PERSONNEL ASSESSMENT AND DECISIONS

Personnel Assessment and Decisions

Volume 5 | Issue 1 Article 3

7-6-2019

Criterion-related Validity of Forced-Choice Personality Measures: A Cautionary Note Regarding Thurstonian IRT versus Classical Test Theory Scoring

Peter A. Fisher
Wilfrid Laurier University, fish0150@mylaurier.ca

Chet Robie
Wifrid Laurier University, crobie@wlu.ca

Neil D. Christiansen
Central Michigan University, chris1nd@cmich.edu

Andrew B. Speer
Wayne State University, andrew.speer2@wayne.edu

Leann Schneider
Plum.io, inc., leann@plum.io

Criterion-Related Validity of Forced-Choice Personality Measures: A Cautionary Note Regarding Thurstonian IRT Versus Classical Test Theory Scoring

Peter A. Fisher[1], Chet Robie[1], Neil D. Christiansen[2], Andrew B. Speer[3], and Leann Schneider[4]

1. Wilfrid Laurier University
2. Central Michigan University
3. Wayne State University
4. Plum.io, Inc.

KEYWORDS

personality, measurement, selection, assessment, classical test theory, item response theory

ABSTRACT

This study examined criterion-related validity for job-related composites of forced-choice personality scores against job performance using both Thurstonian item response theory (TIRT) and classical test theory (CTT) scoring methods. Correlations were computed across 11 different samples that differed in job or role within a job. A meta-analysis of the correlations (k = 11 and N = 613) found a higher average corrected correlation for CTT (mean ρ = .38) than for TIRT (mean ρ = .00). Implications and directions for future research are discussed.

Faking on personality assessments has been a concern for nearly a century, and for as long as there has been faking, psychologists have put significant efforts towards reducing it (cf. Griffith & Robie, 2013; Kelly, Miles, & Terman, 1936). From modifying item design and construction, to varying assessment format and structure, to changing scoring methods, a great deal of academic research has investigated a wide variety of methods of accurately assessing personality while reducing the impact of faking behaviors. Indeed, a recent development in assessment scoring: Thurstonian item response theory (TIRT; Brown and Maydue-Olivares, 2011; 2012; 2013) has been promoted as an overall improvement in traditional scoring based on classical test theory (CTT). Initial investigations of TIRT suggest that it may solve many well-known issues with existing scoring methods; however, little empirical research has been conducted to support these claims. Further, empirical findings from direct comparisons between CTT and TIRT have been mixed. The purpose of the present study is to extend previous empirical research examining CTT and TIRT criterion-related validity scoring comparisons to a range of applied selection contexts.

Personality Faking

In high-stakes situations such as selection testing for employment, assessment takers have strong motivation to portray themselves in a favorable light. Unlike tests of knowledge, skills, or ability, with verifiably correct answers, traditional personality assessments present opportunities for job applicants to portray themselves inaccurately. This is particularly concerning as faking has been found to impact hiring decisions by changing the ranking of applicants (Christiansen, Goffin, Johnston, & Rothstein, 1994), negatively affecting criterion validity (Komar, Brown, Komar, & Robie, 2008), and affecting the psychometric properties of assessments (Schmit & Ryan, 1993).

One modern method of combating faking in personality assessments has been the introduction of forced-choice (FC) personality measures (Christiansen, Burns, & Montgomery, 2005). Contrary to traditional Likert-type, single-stimulus response formats (where respondents indicate the degree

Corresponding author:
Peter A. Fisher
Email: fish0150@mylaurier.ca

to which a single statement describes them), FC response formats present blocks of statements from which applicants must choose from equally desirable self-descriptions. FC measures of personality have been found to reduce applicants' ability to fake as it is more difficult to determine the "correct" response to any given block of statements and increases the cognitive load involved in impression management (Tett & Simonet, 2011).

Traditional classical test theory (CTT) scoring of FC assessments involves adding the inverted rank order of items in their blocks to their respective scales. As a result, a fixed number of points are allocated to an individual within each block, and so the same total number of points are allocated on each assessment. Ultimately, CTT scoring of multidimensional FC measures is, to varying extents, *ipsative*: Trait scores are relative within person rather than absolute on a normative scale. Ipsative scores present a variety of problems in selection testing (Brown & Maydue-Olivares, 2011). In particular, ipsative scores are limited in their ability to make meaningful comparisons between individuals, which is critical to employee selection. Furthermore, construct validity, criterion validity, and reliability estimates are all distorted as scale scores are inherently negatively correlated, regardless of true-score relationships, and measurement errors are not independent (cf. Johnson, Wood, & Blinkhorn, 1988; Meade, 2004). Despite these concerns, the increased difficulty in faking on FC assessments results in comparable to slightly improved criterion-related validity versus single-stimulus assessments (Christiansen et al., 2005; Salgado & Táuriz, 2014).

Thurstonian Item Response Theory

Maydue-Olivares and Böckenhalt (2005) introduced a promising technique to recover normative scores from a forced-choice instrument, thus overcoming the major weakness involved in CTT scoring of forced-choice instruments by allowing for direct between-person comparisons, which is critical to high-stakes assessments and employment selection. Brown and Maydue-Olivares (2011; 2012; 2013) have since named this framework Thurstonian item response theory or TIRT (for a basic introduction to TIRT, which is well beyond the scope of this article, see Dueber, Love, Toland, & Turner, 2019). Work by Brown and Maydue-Olivares (2013) has found measurement properties to improve using TIRT, including increased reliability, positive correlations among scale scores, and a cleaner factor structure. Furthermore, differences in criterion-related validity for employed call center operators were found, with an average .09 difference favoring TIRT over CTT across scales for a measure of personality predicting incentive bonus (an outcome awarded to employees based upon various performance indicators, $N = 219$). On the other hand, P. Lee, S. Lee, and Stark (2018) did not find improvement in the criterion-related validity of TIRT estimates over forced choice CTT estimates in the prediction of nonwork external measures in a sample of university students ($N = 417$).

Although initial work on TIRT is promising, these conflicting results highlight the need for additional research into the use of TIRT scoring in high-stakes assessment situations. Both theoretically and via these empirical findings, it has been shown that TIRT is a potentially promising method that exhibits favorable FC features (i.e., nontransparent items) while avoiding the undesirable psychometric properties that plague traditional CTT FC scoring. However, implementing TIRT scoring in an applied setting can be challenging, given the more stringent data requirements and substantially more complex models necessary to derive scores. To offset the intensive data and modeling needs, TIRT should therefore consistently demonstrate better psychometric properties than the simpler CTT FC method. In employee selection contexts, this would involve improved estimates of criterion-related validity. Unfortunately, there have only been a handful of studies that have examined the criterion-related validity of TIRT-developed scales, which we reference above, and results from these were mixed. Furthermore, only one of these studies was conducted in an actual work context and used actual work criteria. To more confidently support the use of TIRT in high-stakes preemployment settings, research must be conducted to show superiority for TIRT across many work settings and to generalize those findings across different forced-choice scales.

In order to extend the existing research examining CTT and TIRT criterion-related validity, we directly compared criterion-related validity estimates for the two scoring methods for an existing, proprietary personality assessment across 11 concurrent validity sample data sets, meta-analytically corrected for measurement artifacts. Based on the previous research findings and theory presented above, it was expected that TIRT scoring would result in better criterion validity than traditional CTT scoring for selection testing and thus better selection outcomes. Contrary to our expectations, CTT scoring vastly outperformed TIRT scoring for the samples and assessment involved in this study. Below we present our methods and results in detail, and discuss these findings, as well as a potential explanation for our unexpected results.

METHOD

Participants and Procedures

This study included one data set that was used for purposes of calibrating the TIRT model estimates ($n = 12{,}018$) and 11 concurrent validity data sets used as a collective validity sample (total $N = 612$). The organization owning the proprietary FC measure that provided the data does not

collect demographic information. Three of the data sets came from an international marketing firm with jobs titled assistant marketing manager (n = 115), marketing coordinator (n = 54), and marketing executive (n = 75). The remaining eight data sets came from a sales-based organization – although all of the jobs were sales representatives, their roles were deemed different enough by subject matter experts to require different personality competencies: business-to-business (n = 23), energy (n = 48), insurance (n = 28), Internet 1 (n = 112), Internet 2 (n = 22), multiproduct (n = 15), personal security (n = 16), and television (n = 108). Participants were incumbents, instructed to complete the assessment in an unproctored setting for research purposes. Participants were informed that the assessment was meant to help determine hiring criteria for their jobs, and so they had little incentive to distort their responses. Indeed, meta-analytic evidence suggests that incumbents' personality assessment scores tend to be much more consistent with experimental samples instructed to respond honestly than with job applicants or experimental samples instructed to fake (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006).

Measures

Personality. The personality measure used in this study is a five-factor model (FFM)-based (Goldberg, 1992) commercial instrument based on DeYoung, Quilty, and Peterson's (2007) 10-factor (2 facets per FFM) model that uses a partially ipsative, forced-choice methodology (see Salgado, Anderson, & Táuriz, 2015, for additional information on partially vs. fully ipsative FC measures). Facet-level scores were used in the current study because prediction of work outcomes is generally better at a level below the FFM (Christiansen & Robie, 2011), and the instrument was initially constructed with this in mind. Facet dimensions, number of items per facet dimension, test–retest reliabilities (n = 124 from a student sample with a 2- to 4-week retesting interval),[1] broad dimensions, and definitions can be found in Appendix A and B. Assessment takers are presented with 60 items arranged in 20 triplets that have been matched on attractiveness (i.e., social desirability). For each triplet, respondents are asked to choose the statement that is "least like you" and "most like you" (e.g., "I tend to take an interest in other people's lives," "I don't mind taking charge," and "I usually need a creative outlet").

The CTT scoring of the measure was straightforward. If all statements in a triplet were positively keyed, items were scored 2 if chosen as "most like me," 0 if chosen as "least like me," and 1 if not chosen. If all statements in a triplet were negatively keyed, items were scored 0 if chosen as "most like me," 2 if chosen as "least like me," and 1 if not chosen. Scores were then summed across triplets for each facet. No calibration sample was necessary.

The TIRT scoring was more complex, the details of which are beyond the scope of this paper (see Brown & Maydue-Olivares, 2012 for details). TIRT is a model-based scoring methodology with large numbers of parameters to estimate so a calibration sample large enough to reliably estimate said parameters is recommended (Brown & Maydue-Olivares, 2011). The TIRT model fit the 10-factor model in our calibration sample well (RMSEA = .029) using MPlus 7.4 (Muthén, & Muthén, 2015). Model estimates from this calibration sample were used to score personality in the validity sample data sets that contained job performance data.

For jobs within each validity sample, personality composite scores were formed that only included the personality traits deemed relevant to each job (see Tett & Christiansen, 2007). These were formed to reflect the total composite that might be used when making decisions regarding which applicants to hire by combining only those trait scores theoretically linked to performance. The composites, which were formed for both CTT and TIRT scoring methods, aggregated personality facet scores separately for each of the 11 validity sample data sets. Decisions on which facets were relevant to each job were made by subject matter experts who were experts in both personality psychology and knowledgeable of the requirements for each of the jobs.

Job performance. Job performance was measured differently for the marketing versus the sales jobs, both according to the standards of the client organizations. Marketing job performance was provided by the incumbent's supervisor, which was an aggregate of several scales, including an average of key performance indicators for each role, an average proficiency rating of several skills, and an overall subjective rating. Sales job performance was based on a combination of multiple objective performance scores, adjusted for several organization-specific factors such as region and quota.

RESULTS

Criterion-related validity coefficients for the 11 different samples for CTT and TIRT scoring methods are presented in Table 1.[2] Tests of dependent correlations ($|Z|$; Steiger, 1980) were used to compare the correlations across scoring methods. Contrary to expectations, all of the comparisons favored CTT over TIRT; two of these were statistically

1 Please note that test-retest reliability is generally considered to be more appropriate for estimating FC reliability (O'Neill, et al., 2017).

2 Note that the operational validity estimates for CTT scoring may be lower than the ones used for commercial purposes by the test vendor because some personality assessment items were eliminated from the present analyses. These are composed of 50 adjectives, presented in groups of 10, for which each group asks respondents to choose the three adjectives that were "most like you" and three adjectives that were "least like you." CTT-based estimates could be derived from these adjectives, but a TIRT model to score these could not be identified. Thus, analyses in the present study were restricted to the triplet statements so that a fair comparison could be made between scoring methods.

TABLE 1.
Personality Criterion-Related Validity Estimates for CTT and TIRT Scoring Methods

Sample	*n*	CTT	TIRT	|*Z*|
Assistant marketing manager	114	.17*	-.17	2.57*
Marketing coordinator	54	.47**	.04	3.03**
Marketing executive	73	.15	-.12	1.40
Sales (B2B)	23	.23	-.12	1.20
Sales (energy)	48	.21‡	.11	0.50
Sales (insurance)	28	.29‡	-.06	1.36
Sales (Internet 1)	112	.29**	.10	1.93‡
Sales (Internet 2)	22	.32‡	.31‡	0.03
Sales (multiple product lines)	15	.70**	.67**	0.16
Sales (personal security)	16	.18	.28	0.64
Sales (television)	108	.22*	-.07	1.95‡

Note. *$p < .05$. **$p < .01$. ‡$p < .10$. Zero-order correlations were one-tailed. Tests of differences in dependent correlations (|*Z*|) were two-tailed.

significant at conventional levels ($p < .05$) and two more at a relaxed level ($p < .10$). Given the low sample sizes for some of the correlations, psychometric meta-analysis was used to compare aggregated correlations across scoring methods (Schmidt & Hunter, 2015; Schmidt & Le, 2014). The validity generalization meta-analytic framework developed by Schmidt and Hunter (1977) was designed explicitly for this purpose. Meta-analysis sample-size weights the correlations before aggregating to ensure that those derived from larger samples have better representation in the average than those from smaller samples. Omitting correlations from the smaller samples would therefore provide a worse estimate of the population validity coefficient than weighting them appropriately. In fact, after controlling for sample size and other artifacts, there was no remaining variance in the correlations for CTT and very little for TIRT. Validity estimates were corrected for criterion unreliability using an estimate of .52 (Viswesvaran, Ones, & Schmidt, 1996) and average indirect range restriction of .89 (Salgado & Táuriz, 2014).

Meta-analysis of personality criterion-related validity estimates for CTT and TIRT scoring methods can be found in Table 2. The sample size-weighted mean observed correlations were .25 for CTT and .00 for TIRT. The sample-weighted mean corrected correlations were .38 for CTT and .00 for TIRT. Standard deviation of rho was 0 for the CTT scoring method but was .14 the TIRT scoring method. Thus, no variance remained in the criterion-related validity estimates for the CTT scoring method after sample size, criterion unreliability, and indirect range restriction were accounted for, in contrast to the considerable variance remaining for the TIRT scoring method. The confidence intervals for the validity estimates from the two scoring methods did not overlap, indicating that the aggregated criterion-related validity estimate for CTT scoring was higher than that for TIRT.[3]

DISCUSSION

There are many positive aspects to FC assessments in high-stakes testing settings, and TIRT has been promoted as a solution to the problems commonly found when CTT is used to score these measures. However, these findings contrast starkly with what one might expect upon implementing TIRT scoring for a FC employment selection assessment, particularly when considering the theoretical advantages that have previously been proposed. In an applied setting, using real-world incumbents and job performance criteria, TIRT scoring resulted in negligible criterion validity. On the other hand, CTT scoring resulted in acceptable criterion-related validity in the prediction of job performance outcomes. Compared to traditional CTT scoring of the same data, TIRT was clearly inferior and implementation would not have resulted in any benefit in a selection scenario. It is also worth noting that these differences may actually be understated in the results presented: In an attempt to compare the two scoring methods as fairly as possible, no minor modifications were made to the CTT scoring method that are otherwise included in practice by the test vendor. For example, modifications such as differential weighting of specific items that have empirically demonstrated higher reliabilities, or allowing for cross-loading items that have evidenced significant facet overlap, were

3 To help assuage doubts about our use of some of the smaller samples in the meta-analysis, we estimated the sample-weighted mean correlations omitting studies with n < 30 that an anonymous reviewer identified as being potentially problematic: For CTT the sample-weighted mean correlation was .24 (compared to .25 estimate including them), and for TIRT the sample-weighted mean correlation was -.03 (compared to 0 estimate including them). Thus, the substantive conclusion would be unchanged. In fact, omitting small N samples (and pretending they do not exist) will actually bias estimation of the population validity coefficients as compared to including them and weighting appropriately.

TABLE 2.
Meta-Analysis of Personality Criterion-Related Validity Estimates for CTT and TIRT Scoring Methods

Scoring	*N*	*k*	*r*	*SD*r	Rho	*SD*Rho	*SD*Rc	95% CI	80% CrI
CTT	613	11	.25	.11	.38	.00	.16	[.29, .48]	[.38, .38]
TIRT	613	11	.00	.16	.00	.14	.25	[-.15, .15]	[-.17, .17]

Note. *N* = total sample size. *k* = total number of studies. *r* = sample size weighted mean observed correlation. *SD*r = standard deviation of the observed correlation. Rho = mean true score correlation. *SD*Rho = standard deviation of the true score correlation. *SD*Rc = observed standard deviation of the corrected correlations. 95% CI = lower and upper values of the 95% confidence interval. 80% CrI = lower and upper values of the 80% credibility interval.

not implemented when computing the CTT results reported above. As a result, the criterion validity estimates evidenced by CTT scoring would be expected to be slightly higher in practice, further widening the gap between the two scoring methods.

Ultimately, the results presented here are relatively consistent with existing research comparing classical test theory scoring of personality assessments with item response theory counterpart, where IRT-derived scoring does not tend to improve trait estimations (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Ferrando & Chico, 2007; Ling, Zhang, Locke, Li, & Li, 2016; Xu & Stone, 2012), and specifically for selection purposes (Speer, Robie, & Christiansen, 2016). Although TIRT is not without theoretical merits, and assessments constructed with TIRT scoring in mind may be useful for other purposes (e.g., low-stakes, developmental assessments; although more research is certainly required to make that claim), it is clear that applying TIRT scoring to an assessment that was designed to be scored with a CTT methodology may result in inadequate criterion validity. Overall, the CTT scoring method provided adequate and expected levels of criterion-related validity, consistent with the original goal and value proposition of the assessment. Thus, these results serve as a warning against the blind implementation of TIRT scoring over traditional CTT on existing FC assessments without conducting rigorous validation.

Reconciling Results: Trait Retrieval

Although the results we present above favor CTT over TIRT, there are several design factors to consider, particularly with respect to how the mix of response options within an item block can impact trait recovery. In their seminal simulation study, Brown and Maydue-Olivares (2011) demonstrate that blocks of homogeneously keyed (either all positively keyed or all negatively keyed) items merely highlight differences in the latent traits. Thus, TIRT-derived scores for these homogeneously keyed blocks draw conclusions about the relative positions of the underlying traits, rather than absolute, normative locations, similar to CTT scores. When *all* blocks in an assessment are homogenously keyed, trait retrieval may be poor using TIRT, as little information is provided on absolute trait location.

Notably, the proprietary FC assessment involved in this study consists of entirely homogeneously keyed blocks of statements (e.g., "I tend to take an interest in other people's lives," "I don't mind taking charge," and "I usually need a creative outlet"). This was done by the test developers so that the response options could be equated on attractiveness in order to minimize applicant faking of the FC measure. The results above suggest that CTT scoring of such a measure produces acceptable levels of criterion-related validity for use in selection testing. Such a design is useful in contexts where only *some* of the scored traits will be considered important and hence when relative trait standing matters, and negative correlations between scores can be overlooked. The introduction of TIRT scoring, however, creates additional burdens that may not have been considered in the development of many contemporary FC personality assessments. In particular, it is especially difficult to match social desirability of heterogeneously keyed items within blocks, which is critical to maintaining construct validity in applicant contexts where faking is likely. As Heggestad, Morrison, Reeve, and McCloy (2006) noted in their development of a multidimensional FC personality assessment, despite rigorous attempts to match items on social desirability "respondents became reluctant to indicate that a statement indicative of a high standing was 'least like me' and that a statement indicative of a low trait standing was 'most like me' under conditions of faking" (p. 21). Failure to take adequate steps to prevent faking can seriously undermine criterion-related validity when a personality assessment is given to actual job applicants (Tett & Christiansen, 2007).

Interestingly, Lee and colleagues (2018) *did* construct

an FC assessment with TIRT scoring in mind and included heterogeneously keyed blocks of items to ensure that trait retrieval was not a concern. Nevertheless, the findings these authors present continue to cast doubt on the usefulness of TIRT scoring in applied, high-stakes testing, as the average criterion-related validity presented for even nonwork-related criteria tended to be smaller for TIRT, ultimately *in favor of CTT scoring*. In fact, recent theoretical and simulation work being conducted by Bürkner, Schulte, and Holling (2018) bring the authors to a very similar conclusion. Taken together with the current study, this highlights the uncertainty in understanding exactly what variance is being captured by TIRT that is unique from CTT. Because TIRT should better assess the latent trait domain, according to basic validity theory (Binning & Barrett, 1989) it would be expected that TIRT-derived scores should correlate more with theoretically linked outcomes such as job performance. However, at present it is difficult to argue that unique variance captured by TIRT represents true-score variance. Perhaps this setting was one where relative standing across traits was more important than the absolute standing within traits (in which case TIRT could reflect unique true score variance that simply wasn't well-aligned to the performance criteria), or perhaps these findings are specific to the small number of instruments used and studies conducted thus far. Either way, more research on this topic is warranted.

Analysis at the Dimension Level

The analyses presented above are conducted at the practical, composite level, where composites are comprised of various dimensions at the discretion of subject-matter experts. This parallels the scoring system used for decision making in high-stakes testing situations but may make the interpretation of the differences between scoring approaches more difficult from an academic perspective. Three correlation matrices within and between scoring methods at the dimension level, one for each of the calibration, marketing, and sales samples can be seen in the supplemental material. These correlation matrices highlight a possible alternative perspective on the findings we have presented above.[4]

It is apparent that TIRT scoring tended to result in expected positive correlations between personality dimensions, which contrasts starkly with the negative correlations that result from ipsative CTT scoring. However, dimensions that correlate relatively highly will contribute less unique information about a criterion when placed into a composite, which could help to explain the difference in validity between the two scoring methods. As can be seen in Table 3, the average correlations between dimensions within a

TABLE 3.

Meta-Analysis of Personality Criterion-Related Validity Estimates for CTT and TIRT Scoring Methods

	Average correlation between dimensions		Average criterion-related validity of dimensions	
Job	CTT	TIRT	CTT	TIRT
Business-to-business	-.03	.39	.12	-.11
Energy	-.13	.32	.09	.08
Insurance	-.08	.48	.10	-.04
Internet 1	-.10	.29	.12	.07
Internet 2	-.05	.32	.14	.20
Multiproduct	-.15	.31	.21	.48
Personal security	--	--	.18	.28
TV	-.06	.26	.11	-.04
Asst. marketing manager	-.14	.24	.05	-.10
Marketing coordinator	-.07	.43	.19	.04
Marketing executive	-.15	.15	.06	-.08

Note. The composite for personal security contains only a single dimension.

4 We thank an anonymous reviewer for helping us to identify this alternative perspective.

composite for a job tended to be small and negative under CTT scoring, but larger and positive under TIRT scoring, consistent with this line of reasoning. However, a comparison of the average dimension-level criterion-related validity estimates for each scoring method suggest that CTT scores result in better average criterion-related validity, even at the dimension-level, and the same substantive conclusion as above. Thus, the source of the differences in the validity of composite scores appears to be due both to a decline in criterion-related validity of the TIRT dimension scores as well as an increase in the inter-correlation of the dimension scores contributing to the composite. TIRT scoring reduces the amount that choices related to the dimensions identified as relevant in the job analyses increase scores on the composite, relative to simpler CTT scoring; hence, TIRT results in less criterion-related validity (cf. Christiansen et al., 2005).

Concluding Remarks on Implementing TIRT

Theoretically, TIRT has the potential to provide the "best of both worlds" in personality assessments when it comes to reducing applicant faking while also solving issues related to ipsative scores. The results presented here indicate that TIRT scoring should not be blindly implemented to replace CTT scoring on existing FC personality assessments in practice. As demonstrated in the present study, TIRT assessment scoring does not necessarily represent a panacea for high-stakes assessment situations. Assessments that were not originally constructed or validated with TIRT in mind may not be suitable candidates for TIRT scoring. Thus, care in development of FC assessments, as well as rigorous, empirical, concurrent validation should be undertaken before implementing TIRT scoring in applied assessments.

REFERENCES

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. International Journal of Selection and Assessment, 14(4), 317-335. doi: 10.1111/j.1468-2389.2006.00354.x

Brown, A., & Maydue-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. Educational and Psychological Measurement, 71(3), 460-502. doi: 10.1177/0013164410375112

Brown, A., & Maydue-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. Behavioral Research Methods, 44, 1135-1147. doi: 10.3758/s13428-012-0217-x

Brown, A., & Maydue-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. Psychological Methods, 18(1), 36-52. doi: 10.1037/a0030641

Bürkner, P. C., Schulte, N., & Holling, H. (2018). On the statistical and practical limitations of Thurstonian IRT models. Manuscript submitted for publication. Educational and Psychological Measurement. Preprint doi: 10.31234/osf.io/dbwn8

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. doi: 10.1037/1040-3590.19.1.88

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. Human Performance, 5(3), 267-307. doi: 10.1207/s15327043hup1803_4

Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. Personnel Psychology, 47(4), 847–860. doi: 10.1111/j.1744-6570.1994.tb01581.x.

Christiansen, N. D., & Robie, C. (2011). Further consideration of the use of narrow trait scales. Canadian Journal of Behavioural Science, 43(3), 183-194. doi: 10.1037/a0023069

Dueber, D. M., Love, A. M. A., Toland, M. D., & Turner, T. A. (2019). Comparison of single-response format and forced-choice format instruments using Thurstonian item response theory. Educational and Psychological Measurement, 79(1), 108-128. doi: 10.1177/0013164417752782

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. Journal of Personality and Social Psychology, 93(5), 880-896. doi: 10.1037/0022-3514.93.5.880

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.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. Psychological Assessment, 4(1), 26-42. doi: 10.1037/1040-3590.4.1.26

Griffith, R. L., & Robie, C. (2013). Personality testing and the "F-word": Revisiting seven questions about faking. In N. Christiansen, & R. Tett (Eds.), Handbook of personality at work (pp. 253–280). New York: Taylor & Francis

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. Journal of Applied Psychology, 91(1), 9-24. doi: 10.1037/0021-9010.91.1.9

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. Journal of Occupational and Organizational Psychology, 61(2), 153-162. doi: 10.1111/j.2044-8325.1988.tb00279.x

Kelly, E. L., Miles, C. C., & Terman, L. M. (1936). Ability to influence one's score on a typical paper-and-pencil test of personality. Journal of Personality, 4(3), 206–215. doi: 10.1111/j.1467-6494.1936.tb02123.x.

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. Journal of Applied Psychology, 93(1), 140–154. doi: 10.1037/0021-9010.93.1.140.

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. Personality and Individual Differences, 123, 229-235. doi: 10.1016/j.paid.2017.11.031

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. doi: 10.1080/00223891.2015.1077852

Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. Psychological Methods, 10(3), 285-304. doi: 10.1037/1082-989X.10.3.285

Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. Journal of Occupational and Organizational Psychology, 77, 531–552. doi: 10.1348/0963179042596504

Muthén, L. K., & Muthén, B. O. (2015). Mplus user's guide (7th ed.). Los Angeles, CA: Muthén & Muthén.

O'Neill, T. A., Lewis, R. J., Law, S. J., Larson, N., Hancock, S., Radan, J., Lee, N., & Carswell, J. J. (2017). Forced-choice pre-employment personality assessment: Construct validity and resistance to faking. Personality and Individual Differences, 115, 120-127. doi: 10.1016/j.paid.2016.03.075

Salgado, J. F., Anderson, N., & Táuriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: a comprehensive meta-analysis. Journal of Occupational and Organizational Psychology, 88, 797-834. doi: 10.1111/joop.12098

Salgado, J. F., & Táuriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. European Journal of Work and Organizational Psychology, 23(1), 3-30. doi: 10.1080/1359432X.2012.716198

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62(5), 529-540. doi: 10.1037/0021-9010.62.5.529

Schmidt, F. L., & Hunter, J. E. (2015). Methods of meta-analysis (3rd ed.). Los Angeles, CA: Sage.

Schmidt, F. L., & Le, H. (2014). Software for the Hunter-Schmidt meta-analysis methods (Version 2.0). University of Iowa, Department of Management & Organizations, Iowa City, IA.

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. Journal of Applied Psychology, 78(6), 966–974. doi: 10.1037/0021-9010.78.6.966.

Speer, A. B., Robie, C., & Christiansen, N. D. (2016). Effects of item type and estimation method on the accuracy of estimated personality trait scores: Polytomous item response theory models versus summated scoring. Personality and Individual Differences, 102, 41-45. doi: 10.1016/j.paid.2016.06.058

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87(2), 245-251. doi: 10.1037/0033-2909.87.2.245

Tett, R.P. & Christiansen, N.D. (2007). Personality tests at the crossroads: A reply to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt. Personnel Psychology, 60, 267-293. doi:10.1111/j.1744-6570.2007.00098

Tett, R. P., & Simonet, D. V. (2011). Faking in personality assessment: A "multisaturation" perspective on faking as performance. Human Performance, (4), 302–321. http://dx.doi.org/10.1080/08959285.2011.597472.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. Journal of Applied Psychology, 81(5), 557-574. doi: 10.1037/0021-9010.81.5.557

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

RECEIVED 07/27/18 ACCEPTED 01/14/19

Appendix A
Facet Dimensions, Number of Items per Facet Dimension, Test–Retest Reliabilities, Broad Dimensions, and Definitions

Facet dimension	# Items	Test–retest Reliability	FFM dimension	Definition
Compassion	6	.41	Agreeableness	The extent that someone shows empathy, sympathy, and warmth toward others; shows a tendency for being understanding and forgiving of mistakes. It is the degree to which someone is forgiving, helpful, and trusting.
Mannerliness	6	.43	Agreeableness	The extent that someone is pleasant, willing to cooperate, and considerate. It is the degree to which someone is modest, unassuming, and courteous. Mannerliness is sometimes referred to as compliance and politeness.
Industriousness	5	.42	Conscientiousness	The extent that someone maintains high standards, aspires to challenging goals, and is willing to put forth extra effort. It is the degree to which someone is purposeful, efficient, and ambitious.
Orderliness	7	.71	Conscientiousness	The extent that someone acts with deliberation, is focused on quality, and prefers to be organized and have a plan. It is the degree to which someone is thorough, methodical, and organized.
Assertiveness	6	.69	Extraversion	The extent that someone voices their opinions and is comfortable being the center of attention and giving direction to other employees. It is the degree to which someone is influential, persuasive, and self-confident.
Enthusiasm	9	.57	Extraversion	The extent that someone is interested in meeting new people, initiates conversations, and is comfortable in social interactions. It is the degree to which someone is talkative, outgoing, and sociable.
Self-regard	5	.59	Emotional Stability	The extent that someone has a positive self-image, is satisfied with who they are as a person and tends to be self-assured and optimistic. It is the degree to which someone is content, secure, and cheerful.
Stability	5	.53	Emotional Stability	The extent that someone is calm under pressure, even tempered, and resistant to the effects of stress and unexpected changes. It is the degree to which someone is calm, steady, and composed.
Experiential disposition	7	.56	Openness	The extent that someone seeks out new and different experiences, adapts to changes in the workplace, and is tolerant of differences between people. It is the degree to which someone is flexible, unconventional, and reflective.
Intellectual disposition	4	.67	Openness	The extent that someone enjoys learning new things, is interested in different ideas, and tends to imagine how things could be different. It is the degree to which someone is analytical, curious, and imaginative.

Appendix B
Facet Dimensions, Number of Items per Facet Dimension, Test–Retest Reliabilities, Broad Dimensions, and Definitions

Job/Role	Facet (FFM) Dimensions
Business-to-Business	Industriousness (C) Self-Regard (S) Experiential Disposition (O) Intellectual Disposition (O)
Energy	Compassion (A) Mannerliness (A) Industriousness (C) Assertiveness (E)
Insurance	Compassion (A) Assertiveness (E) Enthusiasm (E) Experiential Disposition (O) Intellectual Disposition (O)
Internet 1	Compassion (A) Assertiveness (E) Enthusiasm (E) Stability (S)
Internet 2	Industriousness (C) Self-Regard (S) Stability (S) Intellectual Disposition (O) Experiential Disposition (O)
Multi-Product	Compassion (A) Orderliness (C) Assertiveness (E) Enthusiasm (E) Stability (S)
Personal Security	Orderliness (C)
TV	Assertiveness (E) Enthusiasm (E) Stability (S) Experiential Disposition (O)
Assistant Marketing Manager	Industriousness (C) Orderliness (C) Assertiveness (E) Enthusiasm (E) Stability (S)
Marketing Coordinator	Industriousness (C) Orderliness (C) Assertiveness (E) Self-Regard (S)
Marketing Executive	Industriousness (C) Assertiveness (E) Stability (S) Intellectual Disposition (O)

Note. A = Agreeableness. C = Conscientiousness. E = Extraversion. S = Emotional Stability. O = Openness to Experience.

Supplemental Materials
Calibration Sample Correlations

				CTT									
		M	SD	1	2	3	4	5	6	7	8	9	10
1	COMPASS (CTT)	5.97	2.60										
2	MANNER (CTT)	3.55	2.22	.21									
3	INDUST (CTT)	6.31	1.80	-.10	-.12								
4	ORDER (CTT)	8.11	3.19	-.09	-.01	.22							
5	ASSERT (CTT)	7.09	2.42	-.20	-.44	-.15	-.23						
6	ENTHUS (CTT)	8.82	3.31	-.12	-.32	-.35	-.44	.17					
7	SEFRE (CTT)	3.55	1.75	-.26	-.13	-.10	-.14	.03	-.02				
8	STABLE (CTT)	3.87	2.06	-.19	.08	-.11	-.18	-.28	.07	.18			
9	EXPER (CTT)	7.48	2.23	-.23	-.12	.05	-.22	-.07	-.13	-.14	-.15		
10	INTELL (CTT)	5.25	1.93	-.20	-.03	-.09	-.11	.11	-.24	-.13	-.27	.21	
11	COMPASS (TIRT)	0.02	0.83	.87	.34	-.26	-.04	-.36	.02	-.25	-.20	-.20	-.17
12	MANNER (TIRT)	0.03	0.82	.28	.80	-.22	.17	-.53	-.14	-.04	-.03	-.21	-.14
13	INDUST (TIRT)	0.00	0.71	.00	.10	.68	.49	-.27	-.35	.02	-.30	-.16	-.12
14	ORDER (TIRT)	0.01	0.85	.07	.23	.12	.86	-.33	-.32	-.07	-.19	-.35	-.25
15	ASSERT (TIRT)	0.00	0.77	-.09	-.28	-.15	.08	.55	.24	.23	-.55	-.20	-.03
16	ENTHUS (TIRT)	0.03	0.81	.20	.02	-.48	-.29	-.02	.74	.09	-.04	-.29	-.32
17	SEFRE (TIRT)	0.03	0.71	-.07	.09	-.31	-.05	.03	.28	.60	.12	-.44	-.30
18	STABLE (TIRT)	0.04	0.79	.19	.46	-.23	.11	-.70	-.01	.22	.61	-.28	-.38
19	EXPER (TIRT)	0.02	0.80	.11	.20	-.17	.00	-.25	-.12	-.10	-.29	.47	.16
20	INTELL (TIRT)	0.02	0.75	.38	.38	-.24	-.06	-.25	-.19	-.12	-.40	.20	.31

		TIRT								
		11	12	13	14	15	16	17	18	19
12	MANNER (TIRT)	.61								
13	INDUST (TIRT)	.11	.33							
14	ORDER (TIRT)	.28	.56	.66						
15	ASSERT (TIRT)	.08	.09	.30	.30					
16	ENTHUS (TIRT)	.50	.42	-.06	.09	.49				
17	SEFRE (TIRT)	.21	.47	.21	.33	.62	.69			
18	STABLE (TIRT)	.41	.66	.07	.37	-.30	.35	.49		
19	EXPER (TIRT)	.45	.54	.25	.27	.31	.30	.25	.19	
20	INTELL (TIRT)	.67	.66	.24	.26	.30	.32	.26	.19	.87

Note: n = 12,018; Values $|r| > .02$ are significant at the $p < .01$ level (two-tailed); Values $|r| > .01$ are significant at the $p < .05$ level (two-tailed)

Marketing Sample Correlations

				CTT									
		M	SD	1	2	3	4	5	6	7	8	9	10
1	COMPASS (CTT)	6.28	2.30										
2	MANNER (CTT)	5.01	2.29	.07									
3	INDUST (CTT)	6.74	1.99	-.12	-.19**								
4	ORDER (CTT)	8.02	3.18	.09	.06	.15*							
5	ASSERT (CTT)	5.91	2.55	-.14*	-.46**	-.03	-.23**						
6	ENTHUS (CTT)	8.40	3.47	-.12	-.28**	-.27**	-.52**	.18**					
7	SEFRE (CTT)	3.61	1.89	-.11	-.06	-.05	-.04	-.21**	-.17**				
8	STABLE (CTT)	4.12	2.28	-.28**	.15*	-.05	-.16*	-.42**	-.02	.27**			
9	EXPER (CTT)	7.53	2.32	-.30**	-.15*	.00	-.35**	-.01	.01	-.22**	-.09		
10	INTELL (CTT)	4.38	1.95	-.11	.00	-.27**	-.10	.17**	-.25**	-.13*	-.28**	.22**	
11	COMPASS (TIRT)	0.22	0.76	.83**	.29**	-.32**	.09	-.34**	.03	-.13*	-.22**	-.25**	-.09
12	MANNER (TIRT)	0.59	0.84	.21**	.81**	-.25**	.25**	-.58**	-.16*	.04	.07	-.27**	-.10
13	INDUST (TIRT)	0.28	0.76	.05	.10	.69**	.47**	-.25**	-.36**	.10	-.17**	-.27**	-.25**
14	ORDER (TIRT)	0.27	0.88	.19**	.29**	.08	.86**	-.38**	-.38**	.05	-.10	-.47**	-.24**
15	ASSERT (TIRT)	-0.10	0.75	.08	-.26**	-.06	.14*	.50**	.18**	.09	-.62**	-.24**	.00
16	ENTHUS (TIRT)	0.18	0.80	.20**	.06	-.41**	-.26**	-.10	.71**	.04	-.07	-.24**	-.28**
17	SEFRE (TIRT)	0.24	0.73	.03	.17**	-.20**	.07	-.19**	.12	.60**	.18**	-.49**	-.31**
18	STABLE (TIRT)	0.48	0.90	.10	.47**	-.21**	.16*	-.77**	-.09	.39**	.68**	-.29**	-.35**
19	EXPER (TIRT)	0.25	0.75	.03	.20**	-.18**	.01	-.27**	-.10	-.08	-.21**	.46**	.20**
20	INTELL (TIRT)	0.20	0.71	.34**	.37**	-.29**	.05	-.25**	-.21**	-.07	-.36**	.15*	.39**

		TIRT								
		11	12	13	14	15	16	17	18	19
12	MANNER (TIRT)	.60**								
13	INDUST (TIRT)	.13	.34**							
14	ORDER (TIRT)	.38**	.62**	.64**						
15	ASSERT (TIRT)	.21**	.09	.34**	.32**					
16	ENTHUS (TIRT)	.55**	.43**	-.01	.12	.47**				
17	SEFRE (TIRT)	.33**	.55**	.34**	.45**	.51**	.64**			
18	STABLE (TIRT)	.37**	.66**	.13*	.43**	-.32**	.32**	.60**		
19	EXPER (TIRT)	.42**	.52**	.23**	.26**	.31**	.32**	.27**	.19**	
20	INTELL (TIRT)	.66**	.65**	.23**	.32**	.36**	.33**	.31**	.17**	.85**

Note: $n = 245$; $^{*}p < 0.05$ (2-tailed); $^{**}p < 0.01$ (2-tailed)

Sales Sample Correlations

				CTT									
		M	SD	1	2	3	4	5	6	7	8	9	10
1	COMPASS (CTT)	5.77	2.53										
2	MANNER (CTT)	3.41	2.25	.18**									
3	INDUST (CTT)	5.13	1.90	-.10*	-.23**								
4	ORDER (CTT)	7.25	3.13	-.04	.02	.19**							
5	ASSERT (CTT)	7.74	2.36	-.14**	-.47**	-.01	-.21**						
6	ENTHUS (CTT)	10.74	3.49	-.05	-.25**	-.31**	-.40**	.06					
7	SEFRE (CTT)	4.27	1.95	-.24**	-.05	-.05	-.20**	.01	-.14**				
8	STABLE (CTT)	3.85	2.26	-.25**	.08	-.19**	-.25**	-.25**	.03	.24**			
9	EXPER (CTT)	6.93	2.38	-.31**	-.14**	-.05	-.24**	-.09	-.07	-.14**	-.02		
10	INTELL (CTT)	4.90	2.02	-.17**	-.04	.02	.00	.15**	-.37**	-.13**	-.30**	.15**	
11	COMPASS (TIRT)	0.14	0.83	.85**	.35**	-.29**	.05	-.34**	.13**	-.26**	-.28**	-.28**	-.20**
12	MANNER (TIRT)	0.19	0.88	.26**	.81**	-.29**	.23**	-.53**	-.05	-.04	-.08	-.25**	-.16**
13	INDUST (TIRT)	-0.23	0.76	.01	.04	.67**	.51**	-.13**	-.31**	.04	-.42**	-.29**	-.03
14	ORDER (TIRT)	-0.04	0.89	.11*	.26**	.07	.84**	-.29**	-.22**	-.11*	-.28**	-.40**	-.18**
15	ASSERT (TIRT)	0.42	0.83	.00	-.22**	-.03	.17**	.50**	.18**	.15**	-.60**	-.30**	-.02
16	ENTHUS (TIRT)	0.66	0.85	.24**	.11*	-.44**	-.19**	-.01*	.74**	-.01	-.12*	-.28**	-.41**
17	SEFRE (TIRT)	0.53	0.75	-.02	.17**	-.26**	-.01	-.01	.21**	.58**	.07	-.47**	-.33**
18	STABLE (TIRT)	0.17	0.83	.13**	.51**	-.35**	.06	-.70**	.06	.26**	.60**	-.23**	-.43**
19	EXPER (TIRT)	0.10	0.84	.08	.22**	-.23**	.08	-.25**	-.02	-.13**	-.32**	.38**	.11*
20	INTELL (TIRT)	0.11	0.78	.36**	.40**	-.25**	.09	-.23**	-.14**	-.13**	-.47**	.07	.29**

		TIRT								
		11	12	13	14	15	16	17	18	19
12	MANNER (TIRT)	.63**								
13	INDUST (TIRT)	.12**	.30**							
14	ORDER (TIRT)	.38**	.62**	.66**						
15	ASSERT (TIRT)	.19**	.17**	.50**	.42**					
16	ENTHUS (TIRT)	.58**	.50**	.01	.23**	.49**				
17	SEFRE (TIRT)	.27**	.52**	.28**	.41**	.62**	.65**			
18	STABLE (TIRT)	.38**	.65**	-.05	.33**	-.29**	.36**	.50**		
19	EXPER (TIRT)	.46**	.58**	.25**	.36**	.38**	.39**	.30**	.20**	
20	INTELL (TIRT)	.67**	.69**	.29**	.40**	.40**	.37**	.33**	.18**	.87**

Note: $n = 453$; *$p < 0.05$ (2-tailed); **$p < 0.01$ (2-tailed)