisait un paramètre de discrimination. Cette situation fait en sorte que l'estimateur du niveau d'habileté est tributaire d'une pondération différente pour chacun des items, contrairement à ce qui est sous-jacent à la stratégie de détection selon les erreurs de Guttman. Considérant le faible nombre de répondants, il aurait possiblement été préférable d'utiliser la modélisation de Rasch avec l'indice l_z, en positionnant, d'une part, les 12 sujets en fonction des habiletés et, d'autre part, les items en fonction des difficultés, le tout sur un même continuum.

Examen d'indépendance des items grâce au pseudo-hasard

Une observation supplémentaire liée à l'instrument utilisé est permise ici par la nature de l'expérimentation menée. En présence d'un score moyen légèrement au-dessus du hasard avec un test de lecture semblable, un possible problème d'indépendance des items se pose parfois, soit des questions auxquelles le répondant peut répondre sans avoir lu le texte (p. ex., *Le Titanic transportait de nombreux passagers lorsqu'il a sombré*). Sans ce recours à des personnes ayant répondu de façon pseudo aléatoire, nous pourrions croire en la présence de ce phénomène, car certaines questions avaient été réussies par 98 % ou 99 % des participants. Ces 12 répondants par pseudo-hasard permettent de rejeter cette explication : les items qui ont été réussis par presque tout le monde n'ont pas eu le même succès pour ces 12 participants. À titre d'exemple, les items 49 et 54, qui ont présenté des scores respectifs de 98 % et 96 %, ont été manqués par la majorité des participants à la version modifiée du test : seuls trois participants sur 12 ont réussi l'item 49, tandis que seuls deux participants sur 12 ont réussi l'item 54. Conséquemment, malgré un score presque parfait pour certains items, ces derniers ne peuvent pas être réussis facilement sans que le participant ait lu le texte.

Conclusion

La présente étude visait à comparer l'élimination de participants de recherche de façon intuitive à leur élimination par l'indice l_z, et à examiner le potentiel de l_z à détecter des participants ayant répondu par pseudo-hasard. Un test de compréhension en lecture de l'anglais basé sur quatre textes a donc été soumis à 171 participants, puis une version sans les textes a été soumise à 12 participants supplémentaires afin d'obtenir des réponses par pseudo-hasard. L'indice l_z a permis de découvrir que l'élimination intuitive est fortement influencée par le niveau d'habileté des participants et que, en outre, dans ce cas-ci, elle tire fortement à la baisse l'alpha de Cronbach pour notre test.

L'élimination intuitive de données de recherche est une procédure controversée et peu fiable. En y recourant, le chercheur jette inévitablement un doute sur la valeur de ses données et des conclusions qu'il en tire. Il convient donc de recourir à une méthode statistique comme outil plus objectif d'identification des participants à exclure des analyses.

La réponse au hasard n'est pas facilement détectable. Du moins, elle ne semble pas l'être à l'aide de l_z : le nombre d'erreurs de Guttman ressort comme une stratégie plus efficace que l_z pour l'élimination de données de petits corpus. À cet effet, il importe de prendre certaines précautions à l'égard des données analysées dans le cadre de cette étude. Ainsi, le fait que nous ayons demandé à seulement 12 étudiants de volontairement répondre en l'absence des textes ne signifie pas qu'ils aient produit un patron de réponses typique d'un étudiant ayant répondu au hasard. Un échantillon de taille plus importante de répondants au hasard serait nécessaire pour étudier sérieusement la puissance comparative des deux approches de détection de patrons de réponses aberrants. À cet effet, Meijer (1996) a bien tenté de proposer un patron théorique de réponses au hasard, mais il reste du travail à faire avant de bien comprendre la nature de la réponse au hasard, voire de déterminer si le hasard pur existe vraiment lorsque des personnes font des tests.

Appliquée à des données réelles et non simulées, l'identification des personnes à exclure peut être supportée et confirmée par la connaissance qu'ont les chercheurs de leurs participants. Par exemple, dans notre cas, le chercheur principal savait que l'un des trois étudiants identifiés par l_z et que quatre des 10 étudiants qui affichaient le plus bas nombre d'erreurs de

Guttman n'ont pas l'habitude de faire les choses avec sérieux et qu'ils ont probablement participé à l'étude pour plaire, donc sans fournir d'efforts. Ce dernier point met en lumière l'importance d'utiliser à la fois des considérations quantitatives et qualitatives pour éliminer des données de recherche. Dans le cas d'études sans anonymat où il est possible de relier les scores aux participants, en connaissant les participants, le chercheur peut constater que l'élimination de l_z concorde avec le profil de ceux-ci, ce qui rassure le chercheur sur le fait que les individus à exclure verraient leurs données éliminées.

Toujours au sujet des erreurs de Guttman, le fait que les 12 participants ayant répondu par pseudo-hasard en aient obtenu trois fois plus que les participants réguliers semble soulever la question de la comparabilité du niveau de l'appariement entre les items et les individus à l'étude entre les deux groupes. Toutefois, il reste possible que cette différence soit imputable au faible nombre de répondants au hasard. Un nombre plus élevé de données liées aux réponses par pseudo-hasard permettrait d'élucider ce point.

Néanmoins, il faut garder en tête que l'efficacité de l_z dépend du type de test utilisé. La prudence est donc de mise dans le choix du bon indice. Dans ce cas-ci, les attentes voulaient que l_z soit approprié, en raison des items de difficulté variable du test, car plus il y a de variation de difficulté dans les items (ce qui est le cas ici), plus l_z sera cohérent. Par contre, cet indice s'est avéré moins efficace, probablement parce que le test est trop court et que le nombre de participants au hasard est trop limité. D'autres indices de détection pourraient constituer de meilleures solutions dans des circonstances différentes. Ainsi, il serait pertinent de tenter la même analyse en utilisant l'indice de Snijders (2001), qui corrige la distribution de l'indice l_z. Une autre avenue de recherche pertinente serait de reproduire les analyses en utilisant le processus de purification développé par Magis, Béland et Raîche (2013). Il serait aussi opportun d'appliquer un indice développé par Raîche (Raîche, Magis, Blais & Brochu, 2013) qui est spécifiquement destiné aux réponses au hasard. Dans le même sens, la supériorité du nombre d'erreurs de Guttman souligne l'intérêt d'analyser le test en utilisant d'autres indices de détection qui s'inspirent de cette approche. Enfin, l'élément le plus important de cette étude réside dans le fait qu'utiliser un indice de détection de patrons de réponses inappropriés permet de confirmer ou de corriger le jugement du chercheur.

Réception : 03 février 2016
Version finale : 03 avril 2016
Acceptation : 05 avril 2016

NOTE

1. Le code peut être obtenu en contactant le premier auteur.

RÉFÉRENCES

Bernhardt, E. B. (1991). *Reading development in a second language: Theoretical, empirical, and classroom perspectives*. Norwood, NJ: Ablex.

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.

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, *45*, 122-159. doi: 10.1080/01638530701792826

Borghi A., Glenberg A., & Kaschak, M. (2004). Putting words in perspective. *Memory and Cognition*, *32*, 863-873. Retrieved from http://scalab.cnrs.fr/CNCC09/PuttingWordsInPerspective.pdf

Brassard, P. D. (2011). *Identification des stratégies de sous-classement intentionnel aux tests de classement en anglais, langue seconde, au collégial* (Mémoire de maîtrise non publié). Montréal, Québec : Université du Québec à Montréal. Récupéré de http://www.archipel.uqam.ca/4275/

Cohen, A. D. (1992-1993). Test-taking strategies on language tests. *Journal of English and Foreign Languages*, *10-11*, 90-105.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*, 475-494. doi: 10.1177/001316444600600405

Dodeen, H., & Darabi, M. (2009). Person-fit: Relationship with four personality tests in mathematics. *Research Papers in Education*, *24*, 115-126.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement*, *10*(1), 59-67. doi: 10.1177/014662168601000105

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

Giasson, J. (2007). *La compréhension en lecture*. Paris : De Boeck.

Glenberg, A. M., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., & Buccino, G. (2008). Processing abstract language modulates motor system activity. *Quarterly Journal of Experimental Psychology, 61*, 905-919. doi: 10.1080/17470210701625550

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York, NY: Columbia University Press.

Guasch, M., Sanchez-Casas, R., Ferre, P., & García-Albea, J. E. (2011). Effects of the degree of meaning similarity on cross-language semantic priming in highly proficient bilinguals. *Journal of Cognitive Psychology, 23*(8), 942-961. doi: 10.1080/20445911. 2011.589382

Johnson, E. M. (1998). *A taxonomy of person misfit on affective measures* (Unpublished doctoral dissertation). Denver, CO: University of Denver.

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

Levine, M. B., & 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

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. doi: 10.7202/1024416ar

Marchant, H. G., Royer, J. M., & Greene, B. A. (1988). Superior reliability and validity for a new form of the Sentence Verification Technique for measuring comprehension. *Educational and Psychological Measurement*, *48*, 827-834. doi: 10.1177/0013164488483032

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3-8. doi: 10.1207/s15324818ame0901_2

Meijer, R. R. (2003). Diagnosing item score pattern on a test using item response theory-based person-fit statistics. *Psychological Methods*, *8*(1), 72-87. doi: 10.1037/1082-989X.8.1.72

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

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the l_z person fit statistic. *Applied Psychological Measurement*, *22*, 53-69. doi: 10.1177/01466216980221004

Partchev, I. (2011). Irtoys: Simple interface to the estimation and plotting of IRT models. *R package* (version 0.1.4). Retrieved from http://cRan.R-project.org/package=irtoys

Pichette, F., Béland, S., Jolani, S., & Leśniewska, J. (2015). The handling of missing binary data in language research. *Studies in Second Language Learning and Teaching*, *5*(1), 153-172. doi: http://dx.doi.org/10.14746/ssllt.2015.5.1.8

Pichette, F., Lafontaine, M., & de Serres, L. (2009). *A new tool for measuring L2 reading comprehension ability*. Paper presented at the 20th EuroSLA Conference, Cork, Ireland.

Pothos, E. M., Chater, N., & Ziori, E. (2006). Does stimulus appearance affect learning? *American Journal of Psychology, 119*(2), 277-301. Retrieved from http://www.dectech.co.uk/publications/LinksNick/CategorizationPerceptionAndMemory/Does%20stimulus%20appearance%20affect%20learning.pdf

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

Raîche, G., Magis, D., Béland, S. & Blais, J.-G. (2011). Conditions d'efficacité de la détection des patrons de réponses inappropriés lors de l'administration d'épreuves adaptatives. Dans J.-G. Blais & J.-L. Gilles (dir.), *Évaluation des apprentissages et technologies de l'information et de la communication : le futur est à notre porte* (pp. 339-354). Québec, Québec : Presses de l'Université Laval.

Raîche, G., Magis, D., Blais, J.-G., & Brochu, P. (2013). Taking atypical response pattern into account: A multidimensional measurement model from item response theory. In M. Simon, K. Ercikan & M. Rousseau (Eds.), *Improving large-scale education assessment* (pp. 238-259). New York, NY: Taylor & Francis.

Reise, S. P., & Due, A. M. (1991). Test characteristics and their influence on the detection of aberrant response patterns. *Applied Psychological Measurement*, *15*, 217-226. doi: 10.1177/014662169101500301

Reise, S. P., & Flannerey, W. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education*, *9*(1), 9-26. doi: 10.1207/s15324818ame0901_3

Ro, S. (2001). *Characteristics of a likelihood-based person-fit index under the graded response model* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Royer, J. M. (2004). *Uses for the Sentence Verification Technique for measuring language comprehension*. Amherst, MA: Reading Success Lab. Retrieved from http://www.readingsuccesslab.com/publications/Svt%20Review%20PDF%20version.pdf

Royer, J. M., Hastings, C. N., & Hook, C. (1979). A sentence verification technique for measuring reading comprehension. *Journal of Reading Behavior*, *11*, 355-363.

Scharnagl, T. L. (2005). *The effects of test-taking strategies on students' reading achievement*. (Unpublished doctoral dissertation). University of Michigan, Ann Arbor, MI.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331-342. doi: 10.1007/BF02294437

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology*, *13*(2), 48-67. Retrieved from http://llt.msu.edu/vol13num2/yanguas.pdf

Annexe

Lisez l'histoire suivante lentement et attentivement, une seule fois, en vous concentrant.

A special volunteer

Barkley, our dog, came to me when he was three years old after living with a family that could no longer take care of him.

I took him to visit the school for the blind where I worked as a teacher. He would walk over to the children and wait for a child to pet him.

One day, he started bumping into the walls of our house. When we played ball in the yard, I noticed that he could not catch it. I took him to the veterinarian, who found that he had an eye illness. Barkley had to have several operations. He soon learned to function with his weak eyes.

When he got better, he stood at the door, blocking my way, trying to tell me that he wanted to go to school with me and visit his friends. I started taking him to school again. Everyone was happy. Barkley was the happiest of all.

UNE FOIS LA LECTURE TERMINÉE,
TOURNEZ LA PAGE ET RÉPONDEZ AUX QUESTIONS.
NE REVENEZ PAS À L'HISTOIRE.

Lisez attentivement chacune des phrases suivantes, dont l'ordre peut être différent de celui du texte.

- *Écrivez « YES » si la phrase lue signifie la même chose que dans le texte.*
- *Écrivez « NO » si la phrase a un sens différent du texte, ou si cela n'a pas été dit explicitement dans le texte.*
- *Les mots n'ont pas à être les mêmes.*

1. I took him to visit the old-age home where I worked as a nurse.
2. I received Barkley, our dog, when he was three years old after he lived with people who could not take care of him anymore.
3. The dog would go near the children and wait to be petted.
4. Barkley was so happy that he pulled the leash on the way to school.
5. I took him to the veterinarian, who found that he had an eye illness.
6. One day, he started falling.
7. Barkley was not happy after the operation because he could not see his friends.
8. When we played outside, I realized that he could not catch the ball.
9. He never learned to cope with his sick eyes.
10. The vet had to operate on Barkley several times.
11. That dog was intelligent and eager to please.
12. When he got better, he stood at the door, blocking my way, trying to tell me that he wanted to go to school with me and visit his friends.
13. I started playing with him at school again.
14. Barkley was always very affectionate to the family with whom he lived.
15. The blind children were happiest of all.
16. All the kids were joyful.