Mix-Keying or Desirability-Matching in the Construction of Forced-Choice Measures? An Empirical Investigation and Practical Recommendations

Mengtong Li[1], Bo Zhang[1,2], Lingyue Li[1], Tianjun Sun[3], Anna Brown[4]

[1] *Department of Psychology, University of Illinois Urbana-Champaign*

[2] *School of Labor and Employment Relations, University of Illinois Urbana-Champaign*

[3] *Department of Psychological Sciences, Kansas State University*

[4] *School of Psychology, University of Kent*

Manuscript Status: Accepted by *Organizational Research Methods*, Jan 9th, 2024. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission.

Author Note

An earlier version of this paper was presented at the 38th Annual Conference of the Society for Industrial and Organizational Psychology.

Correspondence concerning this article should be addressed to Bo Zhang, School of Labor and Employment Relations and Department of Psychology, University of Illinois Urbana-Champaign, 504 E Armory Ave, Champaign, IL 61820. Email: bozhang3@illinois.edu.

We thank Dr. Fritz Drasgow for his insightful comments on an earlier version of the manuscript.

Mengtong Li is currently a doctoral candidate in the industrial-organizational psychology program at University of Illinois at Urbana-Champaign. His research mainly focuses on personality measurement, leadership, and research methods. His work advocates for advanced techniques on facilitating automatic test construction and validation.

Bo Zhang is currently an assistant professor at the School of Labor \and Employment Relations and the Department of Psychology at the University of Illinois Urbana-Champaign. His research focuses on personnel selection, personality, and quantitative methods.

Lingyue Li is currently a doctoral candidate in the Department of Psychology at University of Illinois Urbana-Champaign. Her research primarily focuses on personnel selection, individual differences, and quantitative methods.

Tianjun Sun received her Ph.D. in psychology from University of Illinois Urbana-Champaign and is currently an assistant professor of industrial-organizational psychology at Kansas State University. Her research primarily focuses on personnel selection, individual differences, and quantitative methods. Her work advocates for the responsible use of psychometric tools and advanced technology to improve psychological sciences and solve organizational problems.

Anna Brown is currently Professor of Psychometrics at the School of Psychology at University of Kent, United Kingdom. Her research focuses on scaling comparative judgements, detecting and preventing response biases and faking behaviors, and modelling response processes using IRT and SEM frameworks more broadly. Beyond academia, Professor Brown has extensive experience in designing, developing and implementing psychometric testing solutions in the workplace, health settings and education, and provides psychometric advice to several organizations in the private and public sectors internationally.

Abstract

Forced-choice (FC) measures are becoming increasingly popular as an alternative to single-statement (SS) measures. However, to ensure the practical usefulness of an FC measure, it is crucial to address the tension between psychometric properties and faking resistance by balancing mixed keying and social desirability matching. It is currently unknown from an empirical perspective whether the two design criteria can be reconciled, and how they impact respondent reactions. By conducting a two-wave experimental design, we constructed four FC measures with varying degrees of mixed-keying and social desirability matching from the same statement pool and investigated their differences in terms of psychometric properties, faking resistance, and respondent reactions. Results showed that all FC measures demonstrated comparable reliability and induced similar respondent reactions. FC measures with stricter social desirability matching were more faking resistant, while FC measures with more mixed keyed blocks had higher convergent validity with SS measures and displayed similar discriminant and criterion-related validity profiles as the SS benchmark. More importantly, we found that it is possible to strike a balance between social desirability matching and mixed keying, such that FC measures can have adequate psychometric properties and faking resistance. A 7-step recommendation and a tutorial based on the *autoFC* R package were provided to help readers construct their own FC measures.

Keywords: forced-choice; faking; mixed-keying; social desirability matching

Mix-Keying or Desirability-Matching in the Construction of Forced-Choice Measures? An Empirical Investigation and Practical Recommendations

The forced-choice (FC) format has regained popularity in recent years as an alternative to traditional single-statement (SS) format for several reasons. First, properly designed FC measures are substantially more faking-resistant than their SS counterparts (Cao & Drasgow, 2019). Second, compared to SS measures, personality scores derived from FC measures possess higher predictive validity for important work outcomes (Salgado et al., 2015; Speer et al., 2023). Third, FC measures by design are immune to a number of response biases plaguing SS measures, such as acquiescence, extreme responding, and reference bias (Schulte et al., 2021). Finally, recent development of freely accessible R packages for automatic test assembly (Li et al., 2022) and scoring (Bürkner, 2019; Zhang, Tu et al., 2023) lifted many technical barriers. It is not surprising that FC measures are receiving growing attention, especially from those concerned with applicant faking.

However, previous Monte Carlo simulation studies revealed a seeming tension between psychometric properties and faking-resistance (i.e., Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Bürkner, 2022; Lee et al., 2022). On one hand, for the estimation of reliable latent trait scores, the state-of-the-art FC scoring model requires *mixed keying* (Brown & Maydeu-Olivares, 2011), which necessitates the inclusion of FC blocks with a mixture of both positively and negatively keyed statements. On the other hand, faking-resistance requires statements within a block to have similar levels of social desirability, which in most cases means that statements within a block should *not* be mixed keyed. How to strike a balance between mix-keying and social

desirability matching such that we can develop psychometrically sound *and* faking-resistant FC measures is one of the most urgent practical issues to be addressed in FC measurement. While simulation studies are informative for understanding the reliability of trait score estimates under fully controlled ideal conditions, they offer limited insights regarding the fakability of FC measures in real world, as this is an empirical question by nature. Therefore, we argue that it is critical to move beyond simulations and towards empirical studies to investigate the key question of how to strike a balance between mixed-keying and social desirability matching to produce psychometrically sound *and* faking-resistant FC measures.

Another often neglected aspect of FC measurement is respondent reactions, which are critical for respondent engagement, data quality, and recruitment success (Hausknecht et al., 2004). Although some previous examinations of respondent reactions towards FC measures have been conducted (Converse et al., 2008; Dalal et al., 2021; Sass et al., 2020; Zhang, Sun et al., 2020; Zhang, Luo et al., 2023), no evidence is available yet for the potential impact of mixed keying and social desirability matching on respondents' reactions in both honest and motivated faking situations. Again, the question of respondent reactions is purely empirical in nature and thus requires empirical data to answer it.

Therefore, echoing the call for more research on developing psychometrically sound, faking-resistant, and user-friendly personnel selection tools (Van Iddekinge et al., 2023), this study seeks to provide the first piece of *empirical* evidence on how mixed-keyed blocks and social desirability matching may impact FC measures' psychometric properties (i.e., reliability,

convergent validity, discriminant validity, and criterion-related validity), fakability (i.e., rank-order stability and mean score inflation) and respondent reactions (i.e., general and selection-specific). Based on our empirical findings, we also provide a step-by-step recommendation on how to construct good FC measures in the Discussion. While we do not conduct simulation studies to answer the focal *empirical* questions, we do appreciate the value of simulations for FC measure development, such as estimating the reliability of trait scores derived from FC measures in ideal conditions. Therefore, we updated the R package *autoFC* (Li et al., 2022) with several additional functions that users can easily use to run customized simulations, and an associated tutorial in the Online Supplementary Materials. Ultimately, we aim to contribute empirical knowledge and tools to the construction of high-quality FC measures.

A Brief Overview of Forced-Choice Measurement

Noncognitive constructs such as personality and vocational interests have been playing increasingly important roles in organizational research and personnel selection, due to their sizeable predictive validity for important organizational outcomes (He et al., 2019; Nye et al., 2012) and their potential to reduce adverse impact (Cottrell et al., 2015; Jones et al., 2022). Thus, accurate assessment of these constructs has been of key interest. The SS format, which requires respondents to indicate to what extent they agree with each statement on a polytomous scale (e.g., 1 = "Strongly disagree", …, 5 = "Strongly agree"), is no doubt the most widely adopted format to assess noncognitive constructs, due to the relative ease of scale development, administration, and scoring. However, scores derived from SS measures are often contaminated by various response biases and

deliberate faking (Kreitchmann et al., 2019; Wetzel et al., 2021; Zhang, Cao et al., 2020), all of which will, at least to some degree, render such scores less valid for making between-person comparisons and predicting key workplace outcomes (Schulte et al., 2021).

To address the issues of response biases and faking in the SS format, the FC format was introduced as an alternative (Sisson, 1948). In an FC measure, individuals are presented with *blocks*, each containing at least two *statements* ("statements" in FC measures are the same as "items" in SS measures). Respondents are then asked to either (1) choose the statement(s) that are most and/or least descriptive (the MOLE format) of themselves, or (2) rank all the statements in each block from the most descriptive to the least descriptive (the RANK format) of themselves (Cao & Drasgow, 2019). The number of statements per block is called *block size*, which is often constant within an FC measure and typically ranges from 2 to 5. A *block* is called *unidimensional* if the *statements* within that *block* measure the same latent trait, and *multidimensional* if the statements within that block measure different latent traits. Multidimensional blocks are more common than unidimensional ones. In Figure 1, we illustrated examples of multidimensional FC blocks.

Insert Figure 1 here

When the block size n is greater than 2, responses to each block (e.g., Lily chose statement A as "the most like me", statement B as "the least like me", and left statement C in between; A > C > B) will be decomposed into $n(n-1)/2$ pseudo items representing dichotomous outcomes of all unique pairwise comparisons (e.g., AB = 1; AC = 1; BC = 0), each indicating whether the first

statement in a pair is preferred to the second (outcome 1) or not (outcome 0). These pseudo items will serve as indicators of latent factors and be subjected to the Thurstonian Item Response Theory (TIRT; Brown & Maydeu-Olivares, 2011) model, which is a special type of categorical confirmatory factor analysis model for estimating statement parameters and person scores. This explains why larger blocks are more psychometrically informative when everything else is held constant, as larger blocks mean more indicators for latent factors (e.g., responses to block sizes of 2, 3, 4, and 5 corresponds to 1, 3, 6, and 10 pseudo items, respectively, if full ranking is elicited).

Unlike SS measures where participants make an *absolute* judgment regarding their agreement with each statement, FC measures require respondents to decide which statement in the current block describes them *relatively* better than others. Even when all statements within a block describe them with similar accuracy in the *absolute* sense (e.g., "all the statements describe me accurately/inaccurately"), they are still required to make a relative choice. Given the "forced" nature of responding, FC measures are immune to response biases such as acquiescence, extreme responding, halo and leniency bias by design (Schulte et al., 2021). If statements within blocks are further matched on social desirability, FC measures are also substantially more faking-resistant than their SS counterparts (Cao & Drasgow, 2019; Speer et al., 2023).

In sum, the removal of multiple response biases and the faking resistance potential have rendered the FC format a promising alternative to the SS format. However, one challenge in FC measures is that achieving good psychometric properties and faking resistance often seem incompatible with each other. Next, we will elaborate on the rationale for social desirability

matching and mixed keying in constructing FC measures, followed by a discussion on why the two *seem* incompatible. We will further discuss why respondent reactions should be considered when developing FC measures.

Maintaining Faking Resistance with Social Desirability Matching

Many noncognitive measures include both positively and negatively keyed statements for better coverage of the construct continuum (Tay & Ng, 2018). In most cases, positively keyed statements (e.g., "I am hardworking") are substantially more socially desirable than their negatively keyed counterparts (e.g., "I often come to work late"). Participants presented with these statements in the same FC block can often identify and choose the more desirable statement in motivated faking situations (Bürkner et al., 2019) regardless of whether it is truly more descriptive of them than others (Schulte et al., 2021). In this sense, the FC measure is said to be more *fakable*. Following previous practices (Cao & Drasgow, 2019; Hu & Connelly, 2021), we operationalized fakability of a measure in two complimentary ways: (1) as the standardized mean score difference (Cohen's *d*) between latent trait scores obtained in faking versus honest conditions, and (2) as the rank-order stability of latent trait scores across faking and honest situations. A larger standardized mean score difference and a lower rank-order stability both indicate a higher level of fakability. To minimize the opportunities for faking in FC measures, statements within the same block need to be matched on social desirability (Cao & Drasgow, 2019). To achieve this goal, researchers first need to obtain the social desirability values of all statements. Next, an index of similarity for desirability values (and the corresponding cutoff) is determined for the statements to be paired

within a block. More details on how to obtain social desirability values and use them for creating blocks are discussed in the Recommended Steps to Develop FC Measures section in the Discussion.

Improving Psychometric Properties Using Mixed Keying

Aside from fakability, a fundamental requirement for any FC measure is that estimated trait scores are reliable and valid. However, Brown (2016) and Bürkner (2022) mathematically showed that latent trait scores estimated from the FC format will be unreliable if all blocks contain statements keyed in the same direction (all statements have either positive or negative factor loadings). This is because equally keyed blocks provide little information regarding the *sum* of latent trait scores involved in these blocks, which is essential for recovering their *absolute* locations (Brown & Maydeu-Olivares, 2011; Schulte et al., 2021). More precisely, when holding other factors constant, the amount of information provided by a pair is positively related to the absolute difference between factor loadings of the two statements (Brown & Maydeu-Olivares, 2011; Bürkner, 2022), and will drop to zero if the two statements have identical factor loadings. Given that most statements have been selected to possess moderate to high factor loadings, if statements keyed in the same direction are put into the same block, their factor loading difference would be small and thus not psychometrically informative.

Indeed, many simulation studies have confirmed that FC measures with only equally keyed blocks suffered from various psychometric issues, including low model convergence rates, severely biased estimates of statement parameters, unreliable estimates of latent trait scores (Brown & Maydeu-Olivares, 2011), and biased estimates of inter-trait correlations (Bürkner et al.,

2019; Schulte et al., 2021). One theoretical solution is to include a larger number of latent factors (e.g., 30) and/or a large number of statements per trait (e.g., 15) with high factor loadings (e.g., .80 or above) in one FC measure, which has been shown to be effective in simulation studies (Bürkner et al., 2019; Schulte et al., 2021) and empirical settings (Brown & Bartram, 2009). However, this solution is likely too demanding and most often impractical. The easiest and the most effective solution is to include a substantial number of *mixed-keyed blocks* that contain both positively and negatively keyed statements, as it is easier to maintain substantial factor loading difference when one loading is positive and another one is negative, and both are of moderate magnitude. Simulations have shown that FC measures including mixed-keyed blocks consistently outperformed those with only equally keyed blocks in terms of model convergence and parameter recovery accuracy (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Schulte et al., 2021).

Social Desirability Matching vs. Mixed Keying: A Dilemma?

As discussed above, both social desirability matching and mixed keying are important for different aspects of properties of FC measures. However, satisfying both design criteria would place researchers into a dilemmatic position: on one hand, the faking resistance of an FC measure is based on social desirability *matching* that often requires *equally keyed* blocks; on the other hand, accurate score estimation requires a substantial number of *mixed keyed* blocks, which can be equivalent to social desirability *mismatching* because positively keyed statements are often more desirable than their negatively keyed counterparts. It seems that no matter which side we prioritize, the other side will suffer. This leads some researchers to conclude that it is impossible to

simultaneously maintain good psychometric properties and faking resistance for an FC measure (Bürkner et al., 2019; Ng et al., 2021; Schulte et al., 2021).

While we agree that social desirability matching and mixed keying may conflict with each other on some occasions, we argue that it is still possible to find a sweet spot between the two criteria, such that we can develop FC measures that are both sufficiently faking-resistant and psychometrically sound. According to a recent simulation study (Lee et al., 2022), 20% mixed blocks suffice to ensure reliable latent trait scores estimates, as long as the statements are reliable indicators of the latent factors. In addition, the marginal utility of more mixed blocks for reliability gradually decreases and reaches a plateau when the proportion of mixed blocks exceeds 60%, implying that too many mixed keyed blocks are unnecessary. These findings are important as they showed that including a small proportion of mixed-keyed blocks can substantially benefit the psychometric properties of an FC measure, while at the same time *presumably* not affecting the fakability as severely as previously thought, as an FC measure with 80% of its blocks matched on social desirability may still be fairly faking-resistant. It is possible to compromise slightly on both criteria to reach a sweet spot where both *good enough* psychometric properties and fakability are achieved. There is hope!

However, in pursuit of such a sweet spot, we must rely on empirical evidence from real human responses instead of simulations because faking is a complex psychological phenomenon, for which we do not yet have a satisfactory psychometric model. As such, even though Lee et al's. (2022) simulations provide benchmarks for satisfactory reliability, these simulation results tell us

little about the degree to which different proportions of mixed-keying blocks will impact the fakability of an FC measure. Aside from fakability, other important psychometric properties of a measure, such as criterion-related validity, can only be examined in empirical data collected from real human respondents as well. In fact, even previous simulation findings on the reliability of FC scores should also be subjected to empirical tests because all simulations studies are based on untested assumptions and if these assumptions do not hold empirically, findings may be untrustworthy. In sum, it is critical to move beyond simulation studies and use empirical data to investigate whether it is possible to balance social desirability matching and mixed keying, and ultimately, develop FC measures with *good enough* psychometric properties and faking-resistance.

Respondent Reactions to FC Measures

Another important but often neglected issue in developing FC measures is respondent reactions. Respondent reactions refer to respondents' attitudes, affect, or cognitions related to the measurement tool (Hausknecht et al., 2004). Positive respondent reaction can elicit favorable impressions on employers from respondents' perspectives, increase applicants' intention to recommend the employer to other job seekers, and improve data quality through enhanced test motivation (Hausknecht et al., 2004; McCarthy et al., 2017; Sass et al., 2020).

However, studies examining how FC design features can impact respondent reactions are still lacking, except Dalal et al., (2021) and Fuechtenhans and Brown (2022). Although Dalal et al., (2021) examined how different FC designs impacted respondent reactions, their use of a computerized adaptive testing design, where each respondent was presented with different

statements that best matched their latent trait levels, may confound the effect of different statements with the effect of design features. Moreover, in many scenarios, researchers and practitioners would use a static FC measure constructed from a small statement pool. It is thus critical to know how different pairing strategies will impact respondents' reactions when holding the statement pool constant. Fuechtenhans and Brown (2022) used a qualitative study design to examine how statement matching would impact respondents' experience with the FC format. They found that blocks with both desirable and undesirable statements are generally considered as easier and less cognitively demanding than blocks matched on social desirability. Although these findings are valuable, it is important to complement these findings with quantitative estimates from a rigorously designed experiment and provide a more comprehensive coverage of other aspects of respondent reactions (e.g., perceived fakability), which are still lacking in the current literature.

The Present Study

The present study seeks to empirically examine how different levels of social desirability matching and mixed keying influence the (1) psychometric properties of, (2) fakability of, and (3) respondent reactions to FC measures. Answers to these questions would not only complement previous simulation findings but also provide an evidence-based guide to the construction of reliable, valid, faking-resistant, and user-friendly FC measures. To achieve these goals, we constructed 4 different versions of FC measures based on the same set of statements (see the Methods section below for details). To further facilitate the use of the FC format, we provided step-by-step recommendations on how to construct high-quality FC measures in the Discussion. A

tutorial written in R implementing each of these steps using the *autoFC* (Li et al., 2022) package is also provided in the Online Supplementary Materials.

Methods

Participants and procedures

We conducted a two-wave study to examine the psychometric properties of, fakability of, and respondent reactions to four different FC measures with different design features while holding the statement pool constant. Respondents were recruited from the Prolific crowdsourcing platform. The study flow and demographic information in each group can be found in Figure 2. At Time 1 (honest condition), we aimed for 550 respondents per group, and a total of 2,187 respondents were eventually recruited, each paid $3 for participation. After consenting to proceed, participants first responded to demographic questions and were then presented randomly with one of the four FC measure versions. After this, they were immediately asked about their reactions to the FC measure, then followed by the SS measure. Finally, participants completed several criterion measures, which were presented in random order. For both FC and SS measures, participants were instructed to respond as honestly as possible. After excluding responses that failed more than one out of six quality control items, a total of 2,147 usable responses were retained.

Insert Figure 2 here

Three months later (Time 2; fake-good condition), all participants who participated in the Time 1 survey were invited to join a follow-up survey for a $2 reward. Similar to Time 1, participants consented to their participation and provided demographic information. However,

before presenting the focal measures, we simulated a fake-good test situation where participants were instructed to respond as if they were applying for their dream job. To increase the fidelity of the simulation, we first asked respondents to write down the name of their dream organizations and positions. They were then asked to use one or two sentences to explain why they wanted these positions. After the explanation, respondents were asked to imagine that their dream organizations were hiring and the organizations would use a personality measure test to decide who will be invited to fly to the headquarters for the final interview. To seize this opportunity, respondents were asked to try their best to get on the invitation list. Our decision of implementing a "faking for your dream job" scenario was based on two reasons. First, describing their most wanted positions from their dream organization can make the simulated scenario more personally relevant, and can better represent a real job application situation where most respondents would apply for jobs they like. Second, instructing participants to fake for their dream jobs overcomes the limitations brought by differential ability and motivation to fake had they been asked to fake for a predetermined position (Fuechtenhans & Brown, 2022). Following the simulated job application scenario, participants were presented with the same FC personality measure they completed in Time 1. After the completion of the FC measure, participants again indicated their reactions toward the measure, and then responded to the SS personality measure. No criterion measures were presented in Time 2. A total of 1,177 responses were collected at Time 2, resulting in a response rate of 54.82%.

Measures

Demographics. Respondent self-reported their demographic information, including age, gender (1 = Female, 2 = Male, 3 = Non-binary), education level (1 = Primary school, 2 = High school or equivalent, 3 = Some college or equivalent, 4 = Bachelor or equivalent, 5 = Master, 6 = PhD), and annual income before tax (1 = under $10,000, 2 = $10,000-$19,999, 3 = $20,000-$29,999, 4 = $30,000-$39,999, 5 = $40,000-$49,999, 6 = $50,000-$74,999, 7 = $75,000-$99,999, 8 = $100,000-$150,000, 9 = Over $150,000).

HEXACO-60_FC. In this study, we constructed FC measures based on the HEXACO-60 (Ashton & Lee, 2009) and used a triplet format. We chose 20 triplets because (1) we want the FC measures to be sensitive to manipulations, (2) we want to keep the survey to a reasonable length, and (3) 20 triplets for 5-6 latent factors are also quite common (e.g., Brown & Maydeu-Olivares, 2011; Lee et al., 2019; Walton et al., 2020; Wetzel & Frick, 2020). To develop triplet MFC HEXACO measures, we first obtained social desirability ratings (ranging from 1 to 5) for the 60 statements from Anglim et al., (2017). Specifically, social desirability ratings from the applicant sample in their study were used. We then constructed four different FC versions, each with 20 triplets consisting of three statements measuring different dimensions from HEXACO. We constructed these FC measures using an automatic item pairing R package, *autoFC* (Li et al., 2022). The design criteria for these four measures (also see Figure 2) were as follows: (1) For the first two FC measures (FC1 & FC2), statements within a triplet were matched by similar levels of social desirability. More specifically, we operationalized social desirability discrepancy as the maximum difference of social desirability among the three statements in a block, and maintained the mean

discrepancy across all 20 blocks as 0.34 (min = 0.05, max = 0.70) for FC1 and 0.37 (min = 0.13, max = 0.73) for FC2. The numbers of mixed keyed blocks for FC1 and FC2 were set to be 3 and 6, respectively. (2) For the remaining two FC measures (FC3 & FC4), the numbers of mixed keyed blocks were set to be 13 for FC3 and 12 for FC4 to represent cases with more mixed blocks[1]. The mean discrepancy across 20 blocks was 0.86 (min = 0.13, max = 1.85) for FC3 and 1.09 (min = 0.48, max = 2.01) for FC4. (3) Across the four FC versions, we tried our best to ensure that each latent trait was paired with the other five traits for about equal number of times (at least twice) while satisfying all previous constraints. Participants were required to select one statement describing them most, and another one describing them least from each block. Detailed block design for the four FC measures is presented in Table S1 from the Online Supplementary Materials.

In sum, FC1 and FC4 represent two realistic extremes of the compromise between social desirability matching and mixed keying. FC1 has the best match in terms of social desirability but has the fewest mixed-keyed blocks, while FC4 has the majority of blocks being mixed-keyed but is least matched on social desirability. The two FC measures in between (FC2 & FC3) represent attempts to strike a balance between mixed-keying and social desirability matching. Ideally, we would expect FC1 and FC2 to be more faking-resistant and FC3 and FC4 to be superior in measurement precision. By comparing FC1 to FC2, we can examine the effect of increasing the

[1] Ideally, maintaining equal numbers of mixed keyed blocks for FC3 and FC4 would be better (rather than having 13 for FC3 and 12 for FC4). But as shown in Lee et al., (2022), having beyond 12 mixed keyed blocks did not offer substantial psychometric gain. This means that all things being equal, psychometric properties of FC measures with 12 or 13 mixed keyed blocks would most likely be indistinguishable. Hence, when building FC3 and FC4, we were more lenient on the number of mixed keyed blocks but instead focused on manipulating social desirability matching.

number of mixed keyed blocks while maintaining the same degree of social desirability matching; By comparing FC3 to FC4, we examine the effect of relaxing social desirability matching while maintaining a sufficient number of mixed keyed blocks. Finally, by comparing FC2 to FC3, we can investigate the impact of different preferences for the balance between the two design criteria, in which FC2 favors better social desirability matching while FC3 favors more mixed keyed blocks. In sum, different comparisons between the four versions of FC can provide us with a holistic picture of the individual and joint impact of the two criteria.

HEXACO-60_SS. The same 60 statements from HEXACO-60 (Ashton & Lee, 2009) were also used as a single-statement (SS) Likert-type measure. Participants were instructed to indicate the extent to which each item described themselves on a 5-point rating scale. Items were randomly presented for each participant to reduce order effects. The SS measure served as an anchor to evaluate the psychometric properties of different FC measures.

Criterion measures. Details about criterion measures (e.g., reliability, length, rating scales) and their HEXACO correlates based on previous meta-analyses and large-sample primary studies are presented in Table 1. Full items can be found in the Online Supplementary Materials Section 4. Means, SDs, and reliabilities can be found in Table S2 in the Online Supplementary Materials.

Insert Table 1 here

Respondent reactions. Respondent reaction measures at Time 1 focused on general perceptions of the FC measures, while those at Time 2 were tailored to job application contexts. Items were adapted from previous studies (Chan et al., 1998; Dalal et al., 2021; Harris et al., 2021;

Highhouse et al., 2003; Lopez et al., 2019; Macan et al., 1994; Smither et al., 1993; Tonidandel et al., 2002; Zhang et al., 2020, 2023) and self-developed. Assessed facets and example items can be found in Table 1. Complete items can be found in Section 4 of the Online Supplementary Materials.

***Quality control items*.** Six quality control items were embedded, with five in the Likert measures and one in the FC measure. For items embedded in Likert measures, respondents were instructed to endorse a particular response option (e.g., *strongly disagree*). The quality control block in the FC measure required participants to select the first statement from the block as "most like me" and the second statement as "least like me". In all subsequent analyses, we screened out respondents who missed more than one quality control item.

Scoring

All four FC measures were scored using the Thurstonian IRT (TIRT) model (Brown & Maydeu-Olivares, 2011) with the R package *thurstonianIRT* (version 0.12.1; Bürkner, 2019). Specifically, we used the Markov chain Monte Carlo (MCMC) approach with default diffuse priors to estimate the TIRT model. The TIRT model converged well for the four versions with the largest potential scale reduction factor (R hat) less than 1.10. As the SS version of the HEXACO scales was identical across the four groups, we pooled their responses together and scored them by the Multidimensional Graded Response Model (Samejima, 1997) using the R package *mirt* (version 1.33.2; Chalmers, 2012) with the estimator based on Cai's (2010) Metropolis-Hastings Robbins-

Monro (MHRM) algorithm[2]. To ensure comparability across time, Time 2 responses were scored by fixing statement parameters to those obtained at Time 1. Maximum a posteriori (MAP) estimates were obtained for both FC and SS personality measures. For the sake of simplicity, criterion and all other measures were scored using sum scores after reverse coding. For transparency, all data and analysis scripts were made available on the Open Science Framework: https://osf.io/yvpz3/?view_only=08601755f471440b80973194571b60bd.

Results Reporting

For psychometric properties, we reported (a) *empirical reliability*, computed as $\frac{\text{var}(\hat{\theta})}{\text{var}(\hat{\theta})+\text{mean}(\text{SE}(\hat{\theta})^2)}$ (Brown & Maydeu-Olivares, 2018), (b) *convergent validity* between FC scores and their SS counterparts, (c) *discriminant validity* (intercorrelations) between traits, as well as similarity of intercorrelations as indexed by double-entry intra-class correlation (ICC; Furr, 2010) (d) *criterion-related validity* of FC and SS, as well as the profile similarity between FC and SS criterion-related validity profiles as indexed by double-entry ICC. For fakability of FC and SS measures, we reported (e) *rank-order stability* of personality scores between honest and fake-good conditions, and (f) *faking effect* as indexed by Cohen's *d* between trait estimates obtained in honest and fake-good conditions. Respondent reactions were presented for each of the four FC measures

[2] The reason to choose a Bayesian estimator for the TIRT model is because, compared to limited information estimators (e.g., unweighted least square), Bayesian estimator leads to better convergence, more accurate estimates of statement parameters, inter-trait correlations and trait scores (Bürkner et al., 2019; Morillo et al., 2016), and can handle missing data in a way like full information maximum likelihood (FIML) estimators. The reason to choose a FIML estimator for the Graded Response Model, which is not available for the TIRT model, is because it is much faster and produces trait scores that are almost identical to those by Bayesian estimators (Kieftenbeld & Natesan, 2012).

at both time points. We reported descriptive statistics, McDonald's ω, and Cohen's *d* for pairwise comparisons between groups for each of the respondent reaction dimensions. Note that we focused on effect sizes instead of statistical significance.

Results

Psychometric Properties

Reliability. As shown in Table 2, the empirical reliabilities of all six traits for the four FC measures were at least marginally acceptable (> .63). Also, the reliabilities of all four FC measures were consistently lower than those of the SS measure (FC: average reliabilities across traits ranging from .69 to .73 at Time 1 and .67 to .71 at Time 2; SS: average reliabilities across traits equal to .84 at Time 1 and Time 2), regardless of the FC design features or the measurement contexts. Furthermore, for all four FC measures, reliability estimates at Time 2 (fake-good condition) were slightly lower than those at Time 1 (honest condition). Comparing the reliability estimates of each individual factor across different FC measures, the differences were mostly small (less than .10). Notable exceptions were (1) Extraversion between FC1 the other three FC measures, where the reliability for FC1 was .13 lower than FC2, .10 lower than FC3, and .14 lower than FC4 at Time 1, while .10 lower than FC2 at Time 2, and (2) Conscientiousness between FC3 and FC4, where the reliability for FC3 was .10 lower at Time 1. Overall, results from Table 2 showed that the impact of the extent of social desirability matching and mixed keying on FC reliability was limited as long as there were at least three mixed triplets. In addition, empirical reliability did not seem to be substantially compromised even if participants were motivated to fake their responses. Readers

interested in the standard error of measurement for each person score against their estimated latent trait levels can refer to Figures S1 and S2 in the Online Supplementary Materials.

Insert Table 2 here

Convergent validity. Table 3 shows the raw and corrected (for unreliability) convergent validity of all four FC measures with their SS counterpart. To control for the confounding effect of different reliabilities, we interpret the corrected convergent validities. At Time 1, the average correlation between trait scores measured with FC and SS measures was substantially lower for FC1 (but still large in magnitude; M = .81, min = .59, max = 1.00) while the other three versions had very similar and higher convergent validity (FC2: M = .91, min = .79, max = 1.00; FC3: M = .94, min = .88, max = 1.00; FC4: M = .94, min = .85, max = 1.00). This pattern suggests that although the construct validity of all four FC measures was properly retained, FC2-FC4 still fared better than FC1. At Time 2, the convergent validity dropped substantially (FC1: M = .71, min =.53, max = .84; FC2: M = .73, min =.55, max = .84; FC3: M = .79, min =.74, max = .93; FC4: M = .79, min =.73, max = .89). Noticeably, convergent validity for some traits in FC1 and FC2 was substantially worse than that in FC3 or FC4, particularly for Conscientiousness (.53 and .55 for FC1 and FC2 vs. .74 for FC3 and FC4). However, we note that the lower convergent validity with SS scores at Time 2 was likely because SS scores were substantially distorted due to faking.

Insert Table 3 here

Discriminant validity. In Table 4, we reported the model-based latent correlations among the six traits for FC1-FC4 and SS. As we scored Time 2 responses by fixing model parameters

obtained from Time 1 instead of separately estimating a model for each format at Time 2, we only reported discriminant validity information for Time 1. Overall, FC2, FC3, and FC4 demonstrated similarly moderate ICC with those estimated from SS (ICCs = .71, .67, and .64), while the trait intercorrelations for FC1 were vastly different from SS, as shown by the low ICC (.17). For example, the correlations for Emotionality with Honesty-Humility and Extraversion were .06 and -.34 respectively for SS but were .36 and -.03 for FC1. These results suggested that FC1 demonstrated substantially lower construct validity in terms of the intercorrelations between the traits. Besides those point estimates, the standard errors of discriminant validity estimates of FC1 (M = .078) were also about 10% - 20% higher than those for FC2-FC4 (M = .064, .070, and .068).

Insert Table 4 here

Criterion-related validity. Table 5 presents the double-entry ICC between the validity profiles of each FC and the SS measure after correcting for FC and SS reliability (see Table S4 and Table S5 in the Online Supplementary Materials for full corrected and raw correlations). A correction was conducted on these correlations to control for differential reliabilities for FC and SS measures. ICC for each factor and across all HEXACO factors were reported. Generally speaking, FC1 was the least similar to the SS in terms of criterion-related validity (ICC = .54) and FC4 was the most similar (ICC = .96). FC2 and FC3 also had similar validity profiles as the SS (ICCs = .83 and .77 for FC2 and FC3). The patterns when examined trait by trait are also consistent with that revealed by the overall ICCs. Specifically, FC1 displayed validity profiles that were the least similar to the SS (double entry ICCs = -.04, .08, and .23 for Conscientiousness, Openness,

and Honesty-Humility). In contrast, FC4 consistently demonstrated the highest resemblance to the SS for all personality traits. We also observed some trait specificity beyond the general pattern. For example, the validity profiles of Emotionality for FC1, FC2, and FC4 were highly similar to that of the SS (ICCs ranging from .94 to .97) while the ICC for FC3 Emotionality was somehow lower (ICC = .79). Additionally, the ICCs for Openness were consistently among the lower end of the six traits across the four FC measures (ICCs = .08 .40, .66, and .68 for FC1 to FC4), which was likely to be an artifact due to range restriction because openness correlated weakly with all criteria.

We also presented the R^2 of all six personality traits predicting each criterion variable in Table 5. Averaging across all criterion variables, FC4 demonstrated the highest average R^2 (.167) among the four FC measures, which was also close to the one produced by SS (.181). This was followed by FC3 (.129) and FC2 (.123), while FC1 had the lowest average R^2 (.117). If we further differentiate criterion variables that were subjectively assessed by Likert-type measures (e.g., dark personality, OCB, and CWB) from more objectively reported criterion variables (e.g., education, wages), we can see that the SS measure displayed superiority over the FC measures in predicting the former (average R^2s were .144, .152, .145, .209 and .237 for FC1, FC2, FC3, FC4 and SS) but showed no advantages at all in predicting the latter (average R^2s were .072, .075, .099, .095 and .085 for FC1, FC2, FC3, FC4 and SS). Overall, these results showed that the criterion-related validity for FC4 was the best, while for FC1, it was the worst.

Insert Table 5 here

Fakability

***Rank-order stability*.** Raw and corrected correlations between the same personality trait measured in both honest and fake-good conditions are presented in Table 6. To control for the confounding effect of reliability differences, we focused on corrected rank-order stability. A more faking-resistant measure should be better at preserving respondents' rank orders across honest and fake-good conditions. As expected, FC1 was the most faking-resistant (M = .81, min = .72, max = .95) and FC4 was the least faking-resistant (M = .71, min = .54, max = .90). All FC measures except FC4 were more faking-resistant than the SS measure (M = .72, min = .62, max = .89). Averaged across measures, Conscientiousness appeared to be most susceptible to faking, while Openness consistently showed the lowest susceptibility to faking. Correlations between Time 1 SS scores and Time 2 FC scores can be found in Table S7.

Insert Table 6 here

Mean score inflation. Raw and corrected (using formulas from Wiernik & Dahlke, 2020) standardized mean score differences (Cohen's *d*) between the honest and fake-good conditions can also be found in Table 6. Again, we focused on corrected effect sizes to account for the confounding effect of reliability differences. A more faking-resistant test should have smaller mean score inflation. As expected, FC1 was the most faking-resistant (M = 0.25, min = 0.03, max = 0.56) and FC4 was the least faking-resistant (M = 0.70, min = 0.45, max = 1.16). FC2 and FC3 were in between and performed very similarly (FC2: M = 0.37, min = 0.08, max = 0.66; FC3: M = 0.38, min = 0.17, max = 0.75) but showed differential faking effects across traits. For example, participants seemed to inflate their scores more easily on Honesty-Humility and Emotionality in

FC2, while the same trend was observed for Extraversion and Openness in FC3. Inspections by trait yielded the same conclusion as when examining rank-order stability: averaged across FC measures, Conscientiousness consistently showed the highest susceptibility to faking, while Openness consistently exhibited the lowest susceptibility. Interestingly, the observed faking effect sizes in FC4 were higher than those in SS (M = 0.47, min = 0.28, max = 0.65).

Respondent Reactions

As displayed in Table 7, in honest condition, we found almost no meaningful differences in any of the seven respondent reactions. Specifically, after adjusting for multiple comparisons, none of the differences was statistically significant (|*d*s| < .19). Likewise, in the fake-good condition, the majority of the comparisons were statistically non-significant with tiny effect sizes, except for perceived faking resistance between FC2 and FC4, where FC2 was perceived as more faking-resistant than FC4 (adjusted $p < .05$, $d = .30$).

Insert Table 7 here

Discussion

When intended for high-stakes situations, good psychometric properties and strong faking resistance are the two primary yet somewhat contradictory requirements for FC design. Achieving faking resistance often necessitates social desirability *matching*, but good psychometric properties require some *mixed-keyed* blocks which are often inevitably equivalent to social desirability *mismatching*. Besides, respondent reactions are also important in both low- and high-stakes situations because they may impact data quality and selection outcomes. However, no empirical

evidence is yet available regarding the effects of different levels of social desirability matching and mixed keying on the psychometric properties of, faking resistance of, and respondents' reactions to FC measures. To fill in this critical empirical gap, we conducted the first time-lagged experimental study to examine these three issues under different conditions by manipulating the levels of social desirability matching and mixed-keying. Results showed that (1) the impact of social desirability matching and mixed-keying on reliability was small (as long as there are at least 3 mixed blocks), (2) FC measures with more mixed-keyed blocks had substantially higher convergent validity with SS, more similar criterion-related and discriminant validity profile with SS, and can better predict criterion variables, (3) FC measures with better social desirability matching were generally more faking-resistant, and (4) different combinations of mixed keying and social desirability matching had negligible impact on respondents' reactions in both honest and fake-good conditions. These findings demonstrate that it is possible to find a sweet spot between social desirability matching and mixed keying and thus construct a psychometrically sound and faking-resistant FC measure. Based on these findings and our first-hand experience with FC construction, we further provide empirical guidance on how to construct such a measure.

Mixed Keying or Social Desirability Matching?

***Psychometric Properties and Faking Resistance*.** Building upon initial attempts to reach a possible sweet spot between the two design criteria (Lee et al., 2022), we provided the first comprehensive *empirical* investigation on the effects of different mixed keying and social desirability matching combinations. First, as expected, neither solely focusing on social

desirability matching (FC1) nor mixed keyed blocks (FC4) can produce FC measures that are both faking resistant and have good psychometric properties. For FC1, although its high degree of social desirability matching brings notable advantages in faking resistance, the lack of mixed keyed blocks undermines its convergent validity, discriminant validity, and criterion-related validity. This essentially brings into question the construct validity of the scores. On the other hand, for FC4, its extensive focus on more mixed keyed blocks indeed allows its construct validity to be well maintained, consistent with the emphasis on mixed keying from simulation studies. However, relaxing social desirability matching too much also renders it more fakable compared to other FC counterparts. The utility of FC4 under the fake-good condition is hence limited.

These negative outcomes revealed by FC1 and FC4 suggest that for FC measures to be as valid and faking resistant as they are supposed to be, scale developers need to consider designing FC measures within a "middle ground". As such, FC2 and FC3, representing the "middle ground" compromise between mixed keying and social desirability matching, demonstrate a better balance between psychometric properties and faking resistance. Comparing FC1 with FC2, consistent with Lee et al., (2022), psychometric properties of FC measures can be effectively improved and reach an acceptable level with the inclusion of just 3 more mixed keyed blocks (i.e., from 15% to 30%). We further extended their findings by showing that such improvement can even be achieved with a slight compromise in social desirability matching. Comparing FC3 with FC4, we found that faking resistance can be substantially strengthened with a better match in terms of social desirability (from 1.09 to 0.86 in terms of mean block desirability discrepancy), without reducing

the number of mixed keyed blocks. Even more importantly, although FC2 and FC3 differed in terms of social desirability matching and mixed keying, their psychometric properties and faking resistance were largely similar. Admittedly, the tradeoff still exists and a certain amount of loss in desirable psychometric properties is unavoidable, but such a tradeoff seems acceptable for keeping FC as a both valid and faking-resistant measurement tool. We also acknowledge that it is not an easy task to find such a balance manually given a fixed statement pool, because the number of possible combinations can be astronomical. Hence, we recommend researchers use the R package *autoFC* (Li et al., 2022) to automate the search process and find the nearly optimal solutions.

Respondent Reactions. No substantial impact was found for social desirability matching or mixed keying on respondent reactions. This is reassuring because it suggests that test developers do not need to worry a lot about respondent reactions when developing new FC measures.

Forced-Choice vs. Single Statement Measures

Although the primary focus of the present study is on the comparisons across the four FC measures, we believe the comparisons between the FC and the SS measures may also be of interest. By design, FC measures are less susceptible to or even immune from multiple response biases that plagues the SS format (Kreitchmann et al., 2019; Zhang, Luo et al., 2023). Our study further demonstrated that FC measures can be designed to maintain good construct validity. Nevertheless, readers may still be legitimately concerned about the utility of FC measures, given their relatively lower reliability estimates compared to their SS counterparts. The reliability discrepancy between the FC and the SS measures may originate from two sources. First, the FC responses are

dichotomous in nature because respondents are only allowed to choose A or B. In comparison, the SS format allows respondents to indicate their degree of agreement. When holding other factors constant, dichotomous responses provide less information than graded ones, resulting in lower reliability of FC measures (Brown & Maydeu-Olivares, 2018). Fortunately, we can easily add a few more hard-to-fake desirability-matched blocks to FC2/FC3 to make trait scores derived from them as reliable as those from the SS format while maintaining their faking-resistance. However, it is much harder (if possible) to make the SS format as faking-resistant as the FC format. Second, it is well-known that the SS format is susceptible to various response biases, such as acquiescent, midpoint, and extreme response styles (Li et al., 2021; Plieninger & Heck, 2018; Sun et al., 2019; Sun et al., 2022). These systematic but construct-irrelevant biases can inflate reliability estimates. In the Online Supplementary Materials (Tables S8-S12 and Figures S5-S7), we presented additional analysis results where we corrected the SS scores for three common response biases (acquiescence, extreme responding, and midpoint responding) using the method developed by Plieninger and Heck (2018). It turned out that, after correction, the average reliability of the SS scores dropped from .84 to .70, which was very similar to that of FC. Taken together, these additional results suggested that the higher reliability estimates of SS were inflated, at the very least to some extent, by response biases, and that the FC format can mitigate these issues and provide more realistic estimates when carefully designed.

Another important finding is that FC4 demonstrated even greater susceptibility to faking compared to SS. Many papers discuss the FC format as more faking-resistant than the SS format

without properly noting that they have to be thoughtfully designed to be so. Our finding highlighted that the FC format is NOT a panacea for preventing faking. When desirable and undesirable statements are contrasted with each other within the same block, the social desirability difference among statements may become even more salient than when they are presented separately, thus making such blocks more susceptible to faking than their constituting statements (McCloy et al., 2005). If there are a substantial number of such blocks, FC measures can be even more fakable than their SS counterparts. Therefore, we urge users interested in using the FC format to counteract faking to be aware of this issue.

Recommended Steps to Develop FC Measures

Despite all the promises of the FC format, many people still find it difficult to develop a good FC measure due to the lack of guidelines. To promote a wider adoption of the FC format in organizational research and practices, below we provide a step-by-step guideline on how to develop high-quality FC measures based on our research findings and first-hand experience. These recommendations are intended as tentative guidelines that should be updated with more empirical evidence in the future rather than a gold standard.

Step 1. Generate a sufficient pool of high-quality statements for focal traits and obtain statement parameters. Several excellent guidelines have provided detailed discussions on how to write and select high-quality statements (Cao et al., 2015; Clark & Watson, 1995; 2019; Hinkin, 1998; Lambert & Newman, 2022; Worthington & Whittaker, 2006). Readers are encouraged to refer to them for more details. Here we want to emphasize the following considerations in the

context of FC measure development. First, we should avoid the use of extremely worded statements (e.g., “I have never complained about anything”). Avoiding extreme wording can substantially lower the risk of the statement being too socially (un)desirable and hence too difficult to be matched with other statements. Second, it is important to keep a small proportion of negatively keyed statements (2~4 per trait) because we need them for mixed keyed blocks. Third, it is strongly recommended to keep more statements per trait than needed for a target FC measure as this can greatly ease the pairing in subsequent steps. In this step, researchers can also obtain statement parameters that will be used in the following steps. Specifically, if the test-developer adopts a dominance-response-process-based approach (i.e., dominance model), it is recommended to fit a correlated-factor-analysis model to responses to the single statements and record the standardized factor loadings, statement intercepts, variance of statement uniqueness, and latent correlations among traits. If they adopt an unfolding-model-based approach, it is recommended to fit a Multidimensional Generalized Graded Unfolding Model (Tu et al., 2021; Tu et al., 2023; Wang & Wu, 2016) to the dichotomized responses and record statement discrimination, location, and threshold parameters, and latent correlations among traits.

Step 2. Obtain social desirability estimates of statements. There are three approaches to obtaining social desirability estimates for statements developed in Step 1. The first approach is direct rating where a small group of subject matter experts provide their direct ratings of the social desirability of each statement on a Likert scale (e.g., 1 = Very undesirable, 5 = Very desirable; see examples from Vasilopoulos et al., 2006, and Wetzel et al., 2021). Subject matter experts can be

asked to rate the general and/or job-specific social desirability of each statement, depending on the intended use of the measure: If the measure is designed for use in specific jobs or organizations, then job-specific social desirability can be more appropriate; if the measure is intended for selection across jobs/organizations, then general social desirability is preferred. The second approach is to ask respondents to respond to these statements as if they were ideal job candidates (Naemi et al., 2014; Stark et al., 2005). These *fake-good* responses can also be used to operationalize the social desirability of statements. Recently, Hommel (2023) demonstrated that natural language processing techniques can also be used to predict statement social desirability with high accuracy. As of now, we recommend the direct rating approach because it is the most straightforward operationalization of social desirability. *Fake-good* responses may be contaminated by other irrelevant factors such as faking motivation. The natural language processing approach is promising but ignores individual differences in the perception of statement social desirability. Further, we recommend researchers to (1) use at least 30 participants for more reliable estimates of social desirability, (2) ensure each trait has statements spanning a similar range of social desirability levels, and (3) examine inter-rater agreement and prioritize statements whose social desirability was agreed upon by most raters.

Step 3. Determine block size. One of the most important decisions when developing FC measures is block size, which could range from 2 to the total number of statements (full ranking task, which is impractical with any substantial number of statements). When making this decision, researchers need to consider psychometric properties and respondents' cognitive load. Larger

block sizes should demonstrate superior psychometric properties because they produce more pairwise comparisons, but may also impose heavier cognitive load on respondents, potentially leading to compromised respondent reactions and data quality, thereby jeopardizing psychometric properties (Brown & Maydeu-Olivares, 2011). Surprisingly, very few studies have systematically examined the impact of block size on psychometric properties of (but see Frick et al., 2023 for an exception) and respondent reactions to FC measures. Drawing from our own results obtained from three samples with > 4,500 respondents in another ongoing project (results available upon request as we are still writing this manuscript), we found minor differences (Cohen's *d*s = -.16 ~ .18) on perceived difficulty, exhaustion and cognitive load between FC measures with block sizes of three and five when holding statements constant. As such, block sizes ranging from 3 to 5 can all be considered as reasonable for static FC measures (all respondents received identical blocks) because they strike a good balance between psychometric information and respondent reactions. Five is also the up-to-date estimate of the upper limit of working memory capacity for meaningful chunks for adults (Cowan, 2010; Halford et al., 2007). A block size of 2 is recommended for computerized adaptive tests because it is much easier to implement (Stark et al., 2012), but not for static FC measures because it is not very psychometrically efficient. If researchers have specific reasons to maintain a block size of 2, we recommend using the graded FC format. This format allows respondents to indicate their degree of preference, thereby providing more psychometric information (Brown & Maydeu-Olivares, 2018; Zhang, Luo et al., 2023; Zhang, Tu et al., 2023) and fostering more positive respondent reactions (Dalal et al., 2021). It is also recommended that

for multidimensional FC measures, block size should not exceed the number of measured latent traits because we generally want to avoid having more than one statement of the same latent trait in the same block.

Step 4. Determine the number of mixed blocks. After obtaining social desirability and deciding on block size, researchers need to decide on the number of mixed blocks. Previous simulations demonstrated that 20-30% of mixed blocks in a triplet format were sufficient for maintaining satisfactory reliability of trait scores (Lee et al., 2022). Our empirical findings further confirmed that this setting can also maintain sufficient faking-resistance. However, it should be noted that it is hard to recommend an absolute number that universally applies to all FC measures because it depends on block size and the number of latent traits being measured. We also note that what matters for psychometric properties is the number of mixed pairs (recoded pairwise comparisons) and what matters for fakability is the proportion of mixed blocks. Our findings suggest that 6 mixed triplets (6÷20 = 30% mixed blocks) and 14 matched triplets, corresponding to 12 mixed pairs (each mixed triplet has two mixed pairs and one matched pair) and 48 matched pairs (14×3 = 42 matched pairs from matched triplets, and 6×1 = 6 from mixed triplets) when recoded into pairwise comparisons items, suffice for measuring 6 traits. It means that 30% or fewer mixed blocks and 2 mixed pairs per trait without duplication would be a reasonable recommendation. Surely, more matched pairs will be even better as they provide more information without impacting fakability. Let's say three researchers want to measure 12 traits using FC measures, they need to have 24 mixed pairs regardless of the block size. If researcher A plans to

use block size of 3, there should be 12 mixed triplets (24 mixed pairs÷2 mixed pairs per mixed triplet) and 28 ([12 mixed triplets÷30%]×70%) or more matched triplets (28×3 + 12 = 96 matched pairs or more); if researcher B wants to use a block size of 4 and they design the mixed blocks as containing 2 positively keyed statements + 2 negatively keyed statements (4 mixed pairs and 2 matched pairs per mixed block), they need to have 6 (24÷4) mixed quadruplets and 14 ([6÷30%] ×70%) or more matched quadruplets (6×2 + 14×6 = 96 matched pairs or more); if researcher C uses a block size of 5 and they design the mixed blocks as containing 2(3) positively keyed statements + 3(2) negatively keyed statements (6 mixed pairs and 4 matched pairs per mixed block), they need to have 4 (24÷6) mixed quintets and 10 ([4÷30%]×70%, rounded up) or more matched quadruplets (4×4 + 10×10 = 116 matched pairs or more). Furthermore, we recommend that mixed-keyed triplets be composed of 2(1) positively and 1(2) negatively keyed statements, mixed-keyed quadruplets be composed of 2 positively and 2 negatively keyed statements, and mixed-keyed quintets be composed of 3(2) positively and 2(3) negatively keyed statements. These designs allow the maximum number of mixed pairs to appear.

Step 5. Create blocks. While mixed blocks almost inevitably involve bundling desirable and undesirable statements, researchers can still try *some degree of matching* by putting moderately desirable and moderately undesirable statements together instead of putting very desirable and very undesirable statements together. Therefore, we recommend users to construct mixed blocks first so that they have the largest statement pool to choose from. For any FC measures, researchers should try to ensure that (1) statements within the same block measure different latent

traits, (2) each trait should be paired with all other traits for about an equal number of times, (3) each trait should also be involved in at least one mixed pair, and (4) statements in the same block should be matched on social desirability as much as they can. Given all these constraints, it becomes challenging to create optimal blocks manually. Therefore, in the Online Supplementary Materials, we provided a tutorial on how to use the *autoFC* R package (Li et al., 2022) to automatically assemble blocks according to multiple criteria.

Before moving to the next step, we consider two additional issues deserving further attention. The first issue concerns how to use social desirability value for matching. The most popular way is to focus on the mean value for each statement across raters and try to minimize the absolute difference between statements' mean desirability values (the D index; Edwards, 1957; Pavlov, 2022). Statements are said to be matched if the largest D between all possible statement pairs within a block is smaller than a predefined cutoff. We recommend setting the cutoff to be .50 for a 5-point scale based on previous studies (e.g., Vasilopoulos et al. [2006] used .357; Chernyshenko et al. [2009], Drasgow et al. [2012], and Hughes et al. [2021] used .714) and our first-hand experience. For mixed blocks, the cutoff should be relaxed, though less evidence exists on what cutoff should be set. Based on our experience with FC questionnaire construction, 1 to 1.5 on a 5-point scale seem to be a reasonable cutoff for mixed blocks. One potential issue of using mean desirability values is that the variance of social desirability values across raters is ignored. To overcome this issue, Pavlov et al. (2022) proposed the inter-item agreement (IIA) approach, which essentially utilized robust interrater agreement indices, such as Brennan-Prediger index

(Brennan & Prediger, 1981; Gwet, 2014) and AC index (Gwet, 2008, 2014). Statement pairing, in turn, is based on the interrater agreement on social desirability values, rather than differences in mean social desirability values. We believe that the IIA approach is promising for statement matching. Readers interested in this approach can use the *autoFC* R package to execute it.

The second issue concerns the contextual nature of social desirability. While it is common to match statements based on social desirability ratings obtained from SS administration, this practice implicitly assumes that respondents' perception of statement social desirability remains constant when these statements are administered individually versus in pair with other statements (Frick, 2022). However, a statement may become more or less desirable depending on the statements it is paired with (Lin & Brown, 2017). Even two statements with identical social desirability ratings when presented individually can still be perceived as differentially (un)desirable when paired together. As such, after constructing preliminary blocks, test-developers can invite human raters to rate the desirability of multiple statements presented simultaneously in a block. Blocks that may need further revision can be identified by checking the D index computed from the social desirability ratings obtained from block administration. If the D index exceeds the cutoff suggested above, researchers should repeat this process until they find blocks that satisfy the criterion. While the present study focused on dominance-model-based FC measure, the ideal-point-model-based FC measure is also widely used (Drasgow et al., 2012; Boyce et al., 2015). Under the unfolding framework, test developers should match statements on both social desirability and extremity to ensure faking-resistance (Cao & Drasgow, 2019).

Step 6. Examine the reliability of the FC measure using simulated data. It is extremely helpful to have an initial understanding of the reliability of trait scores derived from the FC measure constructed in the previous step under ideal conditions using Monte Carlo simulations. If the reliability does not fare well in these ideal conditions, it is unlikely to be satisfactory in more realistic conditions. In those cases, researchers should go back to the previous step to construct a new FC measure and examine its reliability in ideal conditions again before collecting empirical data. To examine reliability using simulated data, researchers need to simulate FC responses based on the statement parameters obtained in Step 1, assuming that statement parameters are largely invariant across FC and SS (Lin & Brown, 2017; Morillo et al., 2019). Specifically, if test developers adopt the dominance model, FC responses should be generated according to Equation 4 in Brown & Maydeu-Olivares (2013); if the unfolding model is adopted, FC responses can be generated according to Equation 2 in Lee et al., (2019) or Equation 13/14 in Zhang, Tu et al., (2023). Finally, researchers could fit either a dominance (e.g., TIRT) or an unfolding (e.g., GGUM-RANK or GTUM) model to the simulated data depending on the data generation model used, and obtain reliability estimates accordingly.

A natural follow-up question is how to calculate reliability when assembling a new FC measure from calibrated statement banks and empirical data is not yet available. If the FC measure is intended to screen in/out respondents within a certain range of the latent trait continuum, we recommend direct examination of standard errors of measurement within the range of interest because this is the most straightforward way to quantify measurement precision. If the measure is

designed for general purpose and an overall estimate of reliability is needed, following Lin (2022), we recommended test-developers to use the squared correlation between estimated person scores and true person scores as the reliability estimate, because this index relies on the least assumptions and is a straightforward operationalization of reliability under Classical Testing Theory. However, this squared correlation is also the most conservative estimate (Lin, 2022). Thus, we additionally recommend the empirical reliability estimate using the formula $\frac{\text{var}(\hat{\theta})}{\text{var}(\hat{\theta})+\text{mean}(\text{SE}(\hat{\theta})^2)}$ suggested by Brown and Maydeu-Olivares (2018). In this formula, $\text{var}(\hat{\theta})$ refers to the variance of estimated latent trait scores, and $\text{SE}(\hat{\theta})^2$ refers to the squared value of standard error of measurement. But we note that since empirical reliability is likely to be slightly inflated (Lin, 2022), simultaneous consideration of the two reliability indices is recommended. Given the complexity of the TIRT model and simulations in general, we provided a step-by-step tutorial that automates the entire process, implemented using the *autoFC* R package, in the Online Supplementary Materials (Section 1). Users just need to input population parameters mentioned above. The R functions will run the simulation and summarize simulation results automatically. Currently, *autoFC* only covers TIRT-based-dominance models. Other models will be included in future updates.

Step 7. Empirical validation. If the FC measure performs satisfactorily in simulations, researchers can then proceed to empirically test its psychometric properties, fakability and respondent reactions. We believe our study provided a good example of empirical validation design that interested readers can adopt for their own studies. Specifically, we recommend a within-subjects design to contrast honest vs. motivated faking situations to comprehensively estimate the

fakability of an FC measure. Researchers are also recommended to include the SS counterpart as a benchmark to gauge the degree of possible benefits (in terms of psychometric properties and faking resistance) and costs (in terms of respondent reactions) brought by the FC. At this stage, we recommend using empirical reliability or test-retest reliability to quantify measurement precision, as they reflect the measurement accuracy in actual rather than hypothetical samples. We additionally recommend test developers to regularly examine measurement invariance across demographic groups (e.g., gender or racial groups) to identify potentially non-invariant blocks and ensure score comparability across groups. Several techniques for assessing measurement invariance of FC measures have been proposed in recent years (Lee & Smith, 2019; Lee et al., 2021; Qiu & Wang, 2020) and we recommend readers to refer to these approaches.

Limitations and Future Directions

Despite its many strengths (e.g., large sample, experimental design, comprehensiveness, guidelines, and tutorial), the present study is still limited in the following ways. First, we only used FC measures with a block size of three, which provided less information compared to larger block sizes. Future researchers are strongly encouraged to examine whether block size will moderate the effects of mixed-keying and social desirability on psychometric properties, faking resistance, and respondent reactions. Second, although the current study demonstrated negligible impact of FC designs on respondent reactions, it remains possible that respondent reactions are dependent on the construct being measured as well. For example, if respondents are required to choose between statements measuring the dark personality traits (A = “I manipulate people to get what I want”, B

= “I deserve more attention than others”, C = “I enjoy quick and nasty revenge”) that may threaten their self-images, they may have more salient negative reactions (Fuechtenhans & Brown, 2022). In such cases, the design of FC measures may become more relevant. Hence, we believe that a promising future research direction is to examine the effect of construct types on respondent reactions, and whether the impact of FC design on respondent reactions depends on these construct types. Third, we exclusively used self-reported data for measuring both personality and criterion variables. It would be interesting for future studies to explore whether the criterion-related validity of different FC measures would vary in the same manner when predicting other-reported outcomes.

Conclusion

We presented the first piece of comprehensive empirical evidence on the impact of social desirability matching and mixed keying on the psychometric properties of, fakability of, and respondent reactions to FC measures. Most notably, a small compromise on desirability matching in exchange for more mixed keyed blocks is feasible, such that the improvement in psychometric properties does not substantially harm the faking resistance of an FC measure. Also, respondents did not report differential reactions toward different FC designs. All in all, we showed that it is possible to find a middle ground between social desirability matching and mixed keying such that the FC measures can have both good psychometric properties and high faking resistance. We further provided researchers tools for constructing such FC measures.

References

Alarcon, G., Eschleman, K. J., & Bowling, N. A. (2009). Relationships between personality variables and burnout: A meta-analysis. *Work & Stress*, *23*(3), 244-263. https://doi.org/10.1080/02678370903282600

Anglim, J., Horwood, S., Smillie, L. D., Marrero, R. J., & Wood, J. K. (2020). Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychological Bulletin*, *146*(4), 279-323. https://doi.org/10.1037/bul0000226

Anglim, J., Morse, G., De Vries, R. E., MacCann, C., & Marty, A. (2017). Comparing job applicants to non–applicants using an item–level bifactor model on the HEXACO personality inventory. *European Journal of Personality*, *31*(6), 669-684. https://doi.org/10.1002/per.2120

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*(4), 340-345. https://doi.org/10.1080/00223890902935878

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

Boyce, A. S., Conway, J. S., & Caputo, P. M. (2015). *ADEPT-15 technical documentation: Development and validation of Aon Hewitt's Personality Model and Adaptive Employee Personality Test (ADEPT-15)*. Aon Hewitt.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687-699. https://doi.org/10.1177/001316448104100307

Brown, A. & Bartram, D. (2009). Doing less but getting more: Improving forced-choice measures with IRT. In: *Society for Industrial and Organizational Psychology Conference*, 2-4 April 2009, New Orleans. Retrieved from http://kar.kent.ac.uk/44788/

Brown, A. & Maydeu-Olivares, A. (2018). Ordinal Factor Analysis of Graded-Preference Questionnaire Data. *Structural Equation Modeling: A Multidisciplinary Journal, 25(4)*, 516-529. https://doi.org/10.1080/10705511.2017.1392247

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*(1), 135-160. https://doi.org/10.1007/s11336-014-9434-9

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

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

Brown, A., & Maydeu-Olivares, A. (2018). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4),

516-529. https://doi.org/10.1080/10705511.2017.1392247

Bürkner, P. C. (2019). thurstonianIRT: Thurstonian IRT models in R. *Journal of Open Source Software*, *4*(42), 1662-1663. https://doi.org/10.21105/joss.01662

Bürkner, P. C. (2022). On the information obtainable from comparative judgments. *Psychometrika*, *87*, 1439-1472. https://doi.org/10.1007/s11336-022-09843-z

Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, *79*(5), 827-854. https://doi.org/10.1177/0013164419832063

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33-57. https://doi.org/10.1007/s11336-009-9136-x

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, *104*(11), 1347-1368. https://doi.org/10.1037/apl0000414

Cao, M., Drasgow, F., & Cho, S. (2015). Developing ideal intermediate personality items for the ideal point model. *Organizational Research Methods*, *18*(2), 252-275. https://doi.org/10.1177/1094428114555993

Carlo, G., Okun, M. A., Knight, G. P., & de Guzman, M. R. T. (2005). The interplay of traits and motives on volunteering: Agreeableness, extraversion and prosocial value motivation. *Personality and Individual Differences*, *38*(6), 1293-1305. https://doi.org/10.1016/j.paid.2004.08.012

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1-29. https://doi.org/10.186D37/jss.v048.i06

Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology*, *83*(3), 471-485. https://doi.org/10.1037/0021-9010.83.3.471

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. https://doi.org/10.1080/08959280902743303

Chiaburu, D. S., Oh, I.-S., Berry, C. M., Li, N., & Gardner, R. G. (2011). The five-factor model of personality traits and organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 96*(6), 1140–1166. https://doi.org/10.1037/a0024004

Clark, L. A., & Watson D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309-319. https://doi.org/10.1037/1040-3590.7.3.309

Clark, L. A., & Watson D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412-1427. https://doi.org/10.1037/pas0000626

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment*, *16*(2), 155-169. https://doi.org/10.1111/j.1468-2389.2008.00420.x

Converse, P. D., Pathak, J., Quist, J., Merbedone, M., Gotlib, T., & Kostic, E. (2010). Statement desirability ratings in forced-choice personality measure development: Implications for reducing score inflation and providing trait-level information. *Human Performance*, *23*(4), 323-342. https://doi.org/10.1080/08959285.2010.501047

Cottrell, J. M., Newman, D. A., & Roisman, G. I. (2015). Explaining the black-white gap in cognitive test scores: Toward a theory of adverse impact. *Journal of Applied Psychology*, *100*(6), 1713-1736. https://doi.org/10.1037/apl0000020

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why?. *Current Directions in Psychological Science*, *19*(1), 51-57. https://doi.org/10.1177/0963721409359277

Dalal, D. K., Zhu, X. S., Rangel, B., Boyce, A. S., & Lobene, E. (2021). Improving applicant reactions to forced-choice personality measurement: Interventions to reduce threats to test takers' self-concepts. *Journal of Business and Psychology*, *36*(1), 55-70. https://doi.org/10.1007/s10869-019-09655-6

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. https://doi.org/10.1207/s15327752jpa4901_13

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army personnel selection and classification decisions*. Drasgow Consulting Group Urbana IL.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.

Fox, S., Spector, P. E., Bruursema, K., Kessler, S., & Goh, A. (2007). Necessity is the mother of behavior: Organizational constraints, CWB and OCB. In *Meeting of the Academy of Management, Philadelphia, PA*.

Frick, S. (2022). Modeling faking in the multidimensional forced-choice format: the faking mixture model. *Psychometrika*, *87*(2), 773-794. https://doi.org/10.1007/s11336-021-09818-6

Frick, S., Brown, A., & Wetzel, E. (2023). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research*, *58*(1), 1-29. https://doi.org/10.1080/00273171.2021.1938960

Fuechtenhans, M., & Brown, A. (2022). How do applicants fake? A response process model of faking on multidimensional forced-choice personality assessments. *International Journal of Selection and Assessment*. Advanced online publication. https://doi.org/10.1111/ijsa.12409

Furr, R. M. (2010). The double-entry intraclass correlation as an index of profile similarity: Meaning, limitations, and alternatives. *Journal of Personality Assessment*, *92*(1), 1-15. https://doi.org/10.1080/00223890903379134

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29-48. https://doi.org/10.1348/000711006X126600

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences*, *11*(6), 236-242. https://doi.org/10.1016/j.tics.2007.04.001

Harms, P. D., Roberts, B. W., & Wood, D. (2007). Who shall lead? An integrative personality approach to the study of the antecedents of status in informal social organizations. *Journal of Research in Personality*, *41*(3), 689-699. https://doi.org/10.1016/j.jrp.2006.08.001

Harris, A. M., McMillan, J. T., & Carter, N. T. (2021). Test-taker reactions to ideal point measures of personality. *Journal of Business and Psychology*, *36*(3), 513-532. https://doi.org/10.1007/s10869-020-09682-8

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel* P*sychology*, *57*(3), 639-683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

He, Y., Donnellan, M. B., & Mendoza, A. M. (2019). Five-factor personality domains and job performance: A second order meta-analysis. *Journal of Research in Personality*, *82*, 1-24. https://doi.org/10.1016/j.jrp.2019.103848

Highhouse, S., Lievens, F., & Sinar, E. F. (2003). Measuring attraction to organizations. *Educational and Psychological Measurement*, *63*(6), 986-1001. https://doi.org/10.1177/0013164403258403

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*(1), 104-121. https://doi.org/10.1177/109442819800100106

Hu, J., & Connelly, B. S. (2021). Faking by actual applicants on personality tests: A meta-analysis of within-subjects studies. *International Journal of Selection and Assessment*, *29*(3-4), 412-426. https://doi.org/10.1111/ijsa.12338

Hughes, A. W., Dunlop, P. D., Holtrop, D., & Wee, S. (2021). Spotting the “ideal” personality response: Effects of item matching in forced choice measures for personnel selection. *Journal of Personnel Psychology*, *20*(1), 17-26. https://doi.org/10.1027/1866-5888/a000267

Inceoglu, I., & Lin, Y. (2017). Preventing Rater Biases in 360-Degree Feedback by Forcing Choice. *Organizational Research Methods, 20(1),* 121-148. https://doi.org/10.1177/1094428116668036

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: a concise measure of the dark triad. *Psychological Assessment*, *22*(2), 420. https://doi.org/10.1037/a0019265

Jones, K. S., Newman, D. A., Su, R., & Rounds, J. (2022). Vocational interests and adverse impact: How attraction and selection on vocational interests relate to adverse impact potential. *Journal of Applied Psychology*, *107*(4), 604-627. https://doi.org/10.1037/apl0000893

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87*(3), 530–541. https://doi.org/10.1037/0021-9010.87.3.530

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology*, *98*(6), 875. https://doi.org/10.1037/a0033901

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*(5), 399-419. https://doi.org/10.1177/0146621612446170

Kluger, A. N., & Rothstein, H. R. (1993). The influence of selection test type on applicant reactions to employment testing. *Journal of Business and Psychology*, *8*, 3-25. https://doi.org/10.1007/BF02230391

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of Likert items. *Frontiers in Psychology*, *10*, 2309. https://doi.org/10.3389/fpsyg.2019.02309

Kristensen, T. S., Borritz, M., Villadsen, E., & Christensen, K. B. (2005). The Copenhagen Burnout Inventory: A new tool for the assessment of burnout. *Work & Stress*, *19*(3), 192-207. https://doi.org/10.1080/02678370500297720

Lambert, L. S., & Newman, D. A. (2022). Construct development and validation in three practical steps: Recommendations for reviewers, editors, and authors. *Organizational Research Methods*. Advanced online publication. https://doi.org/10.1177/10944281221115374

Lee, H., & Smith, W. Z. (2020). Fit indices for measurement invariance tests in the Thurstonian IRT model. *Applied Psychological Measurement*, *44*(4), 282-295. https://doi.org/10.1177/0146621619893785

Lee, P., Joo, S. H., & Lee, S. (2019). Examining stability of personality profile solutions between Likert-type and multidimensional forced choice measure. *Personality and Individual Differences*, *142*, 13-20. https://doi.org/10.1016/j.paid.2019.01.022

Lee, P., Joo, S. H., & Stark, S. (2021). Detecting DIF in multidimensional forced-choice measures using the Thurstonian item response theory model. *Organizational Research Methods*, *24*(4), 739-771. https://doi.org/10.1177/1094428120959822

Lee, P., Joo, S. H., Stark, S., & Chernyshenko, O. S. (2019). GGUM-RANK statement and

person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement*, *43*(3), 226-240. https://doi.org/10.1177/0146621618768294

Lee, P., Joo, S. H., Zhou, S., & Son, M. (2022). Investigating the impact of negatively keyed statements on multidimensional forced-choice personality measures: A comparison of partially ipsative and IRT scoring methods. *Personality and Individual Differences*, *191*, 111555. https://doi.org/10.1016/j.paid.2022.111555

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. https://doi.org/10.1016/j.paid.2017.11.031

Lee, Y., Berry, C. M., & Gonzalez-Mulé, E. (2019). The importance of being humble: A meta-analysis and incremental validity analysis of the relationship between honesty-humility and job performance. *Journal of Applied Psychology*, *104*(12), 1535. https://doi.org/10.1037/apl0000421

Li, M., Sun, T., & Zhang, B. (2022). autoFC: An R package for automatic item pairing in forced-choice test construction. *Applied Psychological Measurement*, *46*(1), 70-72. https://doi.org/10.1177/01466216211051726

Li, Z., Zhang, B., Cao, M., & Tay, L. (2021). Accounting for item response process and response styles using the Unfolding Item Response Tree (UIRTree) model. Preprint. https://doi. org/10.31219/osf. io/8w36e.

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, *77*(3), 389-414. https://doi.org/10.1177/0013164416646162

Lopez, F. J., Hou, N., & Fan, J. (2019). Reducing faking on personality tests: Testing a new faking-mitigation procedure in a US job applicant sample. *International Journal of Selection and Assessment*, *27*(4), 371-380. https://doi.org/10.1111/ijsa.12265

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, *47*(4), 715-738. https://doi.org/10.1111/j.1744-6570.1994.tb01573.x

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., & Ahmed, S. M. (2017). Applicant perspectives during selection: A review addressing "So what?,""What's new?," and "Where to next?". *Journal of Management*, *43*(6), 1693-1725. https://doi.org/10.1177/0149206316681846

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. https://doi.org/10.1177/1094428105275374

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsoda, V. (2019). The journey from Likert to forced-choice questionnaires: Evidence of the invariance of item parameters. *Revista de Psicología del Trabajo y de las Organizaciones*, *35*(2), 75-83. https://doi.org/10.5093/jwop2019a11

Morillo, D., Ponsoda, V., Leenen, I., Abad, F. J., & Hontangas, P., (2016). *Comparing CFA and Bayesian estimations of forced-choice questionnaires with paired dominance items*. International Test Commission (ITC) 2016 Conference, Vancouver, Canada.

Munyon, T. P., Carnes, A. M., Lyons, L. M., & Zettler, I. (2020). All about the money? Exploring antecedents and consequences for a brief measure of perceived financial security. *Journal of Occupational Health Psychology*, *25*(3), 159-175. https://doi.org/10.1037/ocp0000162

Naemi, B., Seybert, J., Robbins, S., & Kyllonen, P. (2014). Examining the WorkFORCE™ assessment for job fit and core capabilities of FACETS™. *ETS Research Report Series*, *2014*(2), 1-43. https://doi.org/10.1002/ets2.12040

Ng, V., Lee, P., Ho, M. H. R., Kuykendall, L., Stark, S., & Tay, L. (2021). The development and validation of a multidimensional forced-choice format character measure: Testing the Thurstonian IRT approach. *Journal of Personality Assessment*, *103*(2), 224-237. https://doi.org/10.1080/00223891.2020.1739056

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science*, *7*(4), 384-403. https://doi.org/10.1177/1745691612449021

Pavlov, G. (2022). Comparing different approaches for obtaining item desirability ratings [Poster]. Society for Industrial and Organizational Psychology Annual Conference, Seattle, WA, United States.

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2022). Item desirability matching in forced-choice test construction. *Personality and Individual Differences*, *183*, 111114. https://doi.org/10.1016/j.paid.2021.111114

Pletzer, J. L., Oostrom, J. K., Bentvelzen, M., & de Vries, R. E. (2020). Comparing domain-and facet-level relations of the HEXACO personality model with workplace deviance: A meta-analysis. *Personality and Individual Differences*, *152*, 109539. https://doi.org/10.1016/j.paid.2019.109539

Pletzer, J. L., Oostrom, J. K., & de Vries, R. E. (2021). HEXACO personality and organizational citizenship behavior: A domain-and facet-level meta-analysis. *Human Performance*, *34*(2), 126-147. https://doi.org/10.1080/08959285.2021.1891072

Pletzer, J. L., Thielmann, I., & Zettler, I. (2023). Who is healthier? A meta-analysis of the relations between the HEXACO personality domains and health outcomes. *European Journal of Personality*. https://doi.org/10.1177/08902070231174574

Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, *53*(5), 633-654. https://doi.org/10.1080/00273171.2018.1469966

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. https://doi.org/10.1037/0021-9010.88.5.879

Qiu, X. L., & Wang, W. C. (2021). Assessment of differential statement functioning in ipsative

tests with multidimensional forced-choice items. *Applied Psychological Measurement*, *45*(2), 79-94. https://doi.org/10.1177/0146621620965739

Roodt, G. (2004). Turnover intentions. Unpublished document: University of Johannesburg Johannesburg, South Africa.

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management*, *26*(3), 565-606. https://doi.org/10.1016/S0149-2063(00)00041-6

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. https://doi.org/10.1111/joop.12098

Samejima, F. (1997). *Graded Response Model*. In: van der Linden, W.J., Hambleton, R.K. (eds) Handbook of Modern Item Response Theory. Springer, New York, NY. https://doi.org/10.1007/978-1-4757-2691-6_5

Sass, R., Frick, S., Reips, U. D., & Wetzel, E. (2020). Taking the test taker's perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. *Assessment*, *27*(3), 572-584. https://doi.org/10.1177/1073191118762049

Schat, A. C., Kelloway, E. K., & Desmarais, S. (2005). The Physical Health Questionnaire (PHQ): construct validation of a self-report scale of somatic symptoms. *Journal of Occupational Health Psychology*, *10*(4), 363-381. https://doi.org/10.1037/1076-8998.10.4.363

Schreiber, A., & Marcus, B. (2020). The place of the "Dark Triad" in general models of personality: Some meta-analytic clarification. *Psychological Bulletin*, *146*(11), 1021-1041. https://doi.org/10.1037/bul0000299

Schulte, N., Holling, H., & Bürkner, P. C. (2021). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats?. *Educational and Psychological Measurement*, *81*(2), 262-289. https://doi.org/10.1177/0013164420934861

Sisson, E. D. (1948). Forced choice—The new army rating. *Personnel Psychology*, *1*(3), 365-381. https://doi.org/10.1111/j.1744-6570.1948.tb01316.x

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. https://doi.org/10.1111/j.1744-6570.1993.tb00867.x

Spector, P. E. (1985). Measurement of human service staff satisfaction: Development of the Job Satisfaction Survey. *American Journal of Community Psychology*, *13*(6), 693-713. https://doi.org/10.1007/bf00929796

Spector, P. E., Fox, S., Penney, L. M., Bruursema, K., Goh, A., & Kessler, S. (2006). The dimensionality of counterproductivity: Are all counterproductive behaviors created equal?. *Journal of Vocational Behavior*, *68*(3), 446-460. https://doi.org/10.1016/j.jvb.2005.10.005

Speer, A. B., Perrotta, J., & Jacobs, R. R. (2023). Supervised Construct Scoring to Reduce Personality Assessment Length: A Field Study and Introduction to the Short 10. *Organizational Research Methods*. Advanced online publication. https://doi.org/10.1177/1094428122114569

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. https://doi.org/10.1177/0146621604273988

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. https://doi.org/10.1177/1094428112444611

Sun, T., Zhang, B., Cao, M., & Drasgow, F. (2022). Faking detection improved: Adopting a Likert item response process tree model. *Organizational Research Methods, 25*(3), 490-512. https://doi.org/10.1177/10944281211002904

Sun, T., Zhang, B., Phan, W. M. J., Drasgow, F., & Roberts, B. (2019, July). *"Meh!": Examining Midpoint Endorsement Habitude (MEH) in Survey Research*. In Academy of Management Proceedings (Vol. 2019, No. 1, p. 16421). Briarcliff Manor, NY 10510: Academy of Management. https://doi.org/10.5465/AMBPP.2019.227

Tay, L., & Ng, V. (2018). Ideal point modeling of non-cognitive constructs: Review and recommendations for research. *Frontiers in Psychology*, 2423. https://doi.org/10.3389/fpsyg.2018.02423

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*(2), 320-332. https://doi.org/10.1037/0021-9010.87.2.320

Tu, N., Zhang, B., Angrave, L., & Sun, T. (2021). *bmggum*: An R package for Bayesian estimation of the multidimensional generalized graded unfolding model with covariates. *Applied Psychological Measurement*, *45*(7-8), 553-555. https://doi.org/10.1177/01466216211040488

Tu, N., Zhang, B., Angrave, L., Sun, T., & Neuman, M. (2023). Estimating the Multidimensional Generalized Graded Unfolding Model with Covariates Using a Bayesian Approach. *Journal of Intelligence*, *11*(8), 163. https://doi.org/10.3390/jintelligence11080163

Van Iddekinge, C. H., Lievens, F., & Sackett, P. R. (2023) Personnel selection: A review of ways to maximize validity, diversity, and the applicant experience. *Personnel Psychology*. Advanced online publication. https://doi.org/10.1111/j.1468-2389.2008.00420.x

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability?. *Human Performance*, *19*(3), 175-199. https://doi.org/10.1207/s15327043hup1903_1

Walton, K. E., Cherkasova, L., & Roberts, R. D. (2020). On the validity of forced choice scores

derived from the Thurstonian item response theory model. *Assessment*, *27*(4), 706-718. https://doi.org/10.1177/1073191119843585

Wang, W. C., & Wu, S. L. (2016). Confirmatory multidimensional IRT unfolding models for graded-response items. *Applied Psychological Measurement*, *40*(1), 56-72. https://doi.org/10.1177/0146621615602855

Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment*, *32*(3), 239-253. https://doi.org/10.1037/pas0000781

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*, *33*(2), 156-170. https://doi.org/10.1037/pas0000971

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artifacts. *Advances in Methods and Practices in Psychological Science*, *3*(1), 94-123. https://doi.org/10.1177/2515245919885611

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management*, *17*(3), 601-617. https://doi.org/10.1177/014920639101700305

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806-838. https://doi.org/10.1177/0011000006288127

Xu, Y., Beller, A. H., Roberts, B. W., & Brown, J. R. (2015). Personality and young adult financial distress. *Journal of Economic Psychology*, *51*, 90-100. https://doi.org/10.1016/j.joep.2015.08.010

Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2020). The nomological net of the HEXACO model of personality: A large-scale meta-analytic investigation. *Perspectives on Psychological Science*, *15*(3), 723-760. https://doi.org/10.1177/1745691619895036

Zhang, B., Cao, M., Tay, L., Luo, J., & Drasgow, F. (2020). Examining the item response process to personality measures in high-stakes situations: Issues of measurement validity and predictive validity. *Personnel Psychology*, *73*(2), 305-332. https://doi.org/10.1111/peps.12353

Zhang, B., Li, Y. M., Li, J., Luo, J., Ye, Y., Yin, L., Chen, Z., Soto, C. J., & John, O. P. (2022). The Big Five Inventory–2 in China: A comprehensive psychometric evaluation in four diverse samples. *Assessment*, *29*(6), 1262–1284. https://doi.org/10.1177/10731911211008245

Zhang, B., Luo, J., & Li, J. (2023). Moving beyond Likert and traditional forced-choice scales: A comprehensive investigation of the graded forced-choice format. *Multivariate Behavioral Research.* https://doi.org/10.1080/00273171.2023.2235682

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-

statement measures. *Organizational Research Methods*, *23*(3), 569-590. https://doi.org/10.1177/1094428119836486

Zhang, B., Tu, N., Angrave, L., Zhang, S., Sun, T., Tay, L., & Li, J. (2023). The Generalized Thurstonian Unfolding Model (GTUM): Advancing the modeling of forced-choice data. *Organizational Research Methods*. https://doi.org/10.1177/109442812312104

Zimmerman, R. D. (2008). Understanding the impact of personality traits on individuals' turnover decisions: A meta-analytic path model. *Personnel Psychology*, *61*(2), 309-348. https://doi.org/10.1111/j.1744-6570.2008.00115.x

LIST OF TABLES

Table 1. *Details about criterion measures and respondent reactions*

Criterion	Example Item	Rating Scale	α	Theoretical HEXACO Correlates
OCB (10) (Fox et al., 2007)	Helped new employees get oriented to the job.	1 = Never 5 = Every day	.83-.86	H(+), E(/), X(+), A(+), C(+), O(+) (Pletzer et al., 2021; Table 2)
CWB (10) (Spector et al., 2006)	Ignored someone at work.	1 = Never 5 = Every day	.80-.85	H(-), E(-), X(/), A(-), C(-), O(/) (Pletzer et al., 2020; Table 1)
JP (7) (Williams & Anderson, 1991)	I adequately complete assigned duties.	1 = Strongly disagree 5 = Strongly agree	.79-.82	H(+) (Lee, Berry, & Gonzalez-Mulé, 2019; Table 4) E(/), X(/), A(/), C(+), O(+) (Zettler et al., 2020; Table 9)
JS (9) (Spector, 1985)	All in all, how satisfied are you with the pay of your job?	1 = Very dissatisfied 5 = Very satisfied	.85-.88	H(+), E(-), A(+), C(-) (Pletzer et al., 2023; Supplementary Materials Table 9), X(+), O(/) (Judge et al., 2002; Table 1)
BNT (6) (Kristensen et al., 2005)	How often do you feel tired?	1 = Never 5 = Always	.88-.90	H(-), E(+), X(-), A(-), C(-), O(/) (Pletzer et al., 2023; Supplementary Materials Table 9)
TI (6) (Roodt, 2004)	How often have you considered leaving your job?	1 = Always 5 = Never	.77-.82	E(+), X(-), A(-), C(-), O(/) (Zimmerman, 2008; Table 3)
ORG (5) (self-made)	In your current organization, do you have the right to hire people?	1 = Yes, 0 = No	.74-.80	*E(/)*, *X(+)*, *A(/)*, *C(+)*, *O(+)* (Harms et al., 2007; Table 1)
CHAR (3) (self-made)	In the past year, have you volunteered?	1 = Yes, 0 = No	.29-.42	*E(/), X(+), A(+), C(+), O(/)* (Carlo et al., 2005; Table 1)
SWB (5) (Diener et al., 1985)	In most ways my life is close to my ideal.	1 = Strongly disagree 5 = Strongly agree	.87-.90	H(+), E(-), X(+), A(+), C(+), O(+) (Anglim et al., 2020; Table 7)
FS (5) (Munyon et al., 2020)	I have adequate income.	1 = Strongly disagree 5 = Strongly agree	.84-.88	*E(-)* (Munyon et al., 2020; Table 5, 10)
PHQ (14) (Schat et al., 2005)	How often have you experienced headaches?	1 = Not at all 7 = All the time	.85-.87	H(/), E(-), X(/), A(/), C(+), O(/) (Pletzer et al., 2023; Table 3)
NARC (4) (Jonason & Webster, 2010)	I tend to want others to admire me.	1 = Strongly disagree 5 = Strongly agree	.75-.80	H(-), E(-), X(+), A(-), C(/), O(/) (Schreiber & Marcus, 2020; Table 1)
MACH (4) (Jonason & Webster, 2010)	I tend to manipulate others to get my way.	1 = Strongly disagree 5 = Strongly agree	.79-.82	H(-), E(-), X(/), A(-), C(-), O(/) (Schreiber & Marcus, 2020; Table 1)

PSYCH (4) (Jonason & Webster, 2010)	I tend to lack remorse.	1 = Strongly disagree 5 = Strongly agree	.67-.77	H(-), E(-), X(/), A(-), C(-), O(/) (Schreiber & Marcus, 2020; Table 1)

Respondent Reactions (T1)	**Example Item**	**Rating Scale**	**α**
Positive Affect (3) (Adapted from Zhang et al., 2020 + self-made)	This questionnaire is interesting.	1 = Strongly disagree 5 = Strongly agree	.67-.75
Accuracy (3) (Adapted from Dalal et al., 2021 + Self-made)	This questionnaire can accurately measure my personality characteristics.	1 = Strongly disagree 5 = Strongly agree	.66-.74
Utility (3) (self-made)	This questionnaire is useful for personnel selection.	1 = Strongly disagree 5 = Strongly agree	.72-.77
Faking Resistance (3) (self-made)	It is hard to fake on this questionnaire.	1 = Strongly disagree 5 = Strongly agree	.65-.76
Difficulty (3) (Adapted from Zhang et al., 2020 + self-made)	This questionnaire is difficult to answer.	1 = Strongly disagree 5 = Strongly agree	.75-.78
Cognitive Burden (3) (self-made)	Completing this questionnaire makes me exhausted.	1 = Strongly disagree 5 = Strongly agree	.45-.56
Degree of Concentration (3) (Adapted from Zhang et al., 2020 + self-made)	I was concentrated when completing this questionnaire.	1 = Strongly disagree 5 = Strongly agree	.25-.39
Exerted Effort (3) (Adapted from Zhang et al., 2020 + self-made)	How much effort do you have to exert in order to complete this questionnaire as instructed?	0 = Zero effort 10 = All my efforts	NA
Exhaustion (3) (self-made)	How exhausted are you after completing this questionnaire?	0 = Not exhausted at all 10 = Completely exhausted	NA
Energy Level (3) (self-made)	Let's say your energy level was 10 before you start to work on this questionnaire. What's your current energy level after completing this questionnaire?	0 = Zero energy 10 = Full energy	NA

Respondent Reactions (T2)	**Example Item**	**Rating Scale**	**α**
Fairness (2) (Chan et al., 1998; Lopez et al., 2019)	Overall, I believe the test was fair.	1 = Strongly disagree 5 = Strongly agree	.71-.84
Predictive Validity (2) (Kluger & Rothstein, 1993; Macan et al., 1994)	The test measured the skills necessary to perform well on the job.	1 = Strongly disagree 5 = Strongly agree	.82-.85
Satisfaction with Process (2) (Sylva & Mol, 2009; Tonidandel et al., 2002)	I liked taking this type of test.	1 = Strongly disagree 5 = Strongly agree	.69-.77
Organizational Attractiveness (2) (Highhouse et al., 2003)	For me, this company would be a good place to work.	1 = Strongly disagree 5 = Strongly agree	.83-.88
Intent to Accept Job (2) (Highhouse et al., 2003)	I would accept a job offer from this company.	1 = Strongly disagree 5 = Strongly agree	.75-.84
Face Validity (2) (Chan et al., 1998; Macan et al., 1994)	The actual content of the test is clearly related to the job.	1 = Strongly disagree 5 = Strongly agree	.84-.89

Intent to Recommend (2) (Smither et al., 1993; Highhouse et al., 2003)	Based on my experience with the test, I would recommend others to apply to this organization.	1 = Strongly disagree 5 = Strongly agree	.88-.93
Faking Resistance (2) (self-made)	It's hard to fake on this questionnaire.	1 = Strongly disagree 5 = Strongly agree	.87-.91
Accuracy (2) (Harris et al., 2021)	I believe the assessment accurately measured my personality.	1 = Strongly disagree 5 = Strongly agree	.56-.65

Note. Number after each construct name indicates the number of items measuring that construct. H = Honesty-Humility, E = Emotionality (or Neuroticism), X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. For the "Theoretical HEXACO Correlates" column, A "(+)" notation indicates evidence for positive association between the personality trait and the corresponding criterion variable, with an absolute magnitude of ≥.10. A "(-)" notation indicates evidence for negative association between the personality trait and the corresponding criterion variable, with an absolute magnitude of ≥.10. A "(/)" notation indicates the association between the personality trait and the corresponding criterion variable is small with a magnitude of <.10 (regardless of direction). Also for the "Theoretical HEXACO Correlates" column, notation not in italics represents evidence from a published meta-analysis, while italics represent evidence from a primary study. Reliabilities for Exerted Effort, Exhaustion, and Energy Level are not applicable as these constructs were measured by only one item.

Table 2. *FC and SS empirical reliability*

	Time 1 = Honest					Time 2 = Faking				
Trait	FC1	FC2	FC3	FC4	SS	FC1	FC2	FC3	FC4	SS
H	.70	.71	.74	.68	.85	.67	.73	.73	.64	.83
E	.73	.75	.67	.75	.83	.71	.70	.62	.67	.81
X	.66	.79	.76	.80	.87	.68	.78	.73	.76	.87
A	.68	.69	.67	.68	.81	.65	.69	.64	.66	.82
C	.65	.71	.63	.73	.85	.62	.69	.61	.70	.85
O	.73	.70	.67	.68	.83	.71	.67	.67	.66	.84
Mean	.69	.73	.69	.72	.84	.67	.71	.67	.68	.84

Note. H = Honesty-Humility, E = Emotionality, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. FC = Forced-choice measure, SS = Single-statement measure. Empirical reliability for single-statement measure is calculated based on participants from all four study groups.

Table 3. *FC and SS convergent validity*

	Uncorrected								Corrected for unreliability							
	Time 1 = Honest				Time 2 = Faking				Time 1 = Honest				Time 2 = Faking			
	FC1	FC2	FC3	FC4	FC1	FC2	FC3	FC4	FC1	FC2	FC3	FC4	FC1	FC2	FC3	FC4
H	.54	.64	.72	.65	.45	.55	.59	.54	.69	.83	.91	.85	.60	.70	.75	.73
E (Reverse)	.81	.79	.74	.76	.65	.63	.54	.56	1.00	1.00	1.00	.95	.84	.84	.77	.75
X	.57	.79	.76	.84	.52	.60	.64	.62	.75	.95	.94	1.00	.67	.73	.81	.77
A	.67	.72	.64	.71	.57	.57	.54	.55	.89	.95	.88	.96	.77	.75	.76	.75
C	.44	.61	.67	.70	.38	.42	.54	.57	.59	.79	.92	.88	.53	.55	.74	.74
O	.75	.73	.75	.75	.65	.58	.70	.67	.96	.95	1.00	1.00	.84	.78	.93	.89
Mean	.63	.71	.71	.74	.54	.56	.59	.59	.81	.91	.94	.94	.71	.73	.79	.77

Note. H = Honesty-Humility, E = Emotionality, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. FC = Forced-choice measure, SS = Single-statement measure. Time 1 matched sample size: N_{FC1} = 541, N_{FC2} = 528, N_{FC3} = 543, N_{FC4} = 535. Time 2 matched sample size: N_{FC1} = 289, N_{FC1} = 283, N_{FC1} = 302, N_{FC1} = 303. Corrected convergent validity estimates larger than 1.0 were set to 1.0.

Table 4. *Discriminant validity at Time 1*

Trait Pair	**Time 1 Latent Correlations**									
	FC1		FC2		FC3		FC4		SS	
	r	SE	*r*	SE	*r*	SE	*r*	SE	*r*	SE
H-E	.364***	.070	.111	.060	.191**	.067	.216**	.067	.056	.056
H-X	-.130	.076	.174**	.061	-.092	.062	.117	.069	.023	.051
H-A	.250***	.067	.086	.067	.129	.069	.226**	.070	.230***	.044
H-C	.304***	.085	.343***	.065	.188**	.068	.273***	.066	.216***	.029
H-O	.147	.077	.047	.064	.119	.065	.251***	.068	.015	.051
E-X	-.027	.094	-.185**	.059	-.311***	.065	-.153*	.062	-.337***	.044
E-A	.075	.072	-.051	.069	.020	.074	-.127	.069	-.188***	.039
E-C	.162	.095	-.157*	.065	.102	.081	-.115	.068	-.078	.050
E-O	.114	.079	.092	.066	-.012	.071	.125	.066	-.003	.054
X-A	-.003	.079	.442***	.055	.220***	.065	.089	.067	.235***	.048
X-C	-.096	.079	.461***	.052	.047	.075	.369***	.059	.204***	.049
X-O	.114	.073	.087	.063	.086	.066	.149*	.066	.070	.055
A-C	.003	.077	.124	.071	-.078	.076	.151*	.070	.111**	.036
A-O	.024	.071	.069	.071	.118	.073	.227**	.079	-.004	.028
C-O	.159*	.080	.064	.065	.054	.077	.152*	.067	.102*	.052
Mean Absolute Correlation & SE	.131 (.078)		.166 (.064)		.118 (.070)		.183 (.068)		.125 (.046)	
ICC with SS	.17		.71		.67		.64		NA	

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. The SS estimates were based on merged sample from all four conditions, which results in smaller standard errors.

Table 5. *Multiple R^2 of HEXACO traits predicting criterion variables and double entry ICCs of validity profile between FC and SS measures, corrected for reliability of personality traits*

Criterion Variable	FC1	FC2	FC3	FC4	SS
MAC	.181	.137	.268	.328	.377
PSYC	.187	.152	.316	.290	.421
NARC	.276	.135	.175	.275	.217
OCB	.048	.136	.110	.104	.096
CWB	.079	.076	.174	.173	.191
JP	.061	.043	.099	.157	.248
JS	.114	.146	.062	.117	.145
BNT	.232	.279	.133	.333	.311
TI	.104	.138	.088	.115	.162
SWB	.127	.322	.169	.364	.353
FS	.124	.120	.055	.080	.135
PHQ	.189	.135	.096	.168	.186
EDU	.038	.057	.036	.064	.041
WAG	.046	.032	.047	.098	.057
ORG	.163	.112	.063	.093	.089
CHAR	.048	.074	.070	.079	.045
GEN	.142	.225	.365	.172	.217
AGE	.036	.017	.057	.097	.083
TEN	.028	.009	.056	.065	.061
Mean R^2	**.117**	**.123**	**.129**	**.167**	**.181**
FC-SS Validity Profile ICC	**FC1**	**FC2**	**FC3**	**FC4**	**SS**
Honesty-Humility	.23	.67	.88	.96	/
Emotionality	.97	.94	.79	.96	/
Extraversion	.60	.92	.75	.97	/
Agreeableness	.71	.81	.52	.96	/
Conscientiousness	-.04	.70	.72	.95	/
Openness	.08	.40	.66	.68	/
Overall	**.54**	**.83**	**.77**	**.96**	**/**

Note. Multiple R^2 was calculated based on correlations between personality traits and criterion variables. MAC = Machiavellianism, PSYC = Psychopathy, NARC = Narcissism, CWB = Counterproductive work behavior, JS = Job satisfaction, BNT = Burnout, FS = Financial Security, OCB = Organizational citizenship behavior, SWB = Subjective well-being, TI = Turnover intentions, JP = Job performance, PHQ = Physical health, EDU = Education, WAG = Wage, ORG = Organizational status, CHAR = Charity behaviors.

Table 6. *Rank-order stability and mean score differences*

	Rank-Order Stability										Cohen's d (T2-T1)									
	Uncorrected					Corrected for unreliability					Uncorrected					Corrected for unreliability				
	FC1	FC2	FC3	FC4	SS	FC1	FC2	FC3	FC4	SS	FC1	FC2	FC3	FC4	SS	FC1	FC2	FC3	FC4	SS
H	.65	.53	.55	.51	.62	.95	.75	.75	.78	.73	.03	.31	.19	.48	.37	.03	.36	.22	.58	.40
E (Reverse)	.51	.48	.58	.47	.63	.72	.66	.91	.66	.77	.48	.33	.14	.38	.40	.56	.39	.17	.45	.44
X	.56	.49	.55	.49	.54	.83	.62	.74	.63	.62	.21	.32	.49	.68	.61	.25	.36	.56	.77	.65
A	.48	.56	.45	.49	.52	.72	.81	.68	.73	.64	.17	.33	.26	.56	.46	.21	.40	.32	.68	.51
C	.52	.47	.44	.38	.54	.81	.67	.70	.54	.64	.26	.55	.59	.98	.50	.33	.66	.75	1.16	.55
O	.59	.57	.65	.61	.75	.83	.83	.96	.90	.89	.11	.07	.23	.48	.25	.13	.08	.28	.58	.28
Mean	.55	.52	.54	.49	.60	.81	.72	.79	.71	.72	.21	.32	.31	.59	.43	.25	.37	.38	.70	.47

Note. H = Honesty-Humility, E = Emotionality, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. FC = Forced-choice measure, SS = Single-statement measure.

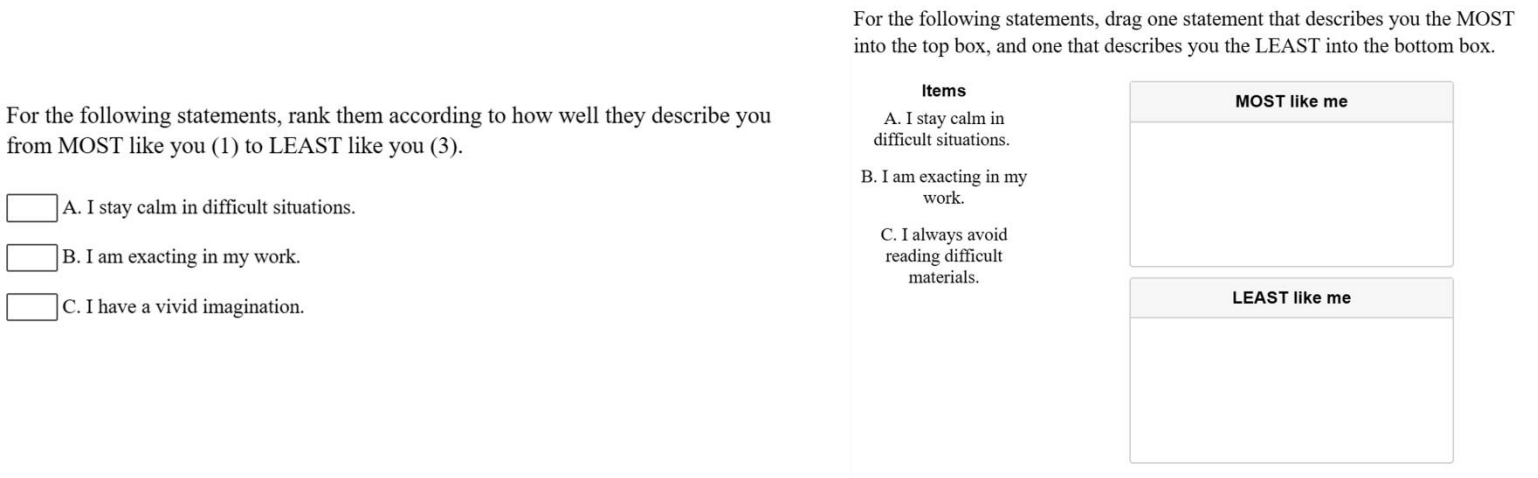
Table 7. *Respondent reactions*

	FC1			FC2			FC3			FC4			Cohen's *d*					
	M	SD	ω	M	SD	ω	M	SD	ω	M	SD	ω	1-2	1-3	1-4	2-3	2-4	3-4
Time 1 Respondent Reactions																		
Affect	4.04	0.70	.75	4.01	0.69	.74	3.99	0.68	.76	4.00	0.65	.68	.05	.08	.06	.03	.01	-.02
Accuracy	3.39	0.75	.68	3.38	0.81	.75	3.34	0.70	.68	3.39	0.74	.69	.02	.07	-.003	.05	-.02	-.07
Utility	3.33	0.80	.76	3.33	0.81	.77	3.32	0.73	.72	3.39	0.73	.72	.00	.02	-.07	.01	-.07	-.09
Faking Resistance	2.93	0.89	.77	2.88	0.84	.68	2.89	0.78	.68	2.87	0.80	.66	.06	.05	.07	-.02	.01	.02
Difficulty	2.47	0.91	.76	2.46	0.93	.78	2.50	0.90	.77	2.49	0.90	.75	.01	-.04	-.02	-.04	-.03	.02
Burden	2.39	0.68	.59	2.38	0.73	.59	2.37	0.66	.52	2.40	0.70	.59	.01	.02	-.02	.01	-.03	-.04
Concentration	3.85	0.62	.53	3.80	0.60	.46	3.75	0.61	.54	3.79	0.60	.52	.07	.15	.10	.08	.02	-.06
Exerted Effort	7.83	2.24	-	7.81	2.29	-	7.43	2.06	-	7.58	2.08	-	.01	.19	.12	.17	.11	-.07
Exhaustion	3.25	2.36	-	3.40	2.48	-	3.51	2.30	-	3.62	2.39	-	-.06	-.11	-.16	-.05	-.09	-.05
Energy Level	9.33	1.66	-	9.19	1.64	-	9.28	1.59	-	9.30	1.64	-	.09	.03	.02	-.06	-.07	-.01
Time 2 Respondent Reactions																		
Fairness	3.22	1.00	.84	3.35	0.95	.81	3.31	0.86	.71	3.36	0.90	.75	-.13	-.10	-.15	.05	-.01	-.06
Validity	2.51	1.06	.85	2.65	1.06	.83	2.48	0.93	.82	2.58	1.08	.85	-.13	.04	-.06	.17	.07	-.10
Satisfaction	3.08	1.01	.77	3.23	0.96	.71	3.15	0.92	.72	3.27	0.92	.69	-.16	-.07	-.19	.09	-.03	-.13
Org Attractiveness	3.66	0.98	.88	3.75	0.87	.86	3.62	0.84	.84	3.72	0.84	.83	-.10	.04	-.07	.16	.04	-.13
Intent to Accept	4.10	0.79	.82	4.16	0.72	.84	4.07	0.69	.75	4.15	0.72	.82	-.08	.04	-.07	.13	.01	-.12
Face Validity	2.67	1.11	.88	2.77	1.07	.89	2.71	1.00	.84	2.78	1.05	.86	-.09	-.04	-.10	.06	-.01	-.07
Intent to Recommend	3.24	1.05	.93	3.33	0.97	.89	3.14	0.89	.88	3.30	0.94	.91	-.09	.11	-.06	.21	.03	-.18
Faking Resistance	2.60	1.13	.91	2.84	1.16	.88	2.65	1.09	.87	2.50	1.12	.88	-.21	-.05	.09	.17	**.30***	.14
Accuracy	3.19	0.91	.65	3.34	0.86	.61	3.20	0.77	.56	3.32	0.88	.65	-.16	-.01	-.14	.17	.02	-.14

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$ after adjusting for multiple comparisons. Significant differences were bolded. FC = Forced-choice measure.

LIST OF FIGURES

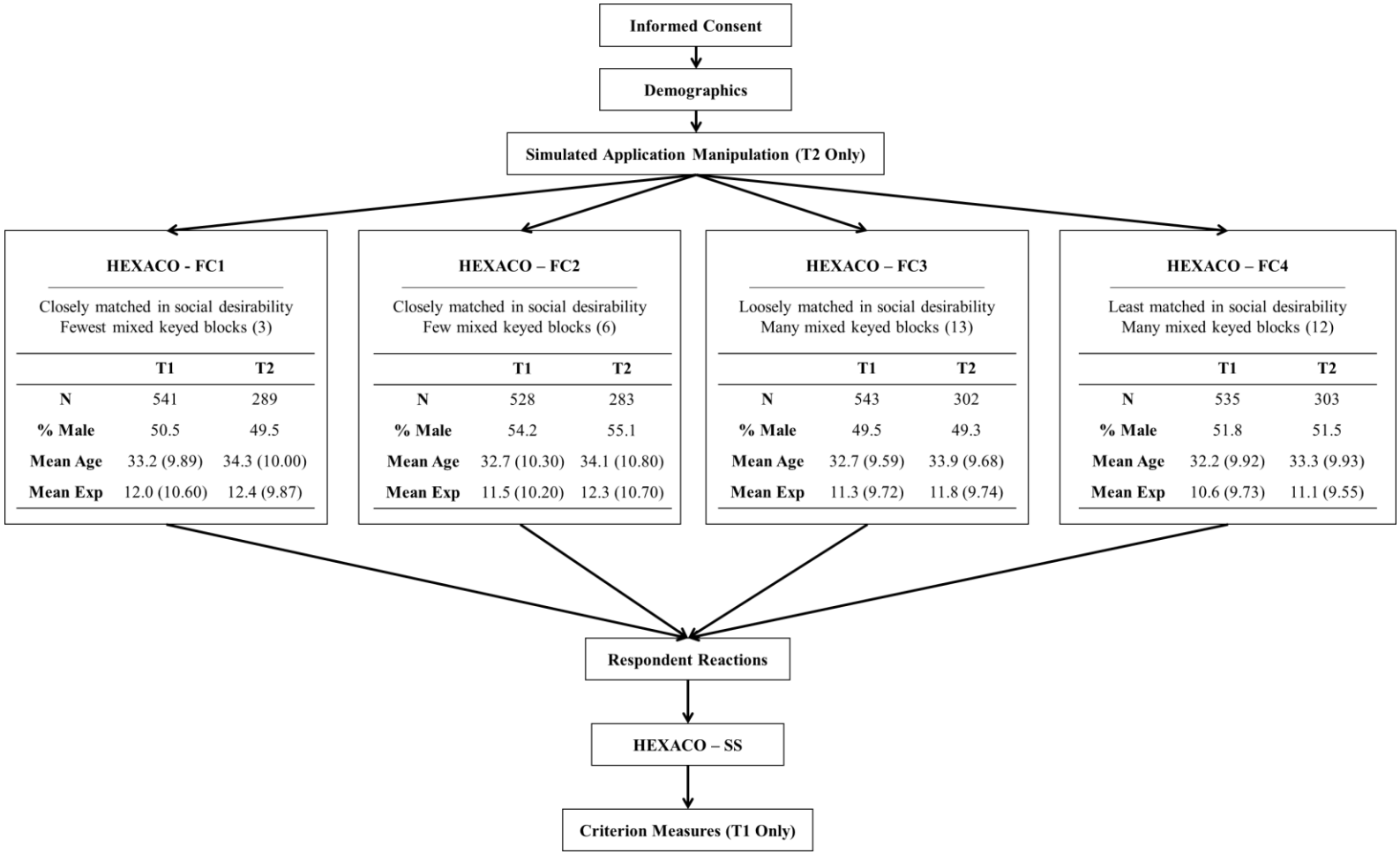
Figure 1. *Examples of multidimensional forced-choice blocks*



RANK format, matched block (Positive) **MOLE format, mixed keyed block**

Note. Multidimensional FC blocks of RANK (left) and MOLE (right) formats are presented. Options A, B, and C are the statements within the block. The block at the left represents a RANK format, where participants provide a full ranking of all statements in the block in terms of descriptiveness. The block at the right represents a MOLE (MOst and LEast like me) format, where participants pick one most descriptive statement and one least descriptive statement.

Figure 2. *Study procedures*



Mix-Keying or Desirability-Matching in the Construction of Forced-Choice Measures?

An Empirical Investigation and Practical Recommendations

SUPPLEMENTARY MATERIALS

Section 1: A Tutorial on Automated FC Scale Construction and Simulations

This section provides a tutorial on (1) how to construct an FC scale using the *autoFC* R package (Li et al., 2022), and (2) how to use Monte Carlo simulations to evaluate the quality of constructed FC scales in ideal conditions. We assume that readers already have the established statements (HEXACO-60; Ashton & Lee, 2009) and the corresponding social desirability values of these statements (please see Steps 1 and 2 in the **Recommended Steps to Develop FC Measures** section in the **Discussion** on how to develop high-quality statements and obtain social desirability values). In the current example, social desirability ratings are obtained from Anglim et al., (2017). Readers can refer to the "**Tutorial.R**" file available at OSF: https://osf.io/yvpz3/?view_only=08601755f471440b80973194571b60bd.

Part 1: Automated FC Scale Construction

Step 3: Determine block size.

Assume we want to use a triplet design (i.e., block size = 3) containing 20 blocks. The objective is hence to construct a 20×3 matrix containing numbers from 1-60 (statement IDs), where each row represents the three statements that should be placed in the same block.

```
## Example objective matrix - each row represents IDs of items in the same block
## Just for illustrative purposes
matrix(sample(1:60, 60), ncol = 3, byrow = TRUE)
```

Step 4. Determine the number of mixed blocks.

Starting from this step, we demonstrate how the *autoFC* R package can be conveniently used to construct the FC scale and provide initial tests of psychometric properties using simulated data. We first load the *autoFC* package (currently in dev version) and a couple of other R packages needed, including *lavaan* (Rosseel, 2012), *MASS* (Venables & Ripley, 2002), *thurstonianIRT* (Bürkner, 2019), *tidyr* (Wickham et al., 2023), and *dplyr* (Wickham, François et al., 2023) as follows. We also load the information of the 60 HEXACO statements:

```
devtools::install_github("tspsyched/autoFC")
library(autoFC)
library(lavaan)
library(MASS)
library(thurstonianIRT)
library(tidyr)
library(dplyr)

item_desirability <- readxl::read_excel("Item_Desirability.xlsx")
```

In the current example, we want 6 mixed blocks and we also want to ensure each trait is paired with other traits for about an equal number of times. We provide a function called `construct_blueprint` for specifying how the 6 mixed blocks should look like, but we note that this method can be applied to equally keyed blocks as well. The `construct_blueprint` function allows you to specify the dimension of the statements within each block and keying of each statement. We call this explicit specification as a "***blueprint***":

```
d_mixed <- 1.25
d_equal <- 0.5
test_bp <- construct_blueprint(N_blocks = 6, block_size = 3,
                               traits = c("H","E","X",
                                          "A","C","O",
                                          "H","C","O",
                                          "X","A","C",
                                          "H","A","X",
                                          "H","E","O"),
                               signs = c(1, -1, -1,
                                         -1, -1, 1,
                                         1, -1, -1,
                                         -1, 1, 1,
                                         -1, 1, 1,
                                         -1, -1, 1))
```

In the above code, the argument `signs` contains information on the direction of keying for each statement, with 1 indicating that the statement is positively keyed, while -1 indicating it is negatively keyed. For example, the first block (highlighted in green) contains a positively

keyed statement measuring Honesty-Humility, a negatively keyed statement measuring Emotional Stability, and another negatively keyed statement measuring Extraversion.

We can additionally add another specification to our blueprint. In our example, we specify a cutoff for the D values (maximum difference of social desirability ratings in a block) of these mixed-keyed blocks to be `d_mixed` (1.25) on a 5-point scale. This is done through adding a new column to the `test_bp` data frame:

```
test_bp$SD_matching <- rep(d_mixed, 18)
```

We note that larger cutoff values for D may be needed when: (1) block size is larger, because finding more statements with similar levels of social desirability is generally more difficult; (2) multiple criteria need to be met when pairing statements (e.g., statements should come from different dimensions; statements need to be mixed keyed). Some statements may have similar levels of social desirability but cannot be paired together due to belonging to the same dimension or not satisfying keying requirements; (3) the set of candidate statements at researchers' disposal for constructing FC scale is limited, especially when all statements need to be used for pairing.

Step 5. Create blocks.

We start by constructing the mixed-keyed blocks. To do so, we use the blueprint `test_bp` constructed in Step 4 and run the function `build_scale_with_blueprint`:

```
range_m <- function(x) {
   return(max(x) - min(x))
}
picked_items <-
build_scale_with_blueprint(
## Your item information data frame
item_df = item_desirability,
## Your blueprint
blueprint = test_bp,
## Which column in test_bp specifies the block number?
bp_block_name = "block",
## Which column in test_bp specifies the item number in each block?
bp_item_nums_name = "item_num",
## Which column in test_bp specifies the desired item traits?
```

```
bp_trait_name = "traits",
## Which column in test_bp specifies the desired item signs?
bp_sign_name = "signs",
## Which column in test_bp specifies your additional matching criterion?
bp_matching_criterion_name = "SD_matching",
## Which column in item_df specifies the item number?
df_item_nums_name = "ID",
## Which column in item_df specifies the item traits?
df_trait_name = "factor",
## Which column in item_df specifies the item signs?
df_sign_name = "Reversed_ES",
## Which column in item_df specifies your variable for calculating the criterion?
df_matching_criterion_name = "SD_rating",
## What function is used to calculate the criterion?
df_matching_function = "range_m",
## If criterion is not met after max_attempts_in_comb attempts,
## how much will it be multiplied for adjustment?
df_matching_adjust_factor = 1.25,
## How many times will we try before adjusting the criterion?
max_attempts_in_comb = 100,
## What is the maximum number of times will be adjust the criterion?
max_attempts_in_adjust = 6)
```

In sum, this function requires users to provide a blueprint (`blueprint`) and information for the original items (`item_df`). It also requires users to specify a couple of column names to tell the function (1) which columns correspond to block number, item number in each block, desired traits for each item, desired keying (signs) for each item, and additional matching criterion (in this case, it is social desirability matching) in `blueprint`; as well as (2) which columns correspond to item number, item traits, item signs, and information for calculating matching criterion in `item_df`.

Next, `build_scale_with_blueprint` starts from the first block in the blueprint, checks all statement combinations that satisfy the trait-sign specification in that block (e.g., positive Honest-Humility + negative Emotional Stability + negative Extraversion), and randomly pick one combination of three statements. Then, it goes to `item_df` and looks for the column specified by `df_matching_criterion_name`, which corresponds to the `SD_rating` column. It then examines the `SD_rating` values of the three statements and calculates the D value using the `range_m` function we provided above, and see if the D value is ***smaller*** than the pre-specified cutoff (i.e., 1.25 - the value given in the `SD_matching`

[specified by `bp_matching_criterion_name`] column in `blueprint`). If so, we consider the block to be successfully constructed consistent with the blueprint. Otherwise, we tried `max_attempts_in_comb` times and see if we have better luck finding another combination with smaller D values. If that fails again, we multiply the cutoff by `df_matching_adjust_factor` and repeats the process of calculating the D value for at most `max_attempts_in_comb` times again, until the cutoff is eventually met or `max_attempts_in_adjust` times of multiplying the cutoff is already done.

We note that if the block cannot be constructed even when we relax the cutoff for `max_attempts_in_adjust` times, this probably signifies that more appropriate cutoff or `df_matching_adjust_factor` values should be used. In this case, a warning message will be produced and the function will return with a partially constructed scale, if applicable:

```
Warning messages:
1: In build_scale_with_blueprint(item_df = item_desirability, blueprint = test_bp,  :
  It seems like we cannot find a match for your blueprint after 6 attempts in relaxing the matching
criteria. Consider increasing max_attempts_in_adjust and/or max_attempts, or adjust your blueprint.
2: In build_scale_with_blueprint(item_df = item_desirability, blueprint = test_bp,  :
  Problem appears when constructing block 2
```

Now we have constructed the first 6 mixed-keyed blocks (stored in a data frame called `picked_items`), we are left with 42 statements to be matched on social desirability, into 14 equally-keyed blocks. This can be achieved by running the `sa_pairing_generalized` function:

```
rest_FC <-
sa_pairing_generalized(block = make_random_block(42, 42, 3),
                       item_chars = item_desirability[-picked_items$ID,c(2,3,4)],
                       r = 0.999,
                       FUN = c("facfun","inn_diff","var"),
                       weights = c(10000,-10,-1000))
```

The first argument, `block`, indicates an initial triplet pairing to start with. As we are left with 42 statements, we construct a 14-block random solution to start the pairing process.

The `item_chars` argument indicates which item characteristics need to be considered for pairing. Here we use the second (factor), third (social desirability rating), and fourth (keying) columns from data frame `item_desirability`, but excluding the statements that have been used for building mixed blocks.

The `r` argument is simply a tuning parameter determining how many iterations to run for automatic pairing. This value should be between 0 and 1, with values closer to 1 indicate more rounds of iteration (and hence more accurate solutions are likely to emerge). We recommend readers to simply use the 0.999 value here, which should be appropriate in most cases. Interested readers can refer to the documentation of the `sa_pairing_generalized` function for more details.

The `FUN` argument tells which function to use for the values within the same block, for each item characteristic column.

- For the second column (factor), we use the function `facfun`, which returns 1 when all passed elements are unique, and 0 otherwise. We use this function to check if the statements in the same block are all coming from different latent factors.
- For the third column (social desirability rating), we use the function `inn_diff`, which returns 1 when the range of the passed elements exceeds a certain cutoff (here we use `d_equal` = 0.5), and 0 otherwise. We use this function to check if the maximum difference in social desirability ratings among the three statements exceeds the pre-determined cutoff.
- For the fourth column (keying), we use the function `var`, which is simply the variance of the passed elements. We use this function to check if statements are

equally keyed. The function will return 0 if statements are all equally keyed, or will return a value larger than 0 if otherwise.

The last argument, `weight`, controls how much weight should we give for the outcomes of each function indicated in `FUN`. We assume that a FC scale with better "fit" should have a more positive score ("energy") and would therefore want positive weights for desirable metrics.

- For `facfun`, it returns 1 when all passed elements are unique, and 0 otherwise. Since statements coming from different factors is almost a must, we will want `facfun` to return 1 as much as possible. As such, we indicate the weight for this function as 10000.
- For `inn_diff`, it returns 1 when the range of the passed elements exceeds a certain cutoff, and 0 otherwise. Since we don't really want statements to be too discrepant in social desirability ratings (since they are now constituting equally keyed blocks), we want `inn_diff` to return 0 as much as possible. As such, we give a negative weight for this function. But this requirement is somehow not as stringent as the requirements for `facfun`, we set -10 as its weight.
- For `var`, as now we are constructing equally keyed blocks, we don't want large positive variance of social desirability ratings. As such, we also give a negative weight for `var` and since mixed blocks are to be avoided, we set a large negative weight of -1000.

After running the previous code, we now have the triplet FC scale set and are ready for further steps:

```
FC3_1 <- matrix(c(picked_items$ID, rest_items$ID), ncol = 3, byrow = TRUE)
```

Alternatively, if readers want to ensure social desirability matching to be as strong as possible while accepting a couple more mixed keyed blocks to pop up, they can adjust the weights accordingly, for example:

```
rest_FC2 <-
sa_pairing_generalized(block = make_random_block(42, 42, 3),
                       item_chars = item_desirability[-picked_items$ID,c(2,3,4)],
                       r = 0.999,
                       FUN = c("facfun","inn_diff","var"),
                       weights = c(10000,-1000,-1))
```

In this case, we still require statements to come from different dimensions (weight for `facfun` being 10000), but now would want to curb social desirability mismatching (weight for `inn_diff` being -1000) and care less about mixed keying (weight for `var` being -1).

We additionally note that if users wish to incorporate the inter-item agreement (IIA) approach proposed by Pavlov et al., (2022) they can add a few arguments in the `sa_pairing_generalized` function as follows:

```
rest_FC3 <-
sa_pairing_generalized(block = make_random_block(42, 42, 3),
                       item_chars = item_desirability[-picked_items$ID,c(2,3,4)],
                       r = 0.999,
                       FUN = c("facfun","inn_diff","var"),
                       weights = c(10000,-1000,-1),
                         use_IIA = TRUE,
                       rather_chars = _______, iia_weights = ________)
```

Where `rather_chars` is a data frame containing the individuals' rating on social desirability for all statements (note that number of rows = number of participants, number of columns = number of statements), `iia_weights` is a vector of length 4 indicating weights given to each IIA metric: Linearly weighted Agreement Coefficients (AC; Gwet, 2008; 2014),

quadratic weighted AC, linearly weighted Brennan-Prediger (BP) index (Brennan & Prediger, 1981; Gwet, 2014), and quadratic weighted BP.

Part 2. Monte Carlo Simulation

Step 6. Examine the reliability of the FC scale using simulations without collecting new data.

After the FC scales are constructed, we want to conduct some initial tests on its psychometric properties. The easiest and most cost-effective way is to use simulations to examine the reliability of trait scores obtained from this FC scale under ideal conditions.

According to the Thurstonian IRT model, when completing a FC block with several statements, participants will evaluate the utility of each statement. They will choose/rank statements within a block by comparing their utilities in a pairwise manner. Statements with higher utilities will be preferred. The utility of each statement *i* is defined as (Brown & Maydeu-Olivares, 2013):

$$t_i = \mu_i + \lambda_i \eta_a + \varepsilon_i$$

where μ_i is the mean utility, λ_i is the loading of statement i on attribute η_a, and ε_i represents the uniqueness factor. Hence, with this formula, the utility of a statement is determined by how the study population on average perceives the statement (μ_i), how strong the statement loads on its corresponding trait η_a (λ_i), and uniqueness of that statement (ε_i). Given that different individuals will have different levels of η_a, such utility value for a specific person will further be determined by that person's trait level of η_a. The matrix notation to represent multiple individuals on multiple traits is as follows:

$$\boldsymbol{t} = \boldsymbol{\mu} + \boldsymbol{\lambda\eta} + \boldsymbol{\varepsilon}$$

Here $\boldsymbol{t}$ is the matrix of utility on multiple statements for multiple individuals, $\boldsymbol{\mu}$ is the statement intercept matrix, $\boldsymbol{\lambda}$ is factor loading matrix, $\boldsymbol{\eta}$ represents individual's standing on multiple traits, and $\boldsymbol{\varepsilon}$ is the uniqueness matrix for each statement.

All in all, four components are needed: $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$, $\boldsymbol{\eta}$, and $\boldsymbol{\varepsilon}$. Among them, $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\varepsilon}$ can be obtained from factor analysis on responses to the SS format obtained at the initial stage of statement development and selection, while $\boldsymbol{\eta}$ for each individual can be generated from a multivariate normal distribution where the covariances between traits are again from that factor analysis. This step assumes that parameters are invariant across the FC and the SS formats, which has been largely supported in previous studies (Lin & Brown, 2017; Morillo et al., 2019).

As such, we first perform traditional confirmatory factor analysis on responses to the SS format and obtain the intercept, loading and uniqueness estimates. The `get_CFA_estimates` function performs the confirmatory factor analysis using the specified model and automatically stores the intercept, loading and uniqueness estimates:

```
SS_model <- paste0("H =~ ", paste0("SS", seq(6,60,6), collapse = " + "), "\n",
                   "E =~ ", paste0("SS", seq(5,60,6), collapse = " + "), "\n",
                   "X =~ ", paste0("SS", seq(4,60,6), collapse = " + "), "\n",
                   "A =~ ", paste0("SS", seq(3,60,6), collapse = " + "), "\n",
                   "C =~ ", paste0("SS", seq(2,60,6), collapse = " + "), "\n",
                   "O =~ ", paste0("SS", seq(1,60,6), collapse = " + "), "\n")
SS_estimates <- get_CFA_estimates(response_data = rating_data,
                                  fit_model = SS_model,
                                  item_names = paste0("SS",c(1:60)))
```

We then convert the CFA estimates into a matrix and generate the utility for the 60 HEXACO statements for 1000 simulated respondents, using the `get_simulation_matrices` function. This function will produce five matrices: Utility($\boldsymbol{t}$),

Mu($\boldsymbol{\mu}$), Lambda($\boldsymbol{\lambda}$), Theta($\boldsymbol{\eta}$), and Epsilon($\boldsymbol{\varepsilon}$). We will build simulated FC responses based on the Utility($\boldsymbol{t}$) matrix (1000 by 60).

```
SS_matrices <- get_simulation_matrices(loadings = SS_estimates$loadings,
                                       intercepts = SS_estimates$intercepts,
                                       residuals = SS_estimates$residuals,
                                       covariances = SS_estimates$covariances,
                                       N = 1000,
                                       N_items = 60,
                                       N_dims = 6,
                                       dim_names = c("H", "E", "X", "A", "C", "O"),
                                       empirical = TRUE)

### Adjust the item order into 1, 2, 3...60
## You can use SS_estimates$loadings to see how it was originally ordered; Should be
consistent with your CFA model
## EDIT THIS LINE ACCORDINGLY BASED ON YOUR CFA MODEL.
SS_matrices$Utility <- SS_matrices$Utility[,c(t(matrix(1:60, ncol = 6)[,6:1]))]
```

The last line in the previous section of code is critical: **You need to reorder the utility values of the columns back to the 1-60 order (they are originally in the order they appear in the CFA model!).**

Now, using the new FC scale we built in Step 5 (`FC3_1`), and the Utility matrix, we construct simulated responses to the FC scale. Note that this step directly produces pairwise or ranked responses which can be directly processed by the *thurstonianIRT* (Bürkner, 2019) package, rather than raw responses to the FC scale.

```
FC3_1_resp <- convert_to_TIRT_response(Utility = SS_matrices$Utility,
                                       block_design = FC3_1,
                                       N_response = 1000,
                                       format = "pairwise",
                                       block_size = 3,
                                       N_blocks = 20)
```

Next, after providing dimension and keying information of each statement, we can convert the pairwise or ranked responses into TIRT data ready for further analysis by the

thurstonianIRT package, using the wrapper function `get_TIRT_long_data`. We then analyze the TIRT data using Mplus (But can also use lavaan or Stan if readers prefer), with the `fit_TIRT_model` function:

```
# See the R file for full annotation of each argument
TIRT_long_FC3_1 <- get_TIRT_long_data(block_info = block_info_FC3_1,
                                      response_data = FC3_1_resp,
                                      response_varname = build_TIRT_var_names("i",
block_size = 3, N_blocks = 20, format = "pairwise"),
                                      partial = FALSE,
                                      format = "pairwise",
                                      direction = "larger",
                                      family = "bernoulli",
                                      range = c(0, 1),
                                      block_name = "Block",
                                      item_name = "ID",
                                      trait_name = "Factor",
                                      sign_name = "Reversed")
### Can also use method = "lavaan" or "stan"
estimate_FC3_1 <- fit_TIRT_model(TIRT_long_FC3_1, method = "mplus")
```

Estimated trait scores and standard errors (not available if estimated using lavaan) will be stored in the `estimate_FC3_1` object.

In item response theory context, the reliability of the trait scores is also dependent on trait level and thus would be different from person to person. To calculate a summary statistic similar to reliability metric in classical test theory, we can calculate the empirical reliability of trait scores (Brown & Maydeu-Olivares, 2018). Thus, now we examine the empirical reliability of these traits and see how they correlate with true Theta (produced using `get_simulation_matrices` function in previous sections):

```
empirical_reliability(dataset = estimate_FC3_1$final_estimates,
                          score_names = paste0("estimate_", c("H", "E", "X", "A", "C",
"O")),
                          se_names = paste0("se_", c("H", "E", "X", "A", "C", "O")))

# In HEXACO order
> 0.6918498 0.6888847 0.7402713 0.6023309 0.6655692 0.6646959
```

The reliability is somehow low considering that we have only 10 statements for each HEXACO dimension. How about correlations with true Theta?

```
diag(cor(FC3_1_traits %>% select(estimate_H, estimate_E, estimate_X, estimate_A,
estimate_C, estimate_O), thetas))

# In HEXACO order
> 0.7855975  0.7875561  0.8263415  0.7297466 -0.7770323  0.7795176
```

Negative correlations with Theta for some traits can appear when the first statement appearing in the FC scale measuring that trait is negatively keyed. In this case, we can manually reverse the trait estimates. Also note that an alternative approach to calculating reliability is to square these correlations with true Theta. Users are recommended to combine this approach with the empirical reliability approach to evaluate the reliability of the resulting FC measure.

Readers can further plot the estimated trait scores or standard errors against Theta values to see at which range of Theta will the trait estimate being the most/least accurate.

```
for (nm in trait_names) {
  ## If estimated score correlates negatively with theta, flip the estimated scores
  flag <- ifelse(cor(FC3_1_traits[,paste0("estimate_",nm)], thetas[,nm]) < 0, -1, 1)

  ## theta-estimated scores
  plot_scores(thetas[,nm], flag * FC3_1_traits[,paste0("estimate_",nm)], xlab = "theta",
ylab = "estimated_trait_score", main = paste0("theta_", nm))

  ## theta-absolute difference
  plot_scores(thetas[,nm], flag * FC3_1_traits[,paste0("estimate_",nm)], xlab = "theta",
ylab = " estimated_difference_from_theta ", main = paste0("abs.diff_", nm), type =
"abs.diff")

  ## theta-standard errors
  plot_scores(thetas[,nm], FC3_1_traits[,paste0("se_",nm)], xlab = "theta", ylab =
"estimated_standard_error", ylim = c(0.4, 0.8), main = paste0("standard_error_", nm))
}
```

Alternatively, they can examine the RMSE at different Theta ranges using the `RMSE_range` function, by passing first the Theta scores, then the estimated trait scores, and the break points that specify the Theta ranges that users wish to examine RMSE on:

```
for (nm in trait_names) {
  ## If estimated score correlates negatively with theta, flip the estimated scores
  flag <- ifelse(cor(FC3_1_traits[,paste0("estimate_",nm)], thetas[,nm]) < 0, -1, 1)

  print(paste0("RMSE for trait ", nm, " (Overall):"))
  print(RMSE_range(thetas[,nm], flag * FC3_1_traits[,paste0("estimate_",nm)]))

  print(paste0("RMSE for trait ", nm, " (Each range):"))
  print(RMSE_range(thetas[,nm], flag * FC3_1_traits[,paste0("estimate_",nm)],
         range_breaks = c(-Inf, -3, -2, -1, 0, 1, 2, 3, Inf)))
  writeLines("")
}
```

Section 2: Supplementary Tables, Figures, and Explanations

Table S1. *Summary for block design of the four FC measures*

Block	FC1						FC2						FC3						FC4					
	Trait/Key			Social Desirability			Trait/Key			Social Desirability			Trait/Key			Social Desirability			Trait/Key			Social Desirability		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	C-	E-	O-	2.42	2.65	2.40	A+	E+	O+	3.17	3.62	3.77	H-	O-	A-	2.18	2.07	2.92	**A+**	**H-**	**E-**	3.44	2.97	2.96
2	E+	O+	X+	3.42	3.55	3.36	H-	A-	C-	2.00	2.11	1.88	E+	O+	A+	2.91	3.55	3.39	**A+**	**C-**	**O-**	3.53	2.42	2.40
3	O-	H-	A-	2.07	2.18	2.68	**E+**	**O-**	**C-**	2.47	2.21	2.29	**C+**	**E+**	**O-**	4.14	3.62	2.40	**C-**	**O+**	**H-**	1.76	3.77	1.76
4	E+	X+	A+	2.87	3.40	3.53	**O-**	**E+**	**A+**	2.4	2.91	2.83	**A+**	**C-**	**O-**	2.83	2.42	2.74	E-	X-	C-	3.22	2.02	2.02
5	X-	C-	H-	2.35	2.29	2.26	O+	A+	X+	3.55	3.44	3.36	A+	O+	H+	3.17	3.80	3.47	X+	H+	E+	4.09	4.23	3.62
6	**H+**	**E-**	**O-**	2.94	3.22	2.74	**O-**	**E+**	**A-**	2.22	2.87	2.63	H-	A-	X-	2.27	2.68	2.02	**H-**	**C-**	**X+**	2.26	2.20	3.40
7	**O+**	**A+**	**E-**	3.77	3.39	3.47	H+	O+	X+	3.47	3.80	3.70	**X+**	**O+**	**C-**	3.70	3.87	2.02	**O-**	**X+**	**A+**	2.74	3.70	3.19
8	X+	C+	E+	3.29	3.44	2.91	H+	X+	C+	4.23	4.09	4.05	H-	X-	C-	1.76	1.89	1.76	E+	A+	O+	2.47	3.39	3.55
9	C-	X-	H-	1.76	1.89	1.76	X+	C+	A+	3.29	3.44	3.39	H-	A-	C-	2.00	2.11	1.88	**X+**	**H-**	**A-**	3.29	2.18	2.68
10	A-	X-	C-	2.11	2.08	2.02	H+	A+	O+	3.65	3.19	3.87	**C+**	**H-**	**O+**	3.20	2.97	3.83	**H+**	**C-**	**O-**	2.94	2.29	2.21
11	C-	H-	X-	1.88	2.00	2.02	E-	O-	X-	2.65	2.07	2.35	**E+**	**C-**	**X-**	2.47	2.20	2.08	E+	O+	H+	2.87	3.83	3.47
12	O-	H-	C-	2.21	2.27	2.20	C-	E-	A-	2.42	2.96	2.92	H+	E+	X+	2.94	2.79	3.29	**H-**	**X-**	**A+**	2.00	1.89	2.83
13	X+	A+	O+	3.70	3.17	3.87	H-	C-	X-	2.26	2.20	2.08	**X+**	**H-**	**A+**	3.40	2.26	3.53	**A-**	**O-**	**X+**	2.63	2.22	3.36
14	**O-**	**E+**	**A-**	2.22	2.47	2.63	**E+**	**A-**	**H-**	2.79	2.68	2.18	**X-**	**O+**	**A+**	2.35	3.77	3.44	**C+**	**H-**	**X-**	3.20	2.27	2.08
15	X+	H+	C+	4.09	4.23	4.05	H-	X-	C-	1.76	1.89	1.76	**X+**	**E-**	**C+**	4.09	3.22	4.05	X+	E+	C+	3.94	3.42	4.05
16	O+	H+	A+	3.83	3.47	3.19	C+	X+	O+	4.14	3.94	3.83	**C-**	**E+**	**O-**	2.29	3.42	2.21	**E-**	**C+**	**A-**	2.65	3.44	2.92
17	C+	X+	H+	4.14	3.94	3.65	**H+**	**E-**	**O-**	2.94	3.47	2.74	**A+**	**H+**	**E-**	3.19	4.23	3.47	H+	C+	E+	3.65	4.14	2.91
18	E+	A+	C+	2.79	2.83	3.20	E+	X+	A+	3.42	3.40	3.53	**X+**	**E-**	**C+**	3.36	2.65	3.44	O+	E+	A+	3.87	2.79	3.17
19	O+	A+	E+	3.80	3.44	3.62	**C+**	**H-**	**E-**	3.20	2.97	3.22	**O-**	**A-**	**E+**	2.22	2.63	2.87	X-	E-	O-	2.35	3.47	2.07
20	H-	E-	A-	2.97	2.96	2.92	X-	H-	C-	2.02	2.27	2.02	**H+**	**E-**	**X+**	3.65	2.96	3.94	**A-**	**O+**	**C-**	2.11	3.80	1.88

Note. H = Honesty-Humility, ES = Emotional stability, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. Blocks containing unequally keyed items were highlighted in bold. Items measuring emotionality were reversed keyed to represent emotional stability before FC scale construction.

Table S2. *Descriptive statistics and reliability of criterion variables*

	FC1				FC2				FC3				FC4			
	N	Mean	SD	α	N	Mean	SD	α	N	Mean	SD	α	N	Mean	SD	α
MAC	535	2.01	0.83	.80	526	2.07	0.83	.79	540	2.06	0.82	.79	531	2.07	0.85	.82
PSYC	540	2.21	0.84	.77	522	2.17	0.75	.67	532	2.25	0.76	.70	527	2.24	0.79	.70
NARC	536	2.40	0.87	.80	517	2.42	0.89	.79	535	2.47	0.83	.75	533	2.40	0.87	.79
OCB	534	2.74	0.75	.86	520	2.78	0.73	.86	533	2.81	0.69	.83	528	2.74	0.72	.85
CWB	536	1.57	0.49	.80	521	1.59	0.50	.82	541	1.62	0.52	.83	525	1.61	0.55	.85
JP	533	4.22	0.50	.80	521	4.17	0.51	.79	536	4.11	0.53	.81	532	4.14	0.53	.82
JS	534	3.39	0.77	.88	524	3.36	0.72	.86	534	3.34	0.72	.87	530	3.33	0.68	.85
BNT	536	2.85	0.81	.88	524	2.85	0.87	.90	538	2.84	0.83	.90	531	2.83	0.83	.89
TI	540	2.94	0.76	.80	527	2.95	0.76	.82	539	2.95	0.69	.77	530	2.97	0.71	.78
SWB	539	3.05	0.90	.88	528	2.97	0.96	.90	540	3.03	0.84	.87	533	3.07	0.88	.89
FS	541	3.18	0.99	.88	526	3.14	0.99	.88	541	3.13	0.88	.84	531	3.10	0.89	.85
PHQ	530	2.98	0.92	.86	519	3.01	0.88	.85	531	2.97	0.92	.86	526	2.97	0.95	.87
EDU	534	3.77	0.98	-	525	3.71	0.92	-	541	3.84	1.05	-	533	3.77	1.11	-
WAG	541	3.92	2.21	-	528	3.85	2.21	-	543	3.15	1.68	-	535	3.13	1.78	-
ORG	541	1.67	1.62	.79	528	1.64	1.66	.80	543	1.51	1.48	.74	535	1.35	1.47	.76
CHAR	541	1.29	0.93	.40	528	1.35	0.93	.42	543	1.38	0.90	.32	535	1.37	0.86	.29
GEN	536	0.49	0.50	-	523	0.53	0.50	-	541	0.49	0.50	-	531	0.51	0.50	-
AGE	541	33.18	9.89	-	528	32.93	10.29	-	543	32.71	9.59	-	535	32.23	9.92	-
TEN	541	12.03	10.59	-	528	11.45	10.19	-	543	11.30	9.72	-	535	10.57	9.73	-

Note. MAC = Machiavellianism; PSYC = Psychopathy; NARC = Narcissism; CWB = Counterproductive work behavior; JS = Job satisfaction; BNT = Burnout; FS = Financial security; OCB = Organizational citizenship behavior; SWB = Subjective well-being; TI = Turnover intentions; JP = Job performance; PHQ = Physical health; EDU = Education; WAG = Wage; ORG = Organizational status; CHAR = Charity behaviors; GEN = Gender, TEN = Tenure.

Table S3. *Discriminant validity (Time 2)*

Trait Pair	**Time 2 Score Correlations**				
	FC1	FC2	FC3	FC4	SS
H-E	.36	.07	.30	.10	-.12
H-X	-.31	.24	-.03	.27	.27
H-A	.34	.13	.25	.41	.42
H-C	.29	.41	.29	.44	.45
H-O	.11	.01	.06	.40	.22
E-X	-.13	-.20	-.28	-.29	-.45
E-A	-.03	-.14	.17	-.32	-.32
E-C	-.02	-.03	.25	-.22	-.24
E-O	.10	.21	.03	-.02	-.14
X-A	-.10	.47	.27	.41	.45
X-C	-.20	.49	.17	.63	.44
X-O	.17	.09	.06	.28	.29
A-C	-.10	.10	.06	.42	.33
A-O	.01	-.02	-.01	.36	.18
C-O	.04	.05	.04	.35	.32
ICC with SS	.02	.68	.28	.91	NA

Table S4. *Raw criterion-related validity at Time 1*

Trait	MAC	PSYC	NARC	OCB	CWB	JP	JS	BNT	TI	SWB	FS	PHQ	EDU	WAG	ORG	CHAR	GEN	AGE	TEN	ICC_SS
H-FC1	-.30	-.11	-.30	-.09	-.10	-.05	-.12	.24	.09	-.17	-.19	.21	-.09	-.12	-.24	-.05	-.12	-.10	-.05	.22
H-FC2	-.24	-.16	-.09	.07	-.16	-.02	.02	-.02	-.04	.04	.04	.02	.07	-.04	.03	.13	-.13	.07	.05	.64
H-FC3	-.38	-.25	-.29	.04	-.24	.08	.02	.06	-.06	.06	-.01	.13	.05	-.04	-.10	.08	-.23	.12	.08	.87
H-FC4	-.42	-.29	-.37	.08	-.21	.17	.05	.05	-.07	.04	.03	.02	.08	.04	.00	.09	-.12	.16	.13	.94
H-SS	-.52	-.39	-.34	.09	-.30	.17	.11	-.08	-.15	.14	.09	-.03	.07	.03	.00	.06	-.12	.18	.14	NA
E-FC1	-.08	-.15	.03	-.09	.07	-.08	-.07	.33	.13	-.10	-.18	.34	-.01	-.11	-.19	-.08	-.30	-.14	-.13	.96
E-FC2	-.05	-.22	.02	-.13	.03	-.04	-.04	.26	.06	-.12	-.10	.24	-.07	-.07	-.12	-.04	-.38	.00	.00	.93
E-FC3	-.25	-.35	-.10	-.10	-.23	.13	.01	.23	.02	-.06	-.11	.21	-.02	-.13	-.17	.01	-.45	-.05	-.02	.81
E-FC4	-.05	-.21	.04	-.02	.07	-.06	-.12	.37	.15	-.22	-.13	.26	-.04	-.22	-.06	.00	-.35	-.13	-.11	.96
E-SS	-.07	-.25	.03	-.08	.01	-.01	-.09	.36	.14	-.16	-.16	.33	-.04	-.14	-.16	-.05	-.39	-.10	-.08	NA
X-FC1	.13	.02	.26	.14	.07	-.03	.10	-.15	-.08	.19	.14	-.10	.04	.06	.19	.15	.08	.04	.04	.55
X-FC2	.10	-.03	.25	.29	.01	.01	.30	-.39	-.30	.48	.26	-.21	.15	.12	.26	.20	.12	-.02	-.02	.92
X-FC3	.11	-.03	.21	.22	.04	.01	.16	-.25	-.14	.33	.11	-.09	.02	.02	.13	.07	.07	.02	.03	.73
X-FC4	-.07	-.16	.04	.19	-.18	.19	.27	-.36	-.23	.48	.19	-.24	.07	.10	.18	.18	.03	.13	.10	.97
X-SS	-.01	-.16	.14	.23	-.12	.16	.30	-.44	-.30	.53	.28	-.30	.12	.16	.23	.14	.07	.16	.15	NA
A-FC1	-.17	-.27	-.20	-.01	-.17	-.08	.15	-.07	-.14	.05	-.01	-.01	-.07	-.04	-.11	-.01	-.01	-.04	-.03	.68
A-FC2	-.07	-.18	.00	.20	-.09	-.01	.23	-.27	-.21	.29	.08	-.17	.04	.07	.12	.10	.09	-.06	-.06	.80
A-FC3	-.08	-.20	-.08	.09	-.10	-.06	.10	-.01	-.06	.09	-.03	.06	.00	-.12	.00	.14	-.01	-.11	-.13	.50
A-FC4	-.29	-.29	-.24	.06	-.24	.14	.13	-.19	-.14	.20	.11	-.16	.06	.04	.00	.02	.09	-.01	-.02	.96
A-SS	-.27	-.33	-.21	.14	-.26	.11	.20	-.22	-.20	.21	.07	-.15	.00	.03	.07	.10	.07	.03	.01	NA
C-FC1	-.06	.06	-.05	.02	.01	.15	-.03	.15	.11	-.15	-.01	.09	.09	.01	.03	.01	-.10	-.07	-.06	-.04
C-FC2	-.03	-.03	.09	.20	-.12	.13	.14	-.21	-.16	.23	.21	-.08	.17	.08	.21	.14	.01	.05	.02	.69
C-FC3	-.17	-.16	-.07	.16	-.18	.22	.06	.04	-.06	.05	.03	.11	.06	.00	-.01	.12	-.19	.06	.08	.67
C-FC4	-.16	-.19	-.09	.24	-.23	.30	.16	-.22	-.16	.23	.16	-.11	.11	.16	.19	.17	-.01	.13	.09	.94
C-SS	-.24	-.25	-.12	.18	-.25	.44	.16	-.18	-.17	.18	.20	-.13	.12	.15	.14	.06	-.12	.16	.14	NA
O-FC1	.09	.17	.10	-.01	.07	-.02	-.14	.13	.15	-.11	-.13	.09	.01	-.09	.00	.07	.03	-.04	-.02	.08
O-FC2	.09	.05	.02	.00	-.02	-.05	-.03	.05	.04	-.05	-.08	.08	.08	-.03	.03	.10	-.07	.00	.00	.40
O-FC3	-.06	-.09	-.06	.04	-.02	.00	-.06	.05	.16	-.06	-.12	.06	.14	-.03	.06	.10	-.02	-.02	-.05	.68
O-FC4	-.06	-.11	-.04	.14	-.06	.05	.03	-.03	-.01	.03	.02	.02	.18	.07	.13	.14	-.05	.02	.04	.71
O-SS	-.02	-.04	-.01	.08	-.06	.12	-.02	.03	.06	-.02	-.04	.03	.11	-.01	.08	.12	-.04	.01	.02	NA

Note. Overall ICC(SS, FC1) = .51, overall ICC(SS, FC2) = .82, overall ICC(SS, FC3) = .75, overall ICC(SS, FC4) = .95. H = Honesty-humility, E = Emotionality, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. FC = Forced-choice measure, SS = Single-statement measure. MAC = Machiavellianism, PSYC = Psychopathy, NARC = Narcissism, CWB = Counterproductive work behavior, JS = Job satisfaction, BNT = Burnout, FS = Financial security, OCB = Organizational citizenship behavior, SWB = Subjective well-being, TI = Turnover intentions, JP = Job performance, PHQ = Physical health, EDU = Education, WAG = Wage, ORG = Organizational status, CHAR = Charity behaviors, GEN = Gender, TEN = Tenure.

Table S5. *Criterion-related validity at Time 1, corrected for reliability in personality traits*

Trait	MAC	PSYC	NARC	OCB	CWB	JP	JS	BNT	TI	SWB	FS	PHQ	EDU	WAG	ORG	CHAR	GEN	AGE	TEN	ICC_SS
H-FC1	-.36	-.13	-.36	-.11	-.12	-.06	-.14	.28	.11	-.20	-.23	.26	-.11	-.15	-.29	-.06	-.14	-.12	-.06	.23
H-FC2	-.28	-.19	-.11	.08	-.19	-.03	.02	-.02	-.05	.05	.05	.03	.08	-.05	.04	.15	-.16	.09	.05	.67
H-FC3	-.45	-.30	-.34	.05	-.28	.09	.02	.07	-.07	.07	-.01	.15	.06	-.05	-.12	.10	-.27	.14	.10	.88
H-FC4	-.51	-.35	-.45	.09	-.26	.21	.07	.06	-.09	.05	.03	.03	.10	.04	.00	.10	-.14	.19	.15	.96
H-SS	-.56	-.42	-.37	.10	-.32	.19	.12	-.08	-.17	.16	.09	-.03	.07	.03	.00	.07	-.13	.19	.15	NA
E-FC1	-.10	-.17	.03	-.10	.09	-.09	-.09	.39	.16	-.12	-.21	.39	-.02	-.13	-.22	-.09	-.35	-.17	-.15	.97
E-FC2	-.06	-.25	.03	-.15	.04	-.05	-.04	.30	.07	-.14	-.12	.27	-.08	-.08	-.14	-.04	-.44	.00	.00	.94
E-FC3	-.30	-.43	-.13	-.12	-.29	.15	.01	.28	.02	-.07	-.14	.26	-.03	-.16	-.21	.01	-.55	-.07	-.02	.79
E-FC4	-.06	-.24	.05	-.03	.09	-.07	-.14	.43	.17	-.26	-.15	.30	-.04	-.25	-.06	.00	-.40	-.15	-.12	.96
E-SS	-.08	-.27	.03	-.09	.01	-.01	-.10	.40	.16	-.17	-.18	.36	-.04	-.15	-.18	-.06	-.43	-.11	-.09	NA
X-FC1	.16	.03	.31	.17	.08	-.03	.13	-.18	-.10	.24	.18	-.13	.05	.08	.23	.19	.09	.05	.05	.60
X-FC2	.11	-.03	.28	.32	.02	.01	.34	-.44	-.34	.54	.29	-.24	.17	.14	.30	.22	.14	-.02	-.03	.92
X-FC3	.13	-.03	.24	.25	.04	.01	.19	-.29	-.16	.38	.12	-.10	.02	.02	.15	.08	.08	.03	.03	.75
X-FC4	-.08	-.18	.04	.21	-.20	.21	.30	-.40	-.26	.54	.21	-.27	.08	.11	.21	.21	.03	.15	.12	.97
X-SS	-.01	-.17	.15	.25	-.13	.17	.32	-.47	-.32	.57	.31	-.32	.13	.17	.25	.15	.07	.17	.16	NA
A-FC1	-.21	-.33	-.25	-.01	-.21	-.10	.19	-.08	-.17	.06	-.02	-.02	-.08	-.05	-.14	-.01	-.01	-.04	-.04	.71
A-FC2	-.08	-.21	.00	.24	-.11	-.01	.28	-.33	-.25	.35	.09	-.21	.04	.08	.14	.12	.11	-.08	-.07	.81
A-FC3	-.10	-.24	-.10	.11	-.12	-.07	.12	-.02	-.08	.11	-.03	.07	.00	-.15	.00	.18	-.01	-.13	-.16	.52
A-FC4	-.35	-.35	-.29	.07	-.29	.17	.16	-.23	-.17	.25	.14	-.19	.07	.05	.00	.03	.11	-.01	-.03	.96
A-SS	-.30	-.36	-.23	.15	-.28	.13	.23	-.24	-.22	.23	.08	-.16	.00	.03	.07	.11	.07	.03	.01	NA
C-FC1	-.08	.07	-.06	.02	.01	.18	-.03	.18	.13	-.19	-.01	.11	.11	.01	.03	.01	-.13	-.09	-.08	-.04
C-FC2	-.03	-.04	.11	.23	-.15	.16	.16	-.25	-.19	.27	.25	-.10	.20	.09	.25	.16	.01	.06	.03	.70
C-FC3	-.21	-.21	-.09	.19	-.23	.27	.07	.05	-.08	.06	.04	.14	.08	.01	-.01	.15	-.24	.08	.10	.72
C-FC4	-.19	-.22	-.11	.28	-.28	.36	.18	-.26	-.19	.27	.19	-.13	.12	.18	.22	.20	-.01	.15	.10	.95
C-SS	-.26	-.27	-.13	.20	-.27	.47	.17	-.20	-.18	.20	.22	-.14	.12	.16	.16	.06	-.13	.17	.15	NA
O-FC1	.11	.20	.12	-.02	.08	-.03	-.17	.15	.17	-.13	-.16	.10	.01	-.10	.00	.08	.03	-.05	-.03	.08
O-FC2	.11	.07	.02	.00	-.03	-.07	-.03	.06	.04	-.06	-.09	.09	.09	-.04	.04	.12	-.08	-.01	.00	.40
O-FC3	-.08	-.12	-.08	.04	-.03	.00	-.08	.06	.19	-.07	-.15	.07	.17	-.03	.07	.13	-.02	-.02	-.06	.66
O-FC4	-.07	-.13	-.05	.16	-.07	.06	.03	-.04	-.01	.04	.03	.02	.22	.08	.16	.17	-.06	.03	.05	.68
O-SS	-.02	-.05	-.01	.08	-.07	.13	-.02	.03	.07	-.02	-.05	.03	.12	-.01	.09	.13	-.04	.01	.03	NA

Note. Overall ICC(SS, FC1) = .54, overall ICC(SS, FC2) = .83, overall ICC(SS, FC3) = .77, overall ICC(SS, FC4) = .96. H = Honesty-Humility, E = Emotionality, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. FC = Forced-choice measure, SS = Single-statement measure. MAC = Machiavellianism, PSYC = Psychopathy, NARC = Narcissism, CWB = Counterproductive work behavior, JS = Job satisfaction, BNT = Burnout, FS = Financial security, OCB = Organizational citizenship behavior, SWB = Subjective well-being, TI = Turnover intentions, JP = Job performance, PHQ = Physical health, EDU = Education, WAG = Wage, ORG = Organizational status, CHAR = Charity behaviors, GEN = Gender, TEN = Tenure.

Table S6. *Uncorrected multiple R^2 of HEXACO traits predicting criterion variables*

Criterion Variable	FC1	FC2	FC3	FC4	SS
Machiavellianism	.119	.101	.178	.219	.310
Psychopathy	.121	.115	.198	.211	.350
Narcissism	.180	.113	.130	.187	.185
Organizational citizenship behaviors	.031	.097	.084	.073	.078
Counterproductive work behavior	.052	.059	.110	.114	.153
Job performance	.038	.030	.060	.108	.206
Job satisfaction	.076	.108	.045	.085	.120
Burnout	.152	.194	.093	.229	.253
Turnover intentions	.067	.102	.064	.080	.133
Subjective well-being	.077	.241	.133	.271	.304
Financial security	.080	.089	.038	.055	.114
Physical health	.127	.092	.059	.118	.150
Education	.025	.041	.025	.042	.034
Wage	.030	.023	.030	.070	.047
Organizational status	.100	.081	.043	.067	.073
Charity behaviors	.033	.056	.050	.056	.037
Organizational tenure	.020	.006	.040	.052	.053
Gender	.100	.163	.232	.129	.183
Age	.026	.012	.042	.075	.071
Mean R^2	.077	.091	.087	.118	.150

Note. Multiple R^2 calculated based on raw data.

Examining Faking Resistance Using Hetero-method Rank-Order Stability

We additionally reported correlations between Time 1 SS scores and Time 2 FC scores in Table S7. A more faking-resistant measure should ideally also maintain respondents' rank orders consistently across both honest and fake-good conditions, as well as across different formats. As shown in Table S7, these hetero-method rank-order stability estimates were smaller than the rank-order stability for each FC measure in Table 6 from the main text. The average corrected hetero-method correlations were .67, .64, .61. and .57 for FC3, FC2, FC4, and FC1, respectively. We note that FC1 might have had the lowest hetero-method correlations across time mainly because of its relatively poor construct validity, while the low correlations for FC4 might have been mainly due to its lowest faking resistance.

Table S7. *Correlation between Time 1 SS scores and Time 2 FC scores*

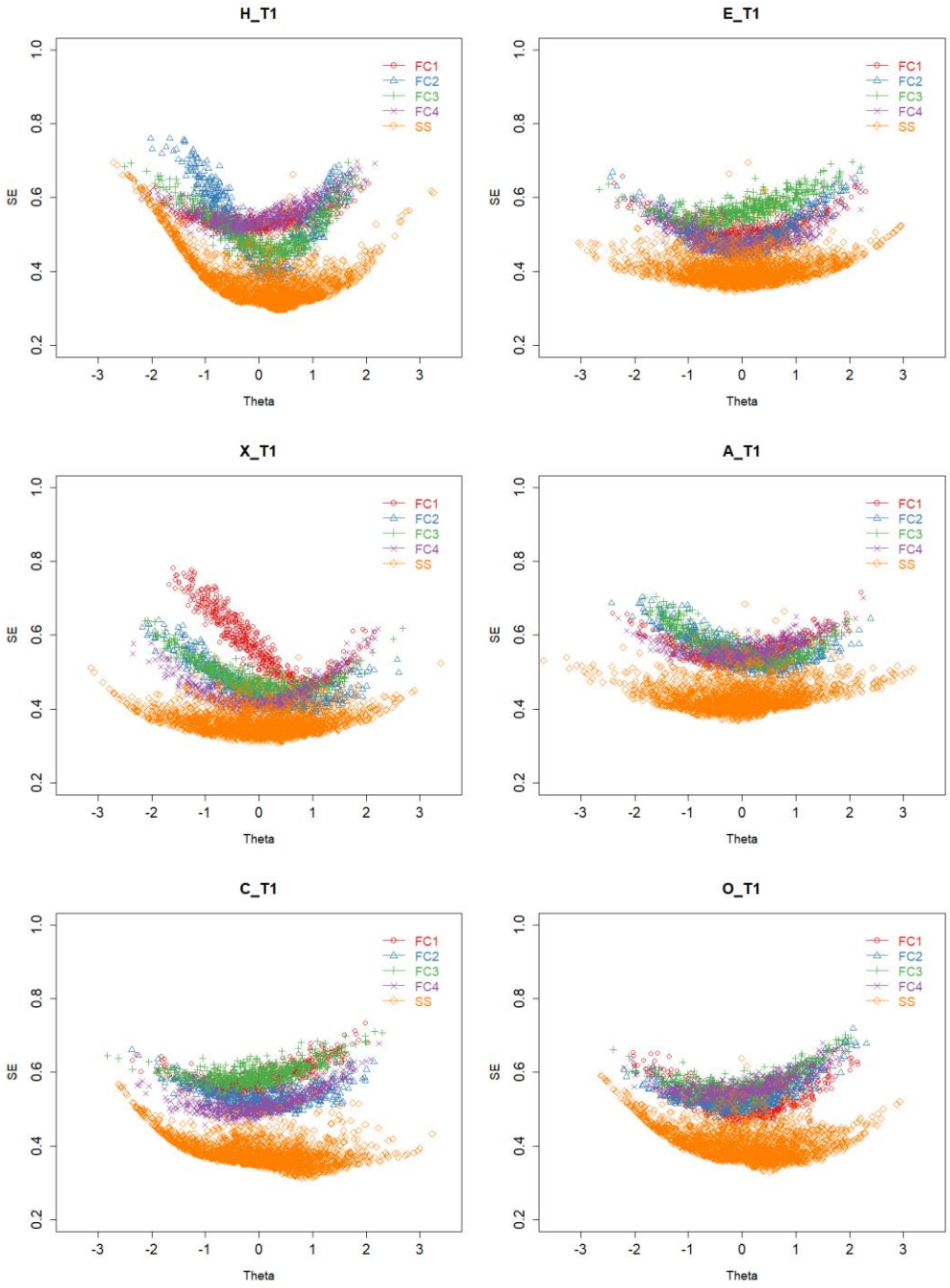
	Uncorrected				Corrected for unreliability			
	FC1	FC2	FC3	FC4	FC1	FC2	FC3	FC4
H	.42	.48	.53	.45	.54	.62	.67	.58
E (Reverse)	.50	.49	.53	.51	.64	.62	.72	.64
X	.44	.50	.52	.48	.57	.60	.64	.58
A	.43	.54	.37	.43	.57	.71	.52	.59
C	.24	.45	.45	.36	.33	.58	.61	.46
O	.58	.56	.65	.60	.74	.73	.87	.81
Mean	.44	.50	.51	.47	.57	.64	.67	.61

Note. H = Honesty-Humility, E = Emotionality, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. FC = Forced-choice measure, SS = Single-statement measure.

Examining Reliability Using Standard Error Plots

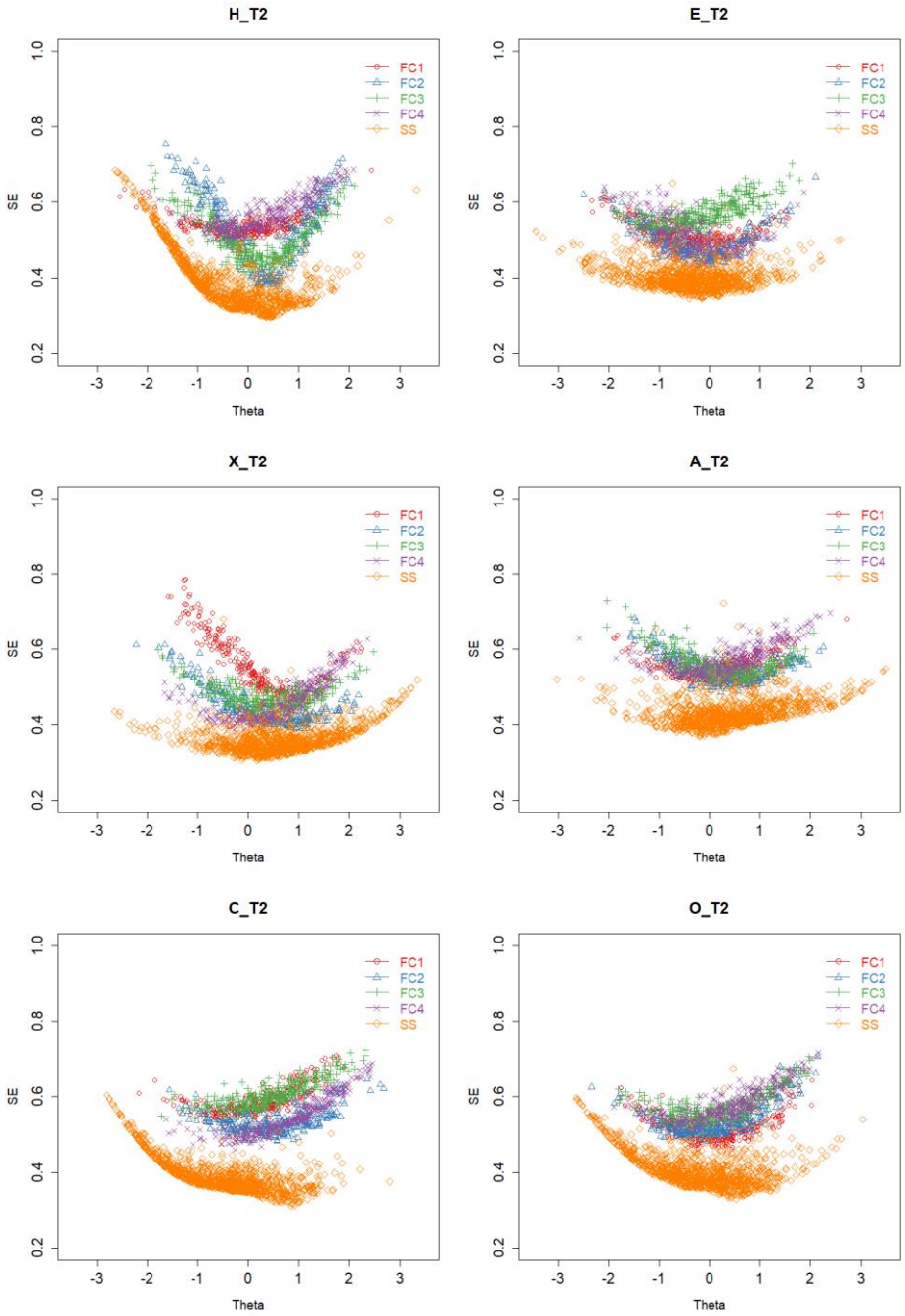
It is well known that empirical reliability is just a general estimate of reliability at the entire sample level and does not capture the measurement precision at the individual level, which varies across different levels of the latent trait within the IRT framework. To gain a deeper understanding of measurement precision at the individual level, we plotted the Time 1 standard error of measurement for each person against their estimated latent trait levels in Figure S1. As can be seen, the differences in standard error were largely minimal across the four FC measures, but two noticeable exceptions appeared for Honesty-Humility and Extraversion continuum. For Honesty-Humility, FC2 and FC3 achieved higher measurement accuracy than FC1 and FC4 within the theta range of (0,1), but FC2 fared worse in accuracy than the other three FC measures along the negative range of the Honesty-Humility trait continuum. For Extraversion, the standard errors for FC1 were remarkably higher than the other three FC measures along the negative side of trait Extraversion. Consistent across traits, the standard errors produced by the four FC measures were consistently larger than those produced by the SS measure. These same patterns were also observed at Time 2 (see Figure S2).

Figure S1. *Standard error of measurement for FC and SS (Time 1)*



Note. Time 1 conditional standard errors for person score estimates obtained from FC1-FC4 and SS are presented. FC = Forced-Choice; SS = Single-Statement; H = Honesty-Humility; E = Emotionality; X = Extraversion; A = Agreeableness; C = Conscientiousness; O = Openness. Red dots represent conditional SEs from FC1; Blue dots represent conditional SEs from FC2; Green dots represent conditional SEs from FC3; Purple dots represent conditional SEs from FC4; Orange dots represent conditional SEs from SS.

Figure S2. *Standard error of measurement for FC and SS (Time 2)*

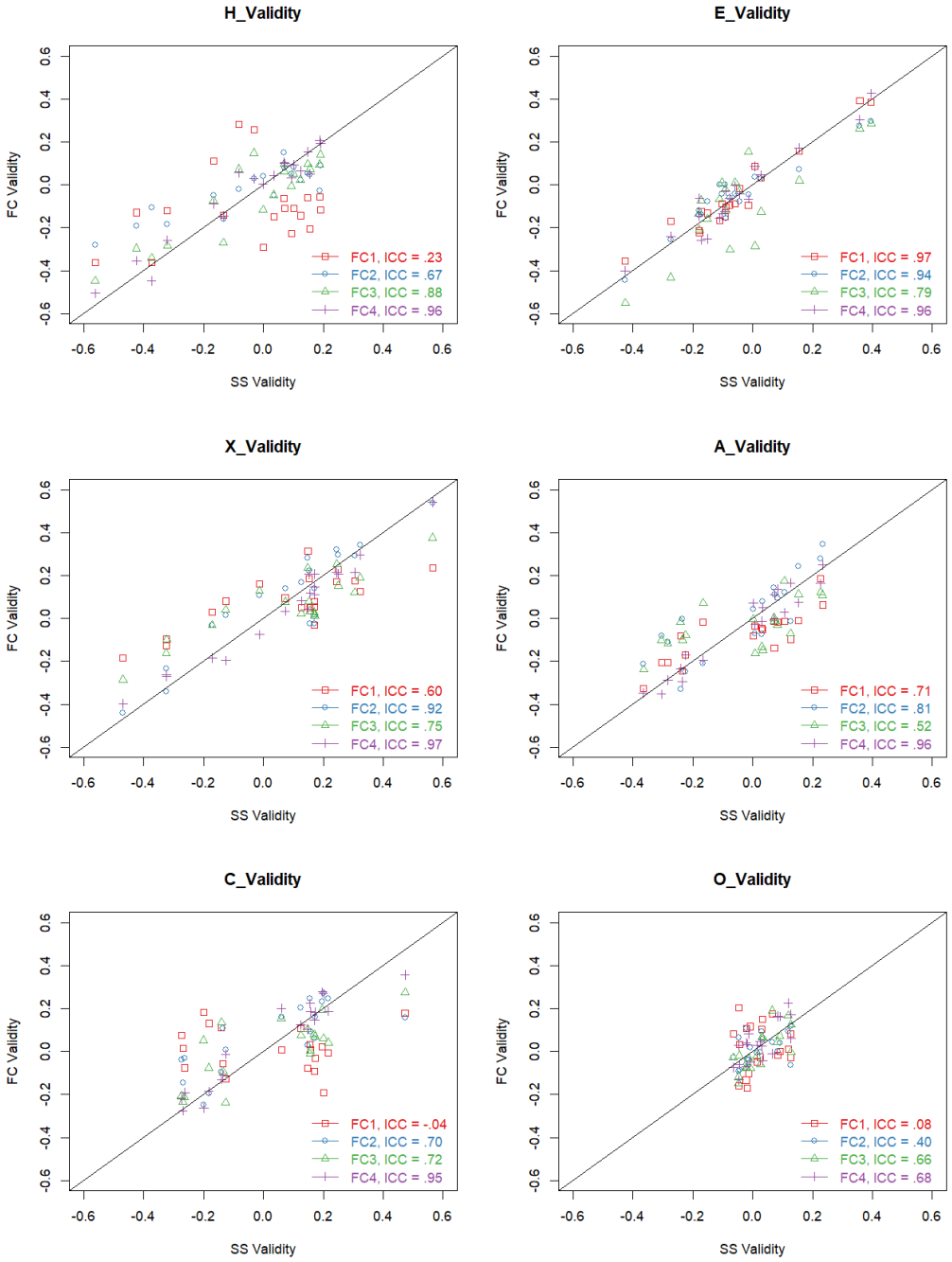


Note. Time 2 conditional standard errors for person score estimates obtained from FC1-FC4 and SS are presented. FC = Forced-Choice; SS = Single-Statement; H = Honesty-Humility; E = Emotionality; X = Extraversion; A = Agreeableness; C = Conscientiousness; O = Openness. Red dots represent conditional SEs from FC1; Blue dots represent conditional SEs from FC2; Green dots represent conditional SEs from FC3; Purple dots represent conditional SEs from FC4; Orange dots represent conditional SEs from SS.

Examining Criterion-Related Validity Using Validity Plots

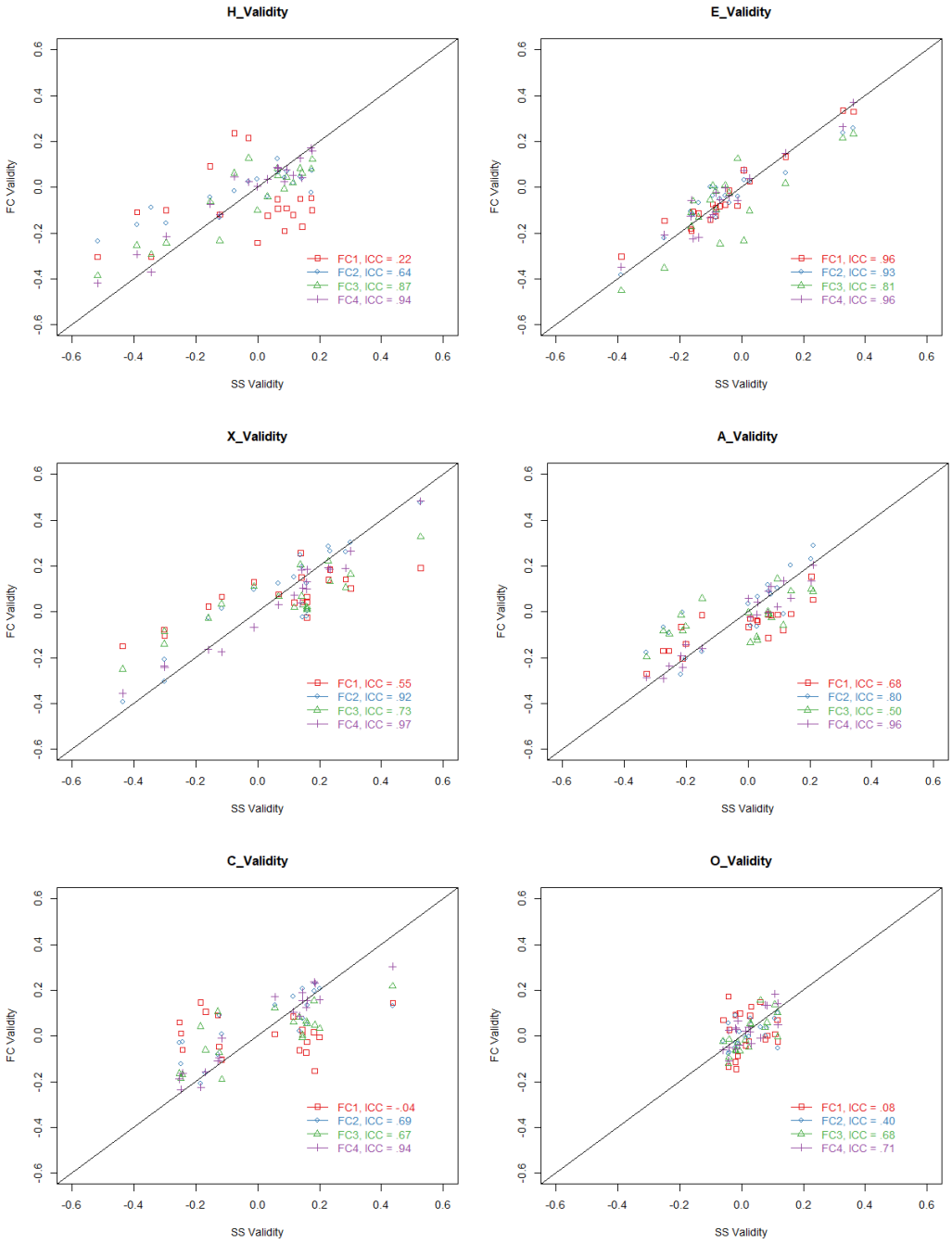
We plotted all validity coefficients of four FC measures against the corresponding SS validity coefficients in Figure S3. In these plots, dots closer to the $y = x$ regression line implied that the corresponding validity coefficients of FC and SS were more similar to each other. Consistent with Table 6 in the main text, the validity coefficients for FC4 were often the closest to those produced by SS, while those from FC1 were the farthest apart. Also, none of the validity coefficients for Openness were larger than .20 in terms of magnitude, which further explained why the ICCs for Openness were consistently low. We also note that plots based on raw validity coefficients (Figure S4) or based on validity coefficients corrected for reliability in personality traits (Figure S3) produced the same pattern.

Figure S3. *Validity coefficients of FC and SS, corrected for reliability in personality traits*



Note. The correlations (after correcting for empirical reliability in personality scores) with each criterion variable for FC1-FC4 (Y axis) and for SS (X axis) are presented. FC = Forced-Choice; SS = Single-Statement; H = Honesty-Humility; E = Emotionality; X = Extraversion; A = Agreeableness; C = Conscientiousness; O = Openness; ICC = Intra-class correlation between the FC validity profile (vector of correlation with criterion variables) and SS validity profile. Red dots represent correlations for FC1; Blue dots represent correlations for FC2; Green dots represent correlations for FC3; Purple dots represent correlations for FC4.

Figure S4. *Raw validity coefficients of FC and SS*



Note. The raw correlations with each criterion variable for FC1-FC4 (Y axis) and for SS (X axis) are presented. FC = Forced-Choice; SS = Single-Statement; H = Honesty-Humility; E = Emotionality; X = Extraversion; A = Agreeableness; C = Conscientiousness; O = Openness; ICC = Intra-class correlation between the FC validity profile (vector of correlation with criterion variables) and SS validity profile. Red dots represent correlations for FC1; Blue dots represent correlations for FC2; Green dots represent correlations for FC3; Purple dots represent correlations for FC4.

Section 3: Correcting SS responses for response biases

As has been discussed in the introduction section, SS measures are often plagued with various response biases that will likely distort reliability and validity estimates. In light of this problem, we attempted modeling techniques capable of accounting for three response biases in SS measures: acquiescence, extreme responding, and midpoint responding (Plieninger & Heck, 2018). This model takes three response biases (midpoint responding, extreme responding, and acquiescence responding) into account and produces trait scores after correcting for these response biases. In this section, we report the psychometric performance of the original SS trait scores verses the SS trait scores corrected for response biases.

Table S8. *Original and corrected SS empirical reliability*

	Time 1		Time 2	
Trait	SS-Original	SS-Corrected	SS-Original	SS-Corrected
H	.85	.66	.83	.61
E	.83	.73	.81	.70
X	.87	.77	.87	.75
A	.81	.70	.82	.68
C	.85	.65	.85	.58
O	.83	.70	.84	.69
Mean	.84	.70	.84	.67

Note. H = Honesty-Humility, E = Emotionality, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. SS-Original = Original trait scores from Single-Statement measure; SS-Corrected = Trait scores from Single-Statement measure, corrected for response biases.

Table S9. *FC and SS convergent validity, with SS scores corrected for response biases*

	Uncorrected								Corrected for unreliability							
	Time 1 = Honest				Time 2 = Faking				Time 1 = Honest				Time 2 = Faking			
	FC1	FC2	FC3	FC4	FC1	FC2	FC3	FC4	FC1	FC2	FC3	FC4	FC1	FC2	FC3	FC4
H	.50	.52	.64	.66	.46	.30	.47	.52	.73	.76	.92	.98	.72	.46	.69	.83
E (Reverse)	.79	.78	.74	.76	.68	.61	.55	.53	1.00	1.00	1.00	1.00	.97	.88	.84	.77
X	.58	.79	.75	.84	.50	.56	.64	.61	.81	1.00	.98	1.00	.69	.73	.87	.81
A	.67	.68	.62	.73	.57	.57	.53	.59	.95	.98	.91	1.00	.86	.84	.81	.89
C	.48	.57	.68	.73	.33	.38	.54	.64	.76	.83	1.00	1.00	.58	.60	.90	.99
O	.72	.68	.69	.75	.55	.49	.62	.66	1.00	.97	1.00	1.00	.79	.72	.90	.97
Mean	.62	.67	.69	.75	.52	.49	.56	.59	.88	.92	.97	1.00	.77	.71	.84	.88

Note. H = Honesty-Humility, E = Emotionality, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. FC = Forced-choice measure. Time 1 matched sample size: N_{FC1} = 541, N_{FC2} = 528, N_{FC3} = 543, N_{FC4} = 535. Time 2 matched sample size: N_{FC1} = 289, N_{FC1} = 283, N_{FC1} = 302, N_{FC1} = 303. Corrected convergent validity estimates larger than 1.0 were set to 1.0.

Table S10. *Discriminant validity at Time 1, with SS scores corrected for response biases*

Trait Pair	**Time 1 Latent Correlations**			
	SS-Original		SS-Corrected	
	r	SE	*r*	SE
H-E	.056	.056	.034	.031
H-X	.023	.051	.006	.031
H-A	.230***	.044	.238***	.031
H-C	.216***	.029	.193***	.033
H-O	.015	.051	.079*	.032
E-X	-.337***	.044	-.268***	.028
E-A	-.188***	.039	-.165***	.030
E-C	-.078	.050	-.013	.031
E-O	.003	.054	-.013	.031
X-A	.235***	.048	.201***	.029
X-C	.204***	.049	.177***	.031
X-O	.070	.055	.112***	.029
A-C	.111**	.036	.018	.034
A-O	-.004	.028	.007	.031
C-O	.102*	.052	.127***	.031
Mean Absolute Correlation & SE	.125 (.046)		.110 (031)	

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. SS-Original = Original trait scores from Single-statement measure; SS-Corrected = Trait scores from Single-Statement measure, corrected for response biases.

Table S11. *Double entry ICCs of validity profile between FC and SS measures, corrected for reliability of personality traits and with SS scores corrected for response biases*

	Double-Entry ICC			
Trait	FC1-SS	FC2-SS	FC3-SS	FC4-SS
H	.25	.65	.86	.95
E	.95	.95	.83	.95
X	.64	.94	.78	.97
A	.71	.78	.47	.97
C	-.08	.74	.73	.96
O	-.03	.39	.62	.79

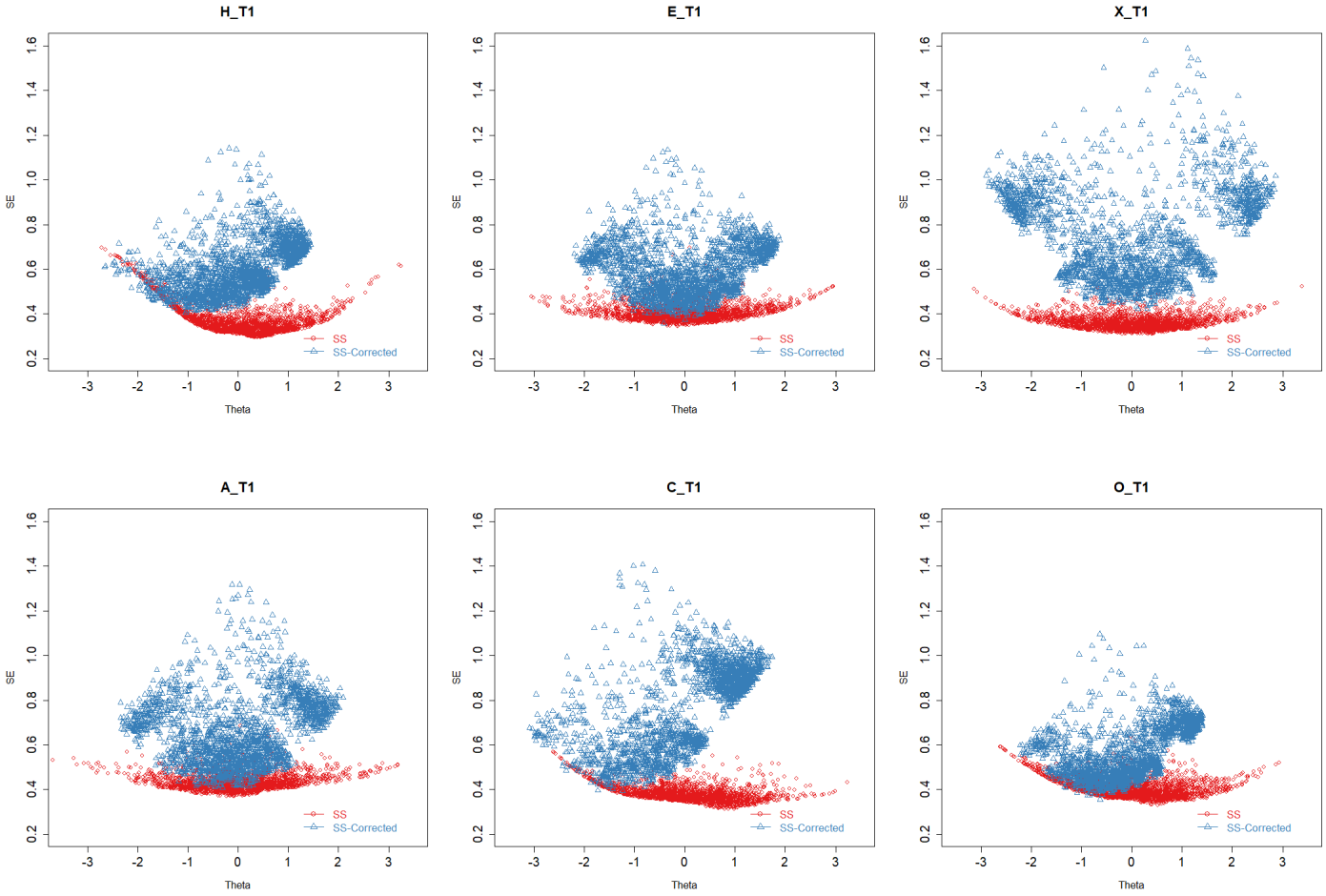
Note. Overall ICC(SS, FC1) = .54, overall ICC(SS, FC2) = .83, overall ICC(SS, FC3) = .78, overall ICC(SS, FC4) = .96. H = Honesty-Humility, E = Emotionality, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness. FC = Forced-choice measure; SS = Single-statement measure, scores corrected for response biases.

Table S12. *Corrected Multiple R^2 of HEXACO Traits Predicting Criterion Variables, with SS scores corrected for response biases*

Criterion Variable	SS-Original	SS-Corrected
MAC	.377	.416
PSYC	.421	.369
NARC	.217	.266
OCB	.096	.102
CWB	.191	.214
JP	.248	.169
JS	.145	.133
BNT	.311	.318
TI	.162	.153
SWB	.353	.357
FS	.135	.137
PHQ	.186	.191
EDU	.041	.068
WAG	.057	.059
ORG	.089	.107
CHAR	.045	.053
GEN	.217	.242
AGE	.083	.109
TEN	.061	.078
Mean R^2	.181	.186

Note. Multiple R^2 was calculated based on correlations between personality traits and criterion variables. MAC = Machiavellianism, PSYC = Psychopathy, NARC = Narcissism, CWB = Counterproductive work behavior, JS = Job satisfaction, BNT = Burnout, FS = Financial Security, OCB = Organizational citizenship behavior, SWB = Subjective well-being, TI = Turnover intentions, JP = Job performance, PHQ = Physical health, EDU = Education, WAG = Wage, ORG = Organizational status, CHAR = Charity behaviors. SS-Original = Original trait scores from Single-Statement measure; SS-Corrected = Trait scores from Single-Statement measure, corrected for response biases.

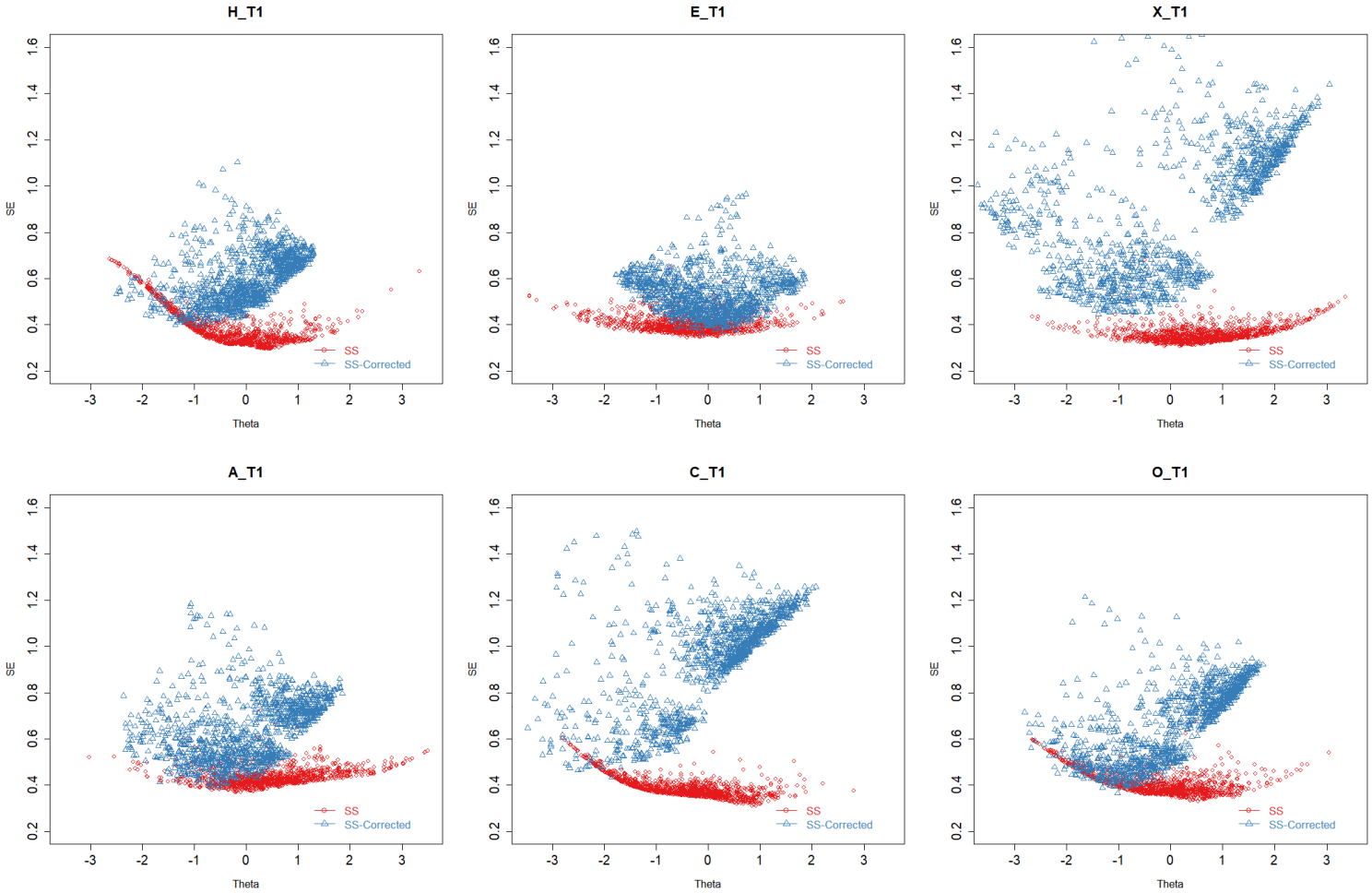
Figure S5. *Standard errors of measurement for FC and SS, with SS scores corrected for response biases (Time 1)*



Note. Time 1 conditional standard errors for person score estimates obtained from SS measures are presented. SS = Raw Single-Statement scores; SS-Corrected = Single-Statement scores, corrected for response biases. H = Honesty-Humility; E = Emotionality; X = Extraversion; A =

Agreeableness; C = Conscientiousness; O = Openness. Red dots represent conditional SEs from SS; Blue dots represent conditional SEs corrected SS.

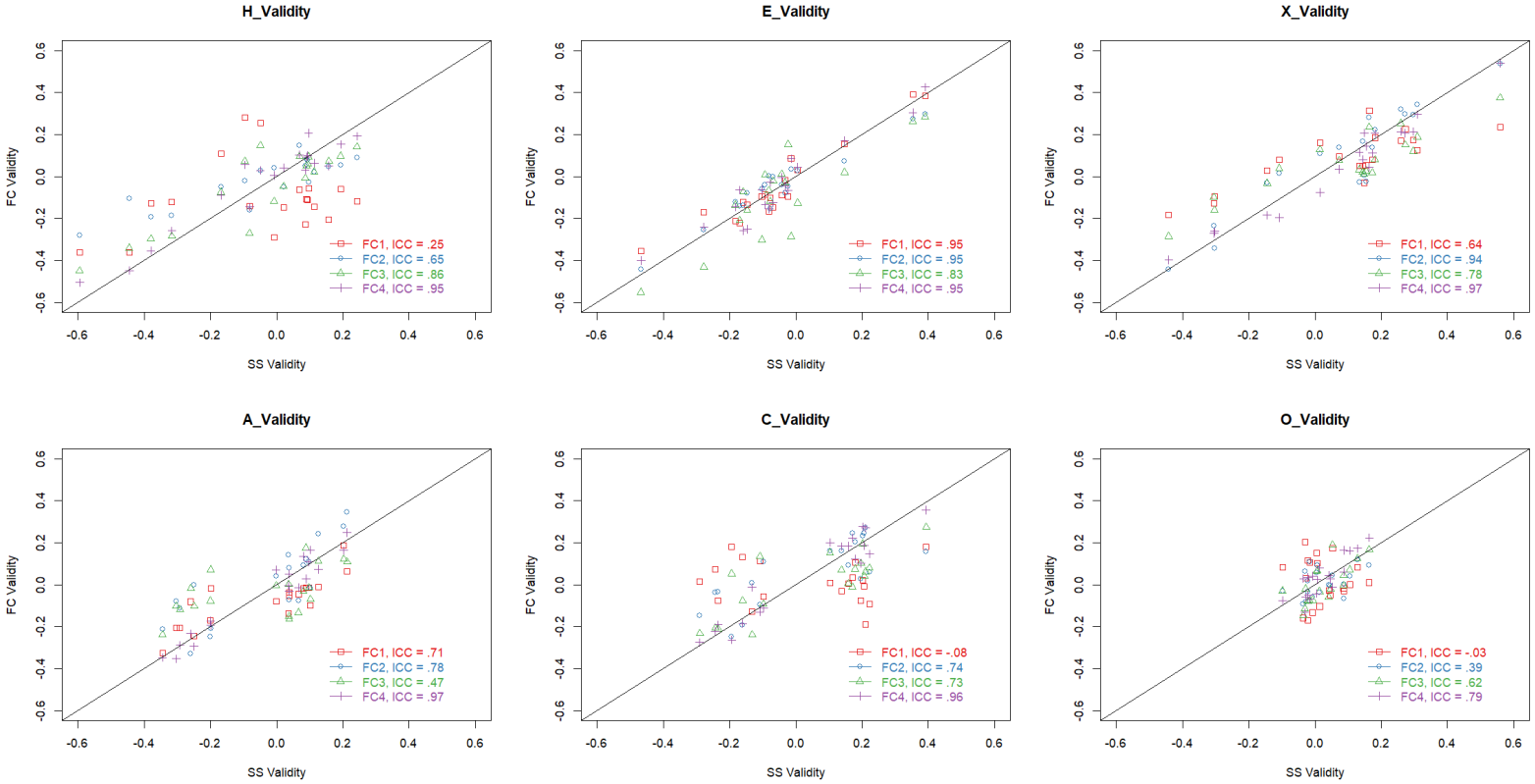
Figure S6. *Standard errors of measurement for FC and SS, with SS scores corrected for response biases (Time 2)*



Note. Time 1 conditional standard errors for person score estimates obtained from SS measures are presented. SS = Raw Single-Statement scores; SS-Corrected = Single-Statement scores, corrected for response biases. H = Honesty-Humility; E = Emotionality; X = Extraversion; A =

Agreeableness; C = Conscientiousness; O = Openness. Red dots represent conditional SEs from SS; Blue dots represent conditional SEs corrected SS.

Figure S7. *Corrected validity coefficients of FC and SS, with SS scores corrected for response biases*



Note. The correlations (after correcting for empirical reliability in personality scores) with each criterion variable for FC1-FC4 (Y axis) and for SS corrected for response biases (X axis) are presented. FC = Forced-Choice; SS = Single-Statement, with response biases corrected; H = Honesty-Humility; E = Emotionality; X = Extraversion; A = Agreeableness; C = Conscientiousness; O = Openness; ICC = Intra-class correlation between the FC validity profile (vector of correlation with criterion variables) and SS validity profile. Red dots represent correlations for FC1; Blue dots represent correlations for FC2; Green dots represent correlations for FC3; Purple dots represent correlations for FC4.

Section 4: List of Criterion Variables and Respondent Reaction Items

1. Organizational Citizenship Behaviors (Fox, Spector, Bruursema, Kessler, & Goh, 2007)

How often have you done the following things at your present job?
1 = Never; 2 = Once or twice; 3 = Once or twice per month; 4 = Once or twice per week; 5 = Every day
1. Purposely wasted your employer's materials/supplies.
2. Complained about insignificant things at work.
3. Told people outside of the job what a lousy place you work for.
4. Played a mean prank to embarrass someone at work.
5. Stayed home from work and said you were sick when you weren't.
6. Insulted someone about their job performance.
7. Made fun of someone's personal life.
8. Ignored someone at work.
9. Insulted or made fun of someone at work.
10. Insulted or made fun of someone at work.

2. Counterproductive Work Behaviors (Spector, Fox, Penney, Bruursema, Goh, & Kessler, 2006)

How often have you done the following things at your present job?
1 = Never; 2 = Once or twice; 3 = Once or twice per month; 4 = Once or twice per week; 5 = Every day
1. Took time to advise, coach, or mentor a co-worker.
2. Helped co-worker learn new skills or shared job knowledge.
3. Helped new employees get oriented to the job.
4. Lent a compassionate ear when someone had a work problem.
5. Offered suggestions to improve how work is done.
6. Helped a co-worker who had too much to do.
7. Volunteered for extra work assignments.
8. Worked weekends or other days off to complete a project or task.
9. Volunteered to attend meetings or work on committees on own time.
10. Gave up meal and other breaks to complete work.

3. Job Performance (Williams & Anderson, 1991)

Please indicate the degree to which you agree with each of the following statements.
1 = *Strongly disagree*; 2 = *Disagree*; 3 = *Neither disagree nor agree*; 4 = *Agree*; 5 = *Strongly agree*
1. I adequately complete assigned duties.

2. I fulfill responsibilities specified in job description.
3. I perform tasks that are expected of myself.
4. I meet formal performance requirements of the job.
5. I engage in activities that will directly affect my performance evaluation.
6. I neglect aspects of the job I am obligated to perform.
7. I fail to perform essential duties.

4. Job Satisfaction (Spector, 1985)

Think of your current job in general, and answer the following questions.
1 = Very dissatisfied; 2 = Dissatisfied; 3 = Neutral; 4 = Satisfied; 5 = Very satisfied
1. All in all, how satisfied are you with the **pay** of your job?
2. All in all, how satisfied are you with the **coworker** of your job?
3. All in all, how satisfied are you with the **supervision** you receive at your work?
4. All in all, how satisfied are you with the **promotion opportunity** of your job?
5. All in all, how satisfied are you with the **benefits** of your job?
6. All in all, how satisfied are you with the **contingent rewards** you receive from your job?
7. All in all, how satisfied are you with the **operating procedures** of your job?
8. All in all, how satisfied are you with the **nature of your work**?
9. All in all, how satisfied are you with the **communication with your organization**?

5. Burnout (Kristensen, Borritz, Villadsen, & Christensen, 2005)

Please indicate how often the following statements describe you or your job.
1 = *Never/Almost never*, 2 = *Seldom*, 3 = *Sometimes*, 4 = *Often*, 5 = *Always*
1. How often do you feel tired?
2. How often are you physically exhausted?
3. How often are you emotionally exhausted?
4. How often do you think: ”I can’t take it anymore”?
5. How often do you feel worn out?
6. How often do you feel weak and susceptible to illness?

6. Turnover Intentions (Roodt, 2004)

For each of the following statements, choose the frequency each statement happens to you, or the extent it describes you.
1. How often do you dream about getting another job that will better suit your personal needs?
1 = *Always*, 2 = *Almost always*, 3 = *Occasionally/sometimes*, 4 = *Almost never*, 5 = *Never*
2. How often are you frustrated when not given the opportunity at work to achieve your personal work-related goals?

1 = *Always*, 2 = *Almost always*, 3 = *Occasionally/sometimes*, 4 = *Almost never*, 5 = *Never*
3. How often have you considered leaving your job?
1 = *Always*, 2 = *Almost always*, 3 = *Occasionally/sometimes*, 4 = *Almost never*, 5 = *Never*
4. How likely are you to accept another job at the same compensation level should it be offered to you?
1 = *Very likely*, 2 = *Likely*, 3 = *Not sure*, 4 = *Unlikely*, 5 = *Very unlikely*
5. To what extent is your current job satisfying your personal needs?
1 = *Not at all*, 2 = *A little bit*, 3 = *Moderately*, 4 = *Fairly well*, 5 = *Completely*
6. How often do you look forward to another day at work?
1 = *Always*, 2 = *Almost always*, 3 = *Occasionally/sometimes*, 4 = *Almost never*, 5 = *Never*

7. Organizational Status

In your current organization, do you have the right to do the following things?
1 = *Yes*, 0 = *No*

1. Hire people
2. Fire people
3. Supervise people
4. Create budgets for the organization
5. Make strategic decision for the company

8. Subjective Well-Being (Diener, Emmons, Larsen, & Griffin, 1986)

For each of the following statements, choose the frequency each statement happens to you, or the extent it describes you.
1 = *Strongly disagree*; 2 = *Disagree*; 3 = *Neither disagree nor agree*; 4 = *Agree*; 5 = *Strongly agree*
1. In most ways my life is close to my ideal.
2. The conditions of my life are excellent.
3. I am satisfied with my life.
4. So far I have gotten the important things I want in life.
5. If I could live my life over, I would change almost nothing.

9. Financial Security (Munyon, Carnes, Lyons, & Zettler, 2020)

Please indicate the degree to which you agree with each of the following statement.
1 = *Strongly disagree*; 2 = *Disagree*; 3 = *Neither disagree nor agree*; 4 = *Agree*; 5 = *Strongly agree*
1. I have adequate income.
2. I have adequate credit.

3. I have financial stability.
4. I have enough savings for an emergency.
5. I have enough assets.

10. Physical Health (Schat, Kelloway, & Desmarais, 2005)

The following items focus on how you have been feeling physically during the past six months. Please indicate the frequency of each item happening to you.
1 = *Not at all*, 2 = *Rarely*, 3 = *Once in a while*, 4 = *Some of the time*, 5 = *Fairly often*, 6 = *Often*, 7 = *All of the time*

1. How often have you had difficulty getting to sleep at night?
2. How often have you woken up during the night?
3. How often have you had nightmares or disturbing dreams?
4. How often has your sleep been peaceful and undisturbed?
5. How often have you experienced headaches?
6. How often did you get a headache when there was a lot of pressure on you to get things done?
7. How often did you get a headache when you were frustrated because things were not going the way they should have or when you were annoyed at someone?
8. How often have you suffered from an upset stomach (indigestion)?
9. How often did you have to watch that you ate carefully to avoid stomach upsets?
10. How often did you feel nauseated ("sick to your stomach")?
11. How often were you constipated or did you suffer from diarrhea?
12. How many times have you had minor colds (that made you feel uncomfortable but didn't keep you sick in bed or make you miss work)?
1 = *0 times*, 2 = *1-2 times*, 3 = *3 times*, 4 = *4 times*, 5 = *5 times*, 6 = *6 times*, 7 = *7 times and more*
13. How many times have you had respiratory infections more severe than minor colds that "laid you low" (such as bronchitis, sinusitis, etc.)?
1 = *0 times*, 2 = *1-2 times*, 3 = *3 times*, 4 = *4 times*, 5 = *5 times*, 6 = *6 times*, 7 = *7 times and more*
14. When you had a bad cold or flu, how long did it typically last (days)?
1 = *0 days*, 2 = *1-2 days*, 3 = *3 days*, 4 = *4 days*, 5 = *5 days*, 6 = *6 days*, 7 = *7 days and more*

11. Charity Behaviors

In the past year, have you done the following things?
1 = *Yes*, 0 = *No*

1. Donate money
2. Donate blood
3. Volunteer

12. Narcissism (Jonason & Webster, 2010)

On the following pages you will find a series of statements about you. Please read each statement and decide how much you agree or disagree with that statement.
1 = *Strongly disagree*; 2 = *Disagree*; 3 = *Neither disagree nor agree*; 4 = *Agree*; 5 = *Strongly agree*
1. I tend to want others to admire me.
2. I tend to want others to pay attention to me.
3. I tend to seek prestige or status.
4. I tend to expect special favors from others.

13. Machiavellianism (Jonason & Webster, 2010)

On the following pages you will find a series of statements about you. Please read each statement and decide how much you agree or disagree with that statement.
1 = *Strongly disagree*; 2 = *Disagree*; 3 = *Neither disagree nor agree*; 4 = *Agree*; 5 = *Strongly agree*
1. I tend to manipulate others to get my way.
2. I have used deceit or lied to get my way.
3. I have used flattery to get my way.
4. I tend to exploit others towards my own end.

14. Psychopathy (Jonason & Webster, 2010)

On the following pages you will find a series of statements about you. Please read each statement and decide how much you agree or disagree with that statement.
1 = *Strongly disagree*; 2 = *Disagree*; 3 = *Neither disagree nor agree*; 4 = *Agree*; 5 = *Strongly agree*
1. I tend to lack remorse.
2. I tend to be unconcerned with the morality of my actions.
3. I tend to be callous or insensitive.
4. I tend to be cynical.

15. Respondent Reactions – Time 1 (Zhang et al., 2023)

Thank you for completing the previous personality questionnaire. We are interested in your attitudes towards and feeling about this questionnaire. Below we present some statements that describe various attitudes and feelings. Please indicate the degree to which you agree with each statement according to how you feel while filling out it. There is no right/favorable or wrong/unfavorable answer. Please be as honest as possible.

1 = Strongly disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly agree

1. **General Positive Affect**

- This questionnaire is very interesting.
- I was annoyed when completing this questionnaire.
- Completing this questionnaire is boring.

2. **Perceived Accuracy**

- This questionnaire can accurately measure my personality characteristics.
- Most items in this questionnaire do not apply to me.
- I would recommend my friends to use this questionnaire if they want to know more about their personality.

3. **Perceived Utility**

- This questionnaire is useful for personnel selection.
- This questionnaire is useful for talent development.
- I do not think this questionnaire has any practical value.

4. **Perceived Faking Resistance**

- It is hard to fake on this questionnaire.
- When completing this questionnaire, I can easily present a personality profile that is different from my true self without being detected.
- I think this questionnaire can effectively resist faking on personality testing.

5. **Perceived Difficulty**

- This questionnaire is difficult to answer.
- There are many questions that I am not sure how to answer.
- It is easy to fill out this questionnaire.

6. **Perceived Cognitive Burden**

- I need to think deeply before making decisions due to the complexity of this questionnaire.
- Completing this questionnaire makes me exhausted.
- Completing this questionnaire is just like going through a difficult exam.

7. **Degree of Concentration**

- I was very concentrated when completing this questionnaire.
- My mind wanders a lot when completing this questionnaire.
- People have to be highly concentrated to successfully complete this questionnaire.

8. **Others**

- How much effort do you have to exert in order to complete this questionnaire as instructed?

  0 (zero effort) ----- 10 (All my efforts)
- How exhausted are you after completing this questionnaire?

0 (Not exhausted at all) ----- 10 (Completely exhausted)

- Let's say your energy level was 10 before you start to work on this questionnaire. What's your current energy level after completing this questionnaire?
  0 (zero energy) ----- 10 (full energy)

16. Respondent Reactions – Time 2

Thank you for completing the previous personality questionnaire. We are interested in your attitudes towards and feeling about this questionnaire. Below we present some statements that describe various attitudes and feelings. Please indicate the degree to which you agree with each statement according to how you feel while filling out it. There is no right/favorable or wrong/unfavorable answer. Please be as honest as possible.
1 = Strongly disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly agree

1. Perceived Test Fairness

Lopez, Hou, & Fan (2019)

- Overall, I believe the test was fair.

Chan, Schmitt, Sacco, & DeShon (1998)

- I feel that using the test to select applicants for the job is fair.

2. Perceived Predictive Validity

Adapted from Kluger & Rothstein (1993)

- I believe that the test can predict how well an applicant will perform on the job.

Adapted from Macan, Avedon, Paese, & Smith (1994)

- The test measured the skills necessary to perform well on the job.

3. Process Satisfaction/Affect Towards the Test

Tonidandel, Quiñones, & Adams (2002)

- I liked taking this type of test.

Adapted from Sylva & Mol (2009)

- Overall, I was satisfied with this employee selection method.

4. Organizational Attractiveness

Highhouse, Lievens, & Sinar (2003)

- For me, this company would be a good place to work.
- This company is attractive to me as a place for employment.

5. Job Acceptance Intentions

Highhouse, Lievens, & Sinar (2003)

- If this company invited me for a job interview, I would go.
- I would accept a job offer from this company.

6. Face Validity

Macan, Avedon, Paese, & Smith (1994)

- The actual content of the test is clearly related to the job.

Chan, Schmitt, Sacco, & DeShon (1998)

- I can see a clear connection between the test and what I think is required by the job.

7. Recommendation Intentions

Adapted from Smither, Reilly, Millsap, Pearlman, & Stoffey (1993)

- Based on my experience with the test, I would recommend others to apply to this organization.

Adapted from Highhouse, Lievens, & Sinar (2003)

- Based on my experience with the test, I would recommend this company to a friend looking for a job.

8. Faking Resistance

Self-developed

- It's hard to fake on this questionnaire.
- I think this test is resistant to applicant faking.

9. Perceived Accuracy

Harris, McMillan, & Carter (2021)

- I believe the assessment accurately measured my personality.
- At least some of the questions on the assessment very accurately described me.

References

Anglim, J., Morse, G., De Vries, R. E., MacCann, C., & Marty, A. (2017). Comparing job applicants to non–applicants using an item–level bifactor model on the HEXACO personality inventory. *European Journal of Personality*, *31*(6), 669-684. https://doi.org/10.1002/per.2120

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*(4), 340-345. https://doi.org/10.1080/00223890902935878

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687-699. https://doi.org/10.1177/001316448104100307

Brown, A. & Maydeu-Olivares, A. (2018). Ordinal Factor Analysis of Graded-Preference Questionnaire Data. *Structural Equation Modeling: A Multidisciplinary Journal, 25(4)*, 516-529. https://doi.org/10.1080/10705511.2017.1392247

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

Brown, A., & Maydeu-Olivares, A. (2018). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 516-529. https://doi.org/10.1080/10705511.2017.1392247

Bürkner, P. C. (2019). thurstonianIRT: Thurstonian IRT models in R. *Journal of Open Source Software*, *4*(42), 1662-1663. https://doi.org/10.21105/joss.01662

Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology*, *83*(3), 471-485. https://doi.org/10.1037/0021-9010.83.3.471

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. https://doi.org/10.1207/s15327752jpa4901_13

Dalal, D. K., Zhu, X. S., Rangel, B., Boyce, A. S., & Lobene, E. (2021). Improving applicant reactions to forced-choice personality measurement: Interventions to reduce threats to test takers' self-concepts. *Journal of Business and Psychology*, *36*(1), 55-70. https://doi.org/10.1007/s10869-019-09655-6

Fox, S., Spector, P. E., Bruursema, K., Kessler, S., & Goh, A. (2007). Necessity is the mother of behavior: Organizational constraints, CWB and OCB. In *Meeting of the Academy of Management, Philadelphia, PA*.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29-48. https://doi.org/10.1348/000711006X126600

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Harris, A. M., McMillan, J. T., & Carter, N. T. (2021). Test-taker reactions to ideal point measures of personality. *Journal of Business and Psychology*, *36*(3), 513-532. https://doi.org/10.1007/s10869-020-09682-8

Highhouse, S., Lievens, F., & Sinar, E. F. (2003). Measuring attraction to organizations. *Educational and Psychological Measurement*, *63*(6), 986-1001. https://doi.org/10.1177/0013164403258403

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: a concise measure of the dark triad. *Psychological Assessment*, *22*(2), 420. https://doi.org/10.1037/a0019265

Kluger, A. N., & Rothstein, H. R. (1993). The influence of selection test type on applicant reactions to employment testing. *Journal of Business and Psychology*, *8*, 3-25. https://doi.org/10.1007/BF02230391

Kristensen, T. S., Borritz, M., Villadsen, E., & Christensen, K. B. (2005). The Copenhagen Burnout Inventory: A new tool for the assessment of burnout. *Work & Stress*, *19*(3), 192-207. https://doi.org/10.1080/02678370500297720

Li, M., Sun, T., & Zhang, B. (2022). AutoFC: An R package for automatic item pairing in forced-choice test construction. *Applied Psychological Measurement*, *46*(1), 70-72. https://doi.org/10.1177/01466216211051726

Lievens, F., De Corte, W., & Brysse, K. (2003). Applicant perceptions of selection procedures: The role of selection information, belief in tests, and comparative anxiety. *International Journal of Selection and Assessment*, *11*(1), 67-77. https://doi.org/10.1111/1468-2389.00227

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, *77*(3), 389-414. https://doi.org/10.1177/0013164416646162

Lopez, F. J., Hou, N., & Fan, J. (2019). Reducing faking on personality tests: Testing a new faking-mitigation procedure in a US job applicant sample. *International Journal of Selection and Assessment*, *27*(4), 371-380. https://doi.org/10.1111/ijsa.12265

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, *47*(4), 715-738. https://doi.org/10.1111/j.1744-6570.1994.tb01573.x

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsoda, V. (2019). The journey from Likert to forced-choice questionnaires: Evidence of the invariance of item parameters. *Revista de Psicología del Trabajo y de las Organizaciones*, *35*(2), 75-83. https://doi.org/10.5093/jwop2019a11

Munyon, T. P., Carnes, A. M., Lyons, L. M., & Zettler, I. (2020). All about the money? Exploring antecedents and consequences for a brief measure of perceived financial security. *Journal of Occupational Health Psychology*, *25*(3), 159-175. https://doi.org/10.1037/ocp0000162

Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, *53*(5), 633-654. https://doi.org/10.1080/00273171.2018.1469966

Roodt, G. (2004). Turnover intentions. Unpublished document: University of Johannesburg Johannesburg, South Africa.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1-36. https://doi.org/10.18637/jss.v048.i02

Schat, A. C., Kelloway, E. K., & Desmarais, S. (2005). The Physical Health Questionnaire (PHQ): construct validation of a self-report scale of somatic symptoms. *Journal of*

Occupational Health Psychology, *10*(4), 363-381. https://doi.org/10.1037/1076-8998.10.4.363

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. https://doi.org/10.1111/j.1744-6570.1993.tb00867.x

Spector, P. E. (1985). Measurement of human service staff satisfaction: Development of the Job Satisfaction Survey. *American Journal of Community Psychology*, *13*(6), 693-713. https://doi.org/10.1007/bf00929796

Spector, P. E., Fox, S., Penney, L. M., Bruursema, K., Goh, A., & Kessler, S. (2006). The dimensionality of counterproductivity: Are all counterproductive behaviors created equal?. *Journal of Vocational Behavior*, *68*(3), 446-460. https://doi.org/10.1016/j.jvb.2005.10.005

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*(2), 320-332. https://doi.org/10.1037/0021-9010.87.2.320

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York.

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4, https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org.

Wickham, H., Vaughan, D., Girlich, M. (2023). *tidyr: Tidy Messy Data*. R package version 1.3.0, https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org.

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management*, *17*(3), 601-617. https://doi.org/10.1177/014920639101700305

Zhang, B., Luo, J., & Li, J. (2023). Moving beyond Likert and Traditional Forced-Choice Scales: A Comprehensive Investigation of the Graded Forced-Choice Format. *Multivariate Behavioral Research*, 1-27. https://doi.org/10.1080/00273171.2023.2235682

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, *23*(3), 569-590. https://doi.org/10.1177/1094428119836486