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.

© 2016 Elsevier Ltd. All rights reserved.

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/© 2016 Elsevier Ltd. All rights reserved.

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

Download English Version:

https://daneshyari.com/en/article/889651

Download Persian Version:

https://daneshyari.com/article/889651

Daneshyari.com