Personality and Individual Differences 102 (2016) 41–45

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Effects of item type and estimation method on the accuracy of estimated personality trait scores: Polytomous item response theory models versus summated scoring

Andrew B. Speer [a,c,*], Chet Robie [b], Neil D. Christiansen [c]

[a] American Family Insurance, 6000 American Parkway, Madison, WI 53704, United States
[b] Lazaridis School of Business & Economics, Wilfrid Laurier University, Waterloo, Ontario N2L 3C5, Canada
[c] Department of Psychology, Central Michigan University, Mount Pleasant, MI 48859, United States

ARTICLE INFO

Article history:
Received 21 April 2016
Received in revised form 21 June 2016
Accepted 22 June 2016
Available online 1 July 2016

Keywords:
Item response theory
Classical test theory
Personality trait estimates

ABSTRACT

Despite an increased use in item response theory (IRT)-based personality testing there is little research documenting whether trait estimations are actually improved over those derived via simply summated scoring according to classical test theory (CTT). In this study personality scale validity was compared using a variety of estimation methods (CTT, adjusted-CTT, SGR, GGUM) and item types (monotonic vs. non-monotonic) for the traits of conscientiousness and extraversion. Regardless of item type or estimation method, trait estimates were highly correlated. Using job performance ratings as an external criterion within the nomological network of these traits, model fit was not related to scale validity, and all estimation procedures resulted in comparable validity coefficients. Implications are discussed.

1. Introduction

Perhaps no other statistical technique has received such a large deal of attention in recent decades as item response theory (IRT), made practically feasible with the advancements in computer processing. The procedure has commonly been applied to large scale cognitive assessments, and within the field of personnel selection is increasingly being used in ability and knowledge testing. Additionally, many test manufacturers and consulting firms offer IRT-scored personality scales, with IRT being advantageous because it allows for computer adaptive testing (CAT), which can reduce testing time and limit concerns over item exposure due to use of large item banks. Furthermore, IRT allows for more accurate scoring of personality items that might not adhere well to dominance-based models (Chernyshenko, Stark, Drasgow, & Roberts, 2007), and it allows for scoring of forced questions without resulting in undesirable score features such as ipsiativity (Brown & Maydeu-Olivares, 2013).

Despite an increase in the usage of IRT-based personality testing there is little research documenting whether estimated trait scores are actually improved over those derived via summated scoring according to classical test theory (CTT). Trait scores that better estimate the latent construct should exhibit stronger correlations with external variables within the construct's nomological network. Considering the resources and time required to calibrate and utilize IRT-based tests, investigating the validity of the method is needed to help guide choice of estimation procedure when scoring personality measures. In pre-employment testing contexts, if IRT-derived trait scores are indeed more accurate representations of the underlying trait they should correlate more strongly with job performance assuming the measured traits are important to job success. Despite this basic assumption, little research has examined this issue in a personnel setting with job performance as the criterion. As such, the current study investigated whether IRT estimates of personality traits result in increased criterion-related validity in the prediction of job performance ratings and under what formats they might be more likely to do so.

1.1. Approaches to scoring personality items

For the greater portions of the 20th century, summated scales dominated the scoring of personality constructs. Operating under CTT, estimation involves simply summing all item scores (e.g., response scores on a Likert scale) into a composite to obtain an estimate of a respondent's trait score. While summated scoring is simple to perform

* Corresponding author at: American Family Insurance, 6000 American Parkway, Madison, WI 53704, United States.
E-mail addresses: speerworking@gmail.com (A.B. Speer), crobie@wlu.ca (C. Robie), chris1nd@cmich.edu (N.D. Christiansen).

http://dx.doi.org/10.1016/j.paid.2016.06.058
0191-8869/

and has been widely used in the field of personality testing, there are limitations to its use, with these being discussed in almost any paper discussing IRT (for a good primer see Hambleton & Swaminathan, 2001). For instance, trait estimations and item parameters are dependent upon one another, CTT assumes a consistent amount of error across the entire trait continuum, and CTT assumes all items are equally good indicators of a given trait. IRT is assumed to overcome these weaknesses.

While many IRT models are capable of scoring personality items, polytomous models most adequately capture item information when response scales have more than two response options. Therefore, for the sake of this paper we focus solely on polytomous IRT models. Within this realm, two polytomous models are frequently used to score personality items. The first, Samejima's Graded Response Model (SGR, Samejima, 1969), assumes item monotonicity, which means that as latent trait scores increase the likelihood of item endorsement also increases (this has also been labeled a *dominance process*). Monotonicity is reflective of CTT-based tests such that items that do not adhere to the monotonic assumption will demonstrate low correlations with other test items and therefore are common culprits for item removal when creating scales. The vast number of developed personality inventories has been created based on the monotonicity assumption (Chernyshenko et al., 2007).

A second common approach is the Generalized Graded Unfolding Model (GGUM). GGUM, an ideal point model, works under Thurstone's (1928) law of comparative judgment where it is assumed that individuals will only endorse an attitude statement to the extent it corresponds to the person's actual level of theta (Roberts, Donoghue, & Laughlin, 2000). Thus, instead of a monotonically increasing response function, single-peaked response functions are possible, in that when the distance between an item's location and a person's theta is zero, respondents will be more likely to agree with a statement. As the distance between an item's location and person's theta increases, individuals will be more likely to disagree with the item, allowing for non-monotonic items. Thus, bell-shaped probability response functions are possible if an item's difficulty is located towards the middle of the theta continuum. Under this scenario, respondents who have very low or very high true trait scores will be less likely to agree with the item because their trait level is more distal from the item's location. This is referred to as item "unfolding."

Most unfolding items use some sort of adverb that attenuates the strength of an item statement. For instance, placing the adverb of "usually" prior to a statement makes for an item that is less definitive in strength. "I like to clean my room" is a stronger statement than "I *usually* like to clean my room." Proponents of ideal point models suggest that respondents who have very high trait scores will not actually agree with the latter question because they *always* would like to clean their room, not only *usually*. Thus, respondents with high true trait levels might disagree with the statement because it is not close enough to their own feelings, whereas a dominance response process would assume that because the respondent's theta is higher than the item, they would be likely to endorse the item.

The few studies that have compared IRT estimation methods for scoring personality items have typically focused on model fit or scale correlations with other self-report measures, in general revealing an inconsistent pattern of results (e.g., Broadfoot, 2008; Chernyshenko, et al., 2007; Kosinski, 2009; Stark, Chernyshenko, Drasgow, & Williams, 2006). Ideal point models have been assumed advantageous over other IRT procedures because they are capable of modeling a greater variety of item functions (i.e. both monotonic and non-monotonic) and should therefore more effectively capture the entire construct domain (e.g., Cherynshenko et al., 2007; Stark et al., 2006). In line with this, Stark et al. (2006) found superiority of GGUM estimates over SGR estimates in terms of model fit. Additionally, Chernyshenko et al. (2007) found GGUM superiority over a two-parameter logistic model, although correlations with external criteria such as other self-report measures were comparable. Kosinski (2009) found scale scores estimated using GGUM had worse fit than when those scales were estimated via SGR, and this occurred even though items were specifically developed according to the ideal point model. Likewise, Broadfoot (2008) demonstrated that GGUM estimates of conscientiousness and agreeableness had comparable fit and correlations with external criteria as a partial credit model. Thus, research does not seem to consistently support superiority of one IRT model over any other.

1.2. Factors affecting estimation accuracy

When comparing IRT to the more traditional CTT-based summated scoring, IRT should theoretically produce more accurate trait estimations because (1) ability estimates are made using a true interval scale (Xu & Stone, 2012), and (2) items are maximally weighted to achieve the best estimate of theta (Ferrando & Chico, 2007). Despite these assumptions, the present literature shows no strong support for the superiority of IRT over CTT estimates as better estimates of latent traits. Better estimates of a trait should correlate more strongly with variables within the construct's nomological network, and yet the correlations of IRT versus CTT estimates with non-personality criteria do not show a consistent difference (Chernyshenko et al., 2007; Ferrando & Chico, 2007; Ling, Zhang, Locke, Li, & Li, 2016; Xu & Stone, 2012) (for an exception see a study using simulated data by Dalal & Carter, 2015). While such is the case, the majority of past studies have not examined this issue in a personnel setting, or essentially when job performance ratings are used as the criterion measure. In the realm of industrial psychology, job performance is unparalleled in its importance as an outcome, and if a test is used for employee selection, prediction of job performance is the focal concern to support test use. If indeed a trait is important to job success, better estimates of that trait should subsequently demonstrate stronger correlations with performance on the job. That no research has compared predictions of performance according to different estimation methods is a shortcoming that must be addressed.

The question then becomes, when might IRT estimates be improved over CTT estimates? It is commonly assumed adequate model fit is a prerequisite of good construct estimation (Ferrando & Chico, 2007; Xu & Stone, 2012). A model's depiction of the relationship between response probability and one's standing on a given construct should correspond to the observed data. SGR assumes monotonicity such that the likelihood of response endorsement should increase as trait levels increases. Monotonic items should thus better assess the latent trait when modeled according to SGR. When items are non-monotonic and unfold, SGR response functions should inaccurately represent actual response patterns and therefore lead to inaccurate trait estimates. GGUM, which is capable of modeling non-monotonic items under the ideal point model, should overcome this and produce accurate trait estimates when non-monotonic items are used within a test.

1.3. Present study

No published study to which we are aware has examined how these methods of personality scoring affect test validity when the tests are used to predict meaningful outcomes such as job performance. If IRT does in fact produce better estimates of latent traits, then those estimates should correspond more closely to constructs assumed to share portions of the construct space. In the case of personnel selection, better measurement methods should result in scores that have higher criterion-related validity when the outcome is job performance ratings.

The current study sought to examine this issue by taking a set of personality scales composed of monotonic and non-monotonic items, scoring them according to both IRT and CTT methods, and then comparing how well the estimations predict employee job performance. Separate scales composed of monotonic and non-monotonic items were taken and trait scores were estimated using SGR, GGUM, traditional CTT, and

CTT weighted by factor loading (labeled adjusted-CTT), which essentially weighs each item by how well it measures the latent construct. This design, although non-experimental, is a 2 (item type) × 4 (estimation method) fully-crossed design, removing the potential confound of using different scoring methods for different scales.

Assuming fit is related to estimation accuracy, IRT estimation should be less accurate when the incorrect model is chosen to score a set of items (Ferrando & Chico, 2007). Thus, only when the data match the assumptions of the scoring model should IRT result in estimations that better reflect the latent construct space. Based on the aforementioned assumptions, we outline a set of propositions in exploratory fashion. Because validity estimates were computed by comparing estimation procedures within scales, minimal differences were expected. Thus, traditional hypothesis testing was not utilized.

1.4. Propositions

Proposition 1. GGUM model fit will be improved over SGR model fit when scales are composed of non-monotonic items.

Proposition 2. Criterion-related validity derived from SGR estimates will be higher than CTT estimates only for scales composed of monotonic items.

Proposition 3. Criterion-related validity derived from GGUM estimates will be higher than CTT and SGR estimates for scales composed of non-monotonic items.

2. Method

2.1. Participants

A sample of 622 incumbent managers from a variety of different organizations responded to a personality inventory as part of concurrent validation projects where the results could also be used for managerial development. All incumbents were rated by their supervisors on their overall job performance.

2.2. Personality scales

Incumbents responded to a personality inventory designed to assess the Five Factor Model traits of Extraversion and Conscientiousness. Of the FFM traits, these are the most predictive of managerial job performance (e.g., Barrick & Mount, 1991). Scale items were designed to measure each of these constructs by reviewing items from the International Personality Item Pool and reviewing the extant literature. Prior to test administration to the study's primary sample, typical CTT-based scale development procedures were undertaken through pilot testing (e.g., examination of corrected item total correlations and *p*-values were performed to remove items) to reduce scale length. The final, operational scales measured each broad trait using 30 items in five-point Likert scale format, with internal consistencies being high for both Extraversion (0.93) and Conscientiousness (0.91).

2.2.1. Monotonic and non-monotonic scales

Even though the inventory was developed according to CTT procedures and therefore any drastically non-monotonic items were likely removed via item analysis during pilot scale creation, we attempted to locate a set of items for each trait that were *more* likely to unfold. To do this, three subject matter experts (SME's) rated all the Extraversion and Conscientiousness items on the degree to which they adhered to the ideal point model. SME's all held at least a master's degree in industrial/organizational psychology and had extensive experience in the field of personality testing. The rating task required judges to rate each item on the degree to which it was expected to unfold using a scale from 1 to 4 where one indicated a purely monotonic item and four indicated the item was non-monotonic. Based on these ratings, for each trait a monotonic composite of items and non-monotonic composite of items were formed. Eleven items were chosen for each scale to maximize the degree of monotonicity or non-monotonicity, with the average non-monotonic rating being 1.87 for the two non-monotonic scales and 1.00 for the two monotonic scales (for items see Supplementary material).

2.3. Trait estimation

2.3.1. Unidimensionality

IRT-analysis requires that a scale of items is unidimensional. Although personality measures are unlikely to ever be completely unidimensional due to response tendencies such as socially desirable responding, this assumption was examined for each of the scales used for the present study. Using principle axis factor analyses for all scales, and in addition comparing communalities of the first two factors using principle components analysis for non-monotonic scales (Roberts et al., 2000), all scales met traditional assumptions of unidimensionality.

2.3.2. Estimation procedures

Trait scores were estimated for all four scales using each of the estimation procedures across the 622 job incumbents. Simple summed composites were used to estimate scores under CTT. Additionally, a set of composites were formed after weighting items by factor loading to determine whether any gains in IRT estimation are simply because items that more effectively measure the latent construct are optimally weighted (adjusted-CTT). Multiplying raw scores by an item's factor loading minimizes measurement error by giving more weight to those items that better measure the latent construct. This procedure should produce near-equivalent estimates to two-parameter model estimates (2PL, Ferrando & Chico, 2007).

IRT estimations were made using SGR and GGUM. SGR was utilized as the basic monotonic model adhering to dominance theory. SGR analyses were conducted using IRTPro with 61 quadrature points and default estimation settings. Expected a posteriori (EAP) was used for trait estimation. Non-monotonic estimates were made using GGUM (Roberts et al., 2000), which uses a marginal maximum likelihood procedure for parameter estimation. For the current study, 50 quadrature points were selected to estimate item parameters, with 250 outer cycles, 10 inner cycles, 30 thresholds and items. A convergence criterion of 0.001 was used, with no lambda constraints. Theta estimates were obtained using EAP.

2.4. Job performance ratings

Single-item performance ratings were made by job incumbents' supervisors on a 1 to 5 scale with higher scores indicating superior performance. These ratings were completed by the employees' supervisors around the time they completed the personality assessment and represent overall job proficiency.

2.5. Correlation analyses

All trait estimates were subsequently correlated with job performance ratings. Although it is typical to calculate IRT parameters and then estimate trait scores on two separate samples, given limitations in sample size, calibration and validation were performed on the same sample of 622 incumbents.

Table 1
Scale descriptive statistics.

	Mean	*SD*	Alpha
Monotonic Extraversion scale	41.21	4.97	0.80
Non-monotonic Extraversion scale	42.04	4.41	0.82
Monotonic Conscientiousness scale	36.55	6.16	0.77
Non-monotonic Conscientiousness scale	41.41	4.55	0.75

Note. $N = 622$. *SD* = standard deviation. Alpha refers to Cronbach's alpha, which is an estimate of internal consistency. Scores represent summated scales using the original 1–5 item scoring format.

3. Results

3.1. Descriptive statistics and scale correlations

Before examining results across the different estimation methods, traditional descriptive statistics were calculated for each of the scales by using summated scoring. Table 1 displays means, standard deviations, and coefficient alphas for each of the four scales of items. As seen, reliability coefficients were adequate for all scales (ranged from 0.75 to 0.82), and scale means and standard deviations were similar for every scale besides the non-monotonic Extraversion scale, which demonstrated a noticeably lower mean and higher standard deviation.

Trait scores were then estimated across all item composites according to each of the estimation procedures. Trait estimates using each of the estimation procedures demonstrated strong convergence (see Supplementary material for full correlation matrix). The average correlation among estimates for the monotonic-Extraversion scale was 0.99, 0.98 for the monotonic-Conscientiousness scale, for the non-monotonic-Extraversion scale 0.97, and for the non-monotonic-Conscientiousness scale 0.99. Based on these scale relationships, and similar to past research (e.g., Fan, 1998), it is clear that the different estimation procedures resulted in very similar trait estimates.

3.2. IRT model fit

IRT model fit was determined by examining fit plots and chi-square values, with emphasis given to the latter. Adjusted chi-square to degree of freedom ratios were used as an index of model fit, with adjusted values lower than 3 indicating good fit (Drasgow, Levine, Tsien, Williams, & Mead, 1995). Ratios of 3 to 4 indicate moderate misfit, and ratios above 4 indicate higher levels of misfit. However, it should be noted that local dependence is not uncommon for measures of personality (Chernyshenko et al., 2007) and thus a higher cutoff may in fact be appropriate.

All models demonstrated good fit according to chi-square singlet values (Table 2), with monotonic scales fitting better according to SGR and non-monotonic scales fitting better according to GGUM, as expected. However, this trend did not occur for chi-square doublets and triplets. In this regard, GGUM estimations resulted in more severe misfit for all types of items (all χ^2/df ratios were above 4.0). The large misfit indicated by doublet and triplet chi-square values is similar to findings provided by Kosinski (2009). When looking at fit plots and item information functions, all scales discriminated better at lower levels of theta, as item information functions peaked at negative item locations and then dramatically decreased as item location moved to upper end of the distribution. Because GGUM model fit was not better for non-monotonic items, limited support was found for Proposition 1.

Table 2
Model fit comparisons.

	Singlet mean χ^2/df	Doublet mean χ^2/df	Triplet mean χ^2/df
Monotonic Extraversion			
SGR	0.03 (0.00)	1.48 (3.32)	1.44 (3.11)
GGUM	0.05 (0.00)	1.80 (4.85)	1.71 (4.40)
Non-monotonic Extraversion			
SGR	0.04 (0.00)	1.56 (3.68)	1.69 (4.22)
GGUM	0.03 (0.00)	1.69 (4.35)	1.74 (4.56)
Monotonic Conscientiousness			
SGR	0.09 (0.00)	1.97 (5.90)	1.99 (5.76)
GGUM	0.09 (0.00)	2.19 (6.96)	2.22 (6.86)
Non-monotonic Conscientiousness			
SGR	0.39 (0.00)	2.05 (6.24)	1.96 (5.61)
GGUM	0.07 (0.00)	2.27 (7.15)	2.21 (6.81)

Note. SGR = Samejima's Graded Response Model. GGUM = Generalized Graded Unfolding Model. Values presented in parentheses are adjusted chi-square to degree of freedom ratios.

3.3. Scale and composite validity

Table 3 displays relationships between personality scales and job performance for each of the estimation methods. We also included composites of the scales as they are often used in decision-making. In general, assumed expectations were not supported, as there was no strong pattern underlying the validity coefficients. For instance, it was expected that GGUM estimates would have higher validity coefficients than SGR estimates when scales of non-monotonic items were examined. As seen, for one of the scales this was the case (Conscientiousness, 0.229 vs. 0.227), although the opposite was true for Extraversion (0.055 vs. 0.059), and in both cases differences were minimal. Validity was not expected to differ between these estimation procedures for monotonic items, but GGUM was better for the Extraversion scale (0.137 vs. 0.131) whereas slightly worse for the Conscientiousness scale (0.281 vs. 0.282). However, once again differences in validity coefficients were quite small.

Summated score (i.e. CTT) estimates were comparable to IRT-estimates for most traits. For all scales except the unfolding-extraversion scale, validity slightly improved from traditional CTT to the adjusted-CTT estimates, which was then minimally improved upon using IRT estimates. Differences between traditional CTT validity and the best IRT estimates were 0.013 for the monotonic-Extraversion scale, 0.015 for the monotonic-Conscientiousness scale, and 0.004 for the non-monotonic-Conscientiousness scale. Thus, IRT demonstrated slight gains in prediction over CTT for these three scales. However, the opposite trend occurred for the non-monotonic-Extraversion scale. For this scale, validity was noticeably strongest for traditional CTT estimates.

Overall, across all scales traditional CTT estimates actually had the highest validity (0.177), followed by the adjusted-CTT estimate (0.176), the GGUM estimate (0.176), and then the SGR estimate (0.175). Thus, there were limited differences in scale validity regardless of the estimation method used. Finally, when examining item format only, non-monotonic scales had lower average validity coefficients ($r = 0.148$) than non-monotonic scales ($r = 0.204$), and this difference was significant ($p < 0.01$). We also found no practically significant differences in validity across estimation methods for the composites.

Table 3
Validity of scale scores and composites by estimation method.

	CTT	Adjusted-CTT	SGR	GGUM
Monotonic				
Extraversion scale	0.126**	0.128**	0.131**	0.137**
Conscientiousness scale	0.267**	0.277**	0.282**	0.281**
Composite (E + C)	0.233**	0.241**	0.245**	0.248**
Non-monotonic				
Extraversion scale	0.088*	0.073	0.059	0.055
Conscientiousness scale	0.225**	0.227**	0.227**	0.229**
Composite (E + C)	0.189**	0.184**	0.176**	0.185**

Note. $N = 622$. Adjusted-CTT refers to the CTT composite created by weighting each item according to its factor loading. Shown are scale correlations with job performance. E + C = composite of Extraversion and Conscientiousness. * $p < 0.05$, ** $p < 0.01$.

4. Discussion

The current study sought to better understand the utility of implementing IRT-based scoring procedures to personality measures. Specifically, methodologically different types of personality items (monotonic and non-monotonic) were scored using different IRT models and also according to CTT. These analyses were conducted using a completely crossed design. Model fit and scale validity were examined under each estimation approach for each set of items, with this being the first known study to examine differences in estimation procedure validity using job performance as an outcome measure.

4.1. Model fit

Although monotonic scales fit better for SGR and non-monotonic scales fit better for GGUM when chi-square singlets were used as a fit index, for all models, chi-square doublets and triplets were worse when using GGUM. These findings coincide with past research (Kosinski, 2009). With that said, it is possible the items did not adhere well enough to an ideal point model and instead reflected a dominance response pattern. Indeed, examination of option response functions revealed that most items were monotonic, with the few items that did exhibit non-monotonicity having extremely high points of unfolding (i.e. high item locations). Because GGUM fit improves as the number of non-monotonic items increases (Stark et al., 2006), different results may have been found if the inventory were not developed according to CTT principles and if indeed the items had been less monotonic. Although such may be the case, given that past research has not found overwhelming support for GGUM's superiority even when used on items developed specifically to unfold (Kosinski, 2009), it is unlikely overwhelmingly large fit improvements would have been found had more items been more non-monotonic.

4.2. Scale and composite validity

This was the first known study to examine the validity of personality scales using a variety of estimation procedures and item types with job performance ratings as the criterion. When looking across IRT models, despite assuming validity would be a function of model fit, no strong trend was found. Validity was very similar irrespective of estimation procedure (SGR vs. GGUM). Additionally, an estimation procedure by item type interaction was also not found. The only clear difference in validity coefficients was for the items themselves, with non-monotonic scales demonstrating significantly lower validity coefficients than monotonic items.

While it is tempting to declare a main effect of item-type as conclusive, findings should be cautiously interpreted, as the non-monotonic items were not explicitly developed according to an ideal point model but rather were rated as most likely to display non-monotonic tendencies out of a large, CTT-developed item pool. Despite this, taking into account the lack of difference between SGR and GGUM and the inferiority of non-monotonic items in this study, it appears some of the proposed benefits of ideal point models are yet to be supported. Chernyshenko et al. (2007) suggest ideal point items broaden the construct domain by providing more information at moderate values of theta. However, if such were the case the increase in obtained construct information would correspond to increased scale correlations with external criteria, which here and in past research (e.g., Broadfoot, 2008; Chernyshenko et al., 2007; Kosinski, 2009) has not been supported. These results raise questions as to whether "unfolding" items really capture unique and useful trait variance, or whether they are simply misinterpreted by respondents, therefore introducing error variance. For example, those with high trait scores might interpret a statement with a "usually" qualifier in different ways, in that some might interpret "usually" in a literal sense and therefore not strongly endorse the item, whereas others might respond in a more dominance-oriented fashion. In this case, a common response function would not accurately model the latent trait for both groups of respondents. More research on this issue is warranted. Our results may have been biased due to the measure being initially developed using CTT. A future study could do the CTT versus IRT comparisons using a measure that was developed using IRT. Additionally, future studies should employ both validation and calibration samples so that chance variability is not capitalized upon.

Finally, this study corroborates the majority of past findings in that the validity of IRT estimations was equivalent to those found with CTT. While for three of the scales there were gradually increasing improvements in validity coefficients from CTT to IRT, the increases were quite small, and it is debatable whether they would be considered meaningful. Additionally, there was quite small, gradual (but not practically significant) improvement in the monotonic composite but this was not evidenced in the non-monotonic composite. All told, results seem to suggest that it is unlikely substantial gains in validity will be found when scoring personality items using IRT methodology.

Appendix A. Supplementary material

Supplementary material for this article can be found online at http://dx.doi.org/10.1016/j.paid.2016.06.058.

References

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1–26. http://dx.doi.org/10.1111/j.1744-6570.1991.tb00688.x.

Broadfoot, A. A. (2008). Comparing the dominance approach to the ideal-point approach in the measurement and predictability of personality. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 3316810.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*, 36–52. http://dx.doi.org/10.1037/a0030641.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*, 88–106. http://dx.doi.org/10.1037/1040-3590.19.1.88.

Dalal, D. K., & Carter, N. T. (2015). Consequences of ignoring ideal point items for applied decisions and criterion-related validity estimates. *Journal of Business and Psychology*, *30*, 483–498. http://dx.doi.org/10.1007/s10869-014-9377-2.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B. A., & Mead, A. D. (1995). Fitting polytomous items response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*, 143–165. http://dx.doi.org/10.1177/014662169501900203.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*, 357–381. http://dx.doi.org/10.1177/0013164498058003001.

Ferrando, P. J., & Chico, E. (2007). The external validity of scores based on the two-parameter logistic model: Some comparisons between IRT and CTT. *Psicológica*, *28*, 237–257.

Hambleton, & Swaminathan (2001). *Item response theory: Principles and applications*. NY: Wiley.

Kosinski, M. (2009). *Application of the dominance and ideal point IRT models to the extraversion scale from the IPIP Big Five Personality Questionnaire.* (Mphil Dissertation) Cambridge University.

Ling, Y., Zhang, M., Locke, K. D., Li, G., & Li, Z. (2016). Examining the process of responding to circumplex scales of interpersonal values items: Should ideal point scoring methods be considered? *Journal of Personality Assessment*, *98*, 310–318. http://dx.doi.org/10.1080/00223891.2015.1077852.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*, 3–32. http://dx.doi.org/10.1177/01466216000241001.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*, 100. http://dx.doi.org/10.1002/j.2333-8504.1968.tb00153.x.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25–39. http://dx.doi.org/10.1037/0021-9010.91.1.25.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554. http://dx.doi.org/10.1086/214483.

Xu, T., & Stone, C. A. (2012). Using IRT trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement*, *72*, 453–468. http://dx.doi.org/10.1177/0013164411419846.