Received: 17 April 2021 | Revised: 17 January 2021 | Accepted: 26 August 2021

DOI: 10.1111/ijsa.12345

RESEARCH ARTICLE

International Journal of SELECTION AND ASSESSMENT WILEY

Introducing a supervised alternative to forced-choice personality scoring: A test of validity and resistance to faking

Andrew B. Speer | Angie Y. Delacruz

Department of Psychology, Wayne State University, Detroit, Michigan, USA

Correspondence
Andrew B. Speer, Department of Psychology, Wayne State University, 5057 Woodward Ave, Detroit, MI 48402, USA.
Email: speerworking@gmail.com

Abstract

This paper examines a new personality assessment scoring approach labeled supervised forced choice scoring (SFCS), which aims to maximize construct validity of forced choice (FC) personality assessments. SFCS maximally weights FC responses to predict or "reproduce" honest, normative, and reliable personality scores using machine learning. In this proof of concept study, a graded response FC assessment was tested across several samples, and SFCS resulted in psychometric improvements over traditional FC scoring. Correlations with aligned single-stimulus trait scores (taken honestly) were strong, both when the FC measure was taken honestly and when taken in induced applicant settings. SFCS scores also exhibited small shifts in average scores between honest and faked conditions and were predictive of organizational citizenship behaviors, employee engagement, and leadership emergence at work. Although SFCS showed merit in this proof of concept study, it is unclear how well results will generalize to new FC measures, and we urge more research on this scoring method.

KEYWORDS
construct validity, faking, forced-choice, machine learning, personality, supervised modeling

Practitioner points

- Forced choice (FC) personality assessments are thought to be more faking resistant than single-stimulus, Likert personality measures. However, there are measurement challenges when scoring FC assessments, and there may be opportunities to improve upon the construct validity of current FC scoring.
- This paper examined a new personality assessment scoring approach labeled supervised forced choice scoring (SFCS), which aims to maximize construct validity of FC personality assessments. Specifically, SFCS maximally weights FC responses to predict or "reproduce" honest, normative, and reliable personality scores using machine learning (ML). The use of ML in this way helps leverage variance from all FC personality responses when estimating each trait, accounts for complex item-trait relationships should they exist, and develops algorithms explicitly designed to account for faking.
- In this proof of concept study, a graded response FC assessment was tested across several samples, and SFCS resulted in psychometric improvements over traditional

[Correction added on 28 September 2021, after first online publication: Citations and reference details of Fisher et al. (2018) has been added in this version.]

FC scoring. Notably, correlations with aligned single stimulus trait scores (taken honestly) were strong, both when the FC measure was taken honestly and when taken in induced applicant settings. SFCS scores also exhibited small shifts in average scores between honest and faked conditions and were predictive of organizational citizenship behaviors, employee engagement, and leadership emergence at work.

- Initial SFCS findings are favorable, suggesting this method of FC scoring should be considered for future use. However, although SFCS showed merit in this proof of concept study, it is unclear how well results will generalize to new FC measures, and we urge more research on this scoring method.

1 | INTRODUCTION

Personality traits are useful predictors of life and work outcomes (e.g., Gatewood et al., 2016) and commonly assessed using single-stimulus (SS) items (e.g., Likert items), which are composed of single statements, items, or questions where respondents make explicit judgments of each statement, item, or question. Although SS personality assessments are useful to quantify individual differences, they are prone to intentional bias in high-stakes settings (e.g., Griffith & Robie, 2013). In other words, SS assessments can be easily faked, resulting in weakening of the construct validity of these measures and the attenuation of their relationships with important criteria (e.g., Christiansen et al., 2017; Holden, 2008; Jeong et al., 2017; Komar et al., 2008; Peterson et al., 2011; Schmit & Ryan, 1993; Topping & Gorman, 1997).

Methods exist to reduce the effects of faking on personality assessments, and one is to design the personality items to be more faking resistant by utilizing what is known as forced choice (FC) measurement (e.g., Salgado & Tauriz, 2014). Because respondents must pick or comparatively evaluate multiple response options, FC measures are thought to be more difficult to fake, and FC measures may also be more resistant to other self-report biases such as acquiescence. However, there are some challenges to scoring using FC assessments. In brief, recent research suggests that FC assessments only produce desirable psychometric properties when certain design features are met, with some of these features conflicting with the goals of faking reduction (e.g., Bürkner et al., 2019; Cao & Drasgow, 2019; Schulte et al., 2021). Additionally, although FC measures are designed to assess specific personality traits, there is considerable variability in whether FC tests strongly correlate with other aligned personality measures composed of the exact same item content (c.f. Zhang et al., 2020), raising the question of what FC scores are measuring in these cases. This problem is exacerbated when the FC measure is taken under conditions of faking, such that the construct validity of FC measures worsens in faked settings (Fisher et al., 2018; Usami et al., 2016). Thus, although FC assessments offer a useful format to mitigate faking and other self-report response biases, there may still be ways to improve upon current FC scoring.

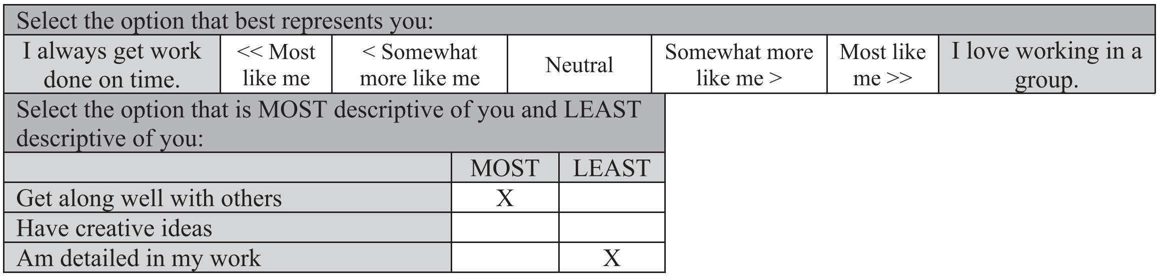
The current paper explores a potential new way to score FC measures (or other self-report formats) in efforts to maximize the degree to which personality scores assess intended construct variance, and especially when taken in contexts of impression management. This new method is labeled supervised construct scoring, and it is explicitly designed to recover honest, normative, and targeted personality scores. When applied to FC tests in particular, we refer to this method as supervised forced choice scoring (SFCS). This method works by training a model to take FC responses as input and to recreate (i.e., predict) honestly derived personality scores that have been established using alternative measurement methods, taken honestly. As we will discuss, we believe this method may circumvent current psychometric issues specific to FC scoring and may produce scores that are more resistant to the effects of impression management.

This paper begins by describing the current state of FC measurement. Then, the SFCS procedure is introduced, and we highlight how this method might potentially improve upon several features of FC scoring. Following this, the SFCS method is applied to a graded preference scale,[1] which is a special version of the FC format. Psychometric evidence is compared versus traditional FC scoring across several samples in terms of faking, reliability, and validity as a proof-of-concept test of this new scoring approach.

2 | FC PERSONALITY ASSESSMENTS

FC personality assessments have respondents choose, rank, sort, indicate higher frequency for, or indicate preference for two or more paired response options. Explicit judgments regarding individual statements, items, or questions are made comparatively within each block of stimuli. Example FC questions can be found in Figure 1, where the bottom example uses the commonly applied triad format. The top example is in a graded response (GR) format (Brown & Maydeu-Olivares, 2018), which is a special case of FC items.

FC assessments result in less score inflation between honest and motivated samples, and this is especially the case when care is made to ensure the statements used in FC item blocks are similar in social desirability (Cao & Drasgow, 2019). When statements are matched in desirability, FC items are thought to be less transparent and more difficult to fake. To be clear, response distortion is not completely



Select the option that best represents you:						
I always get work done on time.	<< Most like me	< Somewhat more like me	Neutral	Somewhat more like me >	Most like me >>	I love working in a group.

Select the option that is MOST descriptive of you and LEAST descriptive of you:		
	MOST	LEAST
Get along well with others	X	
Have creative ideas		
Am detailed in my work		X

FIGURE 1 Example forced choice items

eliminated. The Five Factor Model (FFM) standardized mean differences between induced applicant samples and honest samples (i.e., most similar to our study and also with the largest *k* in the Cao and Drasgow meta-analysis) ranges from small to large effects across FFM traits, depending on whether large samples are removed from the meta-analysis or not. Regardless, it is safe to say that FC scores are inflated to a lesser extent when compared to SS measures (Viswesvaran & Ones, 1999).

Although there may be benefits to FC measurement in high-stakes contexts, the FC measurement format can come at the expense of score interpretability, such that some FC personality measures produce ipsative scores (e.g., Bartram, 2007; Salgado & Tauriz, 2014). Ipsative scores are dependent upon responses to other variables besides the targeted trait of interest, resulting in the total sum of scores being either fully (fully ipsative) or partially (partially ipsative) fixed. As such, ipsative assessments distort intraindividual differences in trait scores, have distorted estimated mean scores, and have distorted intercorrelations among scale scores. This is problematic when interpreting personality scores for several reasons. Most notably, it diminishes the usefulness of the test to make decisions because everyone receives the same (in the case of fully ipsative measures) or very similar (in the case of partially ipsative measures) overall scores. Additionally, estimated trait scores do not reflect absolute standing on the underlying traits, and thus ipsative scores can be challenging to use for development purposes.

To clarify terminology, ipsativity occurs in contrast to normative scoring. Normative scores are independent from other traits measured by the personality assessment; normative personality scores reflect interindividual variance in traits as opposed to intraindividual variance. Most commonly, normative scores are established via SS scales, given these measures adhere to the definition outlined above. However, methods have been developed to uncover normative score estimates from FC assessments using item response theory (IRT). Example IRT methods include Thurstonian IRT (TIRT, e.g., Brown & Maydeu-Olivares, 2011), the multidimensional pairwise-preference model (MUPP; Stark et al., 2005), and the McCloy–Heggestad–Reeve unfolding model (McCloy et al., 2005). IRT scoring has become the dominant method to score FC measures, and when certain design features are in place can accurately recover normative personality scores (Bürkner et al., 2019; Schulte et al., 2021).

3 | AN ALTERNATIVE SCORING METHOD

The purpose of this study was to investigate an alternative FC scoring method that applies machine learning (ML; see Kuhn & Johnson, 2013; Putka et al., 2018; Yarkoni & Westfall, 2017) to FC assessments. Specifically, the introduced method of SFCS creates algorithms to translate FC responses to predict normative, honest, and targeted trait scores from external measures. This requires establishing honest, reliable trait scores collected independently of the FC item responses (e.g., observer ratings of personality, honest FC scores, honest self-report SS scores) and then uses supervised ML to predict or recreate those scores. Supervised ML is a subtype of machine learning that occurs when a set of inputs (e.g., personality item responses) are trained to predict an outcome variable (e.g., existing trait scores). Supervised ML algorithms are trained in a way to maintain strength of outcome prediction in holdout samples, and done by incorporating various ML algorithm features (e.g., regularization, random sampling, ensemble models, parameter dropout), depending on the type of ML algorithm used. Although ML algorithms are often not applied outside of the primary data collection, and therefore the generalizability of some ML models may be unknown, holdout samples are frequently partitioned from the full data set and held out during model training so to later test how well the developed algorithms "perform" (i.e., converge with target-dependent variables) in a portion(s) of the data that was independent from algorithm creation (e.g., James et al., 2017; Kuhn & Johnson, 2013; Putka et al., 2018).

Per the parlance of ML, existing FC scoring methods would be considered unsupervised methods because there is no explicit outcome variable that the FC responses are mapped to during the modeling process. FC responses are gathered and we inherently assume they are caused by unobservable latent constructs. Covariance structures are modeled to estimate latent scores. On the other hand, SFCS uses ML to convert FC responses into targeted and externally measured trait scores that are ideally honest, reliable, and normative. Thus, it is supervised in the sense that a target variable already exists before scaling the FC responses. Although there are challenges to the SFCS approach, which we will elaborate upon in the following sections, there may also be advantages that result in improvements over existing FC scoring methods. In the next three sections we highlight these factors.

3.1 | Maximized use of FC variance

Assuming there is a reliable and valid pre-established set of honest trait scores to use as target variables and adequate sample sizes to perform ML (see Putka et al., 2018), SFCS should maximize usage of FC response data, and in some cases to a greater degree than contemporary FC scoring methods. There are several reasons why this is the case.

First, supervised scoring is capable of automatically adjusting item weights based on how strongly they align to the targeted construct. If an item is rationally aligned to a construct but does not relate to observed construct scores, SFCS will assign less weight for that item and vice versa if the item is strongly related. Contemporary IRT scoring is also capable of accounting for differential item quality by allowing for different factor loadings across items based on assumed relations with the latent trait. Thus, this is a strength of both methods.

Second, SFCS is capable of capturing variance from items that might not be superficially related to a construct but which nonetheless reflect some aspect of construct-related variance. For example, many traditional agreeableness items (e.g., “I love to help others”; J. A. Johnson, 2014) are likely to capture socially oriented variance related to extraversion. SFCS utilizes variance from items that are not linked to rationally targeted traits, but which share some construct overlap, like the example just provided. This allows any item to contribute to construct scores and maximizes information usage from FC responses. Such scoring is also consistent with circumplex models (e.g., Shoss & Witt, 2013), which theoretically allow for combinations of content from different trait domains to form combination trait scores. Thus, empirically allowing items to cross load onto nonfocal constructs is rooted in personality theory. Although traditional factor analytical methods could hypothetically allow all items to cross-load to the same effect, such a model would likely exhibit estimation problems and would drastically overfit the data unless some sort of regularization was incorporated, a feature which most ML algorithms easily deal with.

Lastly, depending on the ML algorithm used, supervised modeling is capable of modeling nonlinearities in the data, should they exist. We refer the reader to James et al. (2017) and Kuhn and Johnson (2013) for friendly introductions to ML methods capable of doing this. In brief, many ML algorithms can model nonlinear relationships with little to no a priori specification by the researcher and do so in a way that prevents overfitting. Assuming best practices are used such that sample sizes are adequate and cross-validation is performed (so to accurately infer validity), this SFCS feature may be beneficial when applied to personality items that use a continuous response scale. For example, empirical scoring may detect nonlinear relationships for SS items (Cucina et al., 2019) or for graded preference questions (Brown & Maydeu-Olivares, 2018), which are a special version of the FC format. Some research suggests that certain personality items exhibit nonlinear relationships with their underlying trait (e.g., Stark et al., 2006). Thus, this scoring feature aligns with existing knowledge of personality and may lead to improved scoring.

3.2 | Clarity in construct validity

A more targeted approach to FC scoring is likely to provide improved clarity as to what exactly FC scores represent. When designing item statements, researchers target specific and well-defined trait domains (e.g., “I am the life of the party”—extraversion), explicitly writing item content that is related and representative of that domain. Frequently, the same statements are used in both SS and FC test formats. However, even in cases when a FC and SS measure contain the exact same items and are taken under the exact same response conditions, the magnitude of the correlations between FC and SS measures can vary substantially. There are some instances where FC and SS measures exhibit strong correlations with one another (e.g., Brown & Maydeu-Olivares, 2011, 2018; Fisher et al., 2018; Lee et al., 2018; Zhang et al., 2020). On the other hand, there are instances when these correlations are also lower on average (e.g., Anguiano-Carrasco et al., 2015; Christiansen et al., 2005; Dueber et al., 2019; Guan, 2015; Ng et al., 2020), or where a number of traits exhibit lower correlations (Brown & Maydeu-Olivares, 2013; Joubert et al., 2015; SHL, 2013). Some of the observed inconsistency is likely due to differences in FC design, affected by factors such as the number of targeted traits and social desirability match within item blocks (Bürkner et al., 2019; Schulte et al., 2021). However, to the extent the FC and SS measures are expected to assess the exact same construct domain and are composed of the exact same items, these two measures might be considered alternative test forms, and therefore correlations between the two should be very high—stronger to the degree that each measure is more strongly influenced by the same latent construct.

The question then becomes if a FC measure and SS measure different things when taken honestly (as reflected by a lack of convergence), then what is the FC assessment uniquely measuring? It is possible that the FC measure assesses some different construct than the SS measure, or it is possible that one of the measures is impacted by bias or random error. The literature does not provide a definitive answer to these considerations. We know that SS variance can be distorted by response biases such as acquiescence or severity (Furr, 2018). On the other hand, the design of FC items results in its own measurement challenges. The ipsative format that makes FC items so appealing by reducing impression management may produce unwanted error in the response process by forcing statement differentiation when no differentiation truly exists for individual respondents. For example, if a respondent finds two statements equally desirable because he or she has high trait levels for both (“hard worker,” “innovative”), the choice of response may be random. Additionally, whereas SS item scores contain information about both absolute and relative trait standings, two paired FC statements with equal positive factor loadings only provides relative information about traits (Brown & Maydeu-Olivares, 2011).[2] This results in higher item precision for SS items, and evidence shows that FC item scores are less reliable than SS item scores (see Brown & Maydeu-Olivares, 2011, 2018; Lee et al., 2018; Zhang et al., 2020). Furthermore, when FC and SS tests are compared in the same honest settings (i.e., apples-to-apples comparison), criterion-related validity with job performance outcomes is typically better for SS assessments than FC assessments (Brown & Maydeu-Olivares, 2013; Christiansen et al., 2005; Conway, 2000; Goffin et al., 2011). That said, it should be noted that because only a handful of studies have compared these test formats on job-related criterion-related validity (while holding other factors constant such as constructs, test content, setting), it is hard to form definitive conclusions based on these results.

In summary, we know (1) that FC and SS scores are not always consistent with one another even when composed of the same item content, (2) it is unclear why exactly this is the case, and (3) SS item scores are likely to be more reliable than FC item scores. Because of this ambiguity regarding FC construct validity, there may be benefits to using SFCS to recreate normative, honest, and better understood trait scores derived from external measures. The challenge is then determining what should be used as the target scores.

The most appropriate method to establish target trait scores is likely debatable. Ideally, trait scores would be reliable, valid, and honest. There are several potential measurement options, including use of observer ratings of personality, honest FC scores, or self-report SS scores taken honestly, among possible others. There are pros and cons to these approaches. In this proof-of-concept study, we chose to focus on honest SS scores.

As mentioned previously, even when responded to honestly there are response biases that may impact SS assessments. This said, SS items are more straightforward to respond to than FC assessments, as respondents only evaluate a single statement at a time and therefore responding is not directly influenced by other statements. This makes responding less complicated,[3] allows for more simplistic factor analytical models, and results in higher item reliability (Brown & Maydeu-Olivares, 2011, 2018; Lee et al., 2018; Zhang et al., 2020). SS items also allow for high levels of variance given the ordinal/interval response scale used for most SS formats. Finally, because in the SFCS design the target measure is taken honestly, scores should not be highly influenced by intentional response distortion (i.e., impression management). Thus, in the current study, we used honest SS scores as estimates of personality traits and trained FC responses to recreate these scores.

3.3 | Reducing faking through *both* item format *and* scoring

Despite the item format being more faking resistant, FC validity is still impacted by faking. This is because respondents are capable of targeting and endorsing job relevant traits more frequently than less relevant traits (e.g., Christiansen et al., 2005), resulting in a mean increase in scores when faked, but also erosion of validity. For example, convergent correlations with honest personality are weakened when the FC is taken under conditions of impression management. Fisher et al. (2018) reported an average correlation of .80 between FC scores taken honestly and SS scores taken honestly, but this dropped to .63 when the FC was taken under conditions of impression management (the SS measure was still taken honestly). Usami et al. (2016) likewise found that validity with honest SS scores was severely affected by faking, dropping from .73 to .39 when the FC measure was faked. FC model fit also worsens when tested in applicant settings (Ng et al., 2020), and recent evidence has shown that IRT-based FC scoring can produce inaccurate personality estimates with certain FC designs that are ironically the most faking resistant (Bürkner et al., 2019; Schulte et al., 2021), with this rendering the criterion-related validity of IRT FC scoring to zero in some cases (Fisher et al., 2019). Thus, even though the FC item format is resistant to faking, current scoring procedures are not explicitly designed to account for faking, which is likely to impact the validity of personality scores when used in conditions of impression management. This raises the question of whether FC responses can be scaled to more explicitly account for faking.

SFCS attempts to more intentionally account for faking by developing scoring that explicitly recreates honestly derived scores. This involves a two-step data collection. In the case of this study, the target SS assessment is administered to a sample of respondents told to respond as honestly as possible. Thus, the target scores reflect variance that should be mostly devoid of impression management. Second, the FC assessment is administered to the exact same respondents under operational response conditions. If the primary intended use of FC assessments is for prehire assessment, the operational conditions would be applicant conditions or simulated applicant conditions. After these data are collected, supervised modeling is then used to recreate the honest SS scores based on the FC responses. Thus, the scoring procedure is explicitly designed to recover estimates of honest personality.

3.4 | Present study

As the previous sections described, there are several reasons why SFCS might result in psychometric improvements when scoring FC measures. The purpose of the present study was to investigate the viability of the SFCS approach. To do so, we applied SFCS to a graded preference scale, which is a special type of FC format (Brown & Maydeu-Olivares, 2018). Across multiple samples and using a variety of psychometric tests, we tested the validity of SFCS scores and compared them to traditional FC scoring. In Study 1, a FC measure was created and cross-validated using *k*-folds cross-validation, examining shifts in mean faking across honest and faked conditions and validity in terms of convergence with honestly taken SS scores. In Study 2, additional psychometric evidence was collected in terms of test–retest reliability, convergence with honestly taken SS scores, and criterion-related validity with work outcomes. Taken together, testing this new scoring approach across multiple samples and according to multiple psychometric features serves as a strong initial test of this new scoring method.

4 | METHODS

4.1 | Participants

4.1.1 | Sample 1

Sample 1 served as the primary data collection and development sample for the GR FC measure used in this study. Participants came from MTurk and Prolific. MTurk is a labor market where people voluntarily perform tasks in exchange for compensation, and it has frequently been used in the social sciences (Cheung et al., 2017; Landers & Behrend, 2015). To participate in the study, MTurk participants had to have an approval rate of 98% or greater and be located in the United States. Prolific is also an online labor market, like Mturk. However, Prolific is specifically used for academic research. Prolific uses good recruitment standards, pays participants well, is well-equipped for longitudinal research, and produces high quality research data (Palan & Schitter, 2018). Participants were required to have an approval rate of 98% or greater and be located in the United States. Collectively, this combined sample is diverse and is composed of people from a variety of industries and employment situations. For both data sources, all responses were screened for purposeful responding using directed response items (e.g., *Choose Strongly Disagree*) and minimal response time cutoffs. The total sample size used for analyses was 1058 respondents, with 452 coming from MTurk and 606 coming from Prolific.[4] Women made up 52.9% of Sample 1. Racially, 70.7% were White. The average age was 36 years old. Of the respondents, 42.3% had below an undergraduate level education, 39.5% had a bachelor's degree, and 17.2% had education beyond a bachelor's degree. Data collection occurred in the summer of 2020 and directly during the COVID-19 pandemic. Twenty-nine percent of the sample was unemployed at the time of data collection. Of those employed, the most common occupations were in management or self-employed (24%), administrative work (16%), customer service positions (10%), tech work (6%), education (6%), healthcare (4%), and manual labor (2%).

4.1.2 | Sample 2

An additional data collection was conducted to further examine SFCS in terms of reliability, convergent validity, and criterion-related validity. All respondents came from Prolific ($N = 271$).[5] To participate, respondents had to be employed within the past several months (work-related outcome measures were later collected), be located in the United States, and have an approval rating greater or equal to 98%. Purposeful responding checks were used to remove inattentive responders (29 respondents, or 9.7%, were removed to arrive at $N = 271$). Women made up 57.2% of the final sample and 71.6% of respondents were White. The average age was 32.16 years old. Of the respondents, 65.3% had at least a bachelor's degree, and 21.8% had an advanced degree. The typical participant had been in their current job for 2 years (median). The median yearly income was $40,000.

4.2 | Measures

4.2.1 | SS measure

A SS measure was created for this study. Details for the creation of this measure can be found in the Supporting Information. The SS assessment measured 15 traits using 7-point Likert items. The measured competencies include: Sociability, Leadership, Risk Seeking, Compassion, Cooperation, Integrity, Industriousness, Self-Efficacy, Detail-Oriented, Composure, Positive Mood, Intellectual, Creativity, Preference for Variety, and Aesthetic Preferences. Table 1 displays definitions for these competencies and rational linkages to other personality scales from popular taxonomies. All participants completed the SS personality measure under honest conditions.

The SS items were evaluated by three PhD IO psychologists to establish content validity, and factor analyses, reliability analyses, and collection of convergent and discriminant correlations were performed to establish validity evidence (see Supporting Information for more details). Table 2 displays intercorrelations among the SS scale scores in Sample 1. A multidimensional confirmatory factor analysis revealed the factor structure generally fit the data (RMSEA = 0.05, CFI = 0.83, SRMR = 0.07), with comparable or better fit than seen for other multidimensional personality inventories (see Hopwood & Donnellan, 2010). Additionally, factor scores all exhibited good to very high levels of reliability, with α's ranging from .78 to .93 and averaging .86: Sociability ($\alpha = .92$), Leadership ($\alpha = .89$), Risk Seeking ($\alpha = .87$), Compassion ($\alpha = .88$), Cooperation ($\alpha = .78$), Integrity ($\alpha = .84$), Industriousness ($\alpha = .87$), Self-Efficacy ($\alpha = .80$), Detail Orientation ($\alpha = .81$), Composure ($\alpha = .90$), Positive Mood ($\alpha = .84$), Intellectual ($\alpha = .83$), Creativity ($\alpha = .93$), Preference for Variety ($\alpha = .82$), and Aesthetic Preferences ($\alpha = .89$). In Sample 2, the SS measure exhibited similarly adequate fit (RMSEA = 0.06, CFI = 0.78, SRMR = 0.08) and similarly high levels of internal consistency ($\bar{\alpha} = .86$). Finally, an additional data collection was performed where a set of convergent personality measures were administered alongside the new SS assessment. Once again, details on these results can be found in the Supporting Information, but the data supported the construct validity of the SS scores.

4.2.2 | FC measure

A GR FC measure was developed for this study that was composed of item blocks containing two statements with a five-option scale between the two statements (see top example Figure 1). GR items still require differentiating between multiple statements but allows for more response variance and results in more favorable reactions, as compared to other FC formats (D. K. Dalal et al., 2019). Respondents select which of the two statements is most like them using this GR format. An example item is shown in the top of Figure 1. This response format was chosen to allow for finer gradations in trait measurement when using item pairs.

A total of 103 item blocks were created for the GR measure. One of these items was later removed because it was unrelated to SS scores, resulting in 102 item blocks. Seventy-two item blocks were multidimensional with paired statements from two different traits. Before pairing, statements were rated for social desirability by the study authors such that judgments were made with a general,

TABLE 1 Traits measured by personality inventory

Trait	Definition	Similar NEO-PI-R constructs	Similar DeYoung et al. constructs	Similar FFM constructs
Sociability	Friendly, prefers to be around others, makes social connections easily, socially confident	Gregariousness, warmth, self-conscientiousness	Enthusiasm	Extraversion
Leadership	Takes charge with others, dominant, assumes leadership	Assertiveness	Assertiveness	Extraversion
Risk Seeking	Takes chances, comfortable with risk, seeks out adventure	Excitement seeking		Extraversion
Compassion	Sympathetic, altruistic, concerned about others, generous	Tender, altruism	Compassion	Agreeableness
Cooperation	Noncombative, cooperative with others, inhibits aggression, polite, mild-mannered	Compliance	Politeness	Agreeableness
Integrity	Honest, rule-following, moral, trustworthy, keeps promises	Straightforwardness, dutifulness		Agreeableness/conscientiousness
Industriousness	Works hard, self-disciplined, driven to succeed, persistent, starts work right away	Achievement, self-discipline	Industriousness	Conscientiousness
Self-Efficacy	Confident about completing work, views self as successful, high self esteem	Competence	Industriousness	Conscientiousness
Detail-Oriented	Deliberate, organized, methodical, focused on accuracy	Order	Orderliness	Conscientiousness
Composure	Not easily frustrated, stays calm in stressful work contexts, copes well to stressful work contexts	Angry hostility (*r*), anxiety (*r*), vulnerability (*r*)	Volatility (*r*)	Emotional stability
Positive Mood	Is energetic and positive at work, joyful, confident	Depression (*r*)	Withdrawal (r), Enthusiasm	Emotional stability
Intellectual	Enjoys problem-solving, comfortable working with abstract concepts	Ideas	Intellect	Openness
Creativity	Imaginative, out of box thinker, develops creative work ideas	Fantasy	Intellect	Openness
Preference for Variety	Enjoys variety at work, reacts well to change, desires doing different types of work each day	Actions		Openness
Aesthetic Preference	Appreciates artistic work, enjoys abstract work	Aesthetics	Openness	Openness

Note: Because the FFM is so broad, linkages are not as conceptually aligned as those with the NEO-PI-R or DeYoung et al. (2007) traits. Even then, many of the listed linkages are not a perfect 1–1 alignment.
Abbreviations: FFM, Five Factor Model; *r*, an expected negative correlation.

TABLE 2 Single-stimulus and graded response scale intercorrelations

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
1. Sociability		.62	.57	.65	−.28	.25	.28	.44	−.01	.51	.76	.30	.44	.63	.30
2. Risk Seeking	.50		.68	.26	−.61	−.10	.16	.41	−.18	.39	.40	.53	.61	.83	.43
3. Leadership	.51	.51		.23	−.68	.00	.39	.62	.09	.42	.36	.59	.61	.72	.37
4. Compassion	.51	.21	.19		.14	.44	.36	.35	.26	.31	.72	.27	.36	.34	.41
5. Cooperation	−.19	−.42	−.56	.18		.41	.03	−.20	.30	−.09	.05	−.42	−.39	−.55	−.19
6. Integrity	.27	−.01	.07	.40	.29		.55	.42	.54	.41	.52	.03	.07	−.01	.12
7. Industriousness	.26	.20	.35	.40	.01	.44		.87	.79	.50	.47	.39	.34	.28	.27
8. Self-Efficacy	.40	.31	.50	.32	−.15	.38	.74		.64	.71	.56	.56	.53	.51	.37
9. Detail-Oriented	.10	.03	.17	.31	.15	.45	.69	.54		.35	.30	.26	.19	−.02	.20
10. Composure	.46	.30	.34	.25	−.03	.39	.40	.58	.28		.68	.33	.34	.42	.22
11. Positive Mood	.67	.38	.33	.61	.06	.44	.45	.49	.32	.59		.26	.38	.42	.33
12. Intellectual	.28	.39	.40	.31	−.21	.13	.37	.45	.28	.34	.27		.74	.66	.69
13. Creativity	.36	.46	.46	.36	−.21	.14	.33	.41	.24	.25	.35	.58		.70	.86
14. Preference for Variety	.50	.60	.49	.28	−.30	.08	.23	.33	.03	.34	.34	.52	.49		.53
15. Aesthetic Preference	.20	.30	.23	.41	−.02	.13	.24	.23	.24	.13	.29	.57	.64	.37	

Note: N = 1058 in Sample 1. $r \geq .08$, $p < .01$. In the lower diagonal are correlations based on single-stimulus scores, and the upper diagonal contains correlations based on forced choice graded response scores.

applicant frame of reference across all jobs. More specifically, the instructions were to "rate each statement in terms of social desirability as a job applicant would perceive the item. Make the ratings from the perspective of a typical applicant (across most jobs)." Both authors met before making these judgments to establish frame of reference and discuss how a typical applicant would respond. Judgments were then made individually, with high agreement between the two raters (ICC 2,2 = .85). Statements were then paired to equate social desirability within item blocks. Item statements from different traits were randomly crossed, and then within the matched traits we identified the items that were closest in social desirability and paired them. Upon matching statements, some items' wording was then edited to increase equivalence of desirability within the block. Additionally, 25 unidimensional item blocks were incorporated into the measure (24% of item blocks are therefore unequally keyed). In line with IRT scoring of FC measures, these were included to help establish initial estimates of absolute trait levels. Lastly, five item blocks were designed to account for impression management. These item blocks were written such that it would be difficult to endorse the more favorable statement if responding honestly (e.g., "I have never broken any rule"). The intent was to naturally incorporate impression management variance into the scoring algorithms, although upon collecting data there was no meaningful psychometric improvements from using these items, meaning inclusion of these items did not improve or worsen construct validity evidence (i.e., convergence with SS measures, standardized mean differences between honest and faked conditions). Nonetheless, we kept the items in the assessment, given it did not negatively impact algorithms and to be consistent with our initial study plans.

4.2.3 | SFCS

SFCS was performed on the GR items. This was done by developing algorithms to link GR responses to honest SS responses. Within Sample 1, a large sample of respondents took the GR measure under induced applicant conditions. They then took the SS measure under honest conditions. Of the total 1058 respondents, 734 were given instructions to respond to the GR measure as if they were job applicants applying for a job they wanted; thus, they were informed to respond like a job applicant and not explicitly to fake, given that the motivations of job applicants are complex (Griffith & Robie, 2013; König et al., 2012, 2017). These instructions were repeated multiple times throughout. Additionally, an independent 324 respondents were instructed to complete the GR measure honestly. This was done to (a) compare validity evidence across faked and honest conditions after the algorithms were formed and (b) to incorporate a wider mix of respondent motivations when developing the final algorithm. Recall, instructed faking conditions likely overestimate the amount of faking that occurs in real applicant settings because respondents tend to use varied response approaches in practice (e.g., Griffith & Robie, 2013); thus, this methodological design was thought to better reflect applicant tendencies.[6] Additionally, it is important to clarify that the SFCS algorithms were trained on the entire 1058 respondents and were completely blind as to what condition a person belonged to (i.e., the only inputs were the GR FC responses). Thus, even though some respondents took the GR measure in the faked condition and some in the honest condition, only one algorithm was formed for each trait, irrespective of what condition a person belonged to. Upon completing the GR section, all respondents were instructed to respond as honestly as possible to the SS measure, as discussed in the previous section.

Sample 1 was used to both develop and test the algorithms via *k*-folds cross-validation. First, the 1058-person development sample was used to test the validity of the GR measure using fivefold cross-validation, and thus to prevent capitalization on chance (e.g., Kuhn & Johnson, 2013; Putka et al., 2018). This approach randomly splits the development sample into five groups. Algorithms are developed using four of the groups, and the fifth is used as the cross-validation holdout sample. Once completed, another of the five groups is selected as the holdout sample, the algorithms are re-estimated using the other four groups, and then cross-validation statistics are calculated. This is repeated until every 20% group has served as the holdout sample once. The result is independent cross-validations performed for a total sample size equal to 1058. After doing this, algorithms developed on the full 1058 person sample were then applied to Sample 2 to establish additional validity evidence.

Several supervised predictive algorithms were considered to score the GR content, including elastic net regression, random forests, and deep neural networks (e.g., Kuhn & Johnson, 2013; Putka et al., 2018). Random forests was ultimately chosen based on cross-validation results from the fivefolds cross-validation, such that it resulted in slightly higher cross-validation than the other investigated methods. Random forests algorithms (e.g., Breiman, 2001) create numerous individual regression trees (typically several hundred to several thousand) to predict some variable or outcome. These trees involve splitting predictor scores based on whether they differentiate respondents in terms of the targeted variable. Because of this, random forests algorithms naturally capture existing nonlinear predictor effects and interactions in the data. This is particularly advantageous for the FC GR scale content, where endorsing a neutral response might indicate a high or low standing on both statements within the item block. Random forests also result in strong prediction because algorithms randomly sample from the potential pool of variables and respondents at each node. This reduces the performance of each individual tree, but when those trees are aggregated into an ensemble, the result is usually a highly predictive model that minimizes variance in new settings.

A separate random forests model was conducted for each of the 15 targeted traits to create the FC scoring algorithms. Optimal random forests settings regarding the number of trees and the number of variables sampled per node were initially established using *k*-folds cross-validation separately *within* the 80% data partitions for each trait, though given the computational burden of this task we chose to simply fix the number of trees to 750 and the number of variables sampled per node at 20 for all traits. All 102 items were candidates for use in the random forests algorithm for each of the 15 traits. Because items that are empirically unrelated to a construct will not improve model performance, items were winnowed before running random forests for each trait. Specifically, using a nested bootstrapped sample with replacement to avoid overfitting, any items that correlated less than .05 with targeted construct scores were removed.

In addition to the *k*-folds cross-validation in Sample 1, the final algorithms developed in Sample 1 were also applied to data from Sample 2. Like Sample 1, a portion of the Sample 2 respondents took the FC assessment honestly (*N* = 128) and a portion took it under induced applicant conditions (*N* = 143), so as to examine the effects of response motivation on assessment properties. This sample was used as an additional cross-validation of the developed SFCS algorithms.

4.2.4 | Scoring via ordinal factor analysis (OFA)

The GR assessment content was also scored using OFA. Brown and Maydeu-Olivares (2018) describe a method that applies OFA to score FC GR items. This is a special version of IRT scoring: "ordinal factor analysis models belong to the general class of item–response theory (IRT) models" (Brown & Maydeu-Olivares, 2018, p. 517) and is consistent with the law of comparative judgments (Thurstone, 1927), such that agreement toward one statement occurs based on the difference of the latent utilities for the paired two statements. Brown and Maydeu-Olivares (2018) outline the architecture for this model, which can be applied to the two-statement item blocks used in this study. The OFA model was fit in R's lavaan[7] using diagonally weighted least squares, with predicted scores formed using *k*-folds cross-validation in Sample 1 to be consistent with the SFCS method. When fit, the items generally exhibited moderate loadings onto their factors, with an average standardized absolute specified loading of .47. Like with SFCS, algorithms derived in Sample 1 were then used to score responses in Sample 2 as an additional validation.

4.3 | Test–retest correlations and criterion-related validity

Sample 2 was also used to establish test–retest and criterion-related validity for the GR scores. After participants responded to the SS and GR measures, they completed a second survey 1–7 days afterwards. This second survey administration was used to examine test–retest reliability and to establish criterion-related validity. In the Time 2 survey, participants completed the GR measure under the exact same conditions as Time 1. Additionally, they responded to a set criterion measures honestly. Of the 143 participants in the Time 1 faking condition, 102 completed the second portion (retention rate = 71%). Of the 128 participants who completed the Time 1 honest condition, 99 completed part 2 (retention rate = 77%).

Four criterion measures were assessed at Time 2: leadership position, engagement, and two types of organizational citizenship behaviors (OCBs). Leadership position was operationalized as whether or not a respondent had ever directly supervised employees as part of formalized job duties. This operationalization was used instead of a leadership effectiveness scale so as to minimize the impact of common method variance; this operationalization has also been used to assess leadership in the past (e.g., Popper et al., 2004; Siegling et al., 2014). It was expected that the GR scale of Leadership would be most strongly related to having experience in leadership roles. Work engagement was assessed using the three item UWES-3 measure taken from Schaufeli et al. (2019). Internal consistency was .84. Most personality traits tend to correlate with engagement, but GR scales related to achievement (Self-Efficacy, Industriousness) and approach-oriented behaviors (Sociability, Composure, Positive Mood) were expected to be most strongly related. OCBs include both OCB-I (i.e., OCBs oriented toward people) and OCB-O (i.e., OCBs oriented toward the organization). Self-reported OCBs exhibit similar psychometric properties to observer-reported OCBs (Carpenter et al., 2014), making them a reasonable criterion when only self-report designs are possible. Each of these facets were assessed using measures from R. S. Dalal et al. (2009). Internal consistency was .85 for OCB-I and .81 for OCB-O. It was expected that socially oriented FC traits (Sociability, Compassion) would be most strongly related to OCB-I, whereas traits related to achievement (Self-Efficacy, Industriousness) would be most strongly related to OCB-O. However, most personality traits were expected to exhibit relationships with OCBs.

5 | RESULTS

5.1 | Descriptive statistics

Intercorrelations for the SS and GR scores can be found in Table 2. As seen, the GR scores generally possess a positive manifold of correlations, with an average intercorrelation of .34. Thus, scores are not ipsative. In comparison, the SS scores had an average intercorrelation of .29.

5.2 | Mean differences between honest and faked conditions

Table 3 provides standardized mean differences between honest and induced applicant conditions for the 15 scales. The average standardized mean difference for SFCS scores was .24 in Sample 1 and .12 in Sample 2. In comparison, the average standardized mean difference for OFA scores was .31 in Sample 1 and .17 in Sample 2. This occurred even though the average correlation between SFCS and OFA aligned scores was .87. Combining across samples, SFCS scores had significantly weaker faking effects than OFA scores (Steiger $z = -2.23$, $p = .013$), indicating SFCS is slightly less susceptible to score inflation. Additionally, the absolute SFCS effect ($d = 0.21$) is considered small per Cohen's benchmarks.

TABLE 3 Standardized mean differences between honest and faked responses by scoring method

	Sample 1		Sample 2		Average	
Forced choice scale	SFCS	OFA	SFCS	OFA	SFCS	OFA
1. Sociability	.37	.42	.29	.25	.36	.39
2. Risk Seeking	.10	.21	.06	.15	.09	.20
3. Leadership	.15	.27	−.02	.06	.11	.23
4. Compassion	.38	.41	.09	.15	.32	.36
5. Cooperation	.08	.03	.14	.08	.09	.04
6. Integrity	.29	.34	.36	.41	.31	.36
7. Industriousness	.28	.41	.08	.18	.24	.36
8. Self-Efficacy	.20	.29	.12	−.04	.18	.22
9. Detail-Oriented	.27	.51	.00	.24	.21	.45
10. Composure	.19	.31	.32	.38	.22	.33
11. Positive Mood	.36	.37	.25	.31	.34	.36
12. Intellectual	.18	.08	−.08	−.11	.13	.04
13. Creativity	.19	.29	.03	.13	.16	.26
14. Preference for Variety	.26	.29	.12	.10	.23	.25
15. Aesthetic Preference	.24	.35	.04	.16	.20	.32
Average	.24	.31	.12	.17	.21	.28

Note: Sample 1, $N = 1058$. Sample 2, $N = 271$. The Average columns represent the sample-weighted averages. Cohen's benchmarks are .2 = small, .5 = moderate, .8 = large, and 1.0 = very large.

Abbreviations: *d*, standardized mean difference for forced choice responses; OFA, ordinal factor analysis; SFCS, supervised forced choice scoring.

5.3 | Correlations with aligned SS scores

Correlations between SFCS scores with honest SS scores are shown in Table 4. The higher the observed correlation, the more the two measures assess the same construct. In Sample 1, the average SFCS correlation across the entire sample was .75 and ranged from .72 to .80. Thus, all scales exhibited strong and acceptable correlations with same construct measures. SFCS correlations with aligned SS scores were particularly strong when the GR measure was taken in honest conditions, correlating on average .81. Additionally, SFCS scores exhibited strong psychometric properties in the induced applicant condition, correlating on average .73. Results were similar for Sample 2, such that the average correlation was .81 in the honest sample and .71 in the induced applicant sample. Taken together, SFCS scores exhibited strong convergence with aligned SS scores when the GR measure was taken either honestly or faked.

As a comparison, OFA scores exhibited weaker convergence with aligned SS scores. The average correlation when the GR measure was taken honestly was .72 and .75 in Samples 1 and 2, and the average correlation was .64 and .64 in Samples 1 and 2

TABLE 4 Correlations between forced choice scores and aligned single-stimulus scores

	Sample 1—Honest		Sample 1—Faked		Sample 2—Honest		Sample 2—Faked	
Scale	SFCS	OFA	SFCS	OFA	SFCS	OFA	SFCS	OFA
1. Sociability	.85	.83	.77	.73	.85	.84	.79	.78
2. Risk Seeking	.72	.70	.73	.69	.79	.78	.80	.75
3. Leadership	.86	.82	.78	.71	.85	.83	.72	.64
4. Compassion	.82	.75	.76	.68	.78	.78	.70	.69
5. Cooperation	.80	.75	.72	.66	.75	.73	.76	.73
6. Integrity	.86	.81	.74	.63	.82	.75	.78	.69
7. Industriousness	.81	.70	.72	.60	.84	.72	.67	.60
8. Self-Efficacy	.76	.49	.71	.53	.80	.64	.64	.59
9. Detail-Oriented	.74	.52	.71	.53	.78	.56	.60	.38
10. Composure	.84	.78	.67	.60	.82	.77	.70	.63
11. Positive Mood	.85	.78	.70	.60	.80	.74	.67	.58
12. Intellectual	.81	.72	.77	.71	.85	.81	.61	.56
13. Creativity	.79	.74	.71	.67	.83	.80	.70	.61
14. Preference for Variety	.79	.77	.75	.73	.82	.80	.74	.70
15. Aesthetic Preference	.80	.67	.70	.58	.79	.69	.74	.60
Average	.81	.72	.73	.64	.81	.75	.71	.64

Note: Sample 1, $N = 1058$. Sample 2, $N = 271$. Correlations are forced choice score correlations with the aligned single-stimulus scores (with the latter always taken under honest conditions).

Abbreviations: OFA, ordinal factor analysis; SFCS, supervised forced choice scoring.

when taken faked. Across both samples, SFCS had an average correlation of .81 in honest samples compared to .73 for OFA, and this difference was significant (Steiger's $z = 9.54$, $p < .001$). Likewise, SFCS exhibited significantly stronger correlations in faked samples (.73) than OFA (.64; Steiger $z = 8.57$, $p < .001$). Thus, SFCS resulted in improved convergence with aligned SS scores in both honest and faked samples.

As a comparison using a more traditional taxonomy, the GR scores were aggregated to the FFM using the rational linkages in Table 1. Specifically, FFM scores were formed based on GR scale scores, equally weighting each scale score when forming the FFM composite. This was done with the SS scores as well, and then scores from the two methods of measurement were correlated. Results for this analysis are shown in Table 5. As seen, aggregation to higher operationalizations resulted in stronger correlations with aligned SS scores, which is to be expected given the composite scales are more reliable. The average correlation in honest conditions was .87 (Sample 1) and .87 (Sample 2) for SFCS scoring, and it was .81 and .82 for OFA scoring. The average correlation in faked settings was .78 and .75 for SFCS scoring, and it was .70 and .68 for OFA scoring. Although not directly comparable because different assessment content was used, the SFCS convergent correlations found here are higher than that found for other FC tests in other studies (e.g., Brown & Maydeu-Olivares, 2011, 2018; Joubert et al., 2015; Lee et al., 2018; SHL, 2013; Usami et al., 2016; Zhang et al., 2020).

5.4 | Evidence of discriminant validity

Both SS scores ($\bar{r} = .29$) and SFCS scores ($\bar{r} = .34$) exhibited similar average intercorrelations within method (i.e., hetero-trait mono-method correlations per multitrait multimethod parlance; Campbell & Fiske, 1959), as shown in Table 2. Table A1 in the Appendix also provides data regarding hetero-trait hetero-method correlations, or the relationships between SFCS scores with nonaligned SS scores. For every SFCS score, the convergent correlation was always stronger than any of the discriminant correlations. The average discriminant correlations for SFCS scores were small and averaged .27, substantially smaller than the average convergent correlations ($r = .75$). Thus, SFCS scores exhibited evidence of discriminant validity.

5.5 | Test–retest reliability

Many researchers estimate FC reliability using a test–retest design. This involves administering the assessment at one sitting and then administering it again later. Random response errors and temporal errors will produce inconsistency in responses, thus attenuating the correlation between Time 1 and Time 2 measurements. The correlation between these measurements is an estimate of reliability. As shown in Table 6, test–retest reliability was strong. The average test–retest reliability in the honest condition was .85 for SFCS

TABLE 5 Forced choice correlations with aligned single-stimulus scores using FFM

Dimension	Sample 1—Honest SFCS	Sample 1—Honest OFA	Sample 1—Faked SFCS	Sample 1—Faked OFA	Sample 2—Honest SFCS	Sample 2—Honest OFA	Sample 2—Faked SFCS	Sample 2—Faked OFA
Extraversion	.87	.84	.81	.75	.87	.87	.80	.75
Agreeableness	.89	.82	.79	.70	.88	.83	.83	.77
Conscientiousness	.86	.75	.77	.65	.88	.75	.71	.62
Emotionally Stability	.87	.83	.71	.63	.84	.78	.70	.63
Openness	.87	.83	.81	.77	.90	.88	.73	.65
Average	.87	.81	.78	.70	.87	.82	.75	.68

Note: Sample 1, $N = 1058$. Sample 2, $N = 271$. Forced choice scores were aggregated to the FFM based on linkages in Table 2. The same was done for single-stimulus scale scores. After aggregating to the FFM, the two sets of scores were correlated with one another.
Abbreviations: FFM, Five Factor Model; OFA, ordinal factor analysis; SFCS, supervised forced choice scoring.

TABLE 6 Test–retest reliability

Forced choice scale	Honest SFCS	Honest OFA	Faked SFCS	Faked OFA
1. Sociability	.88	.90	.80	.83
2. Risk Seeking	.87	.90	.87	.86
3. Leadership	.87	.91	.78	.77
4. Compassion	.81	.80	.84	.82
5. Cooperation	.85	.86	.82	.78
6. Integrity	.85	.79	.70	.76
7. Industriousness	.82	.87	.76	.75
8. Self-Efficacy	.83	.89	.78	.71
9. Detail-Oriented	.81	.81	.81	.81
10. Composure	.85	.85	.79	.77
11. Positive Mood	.88	.88	.84	.82
12. Intellectual	.89	.87	.84	.82
13. Creativity	.89	.87	.85	.85
14. Preference for Variety	.85	.91	.77	.77
15. Aesthetic Preference	.84	.85	.85	.85
Average	.85	.86	.81	.80

Abbreviations: OFA, ordinal factor analysis; SFCS, supervised forced choice scoring.

and .86 for OFA. The average test–retest reliability in the induced applicant condition was .81 for SFCS and .80 for OFA. Thus, for either scoring method and whether taken honestly or faked, test–retest reliability was high.

5.6 | Criterion-related validity

Lastly, Time 1 GR scores were correlated with self-reported work outcomes, as measured during Time 2 in Sample 2. Table 7 contains the correlations between the GR scale scores and the outcome measures. Most SFCS scores were positively related to OCB-I, with the socially oriented scales of Compassion ($r = .48$), Sociability ($r = .47$), and Positive Mood ($r = .46$) being most strongly correlated. Considering all SFCS scores simultaneously, the multiple R with OCB-I was strong using SFCS scoring ($R = .60$, $p < .001$). Although the profile of correlations was similar when using OFA scoring, overall prediction was weaker for IRT (multiple R of .53), and this difference was significant (Steiger $z = 2.33$, $p = .009$).

Many SFCS scores were also positively correlated with OCB-O. The strongest predictors were Industriousness ($r = .42$) and Self-Efficacy ($r = .41$). The multiple R using all SFCS scores was .52 ($p < .001$) for SFCS. Prediction was weaker when using OFA (multiple $R = .46$), and this difference was significant (Steiger $z = 1.91$, $p = .028$).

SFCS Leadership scores were significantly related to whether a person had ever been in a supervisory job ($r = .26$, $p < .001$). The corresponding OFA Leadership score correlated .22 with this outcome, with the SFCS correlation being significantly larger than the OFA correlation (Steiger $z = 1.71$, $p = .044$). SFCS correlations with Leadership Position were also moderate for Self-Efficacy ($r = .20$) and for Intellectual ($r = .18$).

Finally, most SFCS scores were positively related to engagement, with Positive Mood ($r = .43$), Self-Efficacy ($r = .42$), and Industriousness ($r = .38$) having the strongest relationships. Considering all SFCS scores simultaneously, the multiple R with engagement was strong ($R = .57$, $p < .001$). The multiple R for OFA was .55 ($p < .001$), which though trending in the same direction as previous criterion-related findings was not significantly weaker than that found for SFCS (Steiger $z = 0.60$, $p = .274$). Altogether, SFCS scores resulted in meaningful correlations with a number of work-related outcomes.

6 | DISCUSSION

This paper introduced a novel method to score FC assessments, labeled SFCS. The SFCS method is explicitly designed to recover honest scores of normative traits. This paper outlined the logic for

TABLE 7 Criterion-related validity

Forced choice scale	OCB-I		OCB-O		Leadership position		Engagement	
	SFCS	OFA	SFCS	OFA	SFCS	OFA	SFCS	OFA
1. Sociability	.47**	.45**	.28**	.26**	.03	.02	.35**	.35**
2. Risk Seeking	.15*	.20**	.12	.16*	.07	.02	.31**	.35**
3. Leadership	.16*	.15*	.27**	.27**	.26**	.22**	.27**	.30**
4. Compassion	.48**	.48**	.23**	.23**	−.02	−.06	.24**	.23**
5. Cooperation	.13	.01	−.01	−.06	−.16*	−.16*	−.15*	−.20**
6. Integrity	.24**	.18*	.24**	.24**	−.06	−.04	.13	.15*
7. Industriousness	.26**	.11	.42**	.36**	.08	.04	.38**	.30**
8. Self-Efficacy	.28**	.10	.41**	.31**	.19**	.20**	.42**	.32**
9. Detail-Oriented	.09	.28**	.32**	.36**	.10	.05	.20*	.30**
10. Composure	.22**	.18*	.29**	.27**	.03	−.02	.36**	.32**
11. Positive Mood	.46**	.38**	.31**	.31**	−.04	−.03	.43**	.42**
12. Intellectual	.18*	−.03	.28**	.11	.18**	.09	.30**	.11
13. Creativity	.20**	.22**	.31**	.30**	.16*	.11	.33**	.33**
14. Preference for Variety	.19**	.20**	.15*	.20**	.05	.08	.30**	.33**
15. Aesthetic Preference	.22**	.29**	.26**	.35**	.10	.04	.24**	.34**
Multiple *R*	.60**	.53**	.52**	.46**			.57**	.55**

Note: Leadership Position is a dichotomous variable regarding whether a person ever had a job where he or she formally supervised others. Results presented for total sample across honest and induced applicant conditions. *N* = 207 for leadership outcome. *N* = 191 for other outcomes, which removed anyone who was unemployed and not at their current job for at least 1.5 months, at the time of data collection. Shown are zero-order correlations. A multiple *R* value is not provided for Leadership Position because the outcome is dichotomous.

Abbreviations: OCB, organizational citizenship behavior, OCB-I being people focused and OCB-O being organization focused; OFA, ordinal factor analysis; SFCS, supervised forced choice scoring.

*p < . 05; **p < .01.

SFCS and then as a proof-of-concept endeavor tested the SFCS method across several samples. Collectively, support was found for the SFCS method, such that SFCS scores exhibited strong test-retest correlations, exhibited small mean shifts in scores between honest and induced applicant conditions, converged strongly with aligned SS scores, and had meaningful correlations with work outcomes. Furthermore, psychometric evidence was generally stronger for SFCS scores than more traditional FC scoring. Several of these findings will be explored in more detail in the following section.

7 | EXPLORING THE BENEFITS OF SFCS

As outlined in the introduction, there are several likely reasons why SFCS may improve the psychometrics of FC scores. In this section, we will discuss how SFCS maximizes usage of FC response variance and how it does this in several ways. First, SFCS can adjust item weights based on how strongly they align to the targeted construct, therefore allowing for differential item weighting to maximize the relationship with honest personality. This ensures that the items are optimally weighted to reflect the intended construct. Second, by using ML, SFCS more comprehensively leverages all available predictor data. In the parlance of factor analysis, SFCS naturally identifies and then establishes cross-loadings when forming trait scores. Random forests was used as the ML method in this study, and the developed models can be examined to determine the degree to which unaligned items influenced SFCS scores. Specifically, variable importance statistics were examined, which represent the mean decrease in accuracy based on random permutations in out-of-bag samples. Within this study, inspection of the item importance values made clear that nonaligned items were important when estimating trait scores. In some cases, the influence was minor. For example, of the 10 items most strongly contributing to estimating Sociability, only one came from another scale (Positive Mood). However, other trait algorithms were more strongly influenced by other trait content, such as Detail Orientation, where four of the top ten items came from other scales (one from Integrity, three from Industriousness). Thus, SFCS leveraged a greater variety of items when deriving trait estimates, thus making more efficient use of the data.

Third, SFCS also allows for more complex scaling of response patterns. The particular algorithm used in this study (random forests) is capable of modeling nonlinearities in the data, which is likely to be relevant when FC items use a GR format. If the relationship between ordinal item responses and estimated construct scores is nonlinear, SFCS scoring will be beneficial when a nonlinear ML algorithm is used. In this study, most items exhibited at least a modest quadratic effect ($\Delta R^2 > .01$ for quadratic term over linear term). In cases where quadratic effects were stronger ($\Delta R^2 > .05$), this often occurred when the paired statements were similar in content (e.g., Sociability and Leadership) and had similarly high desirability (i.e., both statements desirable if applying for a job). However, it should be noted that the GR format is a unique version of the broader FC format, and thus FC formats that do not allow for nonlinear response patterns (e.g., most/least) might not produce strong gains when SFCS scoring is used because of this.

Additionally, another benefit of SFCS is its targeted approach to recovering honest personality scores. As evidenced, the SFCS method exhibited only modest increases in mean scores from honest to induced applicant conditions, like past research. However, correlations with aligned SS scores remained strong even within the induced applicant condition, supporting the validity of the trait scores. The correlations in this study exceeded those from past research using similar designs (Fisher et al., 2018; Usami et al., 2016). That said, it is important to note that correlations did decrease from honest to induced applicant conditions. Thus, the SFCS method was not invulnerable to the effects of faking. This should not be surprising, as decades of research have yet to completely solve the faking problem. However, the strong psychometric properties found using the SFCS method, across both honest and induced applicant conditions, suggests that this scoring method may be a viable way to maximize FC construct validity, even when respondents are faking.

8 | STUDY LIMITATIONS AND AREAS FOR FUTURE RESEARCH

There are several shortcomings to this study and the SFCS method. For one, the FC measure in this study used a GR format. Little FC research has used the GR format (see Brown & Maydeu-Olivares, 2018), with most FC assessments instead using dyads or most/least item blocks. This is important because GR formats may be more susceptible to response biases like acquiescence than other FC formats, and this might impact convergence with other measures like SS scores. Because other FC formats are more common, this study's results may not generalize well to other FC tests, highlighting a significant limitation of this study. Further research should examine whether SFCS produces validity gains for other FC formats. The comparison of SFCS to other scoring methods might not generalize to other FC tests, and comparing these methods with only one FC assessment is a limitation of the current study.

On the other hand, many of the SFCS benefits discussed in the introduction would still likely apply to other FC formats, including maximizing use of response variance by allowing all items to contribute to trait estimates (likely important with shorter FC scales), allowing for adjustments of item weights based on strength of construct relationship, and explicitly considering the role of faking when creating scoring algorithms. That said, non-GR formats such as triads will not exhibit curvilinear effects, and thus the ability of ML to account for curvilinear effects will not be relevant. Future research should compare SFCS and other FC scoring methods across item formats (e.g., triads) and across other factors that impact FC construct validity, such as the number of measured traits, the consistency in social desirability match within item block, and the absolute magnitude of factor loading of items (e.g., Bürkner et al., 2019; Schulte et al., 2021).

Second, honest SS scores were used as target variables within this study. This is because SS items are straightforward to respond to (only directly influenced by one statement), allow for simplistic factor analytical models to establish validity, have more variance given their ordinal/interval scale, and have higher item reliability than honest FC scores (Brown & Maydeu-Olivares, 2011, 2018; Lee et al., 2018; Zhang et al., 2020). However, although we treated SS scores as our honest, normative, and targeted trait scores, these scores are not perfect reflections of the latent construct. Beyond the random error that affects any measure, SS measures are also affected by other response biases beyond impression management, such as positive self-deception (e.g., Paulhus, 1984) or acquiescence, which cannot be controlled for using SFCS. SFCS minimizes intentional response distortion, or in other words, impression management due to wanting to appear favorable (Paulhus, 1984). This is most relevant to applicant contexts. However, positive self-deception will still impact any self-report measure, including the honest SS scores used for SFCS (also the FC responses, for that matter). Positive-self-deception is defined as the tendency to hold overly positive views of oneself; this occurs subconsciously and is related to healthy personality and self-esteem (e.g., Burns & Christiansen, 2006; Paulhus, 1984). Although positive self-deception is a concern, it is also important to note that Paulhus (1984) recommends that only impression management be controlled for, as positive-self-deception represents important aspects of a person's personality. As such, the SFCS method does not intend to correct for positive self-deception, nor could it without collecting non-self-report data.

Regardless, future research should consider use of other possible target trait variables. Reviewers of this paper repeatedly argued that it is unclear what the best target variables may be, and we concur. Among other alternatives might be observer ratings of personality, which seem particularly well-suited because they are based on observable expressions of behavior. The review team also suggested trying to key responses upon honest FC scores, though ample data did not exist to perform this within the current study. Future research on target variable choice is warranted, and examining psychometric properties across a range of target variables may serve as a cross-validation check of SFCS.

Even though data did not exist to train algorithms that translate FC responses to recreate honest FC scores, it is possible to demonstrate the supervised construct scoring is not limited to FC assessment formats or SS scores as the target variables. At the request of reviewers, we trained algorithms to translate SS responses to recreate honest FC scores as the targets ($N = 324$). There are limitations to this analysis.[8] However, the analysis can provide evidence that empirical scoring via supervised construct scoring can be applied to different predictor formats and target variable formats. Thus, we trained ML algorithms to recreate honest FC scores based on honest SS responses, finding an average convergent correlation of .85 (Sample 1). This is similar to the convergent correlations found using SFCS with honest FC responses ($\overline{r}$ = .81) and larger than when traditionally scored SS and honest FC scores are correlated ($\overline{r}$ = .72). As such, supervised construct scoring can likely be extended to different predictor formats and different target variable formats to increase convergence between assessment scores and target variable scores. This affords researchers more flexibility in choosing target variables based on theoretical and psychometric considerations.

In relation to the limitations just discussed, the review team also wondered whether SFCS might possibly combine the errors from both SS and FC measurement formats when using SS scores as the target variables (i.e., combining the possible bias from SS scores and the possible increased random error in FC scores). It certainly is possible that if the target scores are somehow biased, that SFCS scores could capture those biases. However, that would only occur if the FC responses also contained variance that overlapped with the same biases, as the predictor variance (i.e., FC scores) is ultimately what is used to produce predicted scores. Thus, SFCS would capture SS bias only if that bias already existed in FC scores, in which case the concern is unfounded because the bias was present irrespective of whether SFCS were used or not. It is unclear just how often shared bias overlap exists between SS and FC measures. For example, SS measures are sometimes criticized for susceptibility to acquiescence, but acquiescence bias is not thought to be an issue for FC scores. Thus, although it is possible SFCS may accentuate shared bias between FC and SS measures, we do not see reason to believe this would be a major concern, at least for the design used in this study. That said, an important requirement of SFCS is the use of reliable, well-understood target scores. To the extent target scores are unreliable or biased, SFCS will subsequently be less appropriate for use.

Another limitation is that although this study examined the impact of response distortion, the applicant context was simulated (i.e., induced applicant setting). Ideally, validity could be examined when the assessment is taken by actual applicants and then later correlated with SS scores taken honestly. The same can be said when examining correlations with criterion outcomes. Thus, it is unclear how well the SFCS generalizes to new contexts. Relatedly, it might be ideal to develop SFCS algorithms using actual job applicants who later respond to the SS measure honestly. A concern over SFCS and more generally with ML is that if the training sample is not representative of the broader population, the algorithm could be biased and perform poorly in new contexts.[9] If the calibration sample differs in motivation or the distribution of trait scores, this might affect SFCS algorithm performance. For example, if we trained a SFCS algorithm using honestly responding accountants, would this generalize to a setting for call center applicants? Future research should examine how the SFCS method operates within actual applicant settings, both in terms of calibrating models but also in terms of testing model validity. It is advisable that SFCS scores only be applied to new contexts when those contexts are similar to the calibration setting (e.g., in terms of motivation, type of respondents, job type).

This concern was raised by a reviewer who argued that item invariance should also be considered when training these models. Our approach of using both honest and induced applicant responses in the calibration sample was done to develop algorithms that generalize across various motives and represent the varied motives of applicants within operational contexts (e.g., Griffith & Robie, 2013). Empirical evidence in this study showed that the SFCS algorithms cross-validated well in both honest and induced applicant settings. Nonetheless, this strategy may have masked true differences in contextually dependent response patterns, resulting in suboptimal algorithms not fully customized for their intended context. Thus, future research might consider the role of item invariance (by context) when training SFCS models.

The review team also raised questions about the interpretability of SFCS scores. One of the challenges to ML is lessened interpretability, with some describing these methods as "black box." However, an important factor that distinguishes our ability to interpret algorithmic scores is the job-relevance of the input predictors. If we incorporate a bunch of online meta-data, survey results, and other "digital breadcrumbs" into a ML algorithm to predict an outcome, it will be challenging to understand what is driving prediction. On the other hand, if we only include predictors that are judged to be job-relevant (i.e., content valid), this adds a rational foundation to our inferences. Remember, validity is a unitarian concept (Binning & Barrett, 1989) and on a continuum. When personality scores correlate highly with convergent measures and external criteria, this provides some evidence of validity. This is further bolstered by only including job relevant content from the start (i.e., adopt a "no garbage in" mentality). By only including well-understood predictors, researchers have some concept of what is driving prediction. Furthermore, many ML algorithms commonly used in the social sciences, such as elastic net or random forests, provide variable importance statistics, thus allowing for interpretation of how variables relate to the target variable. Lastly, unlike most applications of ML, which are trained to predict complex criteria (e.g., job performance), SFCS specifically targets well understood trait scores, and the inputs are personality items written specifically to measure an intended trait. Thus, the task itself is designed to create scores that are well understood.

That said, future research is needed to better understand the balance between traditional psychometrics and ML. This is true beyond just FC measures and SFCS. As one example of this, some response sets occur differentially by demographic and cultural groups (T. Johnson et al., 2005; Wetzel et al., 2013), raising concerns over potential aspects of test fairness. Although this concern applies outside of SFCS too, it warrants further consideration regarding validity requirements and understanding exactly what algorithm scores assess. With traditional scoring (e.g., summative scoring), it is clear how item responses are combined to estimate trait scores. With SFCS scoring, this understanding may not be so clear cut. Explainable ML has emerged as a popular field of study (e.g., Samek &

Muller, 2019), with a number of potential procedures to improve understanding of ML models such as local interpretable model-agnostic explanations (LIME) and partial dependency plots (Goldstein et al., 2013). These methods could potentially be useful when performing SFCS, as might using ML algorithms that are generally more interpretable, such as elastic net, where the regression weights of FC items could be examined as to whether they are zero and how large they are.

Another study limitation is the dependent variables examined. This study examined relationships with several work outcomes, including engagement, organizational citizenship behaviors, and leadership position. While these outcomes are important, the gold standard of test validation is to compare test scores to measures of job performance. Future research should collect performance ratings from sources other than those completing the assessment (e.g., from respondents' managers), or use existing performance metrics within organizations (e.g., sales data), to test the criterion-related validity of SFCS assessments. In this study all responses were self-report. As such, common method variance likely did exist. This may explain why SFCS scores correlated more strongly with the work outcomes in this study than OFA scores did, given that the SFCS method was used to reproduce self-report SS scores and the dependent variables were also of a self-report SS format. To truly understand the validity of the SFCS method, traditional criterion-related validation studies using supervisor ratings should be performed.

Finally, the SFCS method may vary in effectiveness based on which ML algorithm is used. Supervised scoring, which is the basis for this method, is defined as linking a set of inputs (e.g., personality item responses) to an outcome variable (e.g., existing trait scores) by way of predictive algorithm. Numerous ML methods fall under this umbrella. This study considered several algorithms when scoring the FC assessment, ultimately landing on random forests. Given the broad scope of this paper, attention was not devoted to the different possible ML methods. However, the choice of ML algorithm can be important; the performance of a ML method is likely to differ by study design and design of the FC measure. For example, if sample sizes had been larger in this study, it is possible deep neural networks might have been the best choice, given their history of strong prediction (e.g., Goodfellow et al., 2016). Alternatively, for different FC formats such as most/least, simpler algorithms such as elastic net might perform well. The predictiveness of different ML algorithms is often context specific. Because of this, we do not restrict SFCS to any single supervised ML algorithm. That said, future research should explore the pros and cons of different ML algorithms based on factors such as study sample size and the format of FC items. Sample size in particular is likely to affect results. There are no strict guidelines for required sample sizes for most ML methods, as it depends on the complexity of the data, model, and the strength of relationships between the inputs and the criterion variable. Thus, like in many domains, more data is always better, and researchers should always be cognizant of whether small sample sizes will attenuate the performance of SFCS models.

9 | CONCLUSION

FC personality testing is an attractive method of assessing personality in high-stakes settings. The current paper introduced a new method of scoring FC tests, explicitly designed to recover honest trait scores. Initial evidence of the SFCS scoring method was promising, as the developed assessment exhibited strong reliability and validity evidence, even under faking conditions. We encourage future research testing the viability of SFCS across different FC assessments and under different conditions.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Andrew B. Speer http://orcid.org/0000-0002-3376-2103

ENDNOTES

[1] Although graded response items are a form of FC, they are less frequently researched than other formats. As highlighted by an expert reviewer, this limits the generalizability of this study and may be a significant limitation of this study.

[2] We thank the review team for articulating this point.

[3] Relatedly, respondents react more negatively to the FC format than the SS format (Bowen et al., 2002; Converse et al., 2008).

[4] A larger number of respondents were removed for nonpurposeful responding using MTurk, with 43.5% of cases removed ($N = 349$) for failing a response check, resulting in a final MTurk $N = 452$. Only 8% ($N = 53$) of respondents were removed from the Prolific sample and using the same response checks, resulting in a final sample of 606 for Prolific.

[5] Note that we conducted power analyses before data collection, with over 80% power.

[6] It should be noted that even despite including honest responders along with instructed fakers, this simulated context still likely does not adequately reflect faking tendencies in real applicant settings.

[7] It is possible that other statistical programs such as MPlus or Stan may have produced slightly different results in estimated scores (Bürkner et al., 2019).

[8] Some limitations include: (1) the paper intent was to develop new scoring for FC assessments and not SS assessments, (2) the SS measure was taken honestly and thus these analyses do not address the issue of faking, and (3) the small sample size prevents robust ML. Because of the smaller sample size, we used elastic net regression instead of random forests. To account for any quadratic effects, we translated SS responses into generative additive model scores (James et al., 2017) formed by incorporating linear and quadratic effects simultaneously using a combined term reflecting these effects as input into the elastic net algorithm. This was performed within nested *k*-folds cross-validation to prevent overfitting.

[9] Though this concern applies to other FC scoring methods as well.

REFERENCES

Anguiano-Carrasco, C., MacCann, C., Geiger, M., Seybert, J. M., & Roberts, R. D. (2015). Development of a forced-choice measure of typical-performance emotional intelligence. *Journal of Psychoeducational Assessment*, *33*, 83–97.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, *15*, 263–272.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*, 478–494.

Bowen, C. C., Martin, B. A., & Hunt, S. T. (2002). A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. *The International Journal of Organizational Analysis*, *32*, 247–256.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

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

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

Brown, A., & Maydeu-Olivares, A. (2018). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling*, *25*, 516–529.

Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of thurstonian IRT models. *Educational and Psychological Measurement*, *79*, 827–854.

Burns, G. N., & Christiansen, N. D. (2006). Sensitive or senseless: On the use of social desirability measures in selection and assessment. In R. L. Griffith, & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 113–148). Information Age.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminate validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

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*, 1347–1368.

Carpenter, N. C., Berry, C. M., & Houston, L. (2014). A meta-analytic comparison of self-reported and other-reported organizational citizenship behavior. *Journal of Organizational Behavior*, *35*, 547–574.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon mechanical turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*, 347–361.

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

Christiansen, N. D., Robie, C., Burns, G. N., & Speer, A. B. (2017). Using item-level covariance to detect response distortion on personality measure. *Human Performance*, *30*, 116–134.

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*, 155–169.

Conway, J. M. (2000). Managerial performance development constructs and personality correlates. *Human Performance*, *13*(1), 23–46.

Cucina, J. M., Vasilopoulos, N. L., Su, C., Busciglio, H. H., Cozma, I., DeCostanza, A. H., Martin, N. R., & Shaw, M. N. (2019). The effects of empirical keying of personality measures on faking and criterion-related validity. *Journal of Business and Psychology*, *34*(3), 337–356.

Dalal, D. K., Zhu, X. S., Rangel, B., Boyce, A. S., & Lobene, E. (2019). Improving applicant reactions to forced-choice personality measurement: Interventions to reduce threats to test takers' self-concepts. *Journal of Business and Psychology*, *36*, 1–16.

Dalal, R. S., Lam, H., Weiss, H. M., Welch, E. R., & Hulin, C. L. (2009). A within-person approach to work behavior and performance: Concurrent and lagged citizenship-counter productivity associations, and dynamic relationships with affect and overall job performance. *Academy of Management Journal*, *52*, 1051–1066.

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

Fisher, P. A., Robie, C., Christiansen, N. D., & Komar, S. (2018). The impact of psychopathy and warnings on faking behavior: A multisaturation perspective. *Personality and Individual Differences*, *127*, 39–43. https://doi.org/10.1016/j.paid.2018.01.033

Fisher, P., Robie, C., Christiansen, N. D., Speer, A. B., & Schneider, L. (2019). Criterion-related validity of forced-choice personality measures: A cautionary note regarding Thurstonian IRT versus classical test theory scoring. *Personnel Assessment & Decisions*, *5*, 49–61.

Furr, R. M. (2018). *Psychometrics: An introduction* (3rd ed.). Sage.

Gatewood, R. D., Field, H. S., & Barrick, M. R. (2016). *Human resource selection* (8th ed.). Cengage Learning

Goffin, R. D., Jang, I., & Skinner, E. (2011). Forced-choice and conventional personality assessment: Each may have unique value in pre-employment testing. *Personality and Individual Differences*, *51*(7), 840–844.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2013). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *arXiv*, *1309*, 6392.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press

Griffith, R. L., & Robie, C. (2013). Personality testing and the "F-word": Revisiting seven questions about faking. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 101–128). Routledge.

Guan, L. (2015). *Personality, faking, and the ability of identify criteria: Can forced choice formats untangle their relationships*? [Unpublished master thesis]. University of Virginia.

Holden, R. R. (2008). Underestimating the effects of faking on the validity of self-report personality scales. *Personality and Individual Differences*, *44*, 311–321.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R* (7th ed.). Springer

Jeong, Y. R., Christiansen, N. D., Robie, C., Kung, M.-C., & Kinney, T. B. (2017). Comparing applicants and incumbents: Effects of response distortion on mean scores and validity of personality measures. *International Journal of Selection and Assessment*, *25*, 311–315.

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78–89.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*, 264–277.

Joubert, T., Inceoglu, I., Bartram, D., Dowdeswell, K., & Lin, Y. (2015). A comparison of the psychometric properties of the forced choice and likert scale versions of a personality instrument. *International Journal of Selection and Assessment*, *23*, 92–97.

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology*, *93*, 140–155.

König, C. J., Jansen, A., & Mathieu, P. L. (2017). What if applicants knew how personality tests are scored? A minimal intervention study. *Journal of Personnel Psychology*, *16*, 206–210.

König, C. J., Merz, A.-S., & Trauffer, N. (2012). What is in applicants' minds when they fill out a personality test? Insights from a qualitative study. *International Journal of Selection and Assessment*, *20*(4), 442–452.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial & Organizational Psychology*, *8*, 142–164.

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.

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*, 222–248.

Ng, V., Lee, P., Ho, M. H. R., Kuykendall, L., Stark, S., & Tay, L. (2020). The development and validation of a multidimensional forced-choice format character measure: Testing the Thurstonian IRT approach. *Journal of Personality Assessment*, *103*, 1–14.

Palan, S., & Schitter, C. (2018). Prolific.ac: A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*, 598–609.

Peterson, M. H., Griffith, R. L., Isaacson, J. A., O'Connell, M. S., & Mangos, P. M. (2011). Applicant faking, social desirability, and the prediction of counterproductive work behaviors. *Human Performance*, *24*, 270–290.

Popper, M., Amit, K., Gal, R., Mishkal-Sinai, M., & Lisak, A. (2004). The capacity to lead: Major psychological differences between leaders and nonleaders. *Military Psychology*, *16*(4), 245–263.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*, 689–732.

Salgado, J. F., & Tauriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, *23*, 3–30.

Samek, W., & Muller, K. R. (2019). Towards explainable artificial intelligence. *arXiv*, *1909*, 1207.

Schaufeli, W. B., Shimazu, A., Hakanen, J., Salanova, M., & De Witte, H. (2019). An ultra-short measure for work engagement: The UWES-3 validation across five countries. *European Journal of Psychological Assessment*, *35*, 577–591.

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

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*, 262–289.

SHL. (2013). *OPQ32r technical manual*. SHL.

Shoss, M., & Witt, L. A. (2013). Trait interactions and other configural approaches to personality. In N. Christiansen & R. Tett (Eds.), *Handbook of personality at work* (pp. 392–418). Routledge.

Siegling, A. B., Nielsen, C., & Petrides, K. V. (2014). Trait emotional intelligence and leadership in a European multinational company. *Personality and Individual Differences*, *65*, 65–68.

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*, 184–203.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25–39.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Topping, G. D., & Gorman, J. G. O. (1997). Effects of faking set on validity of the NEO-FFI. *Personality & Individual Differences*, *23*, 117–124.

Usami, S., Sakamoto, A., Naito, J., & Abe, Y. (2016). Developing pairwise preference-based personality test and experimental investigation of its resistance to faking effect by item response model. *International Journal of Testing*, *16*, 288–309.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurements. *Educational and Psychological Measurement*, *59*, 197–210.

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences*, *34*, 69–81.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.

De Young, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*, 880–896.

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*, 569–590.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Speer, A. B., & Delacruz, A. Y. (2021). Introducing a supervised alternative to forced-choice personality scoring: A test of validity and resistance to faking. *International Journal of Selection and Assessment*, 29, 448–466. https://doi.org/10.1111/ijsa.12345

APPENDIX

TABLE A1 Convergent and discriminant correlations between SFCS scores and single-stimulus scores

	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FC9	FC10	FC11	FC12	FC13	FC14	FC15
1. Sociability	**0.80**	0.52	0.47	0.53	−0.22	0.24	0.21	0.36	−0.02	0.42	0.63	0.25	0.36	0.50	0.23
2. Risk Seeking	0.45	**0.73**	0.51	0.19	−0.46	−0.08	0.13	0.31	−0.12	0.28	0.30	0.39	0.46	0.61	0.33
3. Leadership	0.46	0.54	**0.80**	0.18	−0.57	−0.01	0.31	0.50	0.07	0.32	0.28	0.45	0.49	0.55	0.28
4. Compassion	0.48	0.19	0.16	**0.78**	0.15	0.38	0.29	0.28	0.22	0.22	0.55	0.22	0.28	0.25	0.33
5. Cooperation	−0.18	−0.43	−0.51	0.15	**0.75**	0.34	0.02	−0.14	0.23	−0.02	0.09	−0.28	−0.25	−0.37	−0.09
6. Integrity	0.17	−0.08	0.00	0.35	0.31	**0.78**	0.43	0.34	0.43	0.32	0.38	0.04	0.05	−0.02	0.09
7. Industriousness	0.19	0.12	0.31	0.30	0.03	0.42	**0.75**	0.67	0.63	0.33	0.34	0.30	0.26	0.20	0.19
8. Self-Efficacy	0.29	0.28	0.44	0.26	−0.15	0.32	0.64	**0.72**	0.49	0.49	0.39	0.41	0.36	0.35	0.24
9. Detail-Oriented	0.00	−0.11	0.10	0.21	0.20	0.41	0.56	0.47	**0.72**	0.23	0.22	0.20	0.15	−0.01	0.16
10. Composure	0.34	0.28	0.31	0.21	−0.08	0.31	0.37	0.52	0.26	**0.72**	0.48	0.25	0.21	0.29	0.12
11. Positive Mood	0.55	0.32	0.27	0.55	0.05	0.41	0.35	0.43	0.22	0.51	**0.74**	0.18	0.28	0.29	0.23
12. Intellect	0.23	0.41	0.46	0.25	−0.33	0.07	0.34	0.47	0.22	0.27	0.22	**0.78**	0.58	0.52	0.54
13. Creativity	0.32	0.45	0.47	0.29	−0.28	0.08	0.29	0.42	0.16	0.25	0.30	0.57	**0.74**	0.51	0.65
14. Preference for Variety	0.47	0.65	0.53	0.28	−0.41	0.01	0.21	0.38	−0.03	0.30	0.30	0.52	0.55	**0.76**	0.42
15. Aesthetic Preference	0.19	0.30	0.25	0.33	−0.10	0.07	0.20	0.26	0.18	0.13	0.22	0.51	0.63	0.36	**0.74**
Average discriminant	0.28	0.25	0.27	0.29	−0.13	0.21	0.31	0.38	0.21	0.29	0.34	0.29	0.31	0.29	0.27

Note: $N = 1058$ in Sample 1. Bolded values are convergent correlations. "FC" column headings refer to graded response forced choice, such that the first column pertains for SFCS Sociability scores and the last column SFCS Aesthetic Preference scores.

Abbreviation: SFCS, supervised forced choice scoring.