The Development and Validation of a Multidimensional Forced-Choice Format Character Measure: Testing the Thurstonian IRT Approach

Vincent Ng, Philseok Lee, Moon-Ho Ringo Ho, Lauren Kuykendall, Stephen Stark & Louis Tay

To cite this article: Vincent Ng, Philseok Lee, Moon-Ho Ringo Ho, Lauren Kuykendall, Stephen Stark & Louis Tay (2020): The Development and Validation of a Multidimensional Forced-Choice Format Character Measure: Testing the Thurstonian IRT Approach, Journal of Personality Assessment, DOI: 10.1080/00223891.2020.1739056

To link to this article: https://doi.org/10.1080/00223891.2020.1739056
The Development and Validation of a Multidimensional Forced-Choice Format Character Measure: Testing the Thurstonian IRT Approach

Vincent Ng1, Philseok Lee2, Moon-Ho Ringo Ho3, Lauren Kuykendall2, Stephen Stark4, and Louis Tay5

1Department of Psychology, University of Houston, Houston, Texas; 2Department of Psychology, George Mason University, Fairfax, Virginia; 3School of Humanities and Social Sciences, Nanyang Technological University, Singapore; 4Department of Psychology, University of South Florida, Tampa, Florida; 5Department of Psychological Sciences, Purdue University, West Lafayette, Indiana

ABSTRACT
There has been reemerging interest within psychology in the construct of character, yet assessing it can be difficult due to social desirability of character traits. Forced-choice formats offer one way to address response bias, but traditional scoring methods (i.e., ipsative) associated with this format makes comparing scores between people problematic. Nevertheless, recent advances in modeling item responding (Thurstonian IRT) enable scoring that recovers absolute standing on latent traits and allows for score comparisons between people. Based on recent work in character measurement (CIVIC), we developed a multidimensional forced-choice measure of character (CIVIC-MFC) and scored it using Thurstonian IRT. Initial validation using a sample of 798 participants demonstrated good support for factorial, convergent, and concurrent validity for scores on the CIVIC-MFC, although they did not demonstrate more faking resistance than scores on a Likert-type format version. Potential explanations are discussed.

ARTICLE HISTORY
Received 1 August 2018
Accepted 31 January 2020

Introduction
The past decade has seen a resurgence of interest in the construct domain of character, with many proclaiming that despite early rejection (Allport, 1937) character is not dead within the social sciences (Fleeson, Furr, Jayawickreme, Meindl, & Helzer, 2014; Wright & Goodstein, 2007). Indeed, there have been significant methodological and study design advances in conjunction with innovative reconceptualizations of personality models (e.g., Fleeson, 2001) that have largely undermined the many criticisms leveled against the existence of stable traits in general (Mischel, 1968) and character in particular (e.g., Doris, 1998).

Character includes the morally-valued attributes (e.g., kindness, temperance or self-control, humility) that engender well-being, allow people to live well together as social creatures, and lead to societal flourishing and ethical behavior (Cohen, Panter, Turan, Morse, & Kim, 2014; Dahlsgaard, Peterson, & Seligman, 2005; McCullough & Snyder, 2000; Ng, Tay, & Kuykendall, 2018; Peterson & Seligman, 2004; Robertson, 2013). Research has shown that character traits are not redundant with and increment beyond these “normal” personality traits, as represented by the Big Five and HEXACO taxonomies, in predicting character-relevant outcomes (McGrath, Hall-Simmonds, & Goldberg, 2020; Ng et al., 2018).

People appear to be able to reliably report on their character and manifest it in their behavior. Character traits have shown moderately strong self- and informant-report convergence (Peterson & Seligman, 2004; .29 to .69, median = .40, Ruch et al., 2010), which is typical for self- and other-report convergence on most psychological constructs (Fowers, 2014). It has also been shown that there is moderate inter-informant agreement about perceptions of a target’s moral character (Helzer et al., 2014). In terms of outcomes, character predicts positive engagement in organizational citizenship behaviors (i.e., discretionary effort by employee that benefits the organization; Organ, 1988) and negative engagement with counterproductive work behaviors (i.e., actions that run counter to organizational norms, aims, and goals; e.g., Robinson & Bennett, 1995) in the work context (Cohen et al., 2014), job and contextual performance (Harzer & Ruch, 2014), and well-being (e.g., life satisfaction; Park, Peterson, & Seligman, 2004; see Niemic, 2013). In short, character matters. Consequently, accurately assessing character is important. However, some have raised concerns about the extent to which character can be measured without significant social desirability response bias affecting the scores (e.g., Fowers, 2014).

The goal of this paper is to draw on the past methodology of the multidimensional forced-choice (MFC) item format and combine it with current advances in scoring using Thurstonian item response theory to address potential
issues of response bias. In the following we briefly (a) elucidate key factors unique to character and its assessment that makes addressing the potential for socially desirable response bias particularly salient, (b) review past issues with using the MFC format as a solution to prevent intentional response distortion, and (c) highlight recent advances in MFC format scoring methods that allow recovery of comparable (i.e., normative) scores between respondents. The product of this investigation is the development and validation of, to our knowledge, the first forced-choice format measure of character.

Response biases

Response bias can be defined as introduction of systematic variance to response scores on a measure that is unrelated to the constructs the measure intends to tap (Paulhus, 1991). Response bias comes in three general forms (i.e., socially desirable responding, acquiescence, and extremity bias)—all of which are relevant to character assessment. In the measurement of character, the most salient of these is socially desirable responding (SDR) or the tendency to convey oneself in a flattering way.

Researchers (e.g., Fowers, 2014) have stated that SDR is a primary concern in the accurate assessment of character for several reasons. First, character traits are themselves socially desirable (Peterson & Seligman, 2004). As SDR researchers have mentioned, response bias of this kind is especially problematic when the constructs being measured are value-laden because the evaluative nature of the construct indicates how to respond if one intends to “fake-good” (Rosse, Stecher, Miller, & Levin, 1998). Second, a related concern that would likely interact with value-ladenness of the construct being measured is the degree of item transparency. Some character researchers have acknowledged concerns about how transparent their character trait measure is since it might increase the potential for people to engage in SDR (Duckworth, Peterson, Matthews, & Kelly, 2007), a problem that may be especially relevant in character assessment given how desirable character traits are. Third, many have theorized or found (Goffin & Boyd, 2009; Hogan, Barrett, & Hogan, 2007; McFarland & Ryan, 2000, 2006; Snell, Sydell, & Lueke, 1999) that certain evaluative traits indicative of character (i.e., integrity, Machiavellianism, manipulativeness, social acumen, innovation or creativity) should or do predict ability or motivation to engage in faking. Thus, an issue peculiar to character assessment is that one’s standing on the measured construct may contribute to SDR and thus undermine accurate assessment of that very construct for some participants.

These problems are exacerbated if a character measure is to be used in a high-stakes employment selection context, which is not unlikely, considering the recent special issues devoted to defining and measuring character as it relates to leadership in consulting psychology journals (Leonard, 1997; Thompson & Riggio, 2010) and related interest in management journals (e.g., Wright & Lauer, 2013). The selection context represents an entirely different situation from the research context within which a selection instrument is developed and validated using volunteer participants (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). Applicants’ job need, desire, and aspiration as well as perceived competitiveness of other applicants (Goffin & Boyd, 2009; McFarland & Ryan, 2000) can feed into perceptions about subjective norms surrounding faking and perceived willingness of others to fake (McFarland & Ryan, 2006; Snell et al., 1999). All these situational factors can increase likelihood of “faking good” to attain a desirable outcome like employment.

Conversely, clinical diagnosis of certain psychological conditions (e.g., attention-deficit/hyperactivity disorder) represents a context in which respondents might be motivated to engage in “undesirable responding” or “faking bad” on such character cores as temperance (i.e., self-regulation in pursuit of long-term goals, strategic planning, emotional self-awareness; Ng et al., 2018) in order to procure prescription drugs that may boost mental performance (Sollman, Ranseen, & Berry, 2010).

Multidimensional forced-choice format: Scale construction and scoring

Forced-choice format rating scales were among the earliest methods created to address social desirability response bias and “faking good” on measures (FC; Christiansen, Burns, & Montgomery, 2005). Despite gaining early interest and evidence of validity (Zavala, 1965), the method fell out of favor (Christiansen et al., 2005). First, studies investigating whether FC formats could be faked indicated that respondents were able to do so if given a target job to fake toward (e.g., Dunnette, McCartney, Carlson, & Kirchner, 1962). However, Christiansen et al. (2005) argued that a better criterion for judging the efficacy of the FC technique is not whether respondents can successfully distort responses on the measure at all, but rather if the degree of response distortion is smaller than other methods. Meta-analytic findings have since indeed shown that FC measures lead to smaller mean shifts in scale scores between honest and induced-faking conditions compared to non-FC measures (weighted mean d = .48 vs. .63, respectively; Stanush, 1997, p. 106), even when respondents are told to fake toward a role (e.g., specific job) that would convey information about which characteristics are desirable (e.g., Christiansen et al., 2005; Martin, Bowen, & Hunt, 2002; but c.f. Heggestad, Morrison, Reeve, & Mccloy, 2006). A more recent meta-analysis has shown again that the FC format demonstrates lower group-level mean shifts in scale scores compared to the single stimulus format under faking conditions (Cao & Drasgow, 2019).

Using forced-choice format rating scales for the purpose of selection requires two aims be fulfilled. First, that response biases be negated or at very least, as mentioned above, that they are sufficiently impeded relative to other available rating scale formats. This is to ensure that scores are maximally reflective of the latent traits they are intended to measure and minimally contaminated by systematic error variance. Second, that the scores derived from such a scale
allow for between-person comparisons on the latent trait of interest (i.e., normative scores) since selection entails comparable measurement of persons (Cao & Drasgow, 2019). The former aim is fundamentally a function of scale construction (McCloy, Heggestad, & Reeve, 2005) and the latter can be addressed directly through scale construction or indirectly through scoring methodology (Hicks, 1970).

Regarding response bias and scale construction, in contrast to traditional, single stimulus formats, the basic FC format presents respondents with an item block comprised of two or more statements, each usually representing a different construct (Heggestad et al., 2006), and having respondents either choose between or rank statements in terms of how descriptive they are of the respondent (e.g., three-statement item blocks = triplets; Brown & Maydeu-Olivares, 2011; Chernyshenko et al., 2009, see online supplemental materials, Figure S1). Critically, statements within an item block are selected and matched to be approximately equal in social desirability during scale creation. Since respondents must choose between equally desirable statements within each item block, identifying the most socially desirable option should theoretically be impeded (Lee, Lee, & Stark, 2018; Pavlov, Maydeu-Olivares, & Fairchild, 2019) and faking prevented because they are forced to answer honestly. In other words, forced-choice format scales are potentially well suited to counter the aforementioned response biases because (a) respondents cannot endorse every statement (Brown & Maydeu-Olivares, 2011; Cheung & Chan, 2002) and (b) the respondents are faced with statements that are equally desirable. Thus, the potential of a forced-choice format scale to resist attempts at faking hinges on the statements within item blocks appearing equally socially desirable (McCloy et al., 2005).

Regarding recovering normative scores from FC scales, a major issue emerges when using traditional or classical test theory scoring methodology on scores from scales created using the FC method (Fisher, Robie, Christiansen, Speer, & Schneider, 2019; Hicks, 1970; Johnson, Wood, & Blinkhorn, 1988): ipsativity of data. This was another major reason the FC format fell out of favor (Christiansen et al., 2005). Assuming two statements assess two different constructs within a FC item block, when respondents select one of the two options, they receive a score of one for the option selected and necessarily a score of zero for the other. Regardless of how many statements are contained within an item block, what is clear is that scores for each statement are not only a function of standing on the latent trait they respectively are intended to measure, but also a function of the standing on the other latent trait(s) as well as their associated error terms and other influences (Meade, 2004). This creates intraindividual scoring dependencies such that the score on one construct is a function of the score on the other, but interindividual independencies and incomparability between respondents because the total score on the measure will always sum up to the number of items (Stark, Chernyshenko, & Drasgow, 2005) across all respondents, which is a constant (i.e., ipsative data, Hicks, 1970). This leads to artificial negative correlations between dimensions that distort factor structure (Meade, 2004) and results in rank-order scores that reflect relative standing on latent traits within persons that cannot be compared between persons (Stark, Chernyshenko, & Drasgow, 2012).

It has been proposed in the past that there are ways in which FC measures can be scored or constructed to produce quasi- or partially-ipsative scores where the scores across the entire scale do not sum up to a constant and endorsement of one latent trait statement does not necessarily decrease scores on another latent trait (Hicks, 1970; Salgado, Anderson, & Táuriz, 2014). These scores theoretically have the properties of both ipsative and normative data. This can be achieved through scale construction (e.g., scales within measure have different number of items, instructions to respondents to not rank all statements within item block) or scoring (e.g., not all ranked responses are actually scored, ranked statements are weighted differently). However, counter to expectations they appear to be just as fakable as single stimulus format measures (e.g., Heggestad et al., 2006) and surprisingly more fakable than purely ipsative FC measures (Cao & Drasgow, 2019).

Thus, despite recognition of the strength of the FC format to prevent response distortion (Nederhof, 1985) especially as it relates to the aforementioned issues with character assessment (Fowers, 2014), it has generally not been recommended up until the very recent past (Rothstein & Goffin, 2006) for use where between-person comparisons are made (e.g., high-stakes selection).

The promise and potential pitfall of Thurstonian item response theory scoring

Recent advances in modeling the item response process within an IRT framework for scoring multidimensional FC measures (i.e., MFC measures or those tapping more than one unidimensional construct; e.g., Brown & Maydeu-Olivares, 2013; Brown & Maydeu-Olivares, 2011; McCloy et al., 2005) have addressed some of the past ipsativity issues and allowed for recovery of normative latent construct scores.

Specific to our scale creation is the work by Brown and Maydeu-Olivares (2011, 2013). Briefly, Brown and Maydeu-Olivares’ method involves first recoding each of the FC response ratings endorsing one option over the other as binary outcomes, with multiple pairwise comparisons when there are more than two statements per block. Second, the binary outcomes are modeled within a Thurstonian factor model framework. Thurstone argued that when making pairwise comparisons (1927) or ranking options (1931), respondents engage in a discriminative process that elicits for each object of comparison a latent utility that describes the subjective value (e.g., a psychological value; Brown & Maydeu-Olivares, 2013) underlying those objects; it is assumed that utility is a latent, continuous, normally distributed variable in the population of respondents. As a consequence of this process, respondents choose the object with the greatest utility or subjective value at that point in time. So within the Thurstonian factor model as applied to scores
from an MFC format measure, the binary outcome variables recoded from the MFC item block scores are indicators of the statements’ utilities or subjective value (first-order constructs) that are themselves a function of the latent constructs the measure intends to tap (second-order constructs), which is modeled as a higher-order factor model. The third and final step involves parameterizing the Thurstonian factor model into an algebraically equivalent, first-order Thurstonian IRT model where the binary outcomes are now modeled as having residual error variance and related directly to the latent constructs (now first-order constructs) using the utilities as inputs for the non-linear link function, which allows for estimation and recovery of normative scores for the latent constructs. These recent advances in scoring and the work of FC format researchers have led to a resurgence of interest in FC methodology (Lee et al., 2018; Pavlov et al., 2019).

However, despite this interest in FC methodology brought about by the promise of normative scoring via the Thurstonian IRT model, some are not so sanguine about the pragmatic utility of the Thurstonian IRT approach in an applied context where motivated response biases are likely. This is because of a potential conflict between the two aims of preventing response biases on the one hand and recovering normative scores using the Thurstonian IRT model on the other. A recent meta-analysis has recommended that FC measures intended to combat faking in high-stake situations be a) constructed so as to balance social desirability of statements and b) be scored using some normative scoring approach (Cao & Drasgow, 2019). Indeed, these recommendations directly address the two aforementioned aims to be fulfilled when using FC measures in an applied context.

However, recent theoretical arguments have been made that using Thurstonian IRT scoring creates a burden of sorts due to what the method requires to actually recover normative parameter estimates (Fisher et al., 2019). Simulations suggest that when using Thurstonian IRT scoring, mixing negatively keyed statements – that tap the lower end of the latent continuum of a given trait – with positively keyed items within item blocks is needed for accurate latent trait parameter recovery (Brown & Maydeu-Olivares, 2011) so that between-person comparisons can be made. Simultaneously, it has been argued that the utility of FC measures in combating faking is that item blocks force respondents to choose which equally attractive option to endorse (McCloy et al., 2005), which is why matching statements within item blocks on social desirability is strongly recommended (Brown, Inceoglu, & Lin, 2017): the ability of the FC format to resist faking arguably hinges on this. Yet it has been argued that one cannot both create item blocks that contain negatively and positively keyed statements and also match them within item block on desirability because the negatively keyed statements are either obviously (Bürkner, Schulte, & Holling, 2019) or at least apparently (Wang, Qiu, Chen, Ro, & Jin, 2017) less desirable. Consequently, it is theorized that people who are motivated to fake good will generally choose the seemingly more desirable positively keyed statements within item blocks, yielding little information (Wang et al., 2017) and undermining the Thurstonian IRT approach’s ability to accurately estimate latent trait parameters shown in simulation studies (Bürkner et al., 2019).

In short, fulfilling the aim of producing normative scores using the Thurstonian IRT scoring method requires mixing negatively and positively keyed statements within item blocks. This theoretically undermines the aim of (a) preventing response biases because negatively keyed statements will not be perceived to be as desirable as positively keyed statements, introducing systematic error variance in scores that produce consistent response patterns across respondents that do not accurately differentiate people on the latent traits of interest, which (b) undermines the aim of recovering accurate normative scores.

On the basis of such past results and recent theoretical arguments, we sought to attempt to create and validate an MFC version of an extant character scale (CIVIC; Ng et al., 2018) using the methodology outlined by Brown and Maydeu-Olivares (2011) to create unequally (i.e., including both negatively and positively keyed items) keyed item blocks. We investigate whether such an MFC measure is (a) more resistant to faking in spite of unequally keyed item blocks and (b) whether it is similarly valid to its single-stimulus format counterpart.

Method

Participants

Data was collected using Amazon’s MTurk platform, which allows people (Requesters) to post tasks (HITs or Human Intelligence Tasks) for other people (Workers) to complete in exchange for payment (Amazon Web Services, 2019). Only Workers who lived in the US, were 18 years of age or older, and had at least 95% of HITs they had completed approved as of sufficient quality by Requesters were able to participate. For our Time 1 data collection, the initial pool was compensated $2.00 except those who did not respond to twenty percent of any scale or the overall survey and those who missed more than fifty percent of the attention check items. From this initial pool of 1,013 participants, after data screening for those who had missed more than one attention check item, those who took less than thirty minutes to complete the survey, those who gave the same responses to the non-MFC items, and listwise deletion, a total of 798 participants were retained for data analysis (male = 34.1%, White = 83.3%, Black = 8.0%, Asian = 3.8%, Latin American = 2.5%, others = 2.4%, M age = 40.92, SD age = 12.89).

For Time 2 data collection, approximately fourteen weeks later, participants who passed data screens in the last survey were contacted again through procedures that ensured their anonymity (Wiepand, 2015) and invited to complete a follow-up survey for $4.00. This survey was comprised of an identical set of measures, except using the final and shortened MFC scale and corresponding single statement scale (see below) at the end of the survey. Five hundred twenty-nine people initially agreed to participate. Immediately prior
to the MFC and SS versions of the CIVIC, which were at the end of the survey, participants were randomized into either an honest condition or a faking condition where they were asked to create a positive impression. Eleven did not progress far enough to be randomized and were screened out, leaving 518 respondents with 259 in each condition. Respondents were compensated unless they responded to less than twenty percent of any scale or the overall survey and those who missed thirty percent or more of the attention check items.

For those who were randomly chosen to respond as honestly and accurately as possible, a few data screen out conditions were used. First, those who missed more than one attention check were excluded. Second, those who did not affirm after answering the MFC and single statement character measures that they understood directions to respond honestly to both were excluded. Third, those who left any of the MFC triplets blank were excluded. Fourth, those we were not able to identify so as to align with corresponding Time 1 data were excluded. Finally, those who took less than thirty minutes to complete the assessment were excluded. After data screening, 222 participants (male = 35.6%, White = 84.7%, Black = 8.1%, Asian = 3.2%, Latin American = 2.7%, others = .9%, $M_{age} = 42.93$, $SD_{age} = 12.86$, 1 non-response) remained.

For those in the faking condition, the same data screening procedures were used, but in addition to the manipulation check immediately after the MFC and SS versions of the CIVIC, there were directions immediately prior repeatedly stating that participants should create a positive impression followed by a first manipulation check. Those that failed this by stating they should answer the following two sections honestly were corrected and a shortened set of items were given. Of the 245 that passed the first data screen (see above), only 110 passed both manipulation checks. After the remaining data screening procedures, 105 participants in the faking condition remained. Another 130 participants were collected from the Time 1 pool for the faking condition to remedy the large difference in sample size between the two conditions. After screening these data using the same procedures, 90 participants were asked to create a positive impression. Eleven did not respond to the MFC and SS versions of the CIVIC, which was operationalized as the difference in average scores between honest and faking conditions (see online supplemental materials, Figure S2). This led to the addition of 106 negatively keyed items (seven to ten per trait) to the existing 89 items (four to eleven per character trait) were drawn from the original CIVIC (Ng et al., 2018). Because those items were all positively keyed and the Thurstonian IRT approach relies heavily on including negatively keyed items to recover person and item parameters (Brown & Maydeu-Olivares, 2011) we generated and refined negatively keyed statements for each of the 13 character traits. We subsequently assessed the degree of social desirability for both negatively and positively keyed statements, which was operationalized as the difference in average scores between honest and faking conditions (see online supplemental materials, Figure S2).

Consistent with past works creating an MFC version of an extant measure (e.g., Guenole et al., 2018), items assessing the 13 character traits (89 items, four to eleven per character trait) were drawn from the original CIVIC (Ng et al., 2018). Those items were all positively keyed and the Thurstonian IRT approach relies heavily on including negatively keyed items to recover person and item parameters (Brown & Maydeu-Olivares, 2011) we generated and refined negatively keyed statements for each of the 13 character traits. We listed all possible combinations of the 13 character traits and selected 90 combinations that represented each dimension an approximately equal number of times for initial MFC measure construction. Using these content specifications, 90 triplets were created by grouping statements having high lower-order factor loadings to improve item information (Brown & Maydeu-Olivares, 2011), mixing positively and negatively keyed statements into triplets to enhance parameter estimation (Brown & Maydeu-Olivares, 2011), and matching statements within triplets on their social desirability ratings (see supplemental materials, Figure S2).

Then, 30 triplet-blocks with high psychometric properties (i.e., high factor loadings and similar social desirabilities) were selected from the initial set of 90 triplet-blocks for the final MFC measure. The 30 triplet-block design was well balanced, with each of the eight character dimensions measured by approximately 10-12 item occurrences (i.e., some positively keyed items for each dimension were repeated across triplet blocks; see Table 1). In total, 15 of the items

The analyses of that paper indicated a replicable eight-factor structure of broad character dimensions underlying lower-order character traits. For the current study, it was decided based on simulation studies we ran that reduction to a more manageable number of latent traits was necessary to create the CIVIC-MFC given the Thurstonian IRT model scoring procedures. Consequently, 13 lower-order latent traits were determined to be representative of the eight broad character dimensions on the basis of strength of loadings onto their respective dimensions as well as content representation and breadth: gratitude, love of learning, perseverance, propriety, social perceptiveness, trustworthiness, authenticity, carefulness, foresight, spirituality, meaning and purpose, perspective taking, and openness to evidence.

Of the MFC response formats that have been explored in applied settings, we chose a triplet format based on growing interest in this approach to noncognitive measurement (e.g., Guenole, Brown, & Cooper, 2018) and a ranking response instruction, from most to least "like me" since this has proven to be beneficial for latent trait estimation in simulation studies (Hontangas et al., 2015).

Consistent with past works creating an MFC version of an extant measure (e.g., Guenole et al., 2018), items assessing the 13 character traits (89 items, four to eleven per character trait) were drawn from the original CIVIC (Ng et al., 2018). Because those items were all positively keyed and the Thurstonian IRT approach relies heavily on including negatively keyed items to recover person and item parameters (Brown & Maydeu-Olivares, 2011) we generated and refined negatively keyed statements for each of the 13 character traits. We subsequently assessed the degree of social desirability for both negatively and positively keyed statements, which was operationalized as the difference in average scores between honest and faking conditions (see online supplemental materials, Figure S2). This led to the addition of 106 negatively keyed items (seven to ten per trait) to the existing 89 items (four to eleven per trait) to measure the 13 character traits. We listed all possible combinations of the 13 character traits and selected 90 combinations that represented each dimension an approximately equal number of times for initial MFC measure construction. Using these content specifications, 90 triplets were created by grouping statements having high lower-order factor loadings to improve item information (Brown & Maydeu-Olivares, 2011), mixing positively and negatively keyed statements into triplets to enhance parameter estimation (Brown & Maydeu-Olivares, 2011), and matching statements within triplets on their social desirability ratings (see supplemental materials, Figure S2).

Then, 30 triplet-blocks with high psychometric properties (i.e., high factor loadings and similar social desirabilities) were selected from the initial set of 90 triplet-blocks for the final MFC measure. The 30 triplet-block design was well balanced, with each of the eight character dimensions measured by approximately 10-12 item occurrences (i.e., some positively keyed items for each dimension were repeated across triplet blocks; see Table 1). In total, 15 of the items

Measures

CIVIC-MFC measure development

The development of the CIVIC-MFC was based in the development of the CIVIC (see Ng et al., 2018; Study 1 and 2).1

1Ng et al.’s (2017) Study 3 reduced the number of lower-order character traits from 31 to 29. Perseverance and willpower were combined because they were highly correlated and modeling them as two distinct factors did not significantly improve fit. Propriety was dropped because unidimensionality did not replicate in Study 3.
Table 1. Descriptive statistics and reliability for all study measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th># of Unique Items (# of Item Occurrences)</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appreciation</td>
<td>9</td>
<td>29.57</td>
<td>5.32</td>
<td>9</td>
<td>36</td>
<td>.78</td>
</tr>
<tr>
<td>Intellectual Engagement</td>
<td>7</td>
<td>22.97</td>
<td>3.90</td>
<td>7</td>
<td>28</td>
<td>.79</td>
</tr>
<tr>
<td>Fortitude</td>
<td>9</td>
<td>28.08</td>
<td>5.09</td>
<td>9</td>
<td>36</td>
<td>.81</td>
</tr>
<tr>
<td>Interpersonal Consideration</td>
<td>7</td>
<td>20.36</td>
<td>3.86</td>
<td>7</td>
<td>28</td>
<td>.74</td>
</tr>
<tr>
<td>Sincerity</td>
<td>8</td>
<td>26.85</td>
<td>3.80</td>
<td>10</td>
<td>32</td>
<td>.73</td>
</tr>
<tr>
<td>Temperance</td>
<td>10</td>
<td>31.04</td>
<td>4.98</td>
<td>10</td>
<td>40</td>
<td>.72</td>
</tr>
<tr>
<td>Transcendence</td>
<td>9</td>
<td>24.76</td>
<td>7.44</td>
<td>9</td>
<td>36</td>
<td>.90</td>
</tr>
<tr>
<td>Empathy</td>
<td>10</td>
<td>30.68</td>
<td>4.84</td>
<td>11</td>
<td>40</td>
<td>.79</td>
</tr>
<tr>
<td>CIVIC-MFC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appreciation</td>
<td>9(11)</td>
<td>−.03</td>
<td>.91</td>
<td>−2.91</td>
<td>2.11</td>
<td>.78</td>
</tr>
<tr>
<td>Intellectual Engagement</td>
<td>7(12)</td>
<td>−.03</td>
<td>.90</td>
<td>−3.57</td>
<td>2.31</td>
<td>.81</td>
</tr>
<tr>
<td>Fortitude</td>
<td>9(11)</td>
<td>−.03</td>
<td>.90</td>
<td>−3.58</td>
<td>2.22</td>
<td>.76</td>
</tr>
<tr>
<td>Interpersonal Consideration</td>
<td>7(11)</td>
<td>−.02</td>
<td>.86</td>
<td>−3.11</td>
<td>2.18</td>
<td>.82</td>
</tr>
<tr>
<td>Sincerity</td>
<td>8(10)</td>
<td>−.03</td>
<td>.88</td>
<td>−3.23</td>
<td>2.22</td>
<td>.60</td>
</tr>
<tr>
<td>Temperance</td>
<td>10(12)</td>
<td>−.03</td>
<td>.89</td>
<td>−3.61</td>
<td>2.07</td>
<td>.63</td>
</tr>
<tr>
<td>Transcendence</td>
<td>9(12)</td>
<td>−.01</td>
<td>.95</td>
<td>−2.51</td>
<td>2.73</td>
<td>.60</td>
</tr>
<tr>
<td>Empathy</td>
<td>10(11)</td>
<td>−.02</td>
<td>.87</td>
<td>−2.46</td>
<td>2.11</td>
<td>.60</td>
</tr>
<tr>
<td>Openness</td>
<td>6</td>
<td>14.54</td>
<td>5.12</td>
<td>6</td>
<td>30</td>
<td>.83</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>6</td>
<td>18.43</td>
<td>4.86</td>
<td>6</td>
<td>30</td>
<td>.79</td>
</tr>
<tr>
<td>Extraversion</td>
<td>6</td>
<td>20.11</td>
<td>4.49</td>
<td>6</td>
<td>30</td>
<td>.74</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>6</td>
<td>22.00</td>
<td>4.09</td>
<td>6</td>
<td>30</td>
<td>.73</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>6</td>
<td>22.85</td>
<td>4.44</td>
<td>8</td>
<td>30</td>
<td>.84</td>
</tr>
<tr>
<td>Social Competence Questionnaire</td>
<td>10</td>
<td>25.28</td>
<td>7.26</td>
<td>10</td>
<td>40</td>
<td>.91</td>
</tr>
<tr>
<td>Spiritual Transcendence Scale</td>
<td>9</td>
<td>56.92</td>
<td>16.85</td>
<td>18</td>
<td>90</td>
<td>.94</td>
</tr>
<tr>
<td>Consideration of Future Consequences</td>
<td>12</td>
<td>40.90</td>
<td>6.66</td>
<td>16</td>
<td>56</td>
<td>.89</td>
</tr>
<tr>
<td>Gratitude Questionnaire</td>
<td>6</td>
<td>33.46</td>
<td>7.45</td>
<td>6</td>
<td>42</td>
<td>.92</td>
</tr>
<tr>
<td>Authentic Living Scale</td>
<td>4</td>
<td>24.65</td>
<td>3.34</td>
<td>10</td>
<td>28</td>
<td>.83</td>
</tr>
<tr>
<td>Curiosity and Exploration Inventory II</td>
<td>10</td>
<td>29.15</td>
<td>8.18</td>
<td>10</td>
<td>50</td>
<td>.90</td>
</tr>
<tr>
<td>Grit Scale</td>
<td>6</td>
<td>23.35</td>
<td>4.58</td>
<td>6</td>
<td>30</td>
<td>.88</td>
</tr>
<tr>
<td>Perspective-Taking (IRI)</td>
<td>7</td>
<td>29.08</td>
<td>5.39</td>
<td>8</td>
<td>39</td>
<td>.87</td>
</tr>
<tr>
<td>Self-Deceptive Enhancement (BIDR)</td>
<td>20</td>
<td>5.65</td>
<td>4.04</td>
<td>0</td>
<td>18</td>
<td>.80</td>
</tr>
<tr>
<td>Self Esteem Scale</td>
<td>10</td>
<td>3.08</td>
<td>.69</td>
<td>1</td>
<td>4</td>
<td>.94</td>
</tr>
<tr>
<td>Brief Inventory of Thriving</td>
<td>10</td>
<td>37.31</td>
<td>8.06</td>
<td>10</td>
<td>50</td>
<td>.94</td>
</tr>
<tr>
<td>Behavioral Correlates of Character</td>
<td>35</td>
<td>116.92</td>
<td>23.28</td>
<td>45</td>
<td>191</td>
<td>.86</td>
</tr>
</tbody>
</table>

Note. N = 798. "# of Unique Items” refers to number of unique statements for each scale; "# of Item Occurrences” refers to total number of times items for each scale were presented across all CIVIC-MFC triplets. Reliability estimates for scores on CIVIC-SS and CIVIC-MFC are test-retest reliabilities based on N = 222. IRI = Interpersonal Reactivity Index; BIDR = Balanced Inventory of Desirable Responding.

Convergent/Discriminant Validity Scales

Eight scales were identified as measuring the same or similar broad character dimension and served as convergent validity scales for appreciation (The Gratitude Questionnaire or GQ; McCullough, Emmons, & Tsang, 2002), intellectual engagement (The Curiosity and Exploration Inventory-II or CFI-II; Kashdan et al., 2009), fortitude (Perseverance of Effort Subscale of the Grit Scale or GRIT-PoE; Duckworth et al., 2007), interpersonal consideration (The Social Competence Questionnaire or COMQ; Sarason, Sarason, Hacker, & Basham, 1985), sincerity (Authentic Living subscale of the Authentic Living Scale or ALS-AL; Wood, Linley, Maltby, Baliousis, & Joseph, 2008), temperance (Consideration of Future Consequences Scale or CFC; Strathman, Gleicher, Boninger, & Edwards, 1994), transcendence, (Prayer Fulfillment Subscale of the Spiritual Transcendence Scale or STS-PS; Piedmont, 1999), and empathy (Perspective-Taking subscale of the Interpersonal Reactivity Index or IRI-PT; Davis, 1983). Additionally a measure of Big Five personality (The Five Factor Model Rating Form or FFMRF; Mullins-Sweatt, Jamerson, Samuel, Olson, & Widiger, 2006) was included for discriminant validity.

Criterion-Related Validity Scales

Measures of socially desirable responding (Balanced Inventory of Desirable Responding or BIDR; Paulhus, 1984),

and were negatively keyed2 and 75 item occurrences were positively keyed. We found very high correlations (ranging from .90 to .96) of for corresponding character scores between the 90-triplet and 30-triplet MFC measures across the eight dimensions. Henceforth, we refer to the 30-triplet CIVIC measure as “CIVIC-MFC”3.

Single-statement (SS) CIVIC measure

The unique statements (positively and negatively keyed) used in the final, 30 triplet-block CIVIC-MFC measure were administered individually using an ordered-categorical response format (Very much unlike me, Unlike me, Like me, Very much like me). This measure is henceforth referred to as CIVIC-SS.

Validity-related measures

The number of items in and reliabilities for each scale are presented in Table 1. Details about each scale are included in the online supplementary materials (see Supplement 2).

---

2Due to mislabeling, we discovered in retrospect that a negatively keyed item intended to measure one (i.e., trustworthiness) of two character traits used to represent the broad dimension of sincerity was accidentally used to measure the other (i.e., authenticity). We do not believe this to be a significant issue since scores on both are used to score the same broad dimension.

3Items, associated materials, and data analysis scripts are available upon request from the first author.
self-esteem (The Rosenberg Self-Esteem Scale or RSES; Rosenberg, 1965), general well-being (The Brief Inventory of Thriving or BIT; Su, Tay, & Diener, 2014) and character-relevant behavior (Behavioral Correlates of Character Strengths; Ruch et al., 2010) were identified based on past research (Ng et al., 2018) on likely outcomes of character.

Analytic strategy

Scoring methods

The CIVIC-SS and criterion measures were scored using a conventional summative approach for ordered categorical responses. In the conventional summative approach, negative statements were reverse coded and total scores for each measure were obtained by summing respective item scores. The CIVIC-MFC measure was scored by fitting the Thurstonian Item Response Theory (TIRT) model to the rank responses using the MPLus 7.3 computer program (Muthén & Muthén, 1998-2014). As mentioned, this approach transforms rank-order responses among a set of stimuli into a set of pairwise binary outcomes. A triplet block, [A, B, C] requires three sets of pairwise comparison (i.e., [A, B], [A, C], [B, C]). If the first statement is preferred over the second in paired comparison, the pair would be coded as 1, and 0 otherwise. For example, if a respondent ranks statements as [A = 3rd, B = 1st, C = 2nd], three pairwise comparisons ([A, B], [A, C], [B, C]) would be created. The corresponding binary responses for the pairwise comparisons would be [A, B] = 0; [A, C] = 0; [B, C] = 1. Then, the transformed binary outcomes are analyzed using a two-dimensional standard normal ogive IRT model. The conditional probability of preferring statement s to statement t is obtained as follows,

\[ P_i(\eta_s, \eta_t) = \Phi \left( -\gamma_i + \lambda_s \eta_s - \lambda_t \eta_t \right) \frac{\psi_i^2}{\psi_i^2 + \psi_t^2}, \]

(1)

where \( \gamma_i \) = threshold parameter of a binary outcome; \( \eta_s \) and \( \eta_t \) = latent trait scores for dimensions a and b; \( \lambda_s \) and \( \lambda_t \) = factor loadings of statement s and t on the measured dimensions \( \eta_s \) and \( \eta_t \); \( \psi_i^2 \) and \( \psi_t^2 \) = uniqueness of statement s and t; and \( \Phi(x) \) = cumulative standard normal distribution function evaluated at x. This model was fit to T1 data (N = 798; see online supplemental materials Supplement 3, Table S1, for item parameter estimates) and parameter estimates derived from this calibration sample were used to generate the CIVIC-MFC character trait scores in the smaller T2 honest (N = 222) and faking (N = 195) datasets consistent with past research (Fisher et al., 2019). Please refer to Brown and Maydeu-Olivares (2011) for further technical detail.

Reliability

Two assumptions of coefficient alpha (i.e., consistent responding and independence of errors) are violated by multidimensionality and dependence among responses within MFC blocks (ipsativity) (Brown & Maydeu-Olivares, 2013). First, because items within each block are ranked relative to each other, the same item in two different item blocks could be ranked first and last depending on the utility of the other items within their respective item blocks and receive, inconsistently, maximum and minimum points. Second, rankings by definition lead to within-block item scores that depend on each other and so are not independent ratings with independent errors. Given these concerns and consistent with past research highlighting them (Converse et al., 2008; Heggestad et al., 2006; Salgado & Tauriz, 2014), we conducted test-retest reliability analysis to evaluate reliabilities for scores on the CIVIC-MFC and CIVIC-SS using the subset of participants who participated at T1 and were randomized into the honest response condition at T2 (N = 222).

Validity of measures

To assess convergent validity, (1) monotrait-heteromethod correlations from multitrait-multimethod (MTMM; Campbell & Fiske, 1959) analysis and (2) correlations between CIVIC scores and convergent validity scale scores were examined. To assess discriminant validity, (1) heterotrait-monomethod correlations from multi-trait multi-method (MTMM) analysis and (2) the magnitude of difference between CIVIC convergent validity scale scores and discriminant validity scale scores were examined. Criterion-related validity was also assessed by comparing the correlations of the CIVIC-SS and CIVIC-MFC scores with the criterion variables.

Faking resistance

A predominant method of assessing faking resistance (Pavlov et al., 2019) was used where within-subject correlations between honest and faking scores for each format are compared. For the 195 faking condition participants, T1 (honest) scores were correlated with T2 (faking) scores for both the CIVIC-MFC and CIVIC-SS.

Results

Structural analysis

As a preliminary structural analysis, the TIRT model was fitted to the ranking response data of CIVIC-MFC from T1 data (N = 798). The model successfully converged and the fit was good: \( \chi^2 = 3959.23, \) df = 291; CFI = .93; TLI = .95; RMSEA = .03. Although Brown and Maydeu-Olivares (2011, 2013) do not suggest fit criteria, we note that the results indicate as good or better model fit than previous investigations using similar methods (Guenole et al., 2018; Lee et al., 2018; Morillo et al., 2016) and met cutoffs suggested. The results show that the expected eight-factor structure of CIVIC character was supported. As mentioned above, T2 CIVIC-MFC data for both honest and faking conditions were scored based off of the parameters recovered from fitting the model to T1 data. Subsequent structural analysis of these T2 datasets showed though that whereas fitting the model to the T2 honest condition data
Table 2. Correlations between SS (Single Stimulus) and MFC (Multidimensional Forced-Choice) CIVIC Scores.

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>MFC</th>
<th></th>
<th>SS</th>
<th>MFC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>I E</td>
<td>F</td>
<td>I C</td>
<td>S</td>
</tr>
<tr>
<td>SS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appreciation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intellectual</td>
<td>.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engagement</td>
<td>.43</td>
<td>.52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpersonal</td>
<td>.31</td>
<td>.18</td>
<td>.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consideration</td>
<td>Sincerity</td>
<td>.49</td>
<td>.35</td>
<td>.54</td>
<td>.27</td>
</tr>
<tr>
<td></td>
<td>Temperance</td>
<td>.37</td>
<td>.40</td>
<td>.56</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>Transcendence</td>
<td>.43</td>
<td>.11</td>
<td>.22</td>
<td>.20</td>
</tr>
<tr>
<td></td>
<td>Empathy</td>
<td>.39</td>
<td>.47</td>
<td>.44</td>
<td>.20</td>
</tr>
<tr>
<td>MFC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appreciation</td>
<td>.69</td>
<td>.18</td>
<td>.31</td>
<td>.18</td>
<td>.40</td>
</tr>
<tr>
<td>Intellectual</td>
<td>.26</td>
<td>.64</td>
<td>.41</td>
<td>.06</td>
<td>.24</td>
</tr>
<tr>
<td>Engagement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fortitude</td>
<td>.33</td>
<td>.34</td>
<td>.70</td>
<td>.22</td>
<td>.40</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>.34</td>
<td>.15</td>
<td>.36</td>
<td>.57</td>
<td>.31</td>
</tr>
<tr>
<td>Consideration</td>
<td>Sincerity</td>
<td>.37</td>
<td>.16</td>
<td>.36</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>Temperance</td>
<td>.33</td>
<td>.27</td>
<td>.42</td>
<td>.19</td>
</tr>
<tr>
<td></td>
<td>Transcendence</td>
<td>.37</td>
<td>.04</td>
<td>.16</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>Empathy</td>
<td>.29</td>
<td>.31</td>
<td>.32</td>
<td>.12</td>
</tr>
</tbody>
</table>

Note. N = 798. Bold coefficients indicate monotrait-heteromethod correlations. A = Appreciation; IE = Intellectual Engagement; F = Fortitude; IC = Interpersonal Consideration; S = Sincerity; TE = Temperance; TR = Transcendence; E = Empathy.

demonstrated good fit ($\chi^2 = 119.67$, df = 104, CFI = .96, TLI = .96, RMSEA = .03), fitting the model to T2 faking condition data did not demonstrate good fit ($\chi^2 = 70.06$, df = 22, CFI = .55, TLI = .55, RMSEA = .11).

Descriptive Statistics and Reliability

Table 1 presents descriptive statistics for the CIVIC-SS and CIVIC-MFC measures and the criterion measures. We note that the mean of the CIVIC traits for the MFC measure are approximately zero because multivariate standard normal priors were used for calibration. Test-retest reliability coefficients are also presented in Table 1. The reliabilities were generally lower for scores on the MFC version of the CIVIC measure.

Validity analyses

Table 2 presents the multi-trait multi-method (MTMM) correlations between the CIVIC-SS and CIVIC-MFC scores. Convergent validity correlations between the corresponding SS and MFC constructs are shown in bold. These monotrait-heteromethod correlations ranged from .57 to .83. It is worth noting though the high heterotrait-monomethod correlations for CIVIC-MFC, which relate to discriminant validity. There was at least one heterotrait-monomethod correlation for empathy ($r_s = .63$, .64, .66), temperance ($r_s = .59$, .60, .66), sincerity ($r_s = .64$, .64), and interpersonal consideration ($r = .57$) that was at least as strong as their respective validities (.57, .58, .58, .57). This is intriguing because the heterotrait-monomethod correlations for the CIVIC-SS were much lower and respondents had no incentive to answer in a socially desirable way. More research is needed to investigate possible reasons (e.g., whether how statements were combined in the MFC measure induced a response set that adversely affected discriminant validity).

Nonetheless, evidence for construct validity was generally promising for the CIVIC-MFC. Table 3 shows the correlations between the CIVIC-SS and the CIVIC-MFC with Big Five personality domains, convergent/discriminant validity scales, and criteria measures. We calculated the column-vector correlations that describe how well the pattern of associations between character and Big Five personality dimensions hold across the CIVIC-SS and CIVIC-MFC measures: these ranged from .94 to 1.00 (i.e., .998). This suggests nomological validity (Campbell, 1960) for the CIVIC-MFC; the constructs tapped using it and their respective counterparts as assessed using the CIVIC-SS are similarly situated within the Big Five construct space. Convergent validities for the CIVIC-MFC ranged from .32 to .74 (median = .50), which are lower, yet similar to those of the CIVIC-SS (ranged from .48 to .88; median = .71). For both versions, convergent validities for the eight character dimensions were generally larger than discriminant validities, although there were some exceptions.

Intellectual engagement seemed to be the least discriminable construct for both measures: its convergent validity with curiosity (.52) was only slightly larger than one of its discriminant validities (i.e., correlation with grit, .47) for the CIVIC-SS and a discriminant validity for CIVIC-MFC’s intellectual engagement was actually larger (i.e., correlation with grit, .37) than its convergent validity (.32). One of the discriminant validities (i.e., correlation with gratitude; .38) for interpersonal consideration dimension measured using the CIVIC-MFC was also larger than its convergent validity (.33). For the CIVIC-MFC, intellectual engagement, interpersonal consideration, and sincerity all had discriminant validities that were at least as strong as their convergent validities.

In terms of concurrent validity, all eight character dimensions for both measures were significantly correlated with the outcomes of self-deceptive enhancement (aspect of socially desirable responding), self-esteem, thriving, and behavioral correlates of character, although correlations with criteria were generally stronger for the CIVIC-SS than the CIVIC-MFC.

Faking resistance

Table 3 shows the faking condition participants’ correlations between honest (T1) and faking (T2) scores for each of the eight character dimensions for the MFC and SS formats. Correlations are generally lower for the MFC format except for Temperance. This suggests the MFC version did not prevent faking more than the SS version.

Discussion

When assessing socially desirable characteristics using self-report scales, past research has shown that a multidimensional forced-choice format can offer some protection against motivated faking score inflation while retaining
similar degrees of validity as traditional, Likert-type scale formats. We endeavored to create and validate an MFC measure of character from the CIVIC (Ng et al., 2018) given that the traits assessed are generally considered cross-culturally desirable (Peterson & Seligman, 2004). To the best of our knowledge, we are the first empirical investigation a) comparing the MFC and SS format b) using normative scores for both (employing the TIRT approach for the former) that c) used an MFC scale created specifically to meet best practice recommendations for normative score recovery via TIRT (Brown et al., 2017; Brown & Maydeu-Olivares, 2011) and d) administered both formats under honest and faking conditions to assess how well the MFC format prevents attempts at faking (Wetzel & Frick, 2020).

The results for the CIVIC-MFC under honest conditions indicate some success in creating a valid measure of character. The CIVIC-MFC had high validities from the multitrait-multimethod matrix (.57 − .83), low to moderate heterotrait-heteromethod correlations (.01−.42; see Table 2), and medium to high convergent validities with extant scales measuring similar traits (see Table 3). There were still some issues. For some character dimensions measured using the MFC format, heterotrait-monomethod correlations were stronger than monotrait-heteromethod correlations (i.e., validities). Also, in a few cases the CIVIC-MFC’s discriminant validities were at least as strong as convergent validities. One observation is that the character dimensions for which this occurred (i.e., intellectual engagement, interpersonal consideration, and sincerity) in the CIVIC-MFC were in fact those with the lowest convergent validities (i.e., .32, .33, .39, respectively), whereas the convergent validities for the remaining five character dimensions ranged from .48 to .74. This may indicate that the three aforementioned character dimensions are not adequately measured using the CIVIC-MFC. As a point of comparison, the same three dimensions measured using the CIVIC-SS also had the lowest convergent validities among the CIVIC-SS validities, although they were strong as opposed to medium effect size correlations (Cohen, 1992).

This same pattern was also found with all but two concurrent validity coefficients: correlations with criteria were generally stronger for CIVIC-SS versus CIVIC-MFC scores. However, as mentioned in prior literature, the convergent and concurrent validities may be stronger for the CIVIC-SS than the CIVIC-MFC simply due to two methodological issues: a) common method variance and b) reliability (Heggstad et al., 2006; Wetzel & Frick, 2020). Since the validity scales themselves were comprised of single statement, Likert-type items, this artificially inflates correlations with the CIVIC-SS scores because of the common scale format. And because some of the MFC format character dimension scores evidenced lower test-retest reliability compared to corresponding SS format scores, the formers’ correlations with validity scales will be attenuated to a greater degree relative to the latter’s.

The results also suggest that the MFC version did not prevent faking more than the SS version. It is important to note that the model fit well for the MFC data in the T1 data where everyone responded honestly as well as the T2 subset that responded honestly. In other words, the potential issue of ipsativity generally associated with MFC data at the item level (Meade, 2004) does not seem to be so problematic as to distort factor structure when (a) using TIRT scoring on MFC data (b) from a measure with unequally keyed item blocks (c) when people respond honestly. The dramatic (negative) change in model fit occurred when we (a) used the same TIRT scoring method on MFC data (b) from the same measure with unequally keyed item blocks (c) when people were asked to fake. This suggests like that of past investigations (Heggstad et al., 2006) that it is possible to
recover normative scores from MFC data under honest, but not faking conditions.

How this potentially occurs directly relates to the issue of what these scores capture. Like past studies (e.g., Christiansen et al., 2005), we followed recommendations to include both negatively and positively keyed statements within item blocks, specifically since simulations have shown this is necessary to recover precise person and item parameters when using the TIRT approach (Brown et al., 2017; Brown & Maydeu-Olivares, 2011). Yet recent research has argued this makes respondents who are motivated to “fake good” perceive the negatively keyed statement as the least desirable (Bürkner et al., 2019; Fisher et al., 2019; Wang et al., 2017), one of the likely consequences of which is that item blocks will yield little information (i.e., true score variance) that would accurately differentiate people on the latent constructs of interest (Wang et al., 2017), thereby undermining the recovery of normative scores demonstrated in past simulation work on TIRT (Bürkner et al., 2019; Wang et al., 2017).

Indeed, we found that there were homogenous response patterns for the faking condition data, such as an item block where 161 out of 195 respondents gave the identical ranking pattern: “3-2-1.” In this and many other item blocks where homogenous response patterns emerged, the respondents who were asked to “fake good” at T2 rated the negatively keyed statements as least like them, just as Heggestad et al. (2006) found. Given the aforementioned model fit results and that one of the likely consequences of data saturated with socially desirable responding is factor structure distortion (Schmit & Ryan, 1993; Stark et al., 2001), this supports the recent theoretical arguments that matching unequally keyed statements within item blocks on social desirability, that may not fully address the issue if positively keyed items at least seem more socially desirable to respondents. Further discussion of test-retest reliability, ipsativity, factors contributing to convergent and discriminant validity issues, and the item response process assumptions of Thurstone’s law of comparative judgment (1927, 1931) can be found in the online supplemental materials (see Supplement 4).

Limitations and Future Directions

The current study does shed some light on the applicability of the MFC format, specifically using unequally keyed item blocks for latent trait estimation, and the TIRT scoring approach when attempting to create a valid measure of character that prevents faking. TIRT in theory can be used to derive normative scores from the MFC format that purportedly reduces faking (Dueber, Love, Toland, & Turner, 2019; Fisher et al., 2019; Walton, Cherkasova, & Roberts, 2019; Wetzel & Frick, 2020), but attaining both ends in practice requires scale construction considerations that may be difficult to address. Our initial results raise further questions that point to future directions in research on this topic.

First, as recently recommended for FC scale construction (Brown et al., 2017; Cao & Drasgow, 2019), we matched statements within item blocks on social desirability. Operationalizing social desirability has typically involved conducting preliminary studies in which respondents rate how desirable the items are perceived to be, to job applicants for instance, and taking the average rating as an index of the items’ social desirability (Bürkner et al., 2019; Christiansen et al., 2005; Heggestad et al., 2006). We used a different approach by using the mean difference in item scores from a preliminary study where respondents were randomized to either answer items honestly or to “fake good.” In any case, it is critical to note that the potential of MFC scales to prevent faking argues depends on statements within item blocks appearing to be equal in social desirability in the eyes of respondents (McCloy et al., 2005). As has been theorized, the negatively keyed statements are likely to be perceived as less socially desirable compared to positively keyed statements even if they are not (Wang et al., 2017).

One line of future research is to see if it is even possible to generate negatively keyed statements that are perceived to be as desirable as positively keyed statements. Heggestad et al. (2006) proposed that neutral items or subtle items – those that do not obviously tap something (un)desirable (Worthington & Schollmann, 1986) – could be created as an alternative to the MFC method of combating faking. Relatedly and particularly relevant to creating a character measure that resists faking attempts is past work that has used euphemistic and dysphemistic labels of virtuous and vicious behavior statements to either lower or heighten their levels of perceived social desirability, respectively (Meindl, Jayawickreme, Furr, & Pleesons, 2015). For instance, a statement intended to measure moral courage (“someone who does something they are scared to do”) is changed via a dysphemistic and fictional label (”an ‘emotionally foolhardy’ person is someone who does something they are scared to do”) to sound less desirable. Perhaps these kinds of statements could even be combined with the MFC method. However, the fundamental challenge of (a) ensuring that the items intended to tap the lower end of the latent construct continuum are (b) perceived to be equally socially desirable as those intended to tap the higher end would remain. Indeed, there might be an unavoidable tradeoff in that the very thing that allows for normative score recovery (i.e., items tapping the low end of the latent continuum) – at least under honest response conditions – might also be the very thing that undermines faking resistance when people are motivated to fake (Heggestad et al., 2006). An additional layer of complexity with this approach is that it would be a challenge to ensure that these euphemistic and dysphemistic statements are actually measuring their intended constructs, although this is arguably a potential issue with negatively keyed statements as well (Brown & Maydeu-Olivares, 2011; Dueber et al., 2019).

Second, a related issue is how context, from the micro- to the macro-level, affects the degree of perceived social
desirability of statements. At the micro-level of item block context, the desirability of statements have typically been rated in isolation from other statements, subsequently matched on desirability and combined into item blocks (e.g., Christiansen et al., 2005), yet it is at least uncertain if not unlikely that the statements will be perceived as equally desirable once presented together in item blocks (Bürkner et al., 2019; McCloy et al., 2005). Having statements rated on social desirability when presented together in a test item block and then iteratively (re)matching them on desirability might take several waves of data collection (Bürkner et al., 2019). Matching statements on social desirability may be difficult if not impossible though if the traits they assess differ substantially in how socially desirable they are (e.g., openness to experience versus agreeableness; Wetzel & Frick, 2020). At the meso-level of study design and test conditions context, asking respondents to fake to make a good overall impression is different from asking them to fake good toward a particular job, which might change the patterns of score inflation (Bürkner et al., 2019; Heggestad et al., 2006). At the macro-level of test use and application context, the results of directed faking studies might not generalize to actual applied settings (Heggestad et al., 2006) and past research has shown there is a marked difference between how applicant and non-applicant samples respond to test items (Stark et al., 2001). All these points relate to issues raised decades ago regarding whether social desirability ratings are obtained from a similar sample (e.g., students versus applicants) using a similar response “set” (i.e., honest response, faking to make a good overall impression, faking to indicate possession of qualities desirable for a particular job or activity) as that for which the test is intended to be administered (Waters, 1965): social desirability of the items will depend on insight into what is desirable for a given context and whether and what the sample is faking toward. If these two features are not aligned from scale construction to scale administration, then the studies are arguably not fakability studies at all, but rather generalization or extension studies.

The third related issue is what is an index of faking and thus faking resistance. The more a measure prevents faking, the more correspondence is expected between scores collected from honest and faking conditions, but all score consistency metrics are deficient in some way (Heggestad et al., 2006). Past efforts have generally used tests of mean difference (group-level faking resistance) or correlations (individual-level faking resistance) between honest and faking condition scores (Heggestad et al., 2006; Pavlov et al., 2019). Evidence for the former does not imply evidence for the latter and the latter arguably matters more to demonstrate the extent that the rank ordering of respondents and thus normative trait standing has been preserved under faking conditions (Heggestad et al., 2006) so between-person comparisons, in personnel selection for instance, can be made. But correlations can remain high despite (or are robust against) notable rank order changes of candidates at the tails of the score distribution (Drasgow & Kang, 1984; Heggestad et al., 2006; Mueller-Hanson, Heggestad, & Thornton, 2003; Rosse et al., 1998; Stark, Chernyshenko, Drasgow, & Williams, 2006). Particularly important is change in rank ordering at the top end of the latent continuum under faking conditions since most latent traits of interest for personnel selection are characteristics that are selected for and so fakers will try to inflate their scores. Indeed, there is between-person variability in degree of faking and this results in dramatic changes to rank ordering of respondents at the top end of the score distribution, with increasingly greater negative impact on accuracy of top-down hiring decisions as the selection ratio goes down (i.e., more stringent hiring standards) via a greater a) proportion of fakers in the subset of the pool selected in and b) magnitude of difference in average criterion (e.g., job) performance between fakers and honest high scorers (Mueller-Hanson et al., 2003; Pavlov et al., 2019; Rosse et al., 1998).

A recent development in modeling faking has been a multiple regression-based framework that incorporates the two traditional faking effect estimates and expands on them (Pavlov et al., 2019). They showed that the assumptions underlying those two traditional estimates (i.e., constant magnitude of faking across respondents [mean difference model] and lack of interactive effects between individual differences in faking tendency and honest trait standing [correlational model]) were not supported by including perceived trait desirability and the moderating effect of motivation to fake into their model. Future MFC and faking research should consider using such a model since it incorporates many of the factors that influence degree of faking and is flexible enough to incorporate others (e.g., ability to fake). Additional analyses that compare SS and MFC scale formats by examining their respective and relative individual-level score consistency metrics across honest and faking response conditions, such as average changes in magnitude of percentile ranking in scores or how consistent hiring decisions are assuming top-down selection on a construct at different selection ratios (i.e., decision consistency using kappa coefficient; Heggestad et al., 2006; Rosse et al., 1998) are also worth exploring.

One last issue is whether scores on the CIVIC-MFC demonstrated less faking resistance than those on the CIVIC-SS in part because they were less reliable. Recent research has argued that negatively keyed items may add additional cognitive load and impact reliability (Bürkner et al., 2019). Test-retest reliabilities for CIVIC-MFC scores were generally lower than those found for scores on the CIVIC-SS, but this is consistent with other studies involving MFC reliabilities (e.g., Chernyshenko et al., 2009; Heggestad et al., 2006) that do not use both unequally keyed item blocks and TIRT scoring. The MFC scores’ lower reliability is a consequence of us prioritizing that the same exact item content (i.e., statements) be used across the two formats, trading off making the two formats’ scores comparable in reliability by increasing the number of MFC item blocks or statements contained therein (see Wetzel & Frick, 2020). Alternatively this may be due to the format reducing response consistency bias, which would inflate reliabilities (Wetzel & Frick, 2020).

Like recent research (Walton et al., 2019; Wetzel & Frick, 2020), our results indicate generally favorable evidence for construct validity under honest response conditions, but
some issues with discriminant validity for some traits using the MFC format that were not found using the SS format. It may be a function of or interactions between assessing intercorrelated traits (Dueber et al., 2019; Walton et al., 2019), including negatively keyed items, and the MFC format itself. Indeed, Bürkner et al. (2019) suggested based on past research that including negatively keyed statements within item blocks can result in methodological variance that affects the covariance structure of items. Granted, the CIVIC-SS included the same statements as the CIVIC-MFC, but perhaps doing so is especially problematic in an MFC format when using the TIRT scoring approach.

There are no set rules for creating an MFC measure (Heggstad et al., 2006) and even recent scale construction recommendations are quite general (Cao & Drasgow, 2019). The current study is to our knowledge the first empirical demonstration of the issues associated with using the TIRT scoring approach to combat faking, which entails creating item blocks with unequally keyed statements that are matched on social desirability. It suggests perhaps that including negatively keyed statement in MFC format item blocks creates problems by (a) making the measure more fakable and (b) may contribute in part to lower reliability via methodological variance. This underscores the importance of matching statements on social desirability when constructing FC scales, which is surprisingly rare (Bürkner et al., 2019; Pavlov et al., 2019).

While future research is needed to further examine validity, the current work suggests that our MFC measure demonstrates good support for factorial, convergent, and concurrent validity and thus, although it did not prevent intentional response distortion, can be used as a similarly valid alternative to a SS version when face validity is a concern. Personality-type inventories are viewed as low on face and perceived predictive validity by applicants, which lowers organizational attractiveness and perceptions of fairness (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). In turn, perceived unfairness of selection systems may interact with expectation of successful faking to increase willingness to fake (Snell et al., 1999). The FC format is more face valid than the SS format in a high stakes selection context (Pavlov et al., 2019), so such concerns are eased (and usually considered fairer by test-takers) when using a format that seems less vulnerable to faking. Nevertheless, the MFC format may ultimately deter faking and prevent some response distortion simply because it is more difficult to do so compared to the SS format, thereby reducing motivation rather than ability to fake (O’Neill et al., 2017).

**Funding**

This work was supported by the John Templeton Foundation under Grant #48287.

**ORCID**

Vincent Ng http://orcid.org/0000-0002-9784-7720
Moon-Ho Ringo Ho http://orcid.org/0000-0001-9021-5141
Louis Tay http://orcid.org/0000-0002-5522-4728

**References**

Evaluating the Thurstonian IRT Approach


