Mesure et évaluation en éducation

La méthode de notation d'un questionnaire importe-t-elle ?

Christophe Chénier

Volume 40, Number 1, 2017

URI: https://id.erudit.org/iderudit/1041006ar
DOI: https://doi.org/10.7202/1041006ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

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

Explore this journal

Cite this article

Chénier, C. (2017). La méthode de notation d'un questionnaire importe-t-elle ? *Mesure et évaluation en éducation*, *40*(1), 129–154. https://doi.org/10.7202/1041006ar

Article abstract

Even though several scoring methods exist for scoring questionnaires, few studies have compared the potential effects of scoring methods on the correlations between the scores and other variables. This study seeks to fill this void, by comparing the correlation coefficients between the scores obtained with seven scoring methods from real datasets and, in lieu of available real data, eight randomly generated variables. The results show that the correlations are nearly identical and that no scoring method has a systematic effect on the strength of the resulting correlations. This result reinforces previous results and it is thus recommended that researchers favor the use of simple scoring methods able to handle missing data.

La méthode de notation d'un questionnaire importe-t-elle?

Christophe Chénier
Université du Québec à Montréal

Mots clés : méthodes de notation, scores factoriels, trait latent, score vrai

Bien qu'il existe plusieurs méthodes de notation pour assigner des scores aux répondants d'un questionnaire, peu d'études ont comparé les effets que pourraient avoir les méthodes choisies sur les corrélations entre les scores obtenus et d'autres variables. Cette recherche vise à combler ce manque en comparant les coefficients de corrélation entre les scores générés par sept méthodes de notation à partir de données réelles et, à défaut de données réelles accessibles, huit variables générées aléatoirement. Les résultats montrent que les corrélations sont presque identiques et qu'aucune méthode de notation n'a d'effet systématique sur la force des corrélations obtenues. Ce résultat est conforme aux résultats antérieurs et il est recommandé aux chercheurs de privilégier l'utilisation d'une méthode de notation simple et pouvant être utilisée avec des données manquantes.

Keywords: scoring methods, factor scores, latent trait, true score

Even though several scoring methods exist for scoring questionnaires, few studies have compared the potential effects of scoring methods on the correlations between the scores and other variables. This study seeks to fill this void, by comparing the correlation coefficients between the scores obtained with seven scoring methods from real datasets and, in lieu of available real data, eight randomly generated variables. The results show that the correlations are nearly identical and that no scoring method has a systematic effect on the strength of the resulting correlations. This result reinforces previous results and it is thus recommended that researchers favor the use of simple scoring methods able to handle missing data.

Palavras-chave: métodos de pontuação, pontuações fatoriais, traço latente, verdadeira pontuação

Embora existam vários métodos de pontuação para atribuir pontuações aos respondentes de um questionário, poucos estudos compararam os efeitos que poderiam ter os métodos escolhidos sobre as correlações entre as pontuações obtidas e

as outras variáveis. Esta investigação visa preencher esta lacuna através da comparação dos coeficientes de correlação entre as pontuações geradas por sete métodos de pontuação com base em dados reais e, na ausência de dados reais disponíveis, oito variáveis geradas aleatoriamente. Os resultados mostram que as correlações são quase idênticas e nenhum método de pontuação tem um efeito sistemático sobre a força das correlações obtidas. Este resultado é consistente com resultados anteriores e é recomendado que os investigadores privilegiem a utilização de um método de pontuação simples e que pode ser utilizado com dados ausentes.

Note de l'auteur : La correspondance liée à cet article peut être envoyée à : [chenier.christophe @uqam.ca].

Problématique

Un très grand nombre d'études en éducation ou en psychologie utilisent une méthodologie quantitative qui, peu ou prou, se réduit au schéma suivant : mesurer au moins une variable à l'aide d'un questionnaire psychométrique ou édumétrique et mettre en relation cette variable avec une autre variable d'une nature propre à chaque étude (Cohen, Manion & Morrison, 2007 ; Scheff, 2011 ; Trendler, 2009). Ce peut être une variable sociodémographique, biologique, l'appartenance à un groupe ou autre, mais les techniques statistiques utilisées sont souvent les mêmes (R. D. Howell, 2008). Dans bien des cas, ces techniques, qui appartiennent au modèle linéaire ou logistique généralisé, supposent que les variables sont continues et qu'il est possible de calculer moyenne et variance, les matrices de variance/covariance étant à la base de ces techniques (Tabachnick & Fidell, 2012). Évidemment, les résultats de ces techniques ne valent que si les valeurs des données sur lesquelles portent les calculs sont adéquates et mesurent réellement les quantités d'intérêt. C'est que l'instrument de mesure étant un questionnaire ou un test, les données brutes sont, en fait, des réponses que rien ne lie forcément à une valeur précise (Barrett, 2011 ; Michell, 2002). La méthode de notation de ces instruments est souvent arbitraire et, par conséquent, les valeurs obtenues ne vont jamais de soi en l'absence d'une unité de mesure prédéfinie (DiStefano, Zhu & Mindrila, 2009). De plus, différentes méthodes de notation plus ou moins complexes existent, chacune avec des avantages supposés différentiels (p. ex., Hartig & Höhler, 2008 ; Rotou, Headrick & Elmore, 2002 ; Zhang, 2010).

Il existe trois grandes familles de méthodes de notation : 1) les méthodes utilisant le score brut, pondéré ou non, 2) les méthodes d'estimation des scores factoriels, et 3) les méthodes d'estimation des traits latents. Les chercheurs doivent donc choisir une méthode. Ce choix peut théoriquement avoir une influence sur les résultats des analyses subséquentes puisque ce sont les scores obtenus à l'aide de cette méthode qui seront analysés. En dépit de ce fait, peu d'études justifient la méthode de notation des instruments utilisés. Qui plus est, il semble bien que plusieurs chercheurs n'aient pas les compétences ni l'intérêt requis pour l'utilisation des

processus plus complexes ou exigeant des logiciels spécialisés (Borsboom, 2006 ; Reise & Haviland, 2005 ; Streiner, 2010). Il est donc important de savoir si le choix de la méthode de notation, requis par l'utilisation de tests ou de questionnaires, peut mener à des valeurs relativement différentes. Les valeurs absolues sont bien entendu pratiquement toujours différentes puisqu'elles n'ont pas de réelles unités de mesure au sens strict du terme. Toutefois, puisque les techniques statistiques utilisées avec ces valeurs sont fondamentalement corrélationnelles (elles appartiennent au modèle linéaire général ou au modèle logistique général), ce sont les valeurs relatives qui importent (D. C. Howell, 2008 ; Tabachnick & Fidell, 2012). Il est nécessaire de souligner que cet article cherche à émuler la recherche telle qu'elle est pratiquée. Il ne s'agit donc pas de voir si une méthode de notation est supérieure à une autre, ou si elle se comporte, avec des données réelles, en conformité au modèle théorique sous-jacent, mais bien de reproduire les manières de faire de non-spécialistes afin de voir s'il y a lieu de s'inquiéter d'un large pan des résultats publiés dans les domaines susmentionnés. L'objectif général est, par conséquent, de déterminer si les méthodes de notation ont un impact sur les analyses subséquentes.

Contexte théorique

Bien qu'un grand nombre d'études aient comparé différentes méthodes de notation, c'est souvent avec un objectif différent de celui de ce travail. Ces études ont généralement cherché à comparer divers aspects des modèles théoriques expliquant les interactions entre les répondants et les questionnaires, modèles dont la méthode de notation n'est qu'un aspect parmi d'autres. Ces recherches ne visaient pas à savoir si la méthode de notation retenue a une influence sur les analyses ultérieures, mais plutôt à savoir si les scores obtenus ont certaines propriétés théoriques (invariance ou robustesse). Ainsi, Fan (1998) a comparé, avec des données réelles provenant d'items édumétriques notés de manière dichotomique, deux méthodes de notation provenant de la théorie classique des tests (nombre de bonnes réponses) et de la théorie de la réponse aux items (deux paramètres gradués), le tout afin de voir si les paramètres de personne (niveau d'habileté) étaient semblables. Les résultats obtenus ont montré que les paramètres de personne étaient très similaires et que les estimés étaient stables, même lorsqu'ils étaient obtenus à partir de sous-échantillons spécifiques. MacDonald et Paunonen (2002) ont essentiellement

cherché à faire la même chose, mais avec des données simulées. Leurs résultats sont similaires. Les estimés des paramètres de personne obtenus étaient fortement corrélés et les deux méthodes se sont révélées semblables. Zaman, Kashmiri, Mubarak et Ali (2008) ont cherché à savoir si les deux mêmes méthodes produisent des différences de rang dans le classement des répondants. Bien que les rangs des répondants ne soient pas strictement équivalents, la corrélation de 0,95 montre bien que les deux méthodes produisent des scores comparables.

Ces trois études ont les mêmes limites en ce qu'elles ne s'intéressent qu'aux deux mêmes modèles, et toujours avec des données dichotomiques provenant de tests édumétriques. Qui plus est, l'utilisation subséquente des scores n'est pas étudiée, ce qui diminue l'intérêt de ces travaux pour cette recherche-ci.

Certaines études ont toutefois poursuivi des objectifs similaires à celui de cette recherche. Xu et Stone (2011) ont comparé deux méthodes de notation, l'une issue de la théorie classique des tests et l'autre de la théorie de la réponse aux items, à l'aide de données simulées provenant d'items polychotomiques. L'objectif était de vérifier si les deux méthodes induisaient des différences dans le calcul subséquent d'un coefficient de corrélation avec une variable prédite. Les deux méthodes ont produit des résultats comparables. Notons néanmoins que cette étude ne compare que deux méthodes de notation, ce qui est peu. Ce sont les deux mêmes méthodes qui ont été utilisées par Ferrando et Chico (2007). Une partie de cette étude a utilisé la même méthode que Xu et Stone, soit la comparaison de la corrélation entre les scores produits par la théorie classique des tests ou par le modèle à deux paramètres de la théorie de la réponse aux items et une variable externe, soit les scores provenant d'un questionnaire psychométrique. Dans les deux cas, les corrélations sont pratiquement identiques (différence de 0,01 à 0,03) selon qu'elles incluent les scores provenant d'un modèle théorique ou de l'autre. De même, Dumenci et Achenbach (2008) ont étudié six méthodes de notation pour voir si les scores produits étaient comparables d'une méthode à l'autre. Les six méthodes sont : 1) le score total de la théorie classique des tests, 2-3) les scores factoriels de régression estimés à partir de deux méthodes d'analyse factorielle différentes, 4) des « scores factoriels de régression » estimés à partir d'une analyse en composante principale [*sic*], ainsi que deux modèles de la théorie de la réponse aux items, soit 5) le modèle à crédit partiel

et 6) le modèle de réponse graduée. Les scores ont été estimés à partir de données réelles provenant d'items polychotomiques de trois questionnaires psychométriques. Les résultats ont montré que les scores étaient fortement corrélés à l'intérieur de deux catégories, selon que les méthodes utilisées pour produire les scores tenaient compte ou non de la non-linéarité des données brutes. En revanche, les corrélations de rang entre les six méthodes sont presque parfaites, mais l'article ne mentionne pas le coefficient précis. Si cette recherche a l'avantage de comparer six méthodes de notation, les informations qu'elle donne ne sont pas très claires et une seule méthode d'estimation des scores factoriels est utilisée, ce qui restreint la portée de la recherche.

En résumé, presque toutes les études recensées se limitent à comparer une méthode de notation provenant de la théorie classique des tests (score total ou nombre de bonnes réponses) à une méthode provenant de la théorie de la réponse aux items (modèle de Rasch ou 2PL) et presque toujours avec des données provenant d'un test édumétrique, réel ou simulé. La seule étude ayant comparé davantage de méthodes ne fournit pas toutes les informations pertinentes dans ses résultats. Finalement, une seule étude recensée utilise des données provenant de tests psychométriques, ce qui importe parce que les scores de ces tests sont rarement normaux ; ils sont souvent asymétriques et leptokurtiques (Luo, 2011 ; Micceri, 1989). Les limites des recherches recensées justifient les objectifs spécifiques suivants : utiliser des données provenant de tests psychométriques pour 1) mesurer les corrélations entre les scores produits par des méthodes de notation différentes et 2) mesurer les corrélations entre les scores produits par des méthodes de notation différentes et diverses variables générées aléatoirement.

Méthodologie

Données réelles

Les données réelles de cette étude sont des données secondaires provenant de quatre sources distinctes. Puisqu'elles n'ont aucune importance pour cette étude, aucune information sociodémographique n'est incluse. Le premier ensemble de données provient de 242 étudiants universitaires ayant répondu, en 1994, au questionnaire *Beck's Depression Inventory I* et il est distribué publiquement par la bibliothèque *KernSmoothIRT* du

logiciel R (Mazza, Punzo & McGuire, 2015; Santor, Ramsay & Zuroff, 1994). Le deuxième ensemble est constitué des réponses de 2 800 personnes à 25 items de l'*International Personality Item Pool*, données collectées dans le cadre du projet *Synthetic Aperture Personality Assessment* (Revelle, 2016). Ces données sont distribuées publiquement par la bibliothèque *psych* du logiciel R (Revelle, 2016). Le troisième ensemble provient de 486 étudiants universitaires ayant répondu à un questionnaire mesurant le sentiment d'efficacité personnelle des enseignants (Dubé, Dufour, Chénier & Meunier, 2016; Dufour, Meunier & Chénier, 2014). Le quatrième ensemble de données est constitué de 755 étudiants d'un cégep ayant répondu à un questionnaire visant à mesurer les stratégies d'étude métacognitives et les facteurs de réussite scolaire (Beaulieu, De Sève & Provost, 2015).

Instruments

Le premier ensemble de données a été obtenu avec le questionnaire autodéclaré *Beck's Depression Inventory I*, constitué de 21 items correspondant à divers symptômes dépressifs (Beck, Steer & Garbin, 1988). Les répondants doivent indiquer l'intensité ou la fréquence des symptômes à l'aide d'une échelle de réponse à quatre degrés. Le pourcentage de données manquantes est de 0,17%. Vu le faible pourcentage, aucune imputation n'a été faite. Le deuxième ensemble a été obtenu avec une collection de 25 items de l'*International Personality Item Pool* mesurant les cinq dimensions du modèle théorique de la personnalité *Big Five* (Goldberg, 1999). Chacune des cinq dimensions est mesurée par cinq items et les répondants doivent indiquer leur degré d'accord avec chacun des énoncés à l'aide d'une échelle de réponse à six degrés. Le pourcentage de données manquantes est de 0,73% et aucune imputation n'a été faite. Le troisième ensemble a été collecté avec l'adaptation française du *Ohio State Teacher Efficacy Scale* (Tschannen-Moran & Woolfolk Hoy, 2001), adaptation faite par Ménard, Legault, Nault, St-Pierre, Raîche et Bégin (2011). Ce questionnaire mesurant le sentiment d'efficacité personnelle des enseignants comprend 24 items auxquels les répondants indiquent leur degré d'accord à l'aide d'une échelle de réponse à neuf degrés. Le pourcentage de données manquantes est de 1,17% et aucune imputation n'a été faite. Le quatrième ensemble de données a été obtenu à l'aide d'un questionnaire de 28 items mesurant les stratégies d'étude métacognitives, les facteurs socioaffectifs sous-jacents et les facteurs de réussite scolaire. Chaque

item a une échelle de réponse à quatre degrés avec laquelle le répondant indique la fréquence de comportements ou de pensées (Beaulieu et al., 2015). Il n'y a aucune donnée manquante.

Les données générées aléatoirement proviennent du générateur de pseudo-aléatoires Mersenne Twister du logiciel R. Ces données ont été simulées en fonction de huit distributions différentes : 1) une distribution normale centrée réduite, 2) une normale à moyenne 100 et à écart-type de 15, 3) une distribution dichotomique, 4-7) des distributions uniformes de nombres naturels à 4, 5, 6 et 7 valeurs et, finalement, 8) une distribution de nombres naturels à 9 (de 0 à 8) valeurs ayant une moyenne de 6,15, une médiane de 6, un mode de 7, un écart-type de 1,66, une asymétrie de -0,91 et un kurtosis de 0,43. Cette dernière distribution évoque les réponses à un questionnaire psychométrique typique, où les réponses se concentrent dans les échelons les plus élevés et où il y a un « effet de plafond ».

Déroulement

Le processus de notation et d'analyse de chaque ensemble de données a été fait selon les étapes suivantes : premièrement, une analyse factorielle exploratoire des données a été faite pour établir la structure dimensionnelle et pour sélectionner les items servant à obtenir un score. Ensuite, pour chaque dimension du questionnaire, sept scores ont été calculés ou estimés à l'aide des techniques décrites à la section suivante. Chaque répondant s'est vu assigner, au hasard, huit valeurs pour les variables simulées selon les distributions mentionnées à la section précédente. Il est important de rappeler que la méthodologie utilisée par cette recherche tente de reproduire les méthodes réellement utilisées dans les études publiées (voir p. ex. Fabrigar, Wegener, MacCallum & Strahan, 1999 ; Henson & Kyle Roberts, 2006). Cela explique certains choix autrement discutables. Par exemple, ni l'importance de l'indétermination des scores factoriels, ni les indices développés par Grice (2001) pour quantifier l'indétermination et ainsi évaluer la qualité des scores factoriels ne seront ici pris en compte, car ces éléments sont rarement mentionnés dans les études publiées en éducation ou en psychologie, ne serait-ce que parce que ces indices ne sont pas disponibles dans le logiciel SPSS.

Analyses

Pour les données réelles, les données brutes ont été traitées afin de vérifier le nombre de données manquantes ainsi que le respect des conditions d'utilisation de l'analyse factorielle et de la théorie de la réponse aux items. Pour ce faire, les statistiques descriptives et les représentations graphiques pertinentes ont été établies. Des analyses factorielles exploratoires avec extraction par vraisemblance maximale et rotation *oblimin directe* (3 ensembles de données sur 4) ou *varimax* (données du *Big Five*) ont été faites, et les matrices de structure ou de configuration ont été analysées pour sélectionner les items et les assigner aux dimensions mesurées. Les solutions à 1, 2, 3, 4 et 5 facteurs ont été comparées. Les critères de sélection ont été la cohérence des saturations factorielles et l'élimination des items ne saturant aucun facteur ou des items ayant des saturations croisées. Dans chaque cas, certains items ont été éliminés et les analyses factorielles ont été refaites pour s'assurer que les dimensions obtenues étaient cohérentes. Ensuite, sept méthodes de notation ont été utilisées afin d'obtenir les scores des répondants pour ce questionnaire. Il s'agit du calcul du score brut moyen, du score brut moyen pondéré selon les saturations factorielles, de l'estimation des scores factoriels selon les méthodes de Thurstone, de Bartlett et de ten Berge, de l'estimation du trait latent à l'aide du modèle unidimensionnel de Rasch et de l'estimation des traits latents à l'aide d'un modèle 2PL gradué de la théorie de la réponse aux items multidimensionnelle. Voici les informations pertinentes à chacune des méthodes.

Scores bruts moyens et scores bruts moyens pondérés

Le score brut moyen est la moyenne de la valeur de chaque item composant une dimension. Cette moyenne permet de calculer un score, qu'il y ait ou non des données manquantes, à condition qu'il y ait un nombre minimal d'items ayant une valeur. Le score brut moyen pondéré est calculé de la même manière, mais la valeur de chaque item est préalablement multipliée par la saturation factorielle de cet item (DiStefano et al., 2009).

Scores factoriels : méthode de Thurstone

Cette méthode estime les scores factoriels à l'aide d'une régression linéaire multiple par moindres carrés. La variable prédite est le score factoriel et les prédicteurs sont les valeurs des items de la dimension correspondante. Ces prédicteurs sont modifiés par un coefficient de régression dans lequel entrent les corrélations interitems, les saturations factorielles

et, pour les rotations obliques, les corrélations interfactorielles. Cette méthode maximise la corrélation des scores avec le facteur sous-jacent, mais elle est un estimateur biaisé du paramètre correspondant, soit le score factoriel de la population (Grice, 2001). Cette méthode ne peut pas être utilisée pour des répondants ayant une ou plusieurs données manquantes.

Scores factoriels : méthode de Bartlett

Les valeurs standardisées des items appartenant à un facteur sont pondérées par des coefficients de régression obtenus à partir de la matrice de profil des saturations factorielles et de l'inverse de la matrice de la variance des scores factoriels. Cette méthode est un estimateur non biaisé du paramètre correspondant, mais elle peut produire des scores factoriels corrélés même lorsque la rotation est orthogonale (Grice, 2001). Cette méthode ne peut pas être utilisée pour des répondants ayant une ou plusieurs données manquantes.

Scores factoriels : méthode de ten Berge

Les valeurs standardisées des items appartenant à un facteur sont pondérées par des coefficients de régression obtenus à partir de la matrice de profil des saturations factorielles et de l'inverse de la matrice de la variance des scores factoriels. Le tout est multiplié par une matrice obtenue, entre autres, à l'aide des matrices de valeurs propres et de vecteurs propres. Finalement, le tout est multiplié par la matrice de corrélations interfactorielles, ce qui permet l'application à des cas où la rotation est oblique. Cette méthode produit des scores factoriels préservant exactement les corrélations interfactorielles (Grice, 2001). Elle ne peut pas être utilisée pour des répondants ayant une ou plusieurs données manquantes.

Estimation du trait latent : modèle de Rasch

Le trait latent est estimé à partir d'un modèle de mesure stipulant que la probabilité qu'a un répondant *j* de choisir, pour un item *i*, la catégorie *g* par rapport à la catégorie *g - 1* dépend uniquement du niveau du trait latent de la personne *j* et de la difficulté de l'item *i*. Les valeurs de difficulté des items et des traits latents des répondants sont estimées à l'aide d'un processus itératif visant à trouver les valeurs maximisant la vraisemblance d'obtenir les patrons de réponses présents dans la matrice de données brutes. Comme le modèle est ici unidimensionnel, les traits latents correspondent aux dimensions identifiées à la première étape des analyses et les calculs sont faits avec un sous-ensemble des items (Andrich, 1978).

Estimation des traits latents : modèle multidimensionnel 2PL de la théorie de la réponse aux items

Dans ce cas, chaque répondant a non pas un seul, mais bien plusieurs traits latents, chaque trait correspondant à un construit, à une dimension mesurée par le questionnaire. Les *k* traits latents sont estimés à partir d'un modèle théorique stipulant que la probabilité qu'a un répondant *j* de choisir, pour un item *i*, la catégorie *g* par rapport à la catégorie *g - 1* dépend du niveau des *k* traits latents de la personne *j* et de la difficulté de l'item *i*, le tout multiplié par l'équivalent des saturations factorielles de l'item, soit les discriminations a_k de l'item *i*. Les valeurs des difficultés et des discriminations des items sont estimées à l'aide d'un processus itératif visant à identifier les valeurs maximisant la vraisemblance d'obtenir les patrons de réponses présents dans la matrice de données brutes. Puisque le modèle est ici multidimensionnel, l'ensemble des items est utilisé. Les valeurs des traits latents des répondants sont, elles, estimées par un algorithme de valeurs attendues a posteriori (*expected a posteriori*; Reckase, 2009).

Analyses entre les scores des répondants et diverses variables

Pour chaque dimension de chaque ensemble de données, sept scores ont été calculés ou estimés. Les corrélations linéaires de Pearson et de rang de Spearman ont été calculées entre ces sept scores eux-mêmes, et entre ces sept scores et les valeurs de chacune des huit variables générées aléatoirement. Ainsi, pour chaque dimension d'un ensemble de données, 112 coefficients de corrélation ont été calculés (7 méthodes de notation × 8 variables aléatoires × 2 types de corrélation). Pour chaque variable générée aléatoirement, la différence entre la corrélation maximale et minimale a été calculée. Pour l'ensemble des huit variables aléatoires, la moyenne des différences entre la corrélation maximale et minimale a été calculée, ainsi que l'écart-type de ces différences. Cela permet de quantifier la variance causée par les différentes méthodes de notation et de vérifier si certaines méthodes produisent systématiquement des corrélations plus fortes ou plus faibles avec les variables générées aléatoirement. Les deux types de corrélation ont été choisis en raison de certaines relations entre des paires de variables qui ne sont pas linéaires, mais bien curvilinéaires (p. ex., entre les scores bruts moyens et les scores estimés par le modèle de Rasch). D'autres ont des scores aux extrémités qui gonflent la corrélation linéaire, mais pas la corrélation de rang. Toutes les analyses ont été faites

avec les bibliothèques *psych* (version 1.6.4 ; Revelle, 2016), *mirt* (version 1.17.1 ; Chalmers, 2016) et *eRm* (version 0.15-6 ; Mair, Hatzinger, Maier & Rusch, 2015) du logiciel R, version 3.3.1 (R Core Team, 2016).

Considérations éthiques

Puisque toutes les données sont des données secondaires, aucune considération éthique particulière n'est requise. La permission d'utiliser les données provenant de recherches québécoises récentes a été obtenue auprès des chercheurs concernés. Les données obtenues par l'entremise des bibliothèques *psych* et *KernSmoothIRT* sont publiques et accessibles sans permission. Dans tous les cas, les données sont anonymes.

Résultats

Les résultats sont présentés en quatre sections correspondant aux quatre ensembles de données.

Les données du Beck's Depression Inventory I (BDI-I)

À la suite de plusieurs analyses factorielles exploratoires et de l'élimination de quatre items problématiques, une solution à deux facteurs a été retenue. Le modèle final a deux dimensions, composées de 13 et de 4 items ayant des saturations factorielles univoques, une corrélation interfactorielle de 0,53 et un pourcentage de variance expliquée de 39 %. Des scores ont donc été calculés ou estimés pour ces deux dimensions. Les corrélations linéaires de Pearson et de rang de Spearman entre les scores des deux dimensions sont présentées dans le tableau 1.

Les corrélations sont très élevées pour la première dimension et un peu plus faibles pour la seconde. Il n'y a qu'un seul cas où les deux types de corrélation sont relativement faibles, soit les corrélations entre les estimés du modèle de Rasch et les scores factoriels estimés par la méthode de Thurstone pour la seconde dimension. Il s'agit du seul cas où aucune corrélation n'atteint au moins 0,90.

Tableau 1
Corrélations de rang et linéaires entre les scores du BDI-I selon les méthodes de notation

	Bruts	**Pondérés**	**Thurstone**	**Bartlett**	**ten Berge**	**Rasch**	**TRIM***
Bruts		98/97	96/97	98/99	97/98	1/89	96/93
Pondérés	99/99		93/93	98/98	94/95	98/89	96/94
Thurstone	85/90	84/92		94/96	99/1	96/85	95/93
Bartlett	92/91	93/94	81/92		97/97	98/88	96/92
ten Berge	88/92	87/94	98/99	88/96		97/86	95/93
Rasch	1/99	99/98	85/88	92/87	88/90		96/95
TRIM*	93/94	93/94	92/89	84/85	89/88	93/95	

Note. * = Théorie de la réponse aux items. Les coefficients ont été multipliés par 100 pour faciliter la lecture, sauf les coefficients égaux à 1. Les coefficients au-dessus de la diagonale sont pour la première dimension, tandis que ceux sous la diagonale sont pour la seconde dimension. Les corrélations de rang sont à gauche de la barre oblique et les corrélations linéaires sont à droite.

Les coefficients de corrélation entre les scores et les variables générées aléatoirement n'ayant aucune importance en eux-mêmes, le tableau 2 présente certaines statistiques descriptives sur les écarts entre les coefficients de corrélation les plus élevés et les plus faibles entre les scores de chaque dimension et chacune des huit variables générées aléatoirement. Le tableau présente la différence maximale entre les coefficients obtenus entre les sept scores et une même variable générée aléatoirement, la différence minimale, la moyenne des sept différences et l'écart-type de cette moyenne.

La différence moyenne entre la méthode de notation produisant le coefficient le plus élevé et la méthode produisant le coefficient le plus faible est petite, et la plage des différences est restreinte. De plus, aucune méthode de notation ne produit systématiquement les coefficients les plus élevés ou les plus faibles. Par exemple, pour les scores de la première dimension, sur les 112 coefficients de corrélation calculés, la méthode de Rasch produit quatre fois le coefficient le plus élevé et quatre fois le coefficient le plus faible. Il n'y a donc pas de différence importante entre les coefficients obtenus à partir des scores générés à l'aide des sept méthodes de notation.

Tableau 2
Statistiques descriptives des valeurs de différence entre les coefficients de corrélation des données du BDI-I

	Corrélations de rang	Corrélations linéaires
Différence maximale	0,05/0,07	0,09/0,07
Différence minimale	0,01/0,02	0,02/0,03
Différence moyenne	0,03/0,04	0,05/0,05
Écart-type de la différence moyenne	0,01/0,02	0,03/0,01

Note. Les coefficients à gauche de la barre oblique sont pour la première dimension, tandis que ceux à droite sont pour la seconde dimension.

*Les données de l'*International Personality Item Pool (Big Five)

Il n'y a eu qu'une seule analyse factorielle exploratoire effectuée, et son résultat concorde avec le modèle attendu. Les 25 items ont bien saturé leur dimension respective et il n'y a eu aucune saturation croisée. La rotation varimax a mené à une solution à cinq dimensions expliquant 47% de la variance. Puisque la présentation des scores de cinq dimensions est complexe, voire impossible, les deux dimensions ayant les valeurs propres les plus élevées ont été retenues pour les analyses subséquentes, soit la dimension du névrosisme et de l'extroversion. Des scores ont été calculés ou estimés pour ces deux dimensions. Les corrélations linéaires de Pearson et de rang de Spearman entre les scores des deux dimensions sont présentées dans le tableau 3.

Les corrélations pour la dimension du névrosisme sont très élevées, à l'exception des corrélations linéaires entre les scores obtenus à partir du modèle multidimensionnel de la théorie de la réponse aux items et les autres scores. Pour la dimension de l'extroversion, ce sont les corrélations linéaires entre les scores obtenus à partir du modèle multidimensionnel de la théorie de la réponse aux items et les scores factoriels qui sont particulièrement faibles. Toutefois, cela n'a aucun effet sur les coefficients de corrélation entre les scores et les variables générées aléatoirement, comme l'illustre le tableau 4.

Les coefficients de corrélation entre les scores et les variables générées aléatoirement sont très homogènes, et la méthode de notation n'a aucun effet sur l'ampleur de ces coefficients.

Tableau 3

***Corrélations de rang et linéaires entre les scores de névrosisme et d'extroversion des données de l'*International Personality Item Pool (Big Five)**

selon les méthodes de notation

	Bruts	**Pondérés**	**Thurstone**	**Bartlett**	**ten Berge**	**Rasch**	**TRIM***
Bruts		1/1	96/96	95/95	96/96	1/95	98/89
Pondérés	1/1		98/97	97/97	97/97	1/95	99/90
Thurstone	92/92	92/93		1/1	1/1	96/91	98/89
Bartlett	87/88	88/88	99/99		1/1	95/90	97/88
ten Berge	90/90	90/91	1/1	1/1		96/91	97/89
Rasch	1/94	1/94	92/86	87/82	90/84		98/89
TRIM*	98/85	98/86	91/80	86/76	89/78	98/89	

Note. * = Théorie de la réponse aux items. Les coefficients ont été multipliés par 100 pour faciliter la lecture, sauf les coefficients égaux à 1. Les coefficients au-dessus de la diagonale sont pour la première dimension, tandis que ceux sous la diagonale sont pour la seconde dimension. Les corrélations de rang sont à gauche de la barre oblique et les corrélations linéaires sont à droite.

Tableau 4

***Statistiques descriptives des valeurs de différence entre les coefficients de corrélation des données de l'*International Personality Item Pool (Big Five)**

	Corrélations de rang	**Corrélations linéaires**
Différence maximale	0,03/0,03	0,03/0,03
Différence minimale	0,01/0,01	0,01/0,01
Différence moyenne	0,01/0,01	0,01/0,01
Écart-type de la différence moyenne	0,01/0,01	0,01/0,01

Note. Les coefficients à gauche de la barre oblique sont pour la première dimension, tandis que ceux à droite sont pour la seconde dimension.

*Les données de l'*Ohio State Teacher Efficacy Scale *(sentiment d'efficacité personnelle des enseignants)*

Les analyses factorielles exploratoires ont mené à une solution à deux facteurs composés de 15 et 7 items, 2 items ayant été éliminés. Les deux facteurs retenus expliquent 49% de la variance et la corrélation interfactorielle est de 0,68. Le tableau 5 présente les corrélations entre les scores obtenus à l'aide des sept méthodes de notation utilisées.

Tableau 5

Corrélations de rang et linéaires entre les scores de sentiment d'efficacité personnelle des enseignants de l'*Ohio State Teacher Efficacy Scale *selon les méthodes de notation

	Bruts	**Pondérés**	**Thurstone**	**Bartlett**	**ten Berge**	**Rasch**	**TRIM***
Bruts		99/1	99/99	98/98	98/99	1/95	99/98
Pondérés	1/1		1/1	99/99	99/1	99/95	99/98
Thurstone	99/99	99/99		1/1	1/1	99/95	99/99
Bartlett	98/99	99/99	99/1		1/1	98/94	98/98
ten Berge	99/99	99/99	1/1	1/1		98/95	99/98
Rasch	1/97	1/97	99/96	98/96	99/96		99/97
TRIM*	99/98	99/98	99/98	98/97	99/98	99/98	

Note. * = Théorie de la réponse aux items. Les coefficients ont été multipliés par 100 pour faciliter la lecture, sauf les coefficients égaux à 1. Les coefficients au-dessus de la diagonale sont pour la première dimension, tandis que ceux sous la diagonale sont pour la seconde dimension. Les corrélations de rang sont à gauche de la barre oblique et les corrélations linéaires sont à droite.

Pour cet ensemble de données, toutes les corrélations sont très élevées, soit presque égales à 1. Cette homogénéité se retrouve pour les coefficients de corrélation entre les scores et les variables générées aléatoirement, comme le montre le tableau 6.

Tableau 6
***Statistiques descriptives des valeurs de différence entre les coefficients de corrélation des données de l'*Ohio State Teacher Efficacy Scale**

	Corrélations de rang	Corrélations linéaires
Différence maximale	0,03/0,04	0,04/0,04
Différence minimale	0,01/0,01	0,01/0,01
Différence moyenne	0,02/0,02	0,02/0,02
Écart-type de la différence moyenne	0,01/0,01	0,01/0,01

Note. Les coefficients à gauche de la barre oblique sont pour la première dimension, tandis que ceux à droite sont pour la seconde dimension.

Dans tous les cas, la différence moyenne entre le coefficient le plus élevé et le plus faible est de 0,02, ce qui est minuscule.

Les données sur la métacognition

À la suite de l'élimination de 9 des 28 items initiaux, une solution satisfaisante à trois facteurs a été obtenue. Puisque seulement trois items saturaient la troisième dimension, cette dernière n'a pas été retenue pour les analyses subséquentes. Les deux facteurs retenus sont corrélés à 0,13 et ils expliquent 32% de la variance totale. Le tableau 7 montre les corrélations entre les scores donnés par les sept méthodes de notation.

Tableau 7
Corrélations de rang et linéaires entre les scores de métacognition selon les méthodes de notation

	Bruts	**Pondérés**	**Thurstone**	**Bartlett**	**ten Berge**	**Rasch**	**TRIM***
Bruts		98/98	98/99	99/99	98/99	1/96	98/98
Pondérés	97/98		99/99	99/99	99/99	98/94	97/97
Thurstone	95/96	97/97		1/1	1/1	98/94	99/98
Bartlett	96/96	98/98	98/99		1/1	99/94	99/99
ten Berge	96/96	98/98	1/1	99/1		98/94	99/99
Rasch	1/99	98/97	95/94	96/95	96/95		98/96
TRIM*	94/95	95/95	98/97	96/96	97/97	94/95	

Note. * = Théorie de la réponse aux items. Les coefficients ont été multipliés par 100 pour faciliter la lecture, sauf les coefficients égaux à 1. Les coefficients au-dessus de la diagonale sont pour la première dimension, tandis que ceux sous la diagonale sont pour la seconde dimension. Les corrélations de rang sont à gauche de la barre oblique et les corrélations linéaires sont à droite.

Les corrélations sont encore très élevées, égales ou supérieures à 0,94 dans tous les cas. Cette similarité entre les scores donne lieu à des corrélations très semblables entre les scores obtenus et les variables générées aléatoirement, comme en fait foi le tableau 8.

Tableau 8
Statistiques descriptives des valeurs de différence entre les coefficients de corrélation des données sur la métacognition

	Corrélations de rang	**Corrélations linéaires**
Différence maximale	0,04/0,03	0,04/0,03
Différence minimale	0,00/0,01	0,01/0,01
Différence moyenne	0,02/0,02	0,02/0,02
Écart-type de la différence moyenne	0,01/0,01	0,01/0,01

Note. Les coefficients à gauche de la barre oblique sont pour la première dimension, tandis que ceux à droite sont pour la seconde dimension.

Toutes les différences moyennes entre le coefficient le plus élevé et le plus faible sont de 0,02 et l'écart-type est de 0,01, ce qui est négligeable.

Résultats globaux

Si nous considérons maintenant l'ensemble des 64 valeurs (4 ensembles de données × 2 dimensions × 8 variables aléatoires) de différences entre le coefficient le plus élevé et le plus faible obtenues pour les corrélations linéaires, l'écart moyen entre le coefficient le plus élevé et le plus faible est de 0,03, et l'écart-type est de 0,02. Pour les corrélations de rang, l'écart moyen est de 0,02 et l'écart-type est de 0,01. Non seulement la différence entre le coefficient le plus élevé et le plus faible est, en moyenne, presque nulle, mais il n'y a aucun effet systématique de la méthode de notation sur l'ampleur des corrélations ultérieures. Ainsi, la méthode de notation dont le score donne le coefficient le plus élevé avec la variable QI, par exemple, peut très bien donner le coefficient le plus faible avec la variable uniforme à sept valeurs. Toutes les méthodes de notation ont produit à au moins une reprise la corrélation la plus faible et la plus élevée.

Discussion

Nos résultats montrent que, toutes choses égales par ailleurs, le choix d'une méthode de notation semble sans conséquence, du moins lorsque les scores ainsi obtenus sont uniquement utilisés comme variable dans une analyse statistique subséquente, et ce, en supposant que les scores obtenus représentent également bien les facteurs sous-jacents. Il y a trois raisons à cela.

La première est que l'écart maximal moyen entre les coefficients de corrélation imputable au choix de la méthode de notation est de 0,03 ou de 0,02, selon le type de corrélation, ce qui est vraiment peu. L'interprétation traditionnelle d'un coefficient de corrélation suppose qu'une corrélation est faible à 0,25, moyenne à 0,50 et forte à 0,75. (Ces valeurs sont approximatives, les recommandations changeant quelque peu d'un auteur à l'autre.) Un écart de 0,03 ne changera pas l'interprétation que fera le chercheur des corrélations obtenues.

La deuxième raison est que les valeurs des scores utilisées dans nos analyses sont des estimations ponctuelles qui font fi des erreurs types accompagnant ces estimés. Or, l'une des différences importantes entre les méthodes de notation réside en la manière dont elles traitent les erreurs

types des estimés (Hambleton, Swaminathan & Rogers, 1991 ; Lord & Novick, 1968). Certaines méthodes donnent des erreurs types variant en fonction des scores, alors que d'autres méthodes donnent typiquement une erreur type pour l'ensemble des scores. Cette différence entre les méthodes de notation est toutefois gommée par les modèles linéaires ou logistiques généraux qui ont comme postulat que les variables indépendantes sont réputées sans erreurs[1]. Ainsi, l'utilisation typique des scores générés par l'une ou l'autre des méthodes de notation se fait avec les seules estimations ponctuelles, sans erreur type associée, ce qui rend caduque toute différence entre les méthodes de notation.

La troisième raison est que les méthodes de notation n'ont pas d'effet systématique sur la force des corrélations, une méthode pouvant, à partir d'un ensemble de données, produire le coefficient de corrélation le plus élevé avec une variable aléatoire et le coefficient le plus faible avec une autre variable aléatoire, ces divergences étant inexplicables. Cela dit, ces résultats ne sont pas surprenants puisqu'aucun modèle théorique sous-jacent aux méthodes de notation utilisées ne prédit un avantage différentiel pour la validité prédictive.

Il faut toutefois restreindre la portée de nos résultats. Ainsi, il est possible que certaines méthodes de notation aient des problèmes avec des patrons de réponses inappropriés ou produisant des scores identiques. Certaines méthodes de notation exigent des données complètes (toutes les méthodes d'estimation des scores factoriels), tandis que d'autres s'accommodent de données manquantes. De même, les effets de plancher ou de plafond peuvent survenir avec certaines méthodes (Schaeffer, Henderson-Montero, Julian & Bené, 2002). Il est aussi important de souligner que ces résultats ne démontrent pas une invariance de la mesure ou de la prédiction (Borsboom, 2006 ; Rupp & Zumbo, 2004, 2006 ; Vandenberg & Lance, 2000).

D'autre part, cette étude confirme les nombreux résultats antérieurs obtenus quant aux corrélations très élevées entre les scores obtenus à partir de la théorie classique des tests et des scores estimés par la théorie de la réponse aux items (Adedoyin, Nenty & Chilisa, 2008 ; Courville, 2004 ; Fan, 1998 ; Hernandez, 2009 ; MacDonald et Paunonen, 2002 ; Magno, 2009 ; Zaman et al., 2008). Des corrélations élevées ont été mesurées dans tous les cas entre les scores obtenus par des méthodes relevant de ces théories. Les coefficients de corrélation obtenus ici sont similaires aux coeffi-

cients rapportés dans la littérature, généralement égaux ou supérieurs à 0,95. Ces résultats s'avèrent aussi lorsque les scores sont obtenus par estimation des scores factoriels, ce qui est intéressant si l'on considère les problèmes théoriques liés à l'estimation des scores factoriels, soit l'indétermination des scores ainsi que l'influence des choix liés à l'extraction des facteurs et à la rotation utilisée sur l'estimation des scores factoriels (DiStefano et al., 2009).

Pour terminer, quelle que soit la méthode utilisée, les données brutes contiennent toute l'information nécessaire à la notation — à l'exception des méthodes utilisant un estimateur bayésien, auquel cas l'a priori apporte une information supplémentaire. Que ces données servent seulement à calculer une moyenne ou qu'elles soient multipliées par des coefficients plus ou moins complexes, il n'en demeure pas moins que les méthodes de notation utilisent toutes la même information initiale, soit les données brutes des répondants. Il n'est dès lors guère surprenant de constater que les produits finaux sont très proches dans tous les cas. Si des chercheurs calculent des scores à partir de questionnaires psychométriques ou édumétriques dans le seul but d'utiliser ces scores dans d'autres analyses, ceux-ci devraient alors privilégier une méthode simple pouvant être utilisée avec des données manquantes, par exemple les scores bruts moyens. Nos résultats ne donnent toutefois pas un blanc-seing aux chercheurs désireux de se soustraire aux bonnes pratiques recommandées quant à l'utilisation de questionnaires édumétriques ou psychométriques, en particulier en ce qui concerne l'étude rigoureuse de la méthode de notation utilisée et de la qualité des scores ainsi obtenus.

Conclusion

Étant donné la très grande utilisation des questionnaires psychométriques dans la recherche en éducation et le nombre important de méthodes de notation utilisables pour produire des scores à partir de ces questionnaires, cette étude avait pour objectif de déterminer si les méthodes de notation ont un impact sur les analyses subséquentes faites avec les scores ainsi produits. Pour atteindre cet objectif, sept méthodes de notation différentes ont été comparées. Les scores ainsi produits ont été corrélés à huit variables générées aléatoirement. Les résultats ont principalement démontré que les scores générés par les différentes méthodes étaient fortement

corrélés entre eux et que les différences de corrélation entre les scores d'une même variable et une variable autre étaient négligeables, avec un écart maximal moyen de 0,03 ou 0,02 selon le type de corrélation.

Il y a toutefois de nombreuses limites à cette étude. Premièrement, cette étude ne porte que sur l'utilisation ultérieure des scores dans des analyses corrélationnelles. Elle n'apporte donc aucune information pertinente sur d'autres utilisations possibles des scores, que ce soit en référence à des seuils fixes (*cut-off scores*) ou pour diagnostiquer un état quelconque. L'étude ne tient pas compte, non plus, de la précision des scores obtenus (erreur type ou information) et elle ne concerne pas les autres propriétés potentiellement intéressantes des différentes méthodes de notation, par exemple la robustesse de l'estimation, la facilité logistique impliquée, l'invariance ou l'additivité linéaire des scores ainsi obtenus. Il faut particulièrement mentionner que cette recherche ne concerne pas directement l'invariance prédictive, bien qu'elle puisse être liée à cette problématique (Millsap, 2007). Une autre limite importante est que les scores ont été corrélés à des variables générées aléatoirement, et avec lesquelles les scores ont une corrélation moyenne très faible, voire nulle. Les résultats pourraient différer si les variables générées aléatoirement étaient moyennement ou fortement corrélées aux quatre ensembles de données utilisées pour générer des scores, bien que Xu et Stone (2011) aient obtenu des résultats similaires aux nôtres pour des corrélations positives allant de 0,3 à 0,6 et que nous ayons eu des résultats semblables avec des données réelles positivement corrélées (Chénier, 2015).

Finalement, en cherchant à émuler les pratiques courantes des chercheurs, cette étude a fait certains choix simplistes et a évacué certains aspects importants de l'utilisation de scores factoriels, particulièrement en ce qui concerne les divers indices disponibles pour vérifier la qualité et la fiabilité des scores obtenus (Grice, 2001). Il serait intéressant qu'une étude ultérieure vérifie l'impact de la qualité des scores sur l'utilisation subséquente de ces scores dans des études corrélationnelles afin de pouvoir nuancer les résultats de cette recherche.

Réception : 13 juillet 2016

Version finale : 4 octobre 2016

Acceptation : 4 novembre 2016

NOTES

1. À l'exception des modèles d'erreur de mesure, très rarement utilisés en éducation.

RÉFÉRENCES

Adedoyin, O. O., Nenty, H. J., & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Reviews*, *3*(3), 83-93. Retrieved from www.academicjournals.org/journal/ERR/article-full-text-pdf/326BFD43220

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573. doi: 10.1007/BF02293814

Barrett, P. (2011). Invoking arbitrary units is not a solution to the problem of quantification in the social sciences. *Measurement: Interdisciplinary Research and Perspectives*, *9*(1), 28-31. doi: 10.1080/15366367.2011.558783

Beaulieu, C., De Sève, I. et Provost, C. (2015, mai). *Anxiété et perception négative de ses capacités : obstacles à la réussite en première session du collégial*. Communication présentée au 83e congrès de l'ACFAS, Rimouski, Québec.

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*(1), 77-100. doi: 10.1016/0272-7358(88)90050-5

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440. doi: 10.1007/s11336-006-1447-6

Chalmers, P. (2016). Bibliothèque *mirt* (version 1.17.1) [Logiciel R]. https://github.com/philchalmers/mirt

Chénier, C. (2015, mai). *La méthode de notation a-t-elle une influence sur les opérations statistiques ultérieures?* Communication présentée au 83e congrès de l'ACFAS, Rimouski, Québec.

Cohen, L., Manion, L., & Morrison, K. (Eds.). (2007). *Research methods in education* (6th ed.). London/New York: Routledge/Falmer.

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics.* (Unpublished doctoral dissertation). Texas A&M University, Austin, TX.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1-11. Retrieved from http://pareonline.net/pdf/v14n20.pdf

Dubé, F., Dufour, F., Chénier, C. et Meunier, H. (2016). Sentiment d'efficacité, croyances et attitudes d'enseignants du collégial à l'égard de l'éducation des étudiants ayant des besoins particuliers. *Éducation et francophonie*, *44*(1), 154-172. doi: 10.7202/1036177ar

Dufour, F., Meunier, H. et Chénier, C. (2014). Quel est le sentiment d'efficacité personnelle d'étudiants en enseignement en adaptation scolaire et sociale dans le cadre du stage d'intégration à la vie professionnelle, dernier stage de la formation initiale ? Dans L. Portelance, S. Martineau et J. Mukamurera (dir.), *Développement et persévérance professionnels dans l'enseignement : oui, mais comment ?* (p. 75-92). Québec : Presses de l'Université du Québec.

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment*, *20*(1), 55-62. doi: 10.1037/1040-3590.20.1.55

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299. Retrieved from ww.w.statpower.net/Content/312/Handout/Fabrigar1999.pdf

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357-381. Retrieved from www2.hawaii.edu/~daniel/irtctt.pdf

Ferrando, P. J., & Chico, E. (2007). The external validity of scores based on the two-parameter logistic model: Some comparisons between IRT and CTT. *Psicologica: International Journal of Methodology and Experimental Psychology*, *28*(2), 237-257. Retrieved from https://eric.ed.gov/?id=EJ802623

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt & F. Ostendorf (Eds.), *Personality psychology in Europe 7*. (pp. 7-28). Tilburg, Netherlands: Tilburg University Press.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430-450. doi: 10.1037/1082-989X.6.4.430

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Journal of Psychology*, *216*(2), 89-101. Retrieved from http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=2008-08726-005

Henson, R. K., & Kyle Roberts, J. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3), 393-416. doi: 10.1177/0013164405282485

Hernandez, R. (2009). Comparison of the item discrimination and item difficulty of the Quick-Mental Aptitude Test using CTT and IRT methods. *The International Journal of Educational and Psychological Assessment*, *1*(1), 12-18. Retrieved from http://ecyor.weebly.com/uploads/7/1/1/4/7114954/article2v2.pdf

Howell, D. C. (2008). *Méthodes statistiques en sciences humaines* (6e éd.). Bruxelles, Belgique : De Boeck Université.

Howell, R. D. (2008). Observed variables are indeed more mysterious than commonly supposed. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1-2), 97-101. doi: 10.1080/15366360802121826

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

Luo, H. (2011). *Generation of non-normal data: A study of Fleishman's power method*. [Working paper]. Retrieved from www.diva-portal.org/smash/get/diva2:407995/FULLTEXT01.pdf

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*(6), 921-943. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1001.262&rep=rep1&type=pdf

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *International Journal of Educational and Psychological Assessment*, *1*(1), 1-11. Retrieved from http://files.eric.ed.gov/fulltext/ED506058.pdf

Mair, P., Hatzinger, R., Maier, M. J., & Rusch, T. (2015). Bibliothèque *eRm* (version 0.15-6) [Logiciel R]. http://r-forge.r-project.org/projects/erm/

Mazza, A., Punzo, A., & McGuire, B. (2015). Bibliothèque *KernSmoothIRT* (version 6.1) [Logiciel R]. https://CRAN.R-project.org/package=KernSmoothIRT

Ménard, L., Legault, F., Nault, G., St-Pierre, L., Raîche, G. et Bégin, C. (2011). *Projet de recherche sur l'impact des activités formelles de formation et d'encadrement pédagogiques sur les nouveaux enseignants des cégeps et leurs étudiants*. Rapport de recherche. Montréal : Université du Québec à Montréal.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166. doi: 10.1037/0033-2909.105.1.156

Michell, J. (2002). Stevens's theory of scales of measurement and its place in modern psychology. *Australian Journal of Psychology*, *54*(2), 99-104. doi: 10.1080/00049530210001706563

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*(4), 461-473. doi: 10.1007/s11336-007-9039-7

R Core Team (2016). *R: A language and environment for statistical computing* (version 3.3.1). R Foundation for Statistical Computing, Vienna: Austria.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer Science+Business Media.

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, *84*(3), 228-238. doi: 10.1207/s15327752jpa8403_02

Revelle, W. (2016). Bibliothèque *psych* (version 1.6.4) [Logiciel R]. http://personality-project.org/r/psych

Rotou, O., Headrick, T. C., & Elmore, P. B. (2002). A proposed number correct scoring procedure based on classical true-score theory and multidimensional item response theory. *International Journal of Testing*, *2*(2), 131-141. doi: 10.1207/S15327574IJT0202_3

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement*, *64*(4), 588-599. doi: 10.1177/0013164403261051

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*(1), 63-84. doi: 10.1177/0013164404273942

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, *6*(3), 255-270. doi: 10.1037/1040-3590.6.3.255

Schaeffer, G. A., Henderson-Montero, D., Julian, M., & Bené, N. H. (2002). A comparison of three scoring methods for tests with selected-response and constructed-response items. *Educational Assessment*, *8*(4), 317-340. doi: 10.1207/S15326977EA0804_2

Scheff, T. (2011). The catastrophe of scientism in social/behavioral science. *Contemporary Sociology: A Journal of Reviews*, *40*(3), 264-268. doi: 10.1177/0094306110404513

Streiner, D. L. (2010). Measure for measure: New developments in measurement and item response theory. *Canadian Journal of Psychiatry*, *55*(3), 180-186. doi: 10.1177/070674371005500310

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Upper Saddle River, NJ: Pearson.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, *19*(5), 579-599. doi: 10.1177/0959354309341926

Tschannen-Moran, A., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*, 783-805. Retrieved from http://mxtsch.people.wm.edu/Scholarship/TATE_TSECapturingAnElusiveConstruct.pdf

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70. doi: 10.1177/109442810031002

Xu, T., & Stone, C. A. (2011). Using IRT trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement*, *72*(3), 453-468. doi: 10.1177/0013164411419846

Zaman, A., Kashmiri, A., Mubarak, M., & Ali, A. (2008, November). *Students ranking, based on their abilities on objective type test: Comparison of CTT and IRT.* Paper presented at the EDU-COM International Conference, Perth, Australia. Retrieved from http://researchonline.jcu.edu.au/7620/1/7620_Jones_2008.pdf

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, *27*(1), 119-140. doi: 10.1177/0265532209347363