Personality and Individual Differences 123 (2018) 229–235

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Examining validity evidence for multidimensional forced choice measures with different scoring approaches

Philseok Lee[a], Sunhee Lee[b,*], Stephen Stark[c]

[a] South Dakota State University, United States
[b] Chungnam National University, Republic of Korea
[c] University of South Florida, United States

ARTICLE INFO

Keywords:
Noncognitive construct
Multidimensional forced choice format
Response bias
Scoring methods
Item response theory

ABSTRACT

This study aims to investigate whether the Thurstonian Item Response Theory (TIRT), a recently developed normative scoring method, yields better validity evidence than the traditional partially ipsative scoring methods for multidimensional forced choice measures. For this purpose, we compared the construct- and criterion-related validity evidence of three different scoring methods for MFC measures, including 1) a partially ipsative method based on classical test theory (PI-CTT), 2) an analogous partially ipsative method using the graded item response theory (PI-IRT), and 3) the TIRT method. We also included in our analyses the validity evidence for a single-statement (SS) Big Five personality measures. Overall, the validity evidence for the three types of MFC scoring methods was comparable to the validity evidence for the traditional summative scoring (SS-CTT) method. Interestingly enough, the PI-CTT scoring method, the simplest method for MFC scoring, was about as effective as the more complex TIRT method. We discuss practical implications of the results and offer suggestions for future research.

1. Introduction

Historically, the noncognitive constructs such as personality (Schmidt & Hunter, 1998), emotional intelligence (Van Rooy & Viswesvaran, 2004), and social skills (Semadar, Robins, & Ferris, 2006) have been measured predominantly using Likert-type scales. Likert-type scales provide a set of single statements (SS) and require respondents to indicate their level of agreement with each of the statements using, for example, a 1 (Strongly Disagree) to 5 (Strongly Agree) format. An example of a Likert-type scale is as follows:

Using the 1–5 scale, indicate your agreement with each item.

1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree,

5 = Strongly agree.

- I always turn in my assignments on time. ___

This methodology has been criticized due to its susceptibility to various types of response biases. In particular, faking good responding tends to inflate scale means and intercorrelations, and it can reduce the validity and utility of measures used for high-stakes decision making (Kim, 2011). Likert-type SS scales are also susceptible to rater errors (e.g., leniency, halo) and cultural-specific response biases (e.g., central tendency, extremity or acquiescence), which may inflate cross-dimension correlations (Borman et al., 2001) and attenuate relationships with outcomes in cross-cultural research contexts (He & van de Vijver, 2013).

To deal with these response biases and rater errors, multidimensional forced choice (MFC) measures have been proposed as an alternative to Likert-type scales for noncognitive assessment (Stark, Chernyshenko, & Drasgow, 2005). MFC measures commonly present statements in blocks of two (pair), three (triplet), or four (tetrad). Within the blocks, statements representing different constructs may be matched on social desirability and/or extremity. The respondent's task is to choose the statement in each block that is “most like me”, or to rank the statements in each block from “most like me” to “least like me”. An example of MFC triplet item for rank responses is shown below.

For each block of statements, rank the statements from “most like me (1)” to “least like me (3)”.

(A) I always turn in my assignments on time. (+ C) 3
(B) I generally perform well under pressure. (+ Em) 1

* Corresponding author.
E-mail address: sunhee_lee@cnu.ac.kr (S. Lee).

https://doi.org/10.1016/j.paid.2017.11.031
Received 15 July 2017; Received in revised form 18 November 2017; Accepted 18 November 2017

(C) I enjoy learning about other cultures. (+ O)	2	

Note: A MFC triplet item for rank responses involving positively (+) keyed statements representing Conscientiousness (C), Emotional Stability (Em), and Openness to Experience (O).

In theory, matching on social desirability and/or extremity makes the "best" answers difficult to discern, and by forcing respondents to choose between alternatives, rather than indicating their level of agreement with each statement, response biases and rater errors can be reduced (Brown & Maydeu-Olivares, 2014).

Nevertheless, MFC measures have been criticized because conventional MFC scoring methods lead to ipsativity problems that render scores unsuitable for inter-individual comparisons (Hicks, 1970). However, an advent of partially ipsative scoring methods enables researchers to obtain normative information using a heuristics approach (e.g., Heggestad, Morrison, Reeve, & McCloy, 2006; McCloy, Heggestad, & Reeve, 2005; White & Young, 1998). In addition, meta-analytic research showed that MFC scores based on the partially ipsative scoring methods predict important criteria (e.g., Salgado, Anderson, & Tauriz, 2015; Salgado & Táuriz, 2014). Therefore, MFC measures are surging in popularity and becoming important components of personnel and educational assessment systems.

However, Brown (2015) recently argued that "conclusive validity evidence for forced-choice assessments can only be gained by using model-based measurement" (p.17) such as Thurstonian IRT (TIRT; Brown & Maydeu-Olivares, 2011) model which is designed specifically for comparative judgment. Other researchers also suggested that the model-based MFC IRT scoring methods should perform better than partially ipsative methods in most circumstances (Chernyshenko et al., 2009; Salgado & Táuriz, 2014). Yet, there is still a lack of empirical evidence because model-based IRT advances associated with the MFC format were made only recently. At present, research is needed to answer questions whether a model-based normative scoring method for MFC measures yields better validity evidence than partially ipsative scoring methods do. For this purpose, we compared construct- and criterion-related validities of different scoring methods for MFC versions of Big Five personality measures. In the following section, we briefly describe the scoring issues of MFC measures and the TIRT method.

2. Scoring issues of MFC measures

With MFC measures, simple classical scoring methods produce *ipsative* data that exhibit negative scale-intercorrelations and distorted reliability and validity estimates (Hicks, 1970). In addition, ipsative data support only intra-individual comparisons (Meade, 2004). If one simply assigns points corresponding to the inverted ranks of statements within MFC blocks, the points for each block would sum to a constant, and the sum of the scores would be the same for every examinee, making inter-individual comparisons problematic.

However, by taking steps to introduce variation in scale scores (e.g., by including distractor statements that are not scored or negatively keyed statements in MFC items), it is possible to produce *partially ipsative* scores. In a triplet, for example, if a negatively keyed statement is selected as *least like me* or a positively keyed statement is selected as *most like me*, a score of 2 is assigned to the statement. In contrast, if a negatively keyed statement is selected as *most like me* or a positively keyed statement is selected as *least like me*, a score of 0 is assigned. The second-ranked statements are assigned scores of 1. An example of partial ipsative scoring for triplet rank response is shown below.

For each block of statements that follow, rank the statements from most like you (1) to least like you (3).	RANK	SCORE
(A) I always turn in my assignments on time. (+ C)	3	1
(B) I generally perform well under pressure. (+ Em)	1	2
(C) I tend to say things that hurt other's feelings. (− A)	2	2

Note: A MFC triplet item for rank responses involving positively (+) and negatively (−) keyed statements representing Conscientiousness (C), Emotional Stability (Em), and Agreeableness (A).

Then, the MFC responses are reconstructed by disassembling the triplets and grouping the responses by each dimension scale. Scale scores on each dimension are then analyzed using a CTT (e.g., White & Young, 1998) or unidimensional IRT (e.g., Heggestad et al., 2006) approach as if they were administered by SS Likert-scale measures. The partially ipsative method can yield normative scores that enable researchers to conduct inter-individual analysis, thus it is useful for applications such as personnel screening (Stark et al., 2014).

Although the use of the partially ipsative scoring method seems to circumvent the ipsativity problem, it still has limitations. This method does not follow the comparative judgment process of evaluating statements that comprise MFC items. Furthermore, it does not allow for the computation of endorsement probabilities and estimation of item and person parameters directly from MFC responses. Thus, the quality of MFC items cannot be readily evaluated. In order to remedy such limitation of the partially ipsative scoring methods, the model-based MFC IRT method has been proposed. The model-based MFC IRT method can provide test constructors with more statistical information and a wider range of testing applications such as MFC item analysis, parallel test construction, differential item functioning analysis, or computerized adaptive testing (Brown, 2015; Stark, Chernyshenko, Drasgow, & White, 2012).

3. Thurstonian IRT method

Within the IRT framework, only a few MFC psychometric models have been proposed to yield normative information (e.g., Brown & Maydeu-Olivares, 2011; de la Torre, Ponsoda, Leenen, & Hontangas, 2012; Stark et al., 2005). Among them, the current study focuses on the TIRT approach because it is the most commonly used method using the Mplus program Muthén and Muthén (2014).

The TIRT model can be applied to not only *more like me* judgments associated with pairwise preference items, but also to *most like me*, *most and least like me*, and *rank-order* judgments for items containing three or more statements. The TIRT model assumes Thurstone's (1927) law of comparative judgment. According to Thurstone, statement j is preferred to statement k if the latent utility of j (t_j) is greater than the value of k (t_k). In this case, the statement is coded as 1; otherwise, it is coded as 0. This can be expressed as follows:

$$yl = 1 \text{ if } y_l^* = t_j - t_k \geq 0, \quad \text{and} \quad yl = 0 \quad \text{if } y_l^* = t_j - t_k < 0 \tag{1}$$

where, y_l^* represents the difference in the psychological values of the two statements.

For example, a triplet rank response [A, B, C] requires three sets of pairwise comparisons (i.e., A vs. B, A vs. C, and B vs. C). If a respondent endorses statement A as 1st rank, B as 2nd rank, and C as 3rd rank ([A, B, C] = [1, 2, 3]), the ranking response, [1, 2, 3], is transformed into three sets of paired comparison binary outcomes, [1], [1], [1]. Then, the transformed binary responses are analyzed using a two-dimensional standard normal ogive IRT model under the structural equation modeling (SEM) framework. The conditional probability of preferring statement j to statement k is obtained as follows:

$$P_l(y_l = 1 \mid \eta_a, \eta_b) = \Phi\left(\frac{-\gamma_l + \lambda_j\eta_a - \lambda_k\eta_b}{\sqrt{\psi_j^2 + \psi_k^2}}\right), \tag{2}$$

where $\gamma_l = -(\mu_j - \mu_k)$ is a threshold parameter replacing the difference of utility means; η_a and η_b are the measured attributes a and b; λ_j and λ_k are the loadings on the measured attributes η_a and η_b; ψ_j^2 and ψ_k^2 are the unique variance of the two utilities; and $\Phi(x)$ denotes the