THOUGH FORCED, STILL VALID: PSYCHOMETRIC EQUIVALENCE OF FORCED-CHOICE AND SINGLE-STATEMENT MEASURES

BY

BO ZHANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Advisor:

Professor Fritz Drasgow

ABSTRACT

Forced choice (FC) measures are gaining popularity as an alternative assessment format to single statement (SS) measures due to their potential in reducing the impact of various response styles and faking. However, a fundamental question remains to be answered: do FC and SS instruments measure the same underlying constructs? In addition, FC measures are theorized to be more cognitively challenging, so how would this feature influence respondents' reactions to FC measures compared to SS? Two studies were designed to answer these questions. Study 1 results showed that FC measures scored by the Multi-unidimensional Pairwise Preference Model (MUPP) and SS measures scored with an ideal point model yielded similar factor structures and almost identical criterion-related validity across 12 criteria. Both formats also had similar pattern of marginal reliabilities and test-retest reliabilities. Study 1 findings were replicated in Study 2. In addition, we found strong evidence for convergent validity between the two formats. Though the FC format was perceived to be more difficult, respondents showed no differential preference and expressed similar level of emotional and cognitive reactions to the two formats.

Keywords: forced choice, single statement, equivalence, MUPP, ideal point model

ACKNOWLEDGEMENTS

I would like to thank my research advisor and mentor Dr. Fritz Drasgow for his insightful guidance, incredible patience with me, and the precious research opportunities and full academic autonomy he generously provides. Secondly, I would like to extend my gratitude to Dr. Yihao Liu for sparing his precious time to serve as the second reader of this thesis despite his tight schedule. I would also like to thank Tianjun Sun, Dr. Olexander Chernyshenko, Dr. Christopher Nye, Dr. Stephen Stark, and Dr. Len White for their invaluable help with data collection and invaluable feedback on earlier drafts of the thesis. Finally, I would like express my gratitude to my fellow graduate students and family for their support in this endeavor. Thank you all for letting me know that I am not fighting alone.

TABLE OF CONTENTS

INTRODUCTION

The scientific study of personality, attitudes, motivation, interests, psychopathology and many other psychological attributes heavily relies on the use of self-reported measures. Since its introduction by Likert in 1932, the single statement (SS) rating scale format has no doubt been the most widely used form of measurement (Brown & Maydeu-Olivares, 2013). SS measurement is, however, not without issues. Rating scale errors, such as halo, acquiescent responding, extreme response style, and mid-point response style, lead to serious concerns. In the context of 360 degree performance ratings using SS scales, for example, Yammarino (2003) concluded that "the construct validity of multisource ratings and feedback is faulty or at least highly suspect" (p. 9; see also Brown, Inceoglu, & Lin, 2017). In high stakes settings, some have found that "faking good" may render SS rating scale scores suspect (Donovan, Dwight, & Schneider, 2014) or even virtually useless (Sisson, 1948; White, Young, Hunter, & Rumsey, 2008). SS measures in the judgement and decision making literature have also been criticized (Por & Budescu, 2017) due to their well-known biases (e.g., Tversky & Kahneman, 1974). There have been many attempts to improve SS measures but so little success that Landy and Farr (1980) called for a moratorium on rating scale format research.

Recently, researchers have added yet another criticism of SS measures: namely, that the psychometric model underlying their analysis has been misspecified. Since the time of Likert (1932), dominance models have been used. These models assume a monotonic relationship between the probability of a positive response to a dichotomously scored response or the expected value of a polytomously scored response and the psychological characteristic being assessed. Rather than a dominance model, the new research suggests that ideal point psychometric models should be used; these models assume that a respondent is more likely to endorse an item when its

standing on the latent trait continuum is closer to the individual's trait value. Evidence in support of ideal point models has been found for personality assessment (Cao, Drasgow, & Cho, 2015; Chernyshenko, Stark, Drasgow, & Roberts, 2007; Ling, Zhang, Locke, Li, & Li, 2016; Stark, Chernyshenko, Drasgow, & Williams; 2006), vocational interests (Tay, Drasgow, Rounds, & Williams, 2009), job satisfaction (Carter & Dala, 2010), political efficacy (Maydeu-Olivares, Hernández, & McDonald, 2006), emotional intelligence (Cho, Drasgow, & Cao, 2015), attachment style (Sun, Fraley, & Drasgow, under review) and attitude measures (Roberts & Laughlin, 1996).

An alternative to a traditional SS scale analyzed with dominance methods is a forced-choice (FC) scale analyzed with the multi-unidimensional pairwise preference (MUPP) model (Stark, Chernyshenko, & Drasgow, 2005). The forced-choice format dates back to at least the 1940s (Rundquist, 1946; Sisson, 1948). It has the advantage that some response styles are impossible (e.g., midpoint response style, extreme response style) and some have found improved measurement with this format (Bartram, 2007; Brown, Incelglu, & Lin, 2017; Guenole, Brown, & Cooper, 2016; Por & Budescu, 2017). The traditional scoring of FC responses, however leads to ipsativity, which means "the sum of the scores obtained over the attributes measured for each respondent is constant" (Hicks, 1970, p.169). Consequently, ipsative scoring only allows intra-individual comparisons, and between-person comparisons are technically not appropriate. The MUPP model has solved the ipsativity problem and produces normative scores (Stark, Chernyshenko, Drasgow, & White, 2012; see also Brown & Maydeu-Olivares, 2013, and McCloy, Heggestad, & Reeve, 2005 for another two approaches that overcome ipsativity).

Although it is possible to cast stones at SS measures and dominance models, much of modern psychology is built on findings based on these traditional approaches. For example, in low stakes settings SS personality measures are reliably related to a wide variety of critically important

work and non-work outcomes (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). Rather than arguing that SS measures scored with dominance methods should be replaced, we suggest that FC measures with MUPP scoring may increase the range of settings and applications where reliable and valid assessment can be obtained. Some examples include high stakes pre-employment assessment, multi-country and multi-cultural research, and multi-source performance ratings.

In this paper we examine the comparability of SS and FC measures in the context of response processes modeled by ideal point psychometric models. To the extent that SS and FC measures of the same construct are equivalent under ideal conditions (e.g., low stakes, homogeneous sample), we can build upon SS findings with a more robust measurement method and psychometric model that may provide meaningful results under much wider circumstances.

The single-statement format

Likert's (1932) SS approach avoids items reflecting intermediate trait levels and requires respondents to rate their degree of agreement with each statement on an n-point scale (Likert, 1932). Individual trait scores are then computed as the summation of all their responses after negative items have been reverse-coded. Though Likert did not provide a theoretical model to justify his method, his methods for item selection and scoring are consistent with a dominance response process where the probability of endorsing an item increases monotonically with the individual's trait level (Coombs, 1964; Cho, Drasgow, & Cao, 2015; Drasgow, Chernyshenko, & Stark, 2010). Likert's approach has been popular for scale development and scoring because it is simple and straightforward. Examples of instruments using the Likert format include the Sixteen Personality Factor questionnaire (16PF; Cattell, Eber, & Tatsuoka, 1970), the NEO Personality Inventory (NEO PI; Costa, MacCrae, 1992), the RIASEC Markers (Armstrong, Allison, & Rounds, 2008) and the Beck Depression Inventory (BDI; Beck, Steer, & Brown, 1996).

As noted previously, Likert rating scales are easily contaminated by response style biases. For example, halo rating bias is endemic with SS scales when rating others and results in a lack of differentiation in ratings (Brown & Maydeu-Olivares, 2013). SS data is particularly problematic for cross-cultural research because people from different parts of the word tend to use the response scale differently (Chen, Lee, & Stevenson, 1995). In addition, items using the SS format are very easy to fake if respondents are motivated to do so (Griffith, Chmielowski, & Yoshita, 2007; Ones, Viswesvaran, & Reiss, 1996). When SS scales are used in high-stakes situations (e.g., as a part of hiring decisions), test-takers may fake their responses in self-enhancing ways to present themselves as more desirable candidates for the position (Christiansen, Burns, & Montgomery, 2005; Donovan, Dwight, & Schneider, 2014; McDaniel, Douglas, & Snell, 1997).

The forced-choice format

The FC format was originally developed to meet Army needs (Staugas & McQuitty, 1950). In the 1940s, the U.S Army needed to promote a large number of officers to serve as generals of the rapidly mobilizing forces. The regular performance appraisal, however, was of little help in this process because the traditional ratings did not distinguish among the top fifty percent of this group: Raters apparently manipulated their ratings so that half of the ratees obtained the highest score. The Adjutant General Office then decided to improve the rating system and found that the FC format was resistant to such intentional manipulation and could serve the Army's purposes well (Rundquist, 1946; Sisson, 1948).

In a FC item, respondents are presented with at least two single statements in a block. They are instructed to choose the statement that is "the most like me" (i.e., the PICK format), or in addition to choose the one that is "the most like me", to also choose the statement that is "the least like me" (i.e., the MOLE format), or to rank all statements within a block (i.e., the RANK format;

Brown, 2016a). Some variations to the FC format include the compositional preference task where respondents are required to distribute a fixed number of total points among several statements in a block according to the degree to which these statements describe themselves (Brown, 2016b; Chan, 2003), and the graded preference FC where respondents are asked to indicate how much they prefer statement A to B using a number of ordered categories (Brown & Maydeu-Olivares, 2017). Statements within a block are balanced on social desirability. A typical example of an FC item is given as follows:

Choose the statement that is the most like you:

- I keep my room tidy.
- I like talking to people.

These statements assess the conscientiousness and extraversion dimensions of personality. If a respondent chooses the first statement as "the most like me", the traditional scoring approach would allocate 1 point to conscientiousness and 0 points to extraversion. Trait scores would be computed by adding up scores for all of the items assessing each construct. The Edward Personal Preference Schedule (Edward, 1954), the Kuder Preference Record-Vocational (Kuder, 1960) and the Strong Vocational Interest Blank (Strong, 1959) are some well-known measures administered and scored in this manner.

As a result of the concerns about ipsativity, the popularity of the FC format appears to have declined substantially from its high point in the 1950s. Some contemporary textbooks even suggest not using the FC format (Anastasi & Urbina, 1997). According to Goffin and Christiansen (2003), only one of the fourteen commonly used personality inventories in applied settings uses the FC format.

Recently, however, the tide appears to have turned. Three meta-analyses on the predictive validity and faking-resistance properties of FC measures have indicated substantial promise for this format. Salgado, Anderson and Tauriz (2015) found that quasi-ipsative FC measures showed higher predictive validity for job performance than SS measures of conscientiousness (.38 vs .22), emotional stability (.20 vs .11), openness (.20 vs .05) and agreeableness (.16 vs .08), and similar predictive validity for extraversion (.12 vs .12). Salgado and Tauriz (2014) meta-analyzed the predictive validity of FC measures of the Big Five for academic achievements. Compared with the meta-analytic results on SS measures obtained by Poropat (2009), the quasi-ipsative FC format showed higher predictive validity on three domains and similar validity on the other two domains (openness: .31 vs .12; extraversion: -.21 vs -.01; emotional stability: .10 vs .02; conscientiousness: .21 vs .22; agreeableness: .02 vs .07). As for faking-resistance, Cao (2016) found that the effect size of score inflation (as an indicator of faking) for FC (d = .05) was much smaller than that of the SS format (d = .26).

These meta-analyses provide solid evidence supporting the use of FC. It should be noted that the majority of primary studies meta-analyzed did not use IRT scoring. In fact, Cao (2016) distinguished among scoring methods and found that the faking effect size (i.e., score inflation) for FC was not significantly different from zero when IRT scoring was used. Therefore, we would expect even more favorable results for FC measures when MUPP IRT scoring is adopted.

Psychometric equivalence between FC and SS

After years of vicissitudes, the use of the FC format seems to be on the upswing. However, psychometric models make different assumptions about response processes for FC and SS measures. When answering an SS measure, respondents are assumed to make an absolute judgment. Each respondent's perceived utility of the item itself determines his/her choice. When

answering a FC measure, however, respondents are assumed to make relative judgments among two or more options. They are basically constructing a hierarchy of alternatives. What affects the endorsement is the perceived utility differences among alternatives. When the utility difference is small, this process may require deep contemplation to arrive at a finer differentiation (Meglino & Ravlin, 1998). These theoretical differences raise a fundamental question: Are scales using FC and SS formats measuring the same construct?

Concerns about the construct interpretation of FC measures have been long recognized (Tenopyr, 1988; Usami, Sakomoto, Naito, & Abe, 2016), but little addressed. Although simulation studies have shown that both the MUPP and Thurstonian IRT models can recover person parameters (Brown & Maydeu-Olivares, 2011; Stark et al., 2005), these models have specific assumptions about the underlying response process. If these assumptions do not hold empirically, treating FC estimates and SS estimates as equivalent may be problematic. For example, Guenole, Brown, and Cooper (2016) found that statement factor loadings differed substantially between FC and SS formats, suggesting that respondents may interpret them differently. Much more empirical data are needed to examine their degree of equivalence.

To ensure that lack of convergence is not due to content differences, researchers need to compare FC and SS measures that are constructed from the same statement pool. To our knowledge, however, only a few studies have done this. Among those few studies, most have examined the equivalence of dominance-model-based SS and FC formats and found generally supportive evidence (Brown & Maydeu-Olivares, 2011, 2013; Guenole, Brown, & Cooper, 2016; Lee, Lee, &Stark, 2018). However, as mentioned above, evidence has been accumulating that shows ideal point models more accurately capture the response processes underlying various psychological measures. Thus, it is largely unknown whether ideal-point-based SS and FC formats produce

scores that are equivalent. To date, only one study has compared the marginal reliabilities, convergent validities and criterion-related validities of ideal-point-based SS and FC formats (Chernyshenko, Stark, Prewett, Gray, Stilson, & Tuttle, 2009). This study found that the FC versions showed marginal reliabilities similar to their SS counterparts, good convergent validities ($r = .54 \sim .75$), and similar criterion-related validities.

Rationale and overview for the present study

Although some promising results have been found, these earlier studies have limitations that need to be addressed. First, very little research has examined the comparability of SS and FC measures using ideal point scoring. The single ideal-point-based study that has been conducted only tested format equivalence with three constructs in a homogenous college sample. The generalizability of its results is unknown. Second, earlier studies have generally adopted a within-subjects design in which respondents finished both FC and SS measures at the same session. Such a design is desirable in that it minimizes random error due to sampling differences and allows for estimation of the convergent validity of the FC format with the SS format. However, administering both FC and SS measures in the same session may artificially increase their equivalence due to single-subject response consistency error (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Third, previous studies have not counterbalanced the order in which FC and SS were presented to respondents. Such a fixed order procedure may inflate the consistency between FC and SS responses. Fourth, little empirical evidence on the temporal stability of FC measures has been reported. As hypothesized by Meglino and Ravlin (1998), the cognitive processes underlying FC may be very intricate in that respondents need to make fine differentiations among statements to arrive at their final decision. Such fine differentiations may be much less consistent across time (Höfel & Jacobsen, 2003).

In addition to our studies seeking to confirm and extend the positive findings of earlier research, we examined the important topic of respondent reactions. Previous studies have found that applicants' reactions to test formats and procedures can affect test validity, test motivation, adverse impact, and various other behavioral intentions (Chan & Schmidt, 1997; Chan, Schmidt, DeShon, Clause, & Deebridge, 1997; Hausknecht, Day, & Thomas, 2004; Smither, Reiley, Millsap, Pearlman, & Stoeffy, 1993). Clearly, respondent reactions should be studied and this seems particularly important for the two-alternative forced choice format which may require respondents to make very difficult choices. Bartram and Brown (2004), for example, pointed out that test takers often criticize forced choice measures for their perceived "lack of choice", because the items presented could either all apply or all not apply.

We conducted two studies to carefully examine the degree of equivalence between the FC and SS formats and ascertain respondents' reactions. Study 1 employed a between subject design where a group of respondents completed the FC format and another group completed the SS format. A between subject design allowed us to study score equivalence and circumvent issues related to single subject response consistency. Study 2 adopted a within-subjects design where all respondents completed the FC and SS measures, but with an interval of two days between assessments. The order of administration was counterbalanced. This within-subjects design allowed us to directly study the convergent validity between the two formats and the two-day interval should minimize memory effects without running the risk of personality change.

STUDY 1

Study 1 examined whether ten personality facet scores obtained with FC and SS formats had similar factor structures (i.e., construct validity), nomological networks (i.e., criterion-related validity), and reliabilities (i.e., marginal reliability and test-retest reliability). A between-subjects design was used to obtain accurate estimates free from potential influences of consistency or order effects.

Method

Sample

Two samples were recruited from Amazon Mechanical Turk (MTurk) workers pool through TurkPrime (Litman, Robinson, & Abberbock, 2016). The FC sample completed TAPAS_FC (see below) and other criterion measures. The SS sample rated the same statements that make up the forced choice pairs using a 5-point Likert scale and completed the same criterion measures. Six quality control items were embedded in the FC condition and five in the SS condition. In this study, we deleted data from respondents who responded incorrectly to more than one quality control items. All respondents received monetary compensation. The test-retest interval was 10 days. Respondents were only allowed to enroll in one sample. All the measures were administered twice.

The FC sample consisted of 806 respondents who completed the first assessment. Of this group, 580 participated in the retest, yielding a retention rate of 72%. After screening out those who missed more than one quality control items, 781 respondents for the first assessment (70% female) and 562 for the retest (70% female) remained. Respondents for the first test generally were middle-aged (M = 39.15, SD = 13.01, had an education level between vocational school and bachelor's degree (M = 3.65, SD = 0.96), and had an annual household income about $40,000 to

$50,000 (As shown in the Measures section, income was measured on an 8-point ordinal scale; *M* = 5.03, SD = 2.10). The subset of the sample who completed the retest had similar demographics (M_{age} = 40.12 ± 12.99, M_{edu} = 3.71 ± 0.96, M_{income} = 5.04 ± 2.06).

The SS sample included 811 respondents who completed the first assessment and 567 who participated in the retest (retention rate = 70%). After screening out respondents who missed more than one quality control item, 793 respondents for the first session (68% female) and 553 for the retest (70% female) remained. Respondents for the first assessment were middle-aged (*M* = 40.40, SD = 13.77), had an average education level between vocational school and bachelor's degree (M = 3.61, SD = 0.98), and had an annual household income about $40,000 to $50,000 (*M* = 4.88, SD = 2.12). Participants who completed the retest were similar (M_{age} = 40.61 ± 13.64, M_{edu} = 3.64 ± 1.00, M_{income} = 4.94 ± 2.11).

We conducted independent group t-tests for the four demographic variables to see whether there were any sample differences. Results showed that the two samples were not significantly different from each other with respect to the demographic variables (t = .77 ~ 1.42, p = .16 ~ .44).

Measures

An abridged static version of TAPAS was used in the present study. The original TAPAS is mainly derived from the Big Five framework, with some additional context-specific facets that do not necessarily belong to the Big Five. Within the Big Five, there are 22 facets in total: six facets for Openness, six facets for Conscientiousness, four facets for Extraversion, three facets for Agreeableness, and three facets for Emotional Stability (Drasgow et al., 2012). TAPAS statements were specifically developed for ideal point measurement and, therefore, the statement pools for each facet include both extreme (positive and negative) and intermediate items that cover the entire range of the facet. This abridged version used in the current study included ten facets: Intellectual

Efficiency (IE) and Tolerance (TO) for Openness, Achievement (AC) and Order (OR) for Conscientiousness, Dominance (DO), Sociability (SO) and Physical Conditioning (PC) for Extraversion, Selflessness (SE) for Agreeableness, and Even Tempered (ET) and Optimism (OP) for Emotional Stability. Item pairs were matched on social desirability. A brief description of each facet is shown in Table 1.

TAPAS_SS. There were 16 items for IE, 17 items for TO, 17 items for AC, 16 items for OR, 14 items for DO, 17 items for PC, 18 items for SO, 15 items for SE, 18 items for ET, and 18 items for OP. Since intermediate items were included, Cronbach's α is not an appropriate index of reliability (Cao, Drasgow, & Cho, 2015). Therefore, we used IRT scoring and computed IRT marginal reliabilities.

TAPAS_FC. Statements that made up the FC pairs are the same as those in TAPAS_SS. There were 10% unidimensional pairs and 90% multidimensional pairs. Since Cronbach's α is not an appropriate measure of reliability for FC measures, we computed IRT marginal reliabilities.

Big Five Inventory (BFI). The BFI is a 44-item scale designed to measure the Big Five personality traits (John & Srivastava, 1999). All items begin with "I see myself as someone who…". There are 10 items for Openness (e.g. "*is inventive*"; 2 reverse coded items,), 9 items for Conscientiousness (e.g. "*does a thorough job*"), 8 items for Extraversion (e.g. "*is talkative*"), 9 items for Agreeableness (e.g. "*is generally trusting*"), and 8 items for Emotional Stability (e.g. "*worries a lot*"). Cronbach's α's in both MTurk samples (i.e., the sample completing the TAPAS_FC and the sample completing TAPAS_SS) were .84 and .84 for Openness, .83 and .83 for Conscientiousness, .87 and .86 for Extraversion, .81 and .81 for Agreeableness, .88 and .88 for Emotional stability, respectively.

Subjective Well-being Scale (SWBS). The SWBS is a 5-item scale developed to measure people's general satisfaction with life (Diener, Emmons, Larsen, & Griffin, 1985). An example item is "*In most ways my life is close to my ideal*". Cronbach's α's were .91 in the FC sample and .90 in the SS sample.

Core Self-evaluation Scale (CSES). The CSES is a 12-item scale developed as a direct measure of people's general evaluation of their self-worth and competence (Judge, Erez, Bono, & Thoresen, 2003). An example item is "*I am confident I get the success I deserve in life*". Cronbach's α's were .90 in the FC sample and .86 in the SS sample.

For the sake of consistency, all SS measures were rated on a 5-point Likert scale regardless of their original scales.

Education level. Respondents indicated their education level on a 6-point scale. "1" = primary school; "2" = high school or equivalent; "3" = some college or vocational school; "4" = a bachelor's degree or equivalent; "5" = a master's degree or equivalent; "6" = a doctoral or professional degree.

Income. Respondents indicated their annual household income in US dollars on an 8-point scale. "1" = under 10,000; "2" = 10,000~19,000; "3" = 20,000 ~ 29,000; "4" = 30,000 ~ 39,000; "5" = 40,000 ~ 49,000; "5" = 50,000 ~ 74,999; "6" = 75,000 ~ 99,999; "7" = 100,000 ~ 150,000; "8" = over 150,000.

Subjective health (SH). SH was measured by one item "*In general, how would you describe your health*". Respondents rated themselves on a Likert scale from 1 (excellent) to 5 (bad).

Analytical procedures

Because TAPAS_SS scales were developed based on unfolding model assumptions, we used the software GGUM2004 to obtain trait estimates through expected *a posteriori* (EAP)

estimation (Roberts, Donoghue, & Laughlin, 2000). FC TAPAS was scored using the MUPP model. Trait estimates were obtained using multidimensional Bayes modal estimation. Marginal reliability was computed via the following formula:

$$rel = 1 - \frac{mean(se_{\hat{\theta}}^2)}{var(\hat{\theta})},$$

where $\hat{\theta}$ refers to the trait estimate and $se_{\hat{\theta}}$ refers to the standard error of the estimate. Pearson correlations and average absolute differences were used to quantify the similarity between FC and SS in terms of reliability profile and validity profile.x Except for IRT based trait estimation, all other analyses were performed in R3.2.4 (R Core Team, 2016) using the *psych* package (Revelle, 2016).

Results

Construct validity

Because TAPAS was derived from the Big Five personality framework, exploratory factor analysis was performed on both the FC and SS measures to see whether similar factor solutions would be obtained. Specifically, the ten TAPAS facet scores and five BFI scale scores were subjected to EFA instead of raw item scores. The five BFI scale scores were included in the EFA as anchors to facilitate results interpretation. Both the Big Five theory and parallel analyses suggested that five factors should be retained. Therefore, five factors were extracted using maximum likelihood estimation with oblimin rotation. Results from the first test and retest datasets were very similar. Therefore, we only report results based the first assessment here and full results can be obtained from the first author. Factor loadings are displayed in Table 2.

Table 2 shows that five factors were clearly recovered in both formats: Intellectual Efficiency loaded on the Openness factor; Achievement Motivation and Orderliness loaded on the

Conscientiousness factor; Dominance and Sociability loaded on the Extraversion factor; Even Temper and Optimism loaded on the Emotional Stability factor. One unexpected finding was that Tolerance unexpectedly had its primary loading on Agreeableness. A small number of cross-loadings were present; their pattern was almost identical across the FC and SS formats. Tucker's factor congruence index ranged from .97 ~ .98 for the five factors, indicating a high level of factorial similarity. Importantly, the factor loadings from FC and SS data were very similar in magnitude. The proportions of variance explained by each factor were also very similar across formats (Openness: 8% vs 11%; Conscientiousness: 12% vs 13%; Extraversion: 13 % vs 13%; Agreeableness: 9% vs 11%; Emotional stability: 12% vs 12%).

Correlations among latent factors in both formats are shown in Table 3. Jennrich's (1970) asymptotic χ^2 test was first performed to see whether latent factor correlations estimated from the two formats were equal. Results indicated that latent factor correlations obtained from the two formats were statistically different from each other, $\chi^2 = 36.99$, $p < .001$, though it is known that this method is overly stringent (Revelle, 2016). Correlations from the analysis of SS data tended to be slightly larger (M = .24) than in the FC data (M = .16). The highest correlations in both formats emerged between Conscientiousness and Emotional Stability (r = .45 and .44). Extraversion and Emotional Stability were moderately correlated (r = .28 and .37), as were Agreeableness and Emotional Stability (r = .23 and .25). Openness and Emotional Stability showed the lowest correlation (r = .00 and -.05).

Criterion-related validity

Because validity results based on the first assessment and retest were very similar, only results from the first test dataset are reported in Table 4. Full results can be obtained from the first author.

Two clear patterns emerged. First, for each of the ten facets, the majority of the criterion-related validities of the FC measures and the SS measures are in the same direction. The magnitude of criterion-related validities of the FC measure are relatively close to those of the SS measures, with Tucker congruence coefficients ranging from .93 to 1.00 for eight facets except for Intellectual Efficiency (.84) and Tolerance (.73). The validities of the SS measure tend to be somewhat higher with respect to the subjective criteria (mean of absolute validities: M_{SS} = .24, M_{FC} = .10), perhaps because they have a common response format. The validities with respect to the objective criteria are very similar (mean of absolute validities: M_{SS} = .07, M_{FC} = .10).

Reliability

Test-retest reliability. As shown in Table 5, the average test-retest reliability was .77 (ranging from .69 to .83) for the FC measures, and .85 (ranging from .79 to .91) for the SS measures, both of which are satisfactory. To quantify reliability similarity between the two formats, we computed the Pearson correlation between the two vectors of reliability and also their mean absolute difference (AD). Profile analyses showed that the reliability profiles were quite similar in terms of both rank order (r = .82) and elevation (AD = .08). Further examination of Table 6 revealed that the Achievement Motivation facet showed (relatively speaking) the lowest reliabilities in both formats (r = .69 and .79, respectively, which is still acceptable). The Sociability facet showed the highest test-retest reliabilities in both formats (r = .83 and .89, respectively). Overall, the FC measures showed satisfactory test-retest reliabilities that were generally comparable to the SS versions.

Marginal reliability. Because marginal reliabilities computed from the first and the second assessments were very similar, we only reported the results from the first. Full results can be obtained from the first author. As reported in Table 6, the average marginal reliability was .76

(ranging from .67 to .84) for the FC measures and .85 (ranging from .73 to .92) for the SS version, both of which appear satisfactory. Profile analyses showed that the reliability profiles were quite similar in terms of both rank order (r = .89) and elevation (AD = .10). Table 6 shows that the Selflessness facet was the least reliable facet in both formats (marginal reliability = .67 and .73, respectively). The Physical Conditioning facet presented the highest marginal reliability in both formats (.86 and .92, respectively). Overall, the FC measures showed satisfactory marginal reliabilities that were only slightly lower than the SS measure.

STUDY 2

The purpose of Study 2 was three-fold. First, Study 2 was intended as a partial replication of Study 1. Second, Study 2 was designed to go beyond Study 1 to directly estimate convergent validities between the FC and SS measures. Third, respondent reaction measures were included so that we could examine whether respondents reacted differently to FC and SS measures. A within-subjects design with an interval of two days in between the first and second test administrations was employed. The order of the test administration was counterbalanced to minimize the potential influences of any order effect.

Method

Sample

A sample of 511 respondents was recruited from the MTurk through TurkPrime. To ensure that no respondents from Study 1 were recruited again, we used the "excluding workers" feature in TurkPrime. Half of the sample completed the FC format first and the SS format two days later (FS group), and the other half completed the SS format first and the FC format two days later (SF group). Six quality control items were embedded and we allowed one quality control item to be incorrect. Respondents received monetary compensation.

FS group. A total of 251 participants were randomly assigned to this group, all of whom completed the FC version first, and 193 completed the SS version two days later (retention rate was 76.9%). After screening out those who missed more than one quality control items, 245 respondents from the first assessment (63% female) and 193 from the retest (60% female) remained in the sample. Respondents to the initial assessment were middle-aged on average (M = 37.91, SD = 12.51), had an average education level between vocational school and bachelor's degree (M = 3.68, SD = .88), and had an average annual household income about $40,000 to

$50,000 ($M$ = 4.89, SD = 2.23). The subset who also participated in the retest presented almost the same demographics information (M_{age} = 38.15 ± 12.66, M_{edu} = 3.70 ± .87, M_{income} = 4.98 ± 2.24).

SF group. A total of 260 participants were randomly assigned to this group, all of whom completed the SS version, and among them 210 returned to complete the FC version two days later (retention rate was 80.8%). After screening out participants who missed more than one quality control item, 257 respondents from the first assessment (67% female) and 204 respondents from the retest (65% female) remained. Respondents for the first assessment were middle-aged on average (M = 37.66, SD = 12.73), had an average education level between vocational school and bachelor's degree (M = 3.63, SD = .92), and had an average annual household income about $40,000 to $50,000 ($M$ = 5.00, SD = 2.13). The subset who also participated in the retest showed almost the same demographics (M_{age} = 38.67 ± 12.63, M_{edu} = 3.64 ± .92, M_{income} = 5.11 ± 2.19).

We conducted independent groups t-tests on the four demographic variables to assess their equivalence. No significant results were found (t = -.92 ~ .39, p = .36~.70). Also, no order effect was found for both the FC and SS scores after Bonferroni correction (t = -2.72 ~ .24, p = .07 ~ .84). Therefore, further analyses were performed on the combined sample (N = 381).

Measures

All the measures in Study 1 were included in the present study, with the exception of the BFI being replaced by the BFI-2 (Soto & John, 2017). In addition, we also included several respondent reaction measures and job satisfaction measures.

BFI-2. With growing evidence demonstrating the importance of personality facets, Soto and John (2017) developed the BFI-2, which has three facets for each broad personality domain. Each facet is measured by four items. Specifically, the three Openness facets are Intellectual Curiosity, Aesthetic Sensitivity, and Creative Imagination (α = .77 ~ .82); the three

Conscientiousness facets are Organization, Productiveness, and Responsibility (α = .69 ~ .86); the three Extraversion facets are Sociability, Assertiveness, and Energy (α = .75 ~ .88); the three Agreeableness facets are Compassion, Respectfulness and Trust (α = .69 ~ .74); the three Emotional Stability facets are Anxiety, Depression and Volatility (α = .83 ~ .85). The BFI-2 items use a 5-point Likert rating scale.

Abridged Job Descriptive Index (AJDI). The AJDI is a 38-item measure of job satisfaction (Stanton et al., 2002). It measures an individual's satisfaction with coworker, pay, supervision, work itself, and promotion opportunity with six items for each domain. In addition, it measures an individual's satisfaction with the job in general with eight items. The AJDI was administered with a 3-point rating scale. As suggested in the Scoring Manual (Balzer et al., 1997), "Yes" was scored as 3; "No" was scored as 0; "Cannot decide" was scored as 1. Cronbach's α ranged from .84 to .90.

Positive and Negative Affect Scale (PANAS). The ten-item short form of PANAS (Thompson, 2007) was used to assess an individual's immediate affect, right before and right after the administration of TAPAS. The five adjectives to measure positive affect were *Active, Determined, Attentive, Inspired*, and *Alert* (Cronbach's α = .84). The five adjectives to measure negative affect were *Afraid, Nervous, Upset, Hostile*, and *Ashamed* (Cronbach's α = .87). The PANAS was administered via a 5-point Likert scale.

Vitality. Three items from the Subjective Vitality Scale (Bostic, Rubio, & Hood, 2000) were used to assess an individual's perceived vitality, immediately before and after responding to either format of TAPAS (Cronbach's α = .85). An example item is "*I feel energized right now*". The Vitality scale was administered using a 5-point Likert scale.

Perceived Difficulty (PD). Respondents' perceived difficulty of each format was assessed by the item "*How hard do you feel it is to respond to the previous survey questions*?" immediately

after they finished responding to the SS and FC personality measures. Respondents were instructed to rate their perceived difficulty on a 5-point scale (1 = "Not difficult at all", and 5 = "Very difficult").

Effort. Respondents were also asked to indicate how much effort they had put into responding to these questions on a 5-point scale (1 = "No effort at all", and 5 = "A lot of effort").

Concentration. Two items were used to assess the extent to which respondents concentrated during the process of answering questions. The two items are "*It was hard to keep my mind on the previous survey questions*" and "*During the previous survey, I was bored.*" The two items were administered on a 5-point Likert scale and were reverse coded so that a higher score indicated a higher level of concentration (Cronbach's $\alpha = .63$).

Preference. Two items were used to assess the degree to which each respondent liked each format. The two items were "*How much did you like to respond to the previous survey questions*?" and "*How irritated, stressed, or annoyed were you during the previous survey questions*?" (reverse coded). The two items were administered on a 5-point scale (1 = "Not at all", and 5 = "A lot"; Cronbach's $\alpha = .62$).

Response time. The survey platform, Qualtrics, also recorded respondents' response time in seconds. We transformed these recordings into minutes for the ease of presentation. Although the recorded time was the response time for the whole test battery, the difference for each person between their response times for the two tests can be used as a proxy to indicate response time differences between the FC and SS measures. This is because the only difference between the two assessments at the two administrations was the TAPAS format (FC or SS), and all other measures were exactly the same.

PANAS and Vitality were administered twice within a single test session, immediately before and after each participant responded to either format of the TAPAS. PD, Effort, Concentration, and Preference were administered after each participant completed either format of the TAPAS. All items were worded in the way that specifically targeted the format administered.

Results

Psychometric equivalence

Construct validity. The same EFA procedures as in Study 1 were applied to the current within-subjects data, and results found in Study 1 were well replicated, as shown in Table 6. Specifically, the five-factor structure was successfully recovered: Intellectual Efficiency loaded on the Openness factor; Achievement Motivation and Orderliness loaded on the Conscientiousness factor; Dominance and Sociability loaded on the Extraversion factor; Selflessness loaded on the Agreeableness factor. Even Tempered and Optimism loaded on the Emotional Stability factor. Tolerance still unexpectedly had its primary loadings on Agreeableness. A few cross-loadings were present and were almost identical across formats. For example, Dominance also loaded onto Openness, and Tolerance and BFI_Openness also loaded onto Agreeableness. Tucker's factor congruence index ranged from .91 ~ .98 for the five factors, indicating a high level of factorial similarity. The only exception was that Achievement Motivation had a high cross-loading on Openness (λ = .57) in the SS format but not in the FC format (λ = .29). The proportions of variance explained by each factor were identical across formats.

Correlations among latent factors are shown in Table 7. We again performed Jennrich's (1970) asymptotic χ^2 test as well to see whether latent factor correlations estimated from the two formats were equal or not (albeit their stringency). Results indicated that they were statistically different from each other (χ^2 = 31.23, p < .001). Again, the factor correlation tended to be higher

for the SS format (M=.27) than for the FC format (M=.18) In both formats, Extraversion and Conscientiousness were moderately correlated with Emotional Stability (r = .36 ~ .40 for FC and .43~.45 for SS). Agreeableness and Emotional Stability were also moderately correlated (r = .27 and .30). Openness and Emotional Stability showed low correlations (r = -.00 and .17).

Criterion-related validity. Criterion-related validity results are shown in Table 8. The results are very similar to the pattern observed in Study 1, with each TAPAS facet having its highest correlation with its corresponding BFI-2 domain score. For example, Intellectual Efficiency and Tolerance had their highest correlation with the BFI-2 Openness score. The TAPAS Optimism facet showed some strong correlations, especially with the Subjective Well-Being scale and the Core Self-Evaluation scale. Importantly, the FC and SS versions of TAPAS facets showed very similar patterns of correlation across various criterion measures.

Marginal reliability. The marginal reliabilities are shown in Table 9. The average marginal reliability was .77 (ranging from .66 to .86) for the FC version and .85 (ranging from .72 to .90) for the SS version, both of which are satisfactory. Profile analyses showed that the reliability profiles were similar in terms of both rank order (r_P = .77) and elevation (AD = .08). Overall, we once again found strong evidence showing that FC and SS have similar and reasonably high reliabilities.

Convergent validity. Convergent validity results are also shown in Table 9. We report both raw correlations and correlations corrected for unreliability. The mean raw convergent validity was .72, ranging from .56 ~ .82. A closer inspection revealed that Achievement Motivation and Selflessness had the lowest convergent validities (r = .59 and .63, respectively); those two facets also had relatively low reliabilities. When we corrected the raw convergent validities for unreliability using Spearman's formula (Spearman, 1904), the average convergent validity

reached .89 and Achievement Motivation was the only facet with corrected convergent validity below .80 (r = .79). Overall, there appears to be strong evidence that the FC and SS versions are measuring the same underlying constructs.

Respondent reactions

Positive and negative affect. Four pairwise comparisons were made with the assessment format as the independent variable. Results showed no significant differences in terms of pretest positive affect (M_{FC} = 3.15 ± .87, M_{SS} = 3.15 ± .92, r = .80, p = .97, d = -.001), pretest negative affect (M_{FC} = 1.26 ± .51, M_{SS} = 1.26 ± .51, r = .75, p = .99, d = .00), post-test positive affect (M_{FC} = 3.08 ± .92, M_{SS} = 3.04 ± .99, r = .80, p = .24, d = .04) and post-test negative affect (M_{FC} = 1.20 ± .46, M_{SS}= 1.19 ± .44, r = .71, p = .54, d = .02). Two additional repeated measures ANOVAs were run on post-test affect states with assessment format as the independent variable and pretest affect scores as the covariate. No significant differences were found ($F_{(1,377)}$ = .04 ~ .75, p = .39 ~ .85).

Vitality. Two pairwise comparisons were made with assessment format as the independent variable. Results showed no significant differences in terms of pretest Vitality (M_{FC} = 3.65 ± .72, M_{SS} = 3.60 ± .79, r = .61, p = .17, d = .06) and posttest vitality (M_{FC} = 3.60 ± .83, M_{SS}= 3.61 ± .98, r = .56, p = .86, d = -.006). A repeated measures ANOVA was run on post-test vitality with assessment format as the independent variable and pretest Vitality scores as the covariate. No significant difference was found ($F_{(1,377)}$ = .04, p = .84).

We also performed paired-sample t-test on Perceived difficulty, Effort, Concentration and Preference. No difference was found on Effort and Concentration (ps>.10). Respondents found the FC format harder to complete (M_{FC} = 2.50 ± 1.11, M_{SS} = 1.85 ± 1.07, r = .39, t = 10.56, p

< .001; d = .60), and slightly preferred the SS format to the FC format (M_{FC} = 3.87 ± .71, M_{SS} = 3.99 ± .78, r = .45, t = -2.84, p = .005; d=-.16).

Response time. Independent sample t-test results revealed that respondents in Study 1 took longer to complete the SS version than the FC version (M_{FC} = 19.72 ± .12.76, M_{SS} = 24.85 ± .14.27, t = -7.50, p < .001, d = -.38). The same pattern was found in Study 2 with a paired-sample t-test (M_{FC} = 24.89 ± .14.42, M_{SS} = 31.52 ± .27.59, r = .26, t = -4.68, p < .001, d = -.29). The magnitude of these differences would be classified as "small" by Cohen's criteria.

GENERAL DISCUSSION

The present study used a between-subjects design as well as a within-subjects design to examine the psychometric equivalence of FC and SS versions of TAPAS. As the results illustrated, substantial evidence was found for equivalence of the two formats. Specifically, the theoretical five factor personality structure was recovered with both versions. Validity profiles and reliability profiles between the two formats were highly similar in terms of both shape and elevation. Facet scores obtained from the two formats were highly correlated, providing strong evidence for convergent validity. In addition, the formats had little differential impact on respondents' general reactions except for finding the FC format a bit more difficult.

Both model-based (i.e., IRT marginal reliabilities) and empirical (i.e., test-retest reliabilities) reliabilities were obtained for 10 facets administered in FC and SS formats. All ten facets assessed via the FC format showed moderate to high level of temporal stability across ten days. More importantly, FC test-retest reliabilities were reasonably similar to those of their SS counterparts in terms of both magnitude and rank order. Although the performance of FC measures was found to be similar to that of the SS measures, the FC versions had slightly lower reliabilities. We see two possible explanations for this difference. First, scoring of each FC pair was dichotomous: one statement was selected as "most like me" and the other statement was not. SS scoring was polytomous in nature and appears to offer more psychometric information. For example, respondents can differentiate whether they "Disagree" or "Strongly Disagree". It is therefore not surprising that the SS scale scores were slightly more reliable. A second, more pernicious explanation is that responses to SS scales may be influenced by respondents' focal trait level as well as other stable but irrelevant traits level, such as response styles. Previous studies have found evidence that response styles are stable even across multiple years (Weijters Geuens, &

Schillewaert, 2010; Wetzel, Lüdtke, Zettler, & Böhnke., 2016). Such stable but irrelevant traits can artificially increase the reliability of SS scales. To test the first explanation, future researchers can administer yes/no versions of SS scales and compare their reliabilities with FC versions. As for the second explanation, some advanced statistical models could be applied to partial out the variance due to response styles (Falk & Cai, 2016; Maydeu-Olivares & Coffman, 2006). Note that no such advanced models yet exist for intermediate items and unfolding IRT models.

Studies 1 and 2 are the first empirical evaluations of the underlying factor structure of MUPP-scored FC scales and their SS counterparts. We recovered the Big Five structure underlying TAPAS using both. More importantly, the factor solutions of FC and SS were nearly identical. Tucker's factor congruence indexes were .97~.98 in Study 1 and .91~.98 in Study 2. Lorenzo-Seva1 and ten Berge (2006) suggested that values greater than .95 imply that two factors can be considered equal and values between .85~.94 implied fair similarity.

Though the present study was not designed to directly test whether an FC assessment can mitigate the effects of response styles, we did find some indirect evidence. The first piece of evidence comes from an examination of the validity profiles. Criteria used in this study can be classified into subjective criteria and objective criteria. Subjective criteria included personality, core self-evaluation, well-being, and job satisfaction. Objective criteria were age, gender, education, and income. If the binary nature of FC data is responsible for the slightly lower validity coefficients observed for some criteria, we would expect the FC measure to show lower correlations with both subjective criteria and objective criteria. If it is because the FC format reduces response styles, we would expect systematically lower correlations for FC with subjective criteria but little systematic differences with objective criteria. The rationale for this adjustment is that response styles are expected to be independent of item content for the subjective scale, and

therefore have the same effect on every subjective measure, which would artificially increase correlations between SS scales and subjective measures. However, when the criterion is objective like gender, it is unlikely that response styles would influence respondents' answers. In this case, correlations between TAPAS facets measured by SS and objective criteria would be free from the influence of response styles. Therefore, we would expect similar validity coefficients for SS and FC formats. As shown in Table 4 and Table 8, 95% of the FC validity coefficients (76 out of 80) were smaller than their SS counterparts for subjective criteria in study 1, and 92.8% of the FC validity coefficients (130 out of 140) were smaller in study 2. For objective criteria, 65% to 72.5% of the FC validity coefficients (26~29 out of 40) were smaller. Thus, this suggests some reduction in the effect of response styles for FC scales. A second line of argument comes from the inter-factor correlations. If the TAPAS SS scales are confounded by response styles, we would expect higher factor correlations due to their construct irrelevant variance. Study 1 and 2 had results consistent with this argument: $M_{FC1} = .16$, $M_{SS1} = .24$, $M_{FC2} = .18$, $M_{SS2} = .27$.

Regarding respondents' reactions, it is not surprising that the FC measure was considered harder to complete than the SS measure, which was in line with what Sass, Frick, Reips, and Wetzel (2018) found with think-aloud techniques. Despite the perceived difficulty, respondents showed the same degree of concentration and devoted an equal (and high) level of effort into answering both formats. The two formats also elicited almost identical levels of affective states and subjective vitality before and after their administrations. Although they showed a statistically significant preference for the SS format, the effect size was small (d = .16). A surprising finding is that respondents spent less time answering FC measures than SS measures, which is in direct contrast with previous findings showing that FC version is more time consuming (Bowen, Martin, & Hunt, 2002). One potential explanation for the inconsistencies is the format difference. All

previous studies on respondents' reactions to FC asked respondents to choose "the most like me" and "the least like me" statements from a tetrad. To respond to such a tetrad, three pairwise comparisons per block are needed. However, in our design, only one pairwise comparison is involved in each FC block (pair). Therefore, it is expected that a tetrad design is more time-consuming than a two-alternative design. Such a design difference may also explain why previous studies found negative reactions to FC while we did not. More pairwise comparisons per block would induce more cognitive load, which might result in negative reactions (McLeod, 1988).

Limitations

There are several limitations in our studies that need to be addressed in future work. First, the present studies only used TAPAS as an FC illustration and respondents were only asked to choose from two options. The generalizability of our findings to other instruments or FC formats (such as tetrads) remains to be studied. Future researchers are encouraged to explore the options of employing the RANK or the MOLE design, include more options in a block, and utilize the compositional or graded preference FC task (Brown, 2017). Second, the test-retest intervals used in our studies were relatively short. It would be informative to administer the TAPAS to the same people at multiple times across longer intervals. In this way, we would be able to see how test-retest reliability of FC measures changes across time and compare the change trajectory to that of SS measures.

Implications

The present study provides evidence for the long-held but rarely tested assumption that FC and SS versions of a scale assess the same underlying construct. Additionally, there was indirect evidence that the FC format mitigated construct-irrelevant variance arising from response styles. Response style differences have daunted cross-cultural researchers for some time (Triandis, 1972;

Leung & Bond, 1989). For instance, researchers found cross-cultural differences in endpoint endorsement between blacks and whites (Bacheman & O'Malley, 1984), and Hispanics and non-Hispanics (Hui & Triandis, 1989). It has also been shown that Japanese and Chinese people are more likely to use middle points than North American respondents (Chen, Lee, & Stevenson, 1995). Cultural differences in acquiescent responding are also well-documented (Marin, Gamba, & Marin, 1992; Smith, 2004). These construct-irrelevant differences often obfuscate true trait (non)differences in cross-cultural comparisons. These response style differences cannot occur with the dichotomous FC format, and our findings in the present studies suggest that this format may be used in place of the omnipresent Likert scale with little or no harm.

We also found that respondents did not react more negatively to the FC format than to the SS format, despite the fact that they did find it more difficult to answer. This may be particularly important if FC measurement is used in high-stakes contexts such as personnel selection. According to Hausknecht, Day, and Thomas's (2004) meta-analysis, applicants' perceptions of selection systems were related to their perception of the organization, their intentions to accept an offer, and the likelihood of recommending the employer to others.

Future directions

A first line of research should be directed toward a theory articulating how to design a good FC measure. Brown and Maydeu-Olivares (2011) provided some initial simulation-based guidelines. They found that if items keyed in opposite directions were paired together, such pairs would result in higher measurement precision. However, such item pairs are unlikely to be fake-resistant because respondents would be more inclined to choose the positive statements (Morillo, Leenen, Abad, Hontangas, de la Torre, & Ponsoda, 2016). Future studies should explore optimal item pairing strategies and consider both psychometric desiderata such as IRT item information

(Lee, Joo, & Stark,2018) as well as the psychology of respondents (e.g., resistance to faking by respondents in high stakes settings).

A second line of research should compare and contrast unfolding models and dominance models for FC format. There is growing evidence that the response process for self-report measures is most appropriately modeled as ideal point (Drasgow, Chernyshenko, & Stark, 2010). However, there is a long history of successful use of dominance models. If, where, and when one class of models should be preferred to the other class remains to be studied.

There are many variations of the FC format. A FC measure can take the form of full ranking FC, partial ranking FC, compositional preference FC, or graded preference FC. The block size (the number of statements one block) of a FC measure can vary as well. The best combination of these features is an important topic for future research.

A last but not least line of research asks whether a FC measure functions the same way across populations. For example, FC appears to offer much for cross-cultural studies because it eliminates some response style differences and reduces others (Wetzel et al., 2016; Wetzel, Böhnke, Brown, 2016). However, measurement equivalence must be established before proceeding to group comparisons. IRT techniques for assessing equivalence needed to be extended to the FC format.

CONCLUSION

Several lines of research are converging on the conclusion that FC measurement offers substantial advantages. Salgado and colleagues have demonstrated that FC Big Five personality measures are more predictive of job performance (Salgado, Anderson, & Tauriz, 2015; Salgado & Tauriz, 2014) than SS measures. Cao (2016) provided evidence showing that the FC format is fake-resistant, especially when IRT scoring was adopted. FC formats eliminate some types of response styles such as the mid-point and extreme styles. The current study shows that FC measures and their SS counterparts assess the same underlying constructs. Importantly, it was found that respondents did not have substantially negative reactions to the FC format. Combined, these findings recommend the use of FC formats for the assessment of personality and perhaps other self-report constructs, particularly in high-stakes settings and for culturally diverse groups.

TABLES

Table 1

Brief descriptions of TAPAS facets (Nye, Drasgow, Chernyshenko, Stark, Kubisiak, White, & Jose, 2012)

Big Five Domain	TAPAS Facet (No. of items)	Brief Description
Openness	Intellectual Efficiency (16) Tolerance (17)	Process information and make decisions quickly Interested in other cultures and opinions different from their own
Conscientiousness	Achievement (17) Order (16)	Hard working, ambitious, confident, and resourceful Organize tasks and activities and desire to maintain neat surroundings
Extraversion	Dominance (14) Sociability (18) Physical Conditioning (17)	Domineering, take charge, natural leaders Tend to seek out and initiate social interactions Tend to engage in activities to maintain one's physical fitness
Agreeableness	Selflessness (15)	Generous with one's time and resources
Emotional Stability	Even Tempered (18) Optimism (18)	Tend to be calm and stable Have a positive outlook on life and tend to experience a sense of well-being

Table 2

Factor loadings from exploratory factor analysis (Study 1)

	Openness		Conscientiousness		Extraversion		Agreeableness		Emotional Stability	
	FC	SS	FC	SS	FC	SS	FC	SS	FC	SS
BFI_O	**.42**	**.50**	.09	-.02	.13	.09	**.35**	.25	-.08	-.08
IE	**.70**	**.69**	.01	.08	-.05	-.07	.02	.05	.12	.18
TO	.28	**.32**	-.10	-.12	.07	.05	**.47**	**.51**	-.14	-.07
BFI_C	.00	.07	**.98**	**.88**	.02	-.02	.04	.04	.02	.05
AC	**.34**	**.50**	**.42**	**.53**	-.07	.03	.15	.15	.03	-.07
OR	-.05	-.17	**.64**	**.73**	-.08	.05	-.15	-.10	-.02	.02
BFI_E	-.02	.00	.04	.00	**.97**	**.99**	.02	-.02	-.02	-.02
DO	**.38**	**.52**	-.01	.03	**.46**	**.37**	-.18	-.20	.00	.07
SO	-.02	-.05	-.07	-.01	**.72**	**.79**	.08	.11	.13	.10
BFI_A	-.23	-.17	.14	.17	.14	.07	**.56**	**.66**	.32	.22
SE	.10	.20	.01	-.01	.00	.10	**.68**	**.66**	-.07	-.13
BFI_N	-.05	-.06	-.07	.00	-.12	-.05	.09	.03	**-.83**	**-.96**
ET	-.01	-.06	-.05	-.01	-.21	-.10	.26	.43	**.67**	**.51**
OP	.04	.00	.12	.23	.16	.19	-.12	.04	**.58**	**.54**
PC	-.02	-.03	.24	.29	.12	.21	-.12	.01	.04	.08
Congruence	.98		.98		.98		.97		.97	
Variance	.08	.11	.12	.13	.13	.13	.09	.11	.12	.12

Note. Factor loadings above .30 were marked in bold.IE=Intellectual Efficiency; TO=Tolerance; AC=Achievement; OR=Orderliness; DO=Dominance; PC=Physical conditioning; SO=Sociability; SE=Selflessness; ET=Even Temper.

Table 3

Factor inter-correlations obtained from EFA (Study 1)

	1	2	3	4	5
1. Openness		.27	.25	.25	.00
2. Conscientiousness	.09		.26	.22	.44
3. Extraversion	.14	.23		.13	.37
4. Agreeableness	.09	.1	.03		.25
5. Emotional Stability	-.05	.45	.28	.23	

Note. Values below diagonal were from the FC data and values above diagonal were from the SS data.

Table 4

Criterion-related validity

	Subjective criteria								Objective criteria			
	BFI-O	BFI-C	BFI-E	BFI-A	BFI-N	SWBS	CSES	SH	Gender	Age	Education	Income
IE	.37 / .43	.13 / .36	.08 / .18	-.03 / .16	.11 / .25	-.02/ .1	.08 / .30	.02 / .12	-.05 / -.03	-.05 / .04	.14 / .18	.05 / .15
TO	.38 / .47	-.06 / .11	.06 / .13	.12 / .27	.09 / -.05	-.07 / .09	-.10 / .10	.02 / .12	.07 / .09	-.23 / .07	.08 / .20	.00 / .04
AC	.20 / .40	.47 / .66	.09 / .27	.16 / .32	.21 / .25	.13 / .25	.26 / .42	.14 / .21	.06 / .12	.01 / .12	.07 / .18	.18 / .15
OR	-.06 / 0	.58 / .64	.07 / .17	.09 / .18	.20 / .33	.14 / .27	.25 / .40	.16 / .27	-.03 / .01	.07 / .11	.03 / .05	.15 / .12
DO	.18 / .30	.12 / .25	.48 / .50	-.06 / .00	.17 / .22	.02 / .16	.14 / .25	.14 / .16	-.08 / -.1	-.15 / -.15	.09 / .15	-.04 / .14
PC	.04 / .14	.28 / .33	.17 / .28	.03/ .15	.17 / 0/29	.17 / .28	.21 / .30	.38 / .46	-.13 / -.12	-.08 / .01	.17 / .12	.11 / .09
SO	.11 / .17	.17 / .23	.72 / .80	.28 / .32	.35 / .43	.15 / .23	.26 / .37	.19 / .22	-.01 / .02	.06 / .17	.03 / .06	-.03 / .10
SE	.25 / .33	.07 / .17	.01 / .17	.38 / .46	.00 / -.07	.04 / .10	-.01 / .11	-.03 / -.04	.15 / .21	-.01 / .03	.02 / -.06	.03 / .05
ET	.07 / .11	.25 / .27	-.02 / .13	.47 / .52	.51 / .55	.23 / .31	.32 / .43	.13 / .25	-.11 / -.01	.07 / .13	.03 / .04	.08 / .10
OP	.01 / .12	.42 / .49	.35 / .41	.31 / .39	.62 / .72	.59 / .72	.75 / .82	.33 / .40	-.07 / -.02	.11 / .20	.06 / .07	.17 / .17

Note. Numbers on the left of the slash are the validity coefficients for FC.
Numbers on the right of the slash are validity coefficients for SS.
IE=Intellectual Efficiency; TO=Tolerance; AC=Achievement; OR=Orderliness; DO=Dominance; PC=Physical conditioning; SO=Sociability; SE=Selflessness; ET=Even Temper; OP=Optimism; SWB=Subjective Well-Being; CSES=Core Self-Evaluation; SH=Subjective Health.

Table 5

Test-retest reliability and marginal reliability

	Test-retest reliability		Marginal reliability	
	FC	SS	FC	SS
IE	.73	.79	.69	.79
TO	.77	.83	.75	.83
AC	.69	.79	.69	.83
OR	.83	.85	.84	.89
DO	.80	.91	.80	.92
SO	.83	.89	.77	.90
PC	.79	.90	.86	.92
SE	.73	.79	.67	.73
EV	.77	.84	.70	.84
OP	.77	.86	.80	.89

IE=Intellectual Efficiency; TO=Tolerance; AC=Achievement; OR=Orderliness; DO=Dominance; PC=Physical conditioning; SO=Sociability; SE=Selflessness; EV=Even Temper; OP=Optimism.

Table 6

Factor loadings from exploratory factor analysis (Study 2)

	Openness		Conscientiousness		Extraversion		Agreeableness		Emotional Stability	
	FC	SS	FC	SS	FC	SS	FC	SS	FC	SS
BFI_O	**.43**	**.43**	.04	.02	.09	.02	**.42**	**.39**	.02	-.01
IE	**.74**	**.59**	.03	.02	-.03	-.04	.03	.18	.10	.24
TO	.15	.19	-.18	-.10	.08	.04	**.54**	**.66**	-.11	-.09
BFI_C	.02	.05	**.95**	**.97**	.03	-.03	.05	.03	.09	.04
AC	.29	**.57**	**.41**	**.33**	.02	.03	.17	.27	-.04	.04
OR	-.03	-.07	**.74**	**.78**	-.04	.08	-.13	-.11	-.15	-.06
BFI_E	.00	.10	.04	.06	**.97**	**.93**	.02	-.02	.04	.03
DO	**.37**	**.56**	-.06	-.05	**.55**	**.47**	-.17	-.17	-.06	.02
SO	-.14	-.14	-.07	-.02	**.66**	**.81**	.11	.20	.09	.09
BFI_A	-.20	-.16	.20	.20	.08	.07	**.62**	**.65**	.22	.22
SE	.08	.08	.00	-.03	-.01	.10	**.75**	**.77**	-.08	-.04
BFI_N	-.04	-.04	-.05	-.02	-.06	-.04	.07	.08	**-.96**	**-.98**
ET	.00	-.03	-.08	-.02	-.20	-.17	.27	.29	**.62**	**.63**
OP	-.01	.03	.03	.07	.22	**.30**	-.03	-.01	**.59**	**.62**
PC	-.21	-.03	.08	.16	**.36**	**.36**	.09	.00	.01	.16
Congruence	.91		.93		.98		.98		.98	
Variance	.10	.10	.13	.13	.15	.15	.14	.14	.15	.15

Note. Factor loadings above .30 are marked in bold. IE=Intellectual Efficiency; TO=Tolerance; AC=Achievement; OR=Orderliness; DO=Dominance; PC=Physical conditioning; SO=Sociability; SE=Selflessness; ET=Even Temper; OP=Optimism. Congruence stands for Tucker's factor congruence coefficient.

Table 7

Factor inter-correlations obtained from EFA (Study 2)

	1	2	3	4	5
1. Openness		.17	.35	.23	.14
2. Conscientiousness	-.02		.26	.19	.45
3. Extraversion	.18	.19		.14	.43
4. Agreeableness	.09	.12	.07		.30
5. Emotional Stability	.18	.36	.40	.27	

Note. Values below diagonal were from the FC data and values above diagonal were from the SS data.

Table 8

Criterion-related validity (Study 2)

Criterion	IE	TO	AC	OR	DO	SO	PC	SE	ET	OP
BFI-O	.41/.57	.43/.54	.17/.43	-.02/.05	.15/.26	.10/.22	.04/.12	.33/.38	.16/.16	.09/.16
BFI-C	.08/.29	-.10/.06	.43/.54	.63/.73	.02/.16	.16/.25	.19/.32	.11/.17	.21/.30	.35/.45
BFI-E	.15/.36	.07/.16	.17/.44	.06/.26	.55/.67	.67/.81	.36/.47	.05/.23	.09/.15	.49/.62
BFI-A	-.01/.28	.20/.35	.18/.43	.07/.19	-.12/.04	.28/.39	.13/.22	.47/.56	.48/.54	.31/.43
BFI-N	-.11/-.37	.03/-.07	-.17/-.36	-.10/-.30	-.14/-.27	-.38/-.14	-.22/-.39	-.07/-.18	-.54/-.60	-.69/-.80
SWBS	.01/.18	-.04/.00	.09/.28	.10/.25	.12/.21	.24/.31	.20/.32	.03/.14	.17/.24	.64/.75
CSES	.14/.36	-.02/.05	.24/.44	.16/.35	.20/.35	.35/.50	.24/.43	.05/.17	.32/.38	.73/.86
SH	-.09/.09	-.00/.10	-.03/.13	-.01/.20	.09/.13	.21/.28	.39/.45	-.03/.11	.16/.20	.40/.47
JIG	-.04/-.23	-.01/-.14	-.22/-.33	-.08/-.15	-.13/-.22	-.19/-.28	-.08/-.20	-.12/-.23	-.24/-.22	-.36/-.45
People	.01/-.16	-.18/-.18	-.15/-.25	-.08/-.14	.02/-.07	-.16/-.21	-.02/-.12	-.22/-.28	-.25/-.21	-.20/-.30
Work	-.06/-.20	-.03/-.10	-.21/-.32	-.03/-.17	-.14/-.25	-.22/-.29	-.07/-.20	-.13/-.18	-.15/-.21	-.32/-.43
Payment	-.05/-.15	.06/.01	-.15/-.19	-.07/-.11	-.17/-.25	-.21/-.25	-.17/-.24	.03/.02	-.17/-.14	-.39/-.46
Promotion	.02/-.10	-.05/-.01	-.12/-.16	-.01/-.13	-.16/-.23	-.26/-.27	-.15/-.25	.03/-.06	-.09/-.06	-.31/-.38
Supervision	-.03/-.17	-.03/-.11	-.09/-.28	-.01/-.07	-.06/-.21	-.16/-.24	-.11/-.14	-.04/-.19	-.18/-.11	-.25/-.34
Gender	-.02/-.08	.08/.12	.13/.15	.01/-.01	-.02/-.04	-.05/.00	-.07/-.06	.22/.22	-.13/-.13	-.04/-.01
Age	.06/.09	-.06/.00	.16/.16	.10/.12	-.06/-.06	.22/.28	-.03/.04	.18/.08	.15/.16	.13/.18
Edu	.10/.17	.09/.11	.12/.19	-.02/-.02	.21/.19	.09/.16	.09/.19	.03/.14	.01/.10	.07/.19
Income	.00/.09	-.07/-.05	.04/.06	.00/.04	.13/.11	.13/.16	.09/.15	-.12/-.04	.07/.02	.26/.23

Note. IE=Intellectual Efficiency; TO=Tolerance; AC=Achievement; OR=Orderliness; DO=Dominance; PC=Physical conditioning; SO=Sociability; SE=Selflessness; ET=Even Temper; OP=Optimism; SWB=Subjective Well-Being; CSES=Core Self-Evaluation; SH=Subjective Health. JIG=Job in General; Numbers on the left of the slash are the validity coefficients for FC. Numbers on the right of the slash are validity coefficients for SS.

Table 9

Reliability and convergent validity (Study 2)

Facets	Marginal reliability		Convergent validity (Raw)	Convergent validity (Corrected for unreliability)
	FC	SS		
IE	.74	.78	.68	.89
TO	.77	.83	.76	.95
AC	.66	.84	.59	.79
OR	.85	.87	.74	.87
DO	.81	.92	.76	.88
SO	.78	.91	.76	.90
PC	.86	.91	.81	.92
SE	.67	.72	.63	.90
EV	.73	.80	.72	.94
OP	.82	.90	.78	.91

Note. IE=Intellectual Efficiency; TO=Tolerance; AC=Achievement; OR=Orderliness; DO=Dominance; PC=Physical conditioning; SO=Sociability; SE=Selflessness; ET=Even Temper; OP=Optimism.

REFERENCES

Anastasi, A., & Urbaina, S. (1997*). Psychology testing (7th ed.)*. Upper Saddle River, NJ: Prentice-Hall.

Armstrong, P. I., Allison, W., & Rounds, J. (2008). Development and initial validation of brief public domain RIASEC marker scales. *Journal of Vocational Behavior, 73*(2), 287-299.

Bachman, J. G., & O'MALLEY, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly, 48*(2), 491-509.

Bartram, D. (2007). Increasing validity with forced‐choice criterion measurement formats. *International Journal of Selection and Assessment, 1*5(3), 263-272.

Bartram, D., & Brown, A. (2004). Online testing: Mode of administration and the stability of OPQ 32i scores. *International Journal of Selection and Assessment, 12*(3), 278-284.

Beck, A. T., Steer, R. A., & Brown, G. (1996). *Beck Depression Inventory II manual*. San Antonio, TX: Psychological Corporation.

Bostic, T. J., Rubio, D. M., & Hood, M. (2000). A validation of the subjective vitality scale using structural equation modeling. *Social Indicators Research, 52*(3), 313-324.

Brown, A. (2016a). Item response models for forced-choice questionnaires: A common framework. *Psychometrika, 81*(1), 135-160.

Brown, A. (2016). Thurstonian scaling of compositional questionnaire data. *Multivariate Behavioral Research, 51*(2-3), 345-356.

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods, 20*(1), 121-148.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460-502.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36-52.

Brown, A., & Maydeu-Olivares, A. (2017). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-14.

Cao, M. (2016). *Examining the fakability of forced-choice individual differences measures*. (Doctoral dissertation). University of Illinois at Urbana-Champaign.

Cao, M., Drasgow, F., & Cho, S. (2015). Developing ideal intermediate personality items for the ideal point model. *Organizational Research Methods, 18*(2), 252-275.

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI Work satisfaction scale. *Personality and Individual Differences, 49*(7), 743-748.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16 PF)*. Champaign, IL: Institute for Personality and Ability Testing.

Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika, 30*(1), 99-121.

Chen, C., Lee, S. Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6*(3), 170-175.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*(1), 143-159.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: the relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82(*2), 300-310.

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*(1), 88-106.

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22*(2), 105-127.

Cho, S., Drasgow, F., & Cao, M. (2015). An investigation of emotional intelligence measures using item response theory. *Psychological Assessment, 27*(4), 1241-1252.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267-307.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Costa, P. T., Jr., &McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*(1), 71-75.

Donovan, J. J., Dwight, S. A., & Schneider, D. (2014). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *Journal of Business and Psychology, 29*(3), 479-493.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right!. *Industrial and Organizational Psychology, 3*(4), 465-476.

Edwards, A. L. (1957). *Manual for the Edwards Personal Preference Schedule*. New York, NY: Psychological Corporation.

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods, 21*(3), 328-347.

Guenole, N., Brown, A. A., & Cooper, A. J. (2016). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of thurstonian Item Response Modeling. *Assessment*. Advance online publication. doi:10.1177/1073191116641181.

Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and an initial survey of researchers. *International Journal of Selection and Assessment, 11*(4), 340-344.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*(3), 341-355.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 639-683.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*(3), 167-184.

Höfel, L., & Jacobsen, T. (2003). Temporal stability and consistency of aesthetic judgments of beauty of formal graphic patterns. *Perceptual and Motor Skills, 96*(1), 30-32.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-cultural Psychology, 20*(3), 296-309.

Jennrich, R. I. (1970). An asymptotic χ^2 test for the equality of two correlation matrices. *Journal of the American Statistical Association, 65*(330), 904-912.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and Research*, 2, 102-138.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2003). The core self‐evaluations scale: Development of a measure. *Personnel psychology, 56*(2), 303-331.

Kam, C. C. S. (2016). Further considerations in using items with diverse content to measure acquiescence. *Educational and Psychological Measurement, 76*(1), 164-174.

Kuder, G. F. (1960). *Administrator's manual, Kuder Preference Record, Vocational, Form C*. Chicago, IL: Science Research Associates.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Lee, P., Joo, S.H., Stark, S., & Chernyshenko, O.S. (2018). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement*. Advanced on-line publication.

Leung, K., & Bond, M. H. (1989). On the empirical identification of dimensions for cross-cultural comparisons. *Journal of Cross-cultural Psychology, 20*(2), 133-151.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 1-55.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433-442.

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*(3), 310-318.

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(2), 57-64.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-cultural Psychology, 23*(4), 498-509.

Maydeu-Olivares, A., Hernández, A., & McDonald, R. P. (2006). A multidimensional ideal point item response theory model for binary data. *Multivariate Behavioral Research, 41*(4), 445-472.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344-362.

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods, 8*(2), 222-248.

McLeod, B. (1991). Effects of Eyerobics visual skills training on selected performance measures of female varsity soccer players. *Perceptual and Motor Skills, 72*(3), 863-866.

McDaniel, M. A., Douglas, E. F.,&Snell, A. F. (1997, April). *A survey of deception among job seekers*. Paper presented at the twelfth annual conference of the Society of Industrial and Organizational Psychology, St. Louis, MO.

Meglino, B. M., & Ravlin, E. C. (1998). Individual values in organizations: Concepts, controversies, and research. *Journal of Management, 24*(3), 351-389.

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J., & Ponsoda, V. (2016). A Dominance Variant Under the Multi-Unidimensional Pairwise-Preference Framework: Model Formulation and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement, 40*(7), 500-516.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*(6), 660-679.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903.

Por, H.H., & Budescu, D. V. (2017). Eliciting Subjective Probabilities through Pair‐wise Comparisons. *Journal of Behavioral Decision Making, 30*(2), 181-196.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338.

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment, 25*(4), 1137-1145.

R Core Team (2016): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria .https://www.R-project.org/

Revelle, W. (2016). *How to: Use the psych package for factor analysis and data reduction*. Evanston, IL: Department of Psychology, Northwestern University.

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*(1), 3-32.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2*(4), 313-345.

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*(3), 231-255.

Rundquist, E. A. (1946, September). *The forced-choice technique and rating scales*. Paper presented at American Psychological Association, Philadelphia.

Salgado, J. F., Anderson, N., & Tauriz, 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*(4), 797-834.

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.

Sass, R., Frick, S., Reips, U. D., & Wetzel, E. (2018). Taking the Test Taker's Perspective: Response Process and Test Motivation in Multidimensional Forced-Choice Versus Rating Scale Instruments. *Assessment*. Advanced online publication.

Sisson, E. D. (1948). Forced choice—the new Army rating. *Personnel Psychology, 1*(3), 365-381.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-cultural Psychology, 35*(1), 50-61.

Smither, J. W., Reilly, R. R., Millsap, R. E., AT&T, K. P., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46*(1), 49-76.

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117-143.

Stanton, J. M., Sinar, E. F., Balzer, W. K., Julian, A. L., Thoresen, P., Aziz, S., ... & Smith, P. C. (2002). Development of a compact measure of job satisfaction: The abridged Job Descriptive Index. *Educational and Psychological Measurement, 62*(1), 173-191.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*(3), 184-203.

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology, 26*(3), 153-164.

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15*(3), 463-487.

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*(1), 25-39.

Staugas, L., & McQuitty, L. L. (1950). A new application of forced choice ratings. *Personnel Psychology, 3*(4), 413-424.

Strong, E. K. (1959). *Manual for the Strong Vocational Interest Blanks for men and women*. Palo Alto, CA: Consulting Psychologists Press.

Sun, T., Fraley, C., & Drasgow, F. (under review). When matches are ideal: Fitting measurement models to adult attachment data.

Tay, L., Drasgow, F., Rounds, J., & Williams, B. A. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal?. *Journal of Applied Psychology, 94*(5), 1287-1304.

Tenopyr, M. L. (1988). Artifactual reliability of forced-choice scales. *Journal of Applied Psychology, 73*(4), 749-751.

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-cultural Psychology, 38*(2), 227-242.

Triandis, H. C. (1972). *The analysis of subjective culture*. New York; Wiley.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Usami, S., Sakamoto, A., Naito, J., & Abe, Y. (2016). Developing Pairwise Preference-Based Personality Test and Experimental Investigation of Its Resistance to Faking Effect by Item Response Model. *International Journal of Testing, 16*(4), 288-309.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods, 15*(1), 96-110.

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). *Response biases. The ITC international handbook of testing and assessment* (pp. 349–363). New York: Oxford University Press.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment, 23*(3), 279-291.

White, L. A., Young, M. C., Hunter, A. E., & Rumsey, M. G. (2008). Lessons learned in transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology, 1*(3), 291-295.

Yammarino, F. J. (2003). Modern data analytic techniques for multisource feedback. *Organizational Research Methods, 6*(1), 6-14.