# Examination of the Test–Retest Reliability of a Forced-Choice Personality Measure

ETS RR–19-37

Jacob Seybert
Dovid Becker

*December 2019*

# ETS Research Report Series



Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

# Examination of the Test–Retest Reliability of a Forced-Choice Personality Measure

Jacob Seybert & Dovid Becker

Educational Testing Service, Princeton, NJ

Forced-choice (FC) measures are becoming increasingly common in the assessment of personality for high-stakes testing purposes in both educational and organizational settings. Despite this, there has been relatively little research into the reliability of scores obtained from these measures, particularly when administered as a computerized adaptive test (CAT). This study examined the test–retest reliability of an FC personality CAT, comparing its reliability to the reliability of personality measures using a Likert-type rating scale. Using a relatively large sample ($N = 743$), participants completed multiple personality measures across two time points. The test–retest reliability estimates for the personality dimensions had a mean of .63 with a mean reliability of .73 when formed into Big Five personality trait composites. The FC personality reliabilities were lower than those of the Likert-type scales, though the findings are within the range of those found in meta-analytic studies.

**Keywords** personality; five-factor model; reliability; forced-choice; computerized adaptive testing

doi:10.1002/ets2.12273

There is a growing interest in the use of assessments of noncognitive constructs, such as *personality*, in educational and organizational applications (Kyllonen, Lipnevich, Burrus, & Roberts, 2014; Ryan et al., 2015). These constructs have shown utility in the prediction of relevant outcomes and behaviors, often beyond cognitive ability alone, with personality of particular interest to researchers and practitioners alike (Hough & Dilchert, 2010; Richardson, Abraham, & Bond, 2012). The use of personality is not without its difficulties. Measurement of these constructs typically relies on self-report questionnaires, which most often ask individuals to indicate their level of agreement with a range of statements on a Likert-type scale. These measures are susceptible to both intentional and unintentional response biases and distortions that may reduce their validity (Peterson, Griffith, Isaacson, O'Connell, & Mangos, 2011; van Herk, Poortinga, & Verhallen, 2004). Additionally, personality typically exhibits lower reliabilities than those observed with more traditional cognitive constructs, potentially limiting its usefulness in high-stakes contexts (Gnambs, 2014; Viswesvaran & Ones, 2000).

One increasingly common way to address issues of response distortion in obtaining self-report personality data is the use of forced-choice (FC) response types (Böckenholt, 2004; Brown & Maydeu-Olivares, 2011). These FC items require an individual to evaluate two or more statements and select which from among those options are most like him or her, or to rank order the responses. Application of FC methods has included efforts addressing constructs such as personality (Stark, Chernyshenko, & Drasgow, 2005; White & Young, 1998), vocational interest (SHL Group, 2006), and supervisor ratings of performance (Borman et al., 2001). Empirical examinations have supported the ability of these measures to reduce response biases and distortions (O'Neill et al., 2017; Salgado & Tauriz, 2014), though this mitigation may be somewhat moderate (Niessen, Meijer, & Tendeiro, 2017).

Despite the increased use of FC response formats, there has been very limited empirical investigation of the reliability of these measures. Instead, the majority of work to date has focused on the validity of the scores. These studies have found that the use of FC response formats may generally decrease faking and increase validity relative to measures using traditional Likert-type scales. Given that the use of FC measures is relatively new and that they are increasingly used for high-stakes purposes, there is a need for research into the stability of those scores and their relationship to scores from other response formats. This study contributes to this need by evaluating the reliability and construct validity of an FC personality measure administered via a computerized adaptive test (CAT) format.

*Corresponding author*: J. Seybert, E-mail: jseybert@imbellus.com

## Forced-Choice Personality Measurement

There are a number of item types and response instructions associated with FC measures, notably item types of statement pairs, triplets, and tetrads. Response instructions for these items generally range from instructing the respondent to select the statement that is most like him or her all the way to requiring complete ranking of every statement comprising the item. Of these, the use of pairs with the selection of the most-like statement may be most common. For example, a respondent may be presented with the following pair of statements, the first representing a high level of agreeableness and the second a high level of conscientiousness:

- I get along well with others.
- I always arrive to meetings on time.

For each item pair, the respondent is instructed to select the statement that is "most like" him or her. Given responses to a sufficiently large number of these items, scores for each of the studied dimensions can be estimated.

Early research with FC methods relied on the use of ipsative or quasi-ipsative scores that produced scores with some validity but questionable psychometric properties (Hicks, 1970; Salgado & Tauriz, 2014). Over the last two decades, there have been advances in strategies for scoring FC measures in ways that produce normative scores and for estimating item parameters using factor-analytic and item response theory (IRT) approaches (Brown & Maydeu-Olivares, 2011; Lee, Joo, Stark, & Chernyshenko, 2018; Stark et al., 2005).The current investigation uses the multiunidimensional pairwise preference (Stark et al., 2005) IRT model for evaluating the FC measure. The development of such models has provided opportunity to expand the way that these FC items are administered and scored. The dichotomous response format used provides relatively low information for a specific item, so the use of IRT models has led to the use of CAT for their administration (Boyce, Conway, & Caputo, 2014; CEB, 2014; Naemi, Seybert, Robbins, & Kyllonen, 2014; Stark, Chernyshenko, Drasgow, & White, 2012). These CAT algorithms tailor the test to each response, providing more information at each response than would be observed in a static test form.

For the administration of the CAT, an item bank is developed that consists of individual personality statements describing each of the personality dimensions being assessed. During administration, the test is targeted to an examinee at whatever the estimated personality estimate is at that time. For item pairs, each of the two statements is selected and paired so that the constructed item provides a high level of information at the trait level of the individual. Because the test is tailored to each individual, the test is theoretically unique to each test taker and even unique to the same test taker across two administrations. Consequently, evaluation of test properties for CAT is difficult, as there is not a single form from which to draw conclusions.

## Evaluating Reliability

The estimation of reliability seeks to determine how consistent scores from a measure are and to separate out the sources of variation that may be classified as error. When selecting a strategy for estimating reliability, a choice exists as to whether there will be one or two occasions of administration and whether there will be a single test form. The most common strategy for estimating reliability is to evaluate internal consistency estimates based on a single occasion using a single form. Using estimates such as coefficient alpha (Cronbach, 1951), this approach describes reliability based on the level of intercorrelation among the items in the measure. These estimates are convenient, as they only require a single data collection for a single measure. Unfortunately for CAT measures, there is no readily available formula for computing a reliability estimate based on the observed data or scores from a single occasion.

Another strategy for estimating reliability is through the use of an alternate forms approach, where two parallel measures are administered on a single occasion. Sources of error for this approach include fatigue from extended testing and variation in item content, which may result in differences in responding between the two forms (Traub, 1994).

Test–retest reliability, on the other hand, is a strategy for estimating reliability through administering the same measure across two occasions. There are more sources of error for this approach compared to those on a single occasion, as responses on the first occasion may impact responses on the second, variations in mood or the environment may affect scores, and true changes in the construct of interest may have occurred between the two occasions. Finally, administering a parallel form, rather than the same form, at the second occasion is an alternate form test–retest approach. This strategy is susceptible to the most sources of error, as it has the same limitations as those for test–retest combined with the additional impact of content changes between the forms affecting scores (Traub, 1994).

As single-occasion, single-form approaches such as coefficient alpha are not readily available for FC personality CAT, two administrations of these assessments are needed to estimate reliability. During administration of a CAT, the items are selected based on not only the estimated ability of the respondent at each point, but also on any selection constraints and exposure controls (Revuelta & Ponsoda, 1998). The result of this is that an individual is unlikely to see the same items across two occasions, making the retesting of CAT most analogous to an alternate form test–retest approach in terms of sources of error. Consequently, interpreting the test–retest reliability of a CAT may best be accomplished through a comparison to analogous measures and approaches.

## Interpreting Personality Measure Reliability

With the growth in use of personality for high-stakes assessments, numerous primary studies have examined their reliability, with meta-analytic findings providing an overall summary for both specific measures (e.g., the NEO; Caruso, 2000) and across measures of the same construct (Gnambs, 2014; Viswesvaran & Ones, 2000). Summarizing the findings of the extant literature typically focuses on constructs of the Big Five framework, which has been widely accepted as a broad structure to describe personality (Barrick & Mount, 1991). Examining internal consistency estimates, Viswesvaran and Ones (2000) found unit-weighted meta-analytic estimates of reliability for the Big Five to range from .73 to .78. Comparable estimates for test–retest reliability were markedly lower, ranging from .69 to .76.

The range of test–retest reliability estimates reported by Viswesvaran and Ones (2000) did not take into account the interval between the first and second administrations. For example, the average number of days between administrations for agreeableness was 440.78, with a standard deviation of 1,623.31. Gnambs (2014) found that the interval between administrations had a significant impact on test–retest reliability estimates, with longer intervals resulting in lower reliability estimates. For an interval of approximately 4 weeks, Gnambs found test–retest reliability estimates higher than those from Viswesvaran and Ones, ranging from .77 to .82. There were also differences across popular measures of the Big Five, illustrating that content and format play a role in determining reliability.

For examining a personality CAT that uses an FC response format, it would be advantageous to have comparisons from which to draw conclusions about the reliability of the measure. Unfortunately, only limited instances exist in the available literature of both CAT and FC measures being evaluated. Although CAT has been used by a number of operational testing programs, no public-facing data could be found on the test–retest reliability of these measures. For example, Yang, Bontya, and Moses (2011) reported test–retest correlations of .72 and .74 for the verbal and quantitative scores, respectively, for a paper-and-pencil graduate admissions exam, but no comparable results for scores from a CAT format. In one available research example, Barker (2008) developed a CAT to assess the knowledge of computer science undergraduates, finding a test–retest reliability of .62 for the 133 participants. Similarly, with the relatively recent use of FC measures for high-stakes testing, we found limited information on their test–retest reliability. The only close example that could be located was noted by Bartram (2013), reporting a median correlation of .86 between two-language versions of the Occupational Personality Questionnaire (OPQ32; SHL Group, 2006).

Although the use of test–retest reliability is the most appropriate for constructs such as personality, where the consistency in scores across short time intervals is of primary interest (McCrae, Kurtz, Yamagata, & Terracciano, 2011), the relative scarcity of evaluation of test–retest reliability has led some researchers to call for a renewed interest in these studies (Schmidt, Le, & Ilies, 2003). Watson (2004) called attention to the limited number of high-quality test–retest studies. In particular, Watson noted that sample sizes for the studies typically were less than 100, that the time between occasions was often far too long, and that most studies only examined a single instrument. Despite these limitations, Watson noted that researchers almost always determine that their test–retest reliability estimates are sufficient, regardless of the size of the estimate.

To provide a context within which to evaluate the reliability of a measure when conducting a test–retest study of personality, Watson (2004) recommended three important considerations for planning the study. First, the retest interval should be relatively small, to limit true change in the construct, so that score differences are more easily attributed to measurement error. Second, a large sample size should be used so as to provide an estimate with a high level of precision. Third, benchmark scales should be included in the study to compare the stability of the primary measure of interest. The additional scale or scales might include parallel measures of the same constructs or other construct validity measures for comparison.

## The Current Study

Following the advice given by Watson (2004), this study seeks to examine the test–retest reliability of an FC personality CAT by comparing the reliability of the primary measure of interest to other reliability estimates from the studied sample. Specifically, this study will provide comparison reliability estimates using traditional Likert-type rating scales in addition to those obtained using the FC CAT. Through the inclusion of a widely used measure of the Big Five across two occasions, comparison test–retest estimates can be calculated. Additionally, the inclusion of a second Likert-type scale, with parallel content, will provide the opportunity to estimate alternate form reliability as well as alternate form test–retest reliability. This series of benchmark scales will not only provide a comparison context within which to interpret the reliability of the FC CAT, but also allow an examination of construct validity for the measure.

## Method

### Sample and Procedure

Study participants were recruited from Amazon Mechanical Turk and agreed to participate in this study in exchange for compensation for their time. A total of 1,162 participants were recruited at the start of the study under the expectation of returning after approximately 2 weeks for completion of the second phase of the study. At Time 1, participants completed the FC personality CAT and the Big Five Inventory Two (BFI2; Soto & John, 2017) Likert-type personality measure. Two weeks after the conclusion of Time 1, participants were contacted and asked to complete the second phase of the study. At Time 2, 789 participants returned and again completed the FC personality CAT and the BFI2 Likert-type measure along with the International Personality Item Pool (IPIP; Goldberg et al., 2006) Likert-type personality measure. The interval between Time 1 and Time 2 ranged from 19 to 45 days, with an average of 22.72 ($SD = 4.00$).

Following the elimination of participants who exhibited response patterns that indicated failure to take the study seriously (e.g., selecting the first statement in every pair), complete data were available for a total of 743 individuals. Participants were 47% male, 69% White, 92% spoke English as their first language, and had an average age of 36. Complete sample demographic information is provided in Table 1.

## Measures

### Forced-Choice Personality Computerized Adaptive Test

A 104-item CAT was administered at each time point, assessing 13 personality dimensions. These dimensions were selected from among a set of 21 lower order personality dimensions identified to describe the Big Five (Drasgow et al., 2012). As the 13 dimensions examined here were not reflective of the entire construct map, gaps in describing any specific Big Five construct were expected. A summary of the 13 dimensions, their relationship to the Big Five, and a definition

**Table 1** Sample Demographic Information

| Demographic | Category | *N* | % |
|---|---|---|---|
| Gender | Male | 347 | 46.7% |
| | Female | 391 | 52.6% |
| | Other | 5 | 0.7% |
| Race | African American | 76 | 10.2% |
| | Asian | 90 | 12.1% |
| | Hispanic/Latino | 39 | 5.2% |
| | Native American | 8 | 1.1% |
| | White | 518 | 69.7% |
| | Other | 12 | 1.6% |
| Education level | High school | 91 | 12.2% |
| | Some college/associate's | 262 | 35.3% |
| | Bachelor's degree | 292 | 39.3% |
| | Advanced degree | 97 | 13.1% |

*Note*. Age = 18–73; $M = 36.14$; $SD = 10.82$. The high school education level includes those with an equivalency degree.

**Table 2** Personality Dimensions and Definitions for Forced-Choice Personality CAT Measure

| Big Five construct | Lower-order dimension name | Dimension description | Time 1 | | Time 2 | |
|---|---|---|---|---|---|---|
| | | | *M* | (*SD*) | *M* | (*SD*) |
| Agreeableness | Cooperation | Behaviors and intentions centered around a desire to work or act with others for a common benefit. | −.21 | (.45) | −.21 | (.47) |
| | Generosity | Behaviors associated with activities such as helping and doing things for others, giving to charity and volunteering for community improvement. | −.17 | (.44) | −.17 | (.47) |
| Conscientiousness | Achievement | Feelings and behaviors associated with working toward goals and other positive outcomes. | .04 | (.55) | .03 | (.55) |
| | Order | Behaviors and intentions related to the ability to plan and organize tasks and activities. | −.12 | (.62) | −.10 | (.61) |
| | Responsibility | Feelings and actions related to a sense of duty or being answerable for one's behavior. | .00 | (.44) | −.01 | (.44) |
| | Self-control | Thoughts and behaviors centered around impulsiveness, the ability to focus on tasks without distraction, and the consideration of consequences before taking action. | .05 | (.50) | .06 | (.48) |
| Emotional stability | Adjustment | Feelings and behaviors associated with various degrees of insecurity and anxiety. | −.29 | (.67) | −.28 | (.68) |
| | Well-being | Thoughts and behaviors associated with an individual's general emotional tone and world outlook. | −.15 | (.69) | −.14 | (.66) |
| Extraversion | Dominance | Behaviors associated with being direct and decisive. | −.44 | (.69) | −.43 | (.71) |
| | Sociability | Interest in engaging in friendly social interactions. | −.46 | (.67) | −.47 | (.67) |
| Openness | Curiosity | Interest and behaviors directed toward understanding how the world around us works. | .21 | (.61) | .22 | (.61) |
| | Ingenuity | Thoughts and behaviors associated with imagination and original thinking. | .13 | (.60) | .14 | (.61) |
| | Intellectual efficiency | Interest in and comfort with intellectual and conceptual matters. | .17 | (.58) | .14 | (.55) |

*Note*. $N = 743$.

for each are provided in Table 2. Each dimension was measured by two unidimensional pairs and 12 multidimensional pairs (where one of the two statements is reflective of that dimension), so that each dimension was represented by 16 statements in total across the assessment. The CAT dynamically creates each item pair, selecting two statements based on the test specifications that provide a high level of information for each dimension at each instance and balance additional specifications, such as the social desirability of each statement.

## BFI2 Likert-Type Personality Scale

The BFI2 is an improvement upon the well-established BFI scale (Soto & John, 2009), constructed by Soto and John (2017) as a 60-item measure of 15 dimensions of personality that are hierarchically organized into the overarching Big Five framework. Each dimension is represented by four items, and each Big Five construct comprises three of the dimensions. A summary of the 15 BFI2 dimensions and the relationship of each to the Big Five is provided in Table 3. Responses to each of the items were obtained via a 5-option Likert-type response rating (*disagree strongly*, *disagree a little*, *neutral*, *agree a little*, and *agree strongly*).

## IPIP Likert-Type Personality Scale

To examine the alternate form test – retest reliability of the Likert-type format, a second scale was constructed by selecting items from the IPIP (Goldberg et al., 2006) with content parallel to that of the BFI2. Responses to each of these items were also obtained via the same instructions and 5-option Likert-type response rating scale as the BFI2.

**Table 3** Personality Dimensions and Definitions for Likert-Type Personality Scales

| Big Five construct | Lower-order dimension name | Dimension items | BFI2 Time 1 | | BFI2 Time 2 | | IPIP Time 2 | |
|---|---|---|---|---|---|---|---|---|
| | | | *M* | (*SD*) | *M* | (*SD*) | *M* | (*SD*) |
| Agreeableness | Trust | Tends to find fault with others(R); Has a forgiving nature; Is suspicious of others' intentions(R); Assumes the best about people; | 3.34 | (.94) | 3.35 | (.95) | 3.66 | (.77) |
| | Compassion | Is compassionate, has a soft heart; Feels little sympathy for others(R); Is helpful and unselfish with others; Can be cold and uncaring(R); | 3.81 | (.84) | 3.82 | (.85) | 3.70 | (.74) |
| | Respectfulness | Is respectful, treats others with respect; Starts arguments with others(R); Is sometimes rude to others(R); Is polite, courteous to others; | 4.10 | (.76) | 4.08 | (.77) | 3.86 | (.73) |
| Conscientiousness | Productiveness | Tends to be lazy(R); Has difficulty getting started on tasks(R); Is efficient, gets things done; Is persistent, works until the task is finished; | 3.90 | (.83) | 3.88 | (.85) | 3.71 | (.85) |
| | Organization | Tends to be disorganized(R); Is systematic, likes to keep things in order; Keeps things neat and tidy; Leaves a mess, does not clean up(R); | 3.81 | (.92) | 3.81 | (.94) | 4.04 | (.70) |
| | Responsibility | Is dependable, steady; Can be somewhat careless(R); Is reliable, can always be counted on; Sometimes behaves irresponsibly(R); | 3.90 | (.81) | 3.95 | (.78) | 3.92 | (.77) |
| Emotional Stability | Emotional volatility | Is moody, has up and down mood swings(R); Is emotionally stable, not easily upset; Keeps their emotions under control; Is temperamental, gets emotional easily(R); | 3.58 | (1.00) | 3.60 | (1.00) | 3.55 | (.95) |
| | Depression | Stays optimistic after experiencing a setback; Feels secure, comfortable with self; Often feels sad(R); Tends to feel depressed, blue(R); | 3.61 | (1.07) | 3.65 | (1.07) | 3.57 | (1.11) |
| | Anxiety | Is relaxed, handles stress well; Can be tense(R); Worries a lot(R); Rarely feels anxious or afraid; | 3.11 | (1.08) | 3.18 | (1.10) | 3.37 | (.97) |
| Extraversion | Assertiveness | Is dominant, acts as a leader; Is dominant, acts as a leader; Finds it hard to influence people(R); Prefers to have others take charge(R); | 3.05 | (.94) | 3.03 | (.98) | 2.91 | (.97) |
| | Sociability | Is outgoing, sociable; Tends to be quiet(R); Is sometimes shy, introverted(R); Is talkative; | 2.79 | (1.11) | 2.76 | (1.13) | 3.21 | (1.03) |
| | Energy level | Is less active than other people(R); Is less active than other people(R); Is full of energy; Shows a lot of enthusiasm; | 3.37 | (.93) | 3.34 | (.95) | 3.03 | (.93) |
| Openness | Intellectual curiosity | Is curious about many different things; Avoids intellectual, philosophical discussions(R); Is complex, a deep thinker; Has little interest in abstract ideas(R); | 3.90 | (.84) | 3.93 | (.82) | 3.78 | (.82) |
| | Creative imagination | Is inventive, finds clever ways to do things; Has little creativity(R); Has difficulty imagining things(R); Is original, comes up with new ideas; | 3.85 | (.84) | 3.83 | (.88) | 3.81 | (.76) |
| | Aesthetic sensitivity | Has few artistic interests(R); Is fascinated by art, music, or literature; Values art and beauty; Thinks poetry and plays are boring(R); | 3.70 | (.97) | 3.69 | (.98) | 3.78 | (.91) |

*Note.* $N = 743$; (R) = item is reverse coded; BFI2 = Big Five Inventory Two; IPIP = International Personality Item Pool. Emotional stability dimensions for BFI2 are often negatively valenced. For consistency with other measures in this study, they have been coded into a positive direction.

## Analysis

To examine the test–retest reliability of the FC personality measure and the Likert-type format scales, scores were first obtained for each of the studied attributes and Big Five constructs. The FC personality CAT trait scores were estimated during administration of the assessment using a maximum a posteriori likelihood estimation method. To form Big Five–related composites, the scores for each dimension were transformed to $z$ scores and combined based on their respective relationship to each Big Five construct (although not expected to fully describe each). Scores for the BFI2 and IPIP scales were obtained by first reverse-coding negatively worded items and averaging each of the item responses for the respective scale. A total score approach rather than an IRT scoring for the Likert-type items was used as the reliability, and equivalences of these scores are nearly identical for well-developed scales such as those used here (Culpepper, 2013).

Test–retest reliability estimates were calculated as the correlation between scores at each time point. To examine the impact of demographic characteristics, reliability estimates were also calculated for gender, race, and education subgroups. For the Likert-type response format, alternate form reliability was calculated as the correlation between the BFI2 and IPIP scores at Time 2, and alternate form test–retest reliability was calculated as the correlation between BFI2 scores at Time 1 and IPIP scores at Time 2. These reliability estimates provide comparison context in which to interpret the test–retest reliability estimates obtained from the FC personality measure.

To examine the potential impact of demographic characteristics on reliability estimates for the FC personality measure, three variables were examined based on sufficiently available data. Specifically, reliability was calculated separated by gender (male, female), race (White, non-White), and education level (high school, some college/associate's degree, bachelor's degree, advanced degree). The significance of the differences between the reliability estimates was calculated for each characteristic using an appropriate difference test.

Given the availability of three measures of personality, construct validity evidence for the FC personality scores can also be examined. There was not complete concordance between each lower order personality dimension, so instead the relationship between the FC scores and the Big Five scores for the BFI2 and IPIP scales at each time point can be used to explore validity.

## Results

Scores were obtained for each of the personality dimensions at each time point. Tables 2 and 3 provide the mean and standard deviation for the dimension scores. It can be seen that the dimension means and standard deviations appear consistent across the two administrations for each scale for both the FC personality CAT and BFI2. The pattern of the score means across occasions indicated a basic level of consistency across the two time periods and the probability that maturation effects were unlikely to occur between the two study periods. There also were only minor differences in the descriptive statistics between the BFI2 and the IPIP, providing some evidence of the similarity of the scales in their measurement properties.

## Reliability Estimates

The test–retest reliability estimates for the FC personality CAT measure are provided in Table 4. The dimension-level estimates range from .45 to .77, with a mean reliability of .63. These estimates are on the low end of what was expected as compared to the meta-analytic estimates reported by Gnambs (2014) and Viswesvaran and Ones (2000), but they reflect more narrow traits than the broader Big Five constructs. For the composite scores associated with the dimensions from each Big Five construct, the reliability ranged from .63 to .81, larger than those of the dimension scores and comparable to the meta-analytic estimates reported by Viswesvaran and Ones but smaller than those found by Gnambs for such a short retest interval.

The reliability estimates for the Likert-type personality scales are provided in Table 5. The dimension-level estimates of test–retest reliability for the BFI2 from Time 1 and Time 2 range from .74 to .88, with a mean reliability of .83. These estimates are larger than those for the FC personality CAT, despite being represented by few items. The test–retest reliability for the Big Five composites ranged from .87 to .92, which is very similar to those meta-analytic estimates reported by Gnambs (2014). Alternate form reliability was estimated as the correlation between the BFI2 and the IPIP at Time 2. The dimension-level estimates of alternate form reliability ranged from .57 to .90, with a mean reliability of .75. These estimates overall were smaller than those for test–retest reliability, indicating that content changes across scales, even within

**Table 4** Reliability Estimates for the Forced-Choice Personality Computerized Adaptive Test Measure

| Big Five construct | Dimension | Test–retest reliability |
|---|---|---|
| Agreeableness | Cooperation | .56 |
| | Generosity | .53 |
| Conscientiousness | Achievement | .60 |
| | Order | .68 |
| | Responsibility | .53 |
| | Self-control | .45 |
| Emotional stability | Adjustment | .65 |
| | Well-being | .73 |
| Extraversion | Dominance | .77 |
| | Sociability | .72 |
| Openness | Curiosity | .58 |
| | Ingenuity | .71 |
| | Intellectual efficiency | .64 |
| *Mean* | | *.63* |
| Big Five–related composite | Agreeableness | .63 |
| | Conscientiousness | .67 |
| | Emotional stability | .78 |
| | Extraversion | .81 |
| | Openness | .77 |

*Note*. $N = 743$.

**Table 5** Reliability Estimates for the Likert-Type Personality Scales

| Big Five construct | Dimension | Test–retest reliability | Alt. form reliability | Alt. form test–retest reliability |
|---|---|---|---|---|
| Agreeableness | Trust | .86 | .81 | .75 |
| | Compassion | .74 | .61 | .63 |
| | Respectfulness | .79 | .77 | .72 |
| Conscientiousness | Productiveness | .83 | .81 | .76 |
| | Organization | .84 | .64 | .60 |
| | Responsibility | .81 | .74 | .69 |
| Emotional stability | Emotional volatility | .86 | .85 | .81 |
| | Depression | .87 | .90 | .85 |
| | Anxiety | .87 | .82 | .78 |
| Extraversion | Assertiveness | .82 | .63 | .62 |
| | Sociability | .88 | .83 | .78 |
| | Energy level | .83 | .66 | .64 |
| Openness | Intellectual curiosity | .80 | .81 | .77 |
| | Creative imagination | .81 | .57 | .55 |
| | Aesthetic sensitivity | .86 | .78 | .77 |
| *Mean* | | *.83* | *.75* | *.72* |
| Big Five composite | Agreeableness | .87 | .88 | .84 |
| | Conscientiousness | .89 | .87 | .81 |
| | Emotional stability | .91 | .93 | .85 |
| | Extraversion | .92 | .89 | .90 |
| | Openness | .87 | .83 | .81 |

*Note*. $N = 743$.

a single occasion, resulted in a substantial decrease in reliability. Looking at the Big Five composites, the alternate form reliability ranged from .83 to .93, again similar to the high end of the meta-analytic estimates.

Finally, alternate form test–retest reliability was estimated as the correlation between the BFI2 at Time 1 and the IPIP at Time 2. The dimension-level estimates of alternate form test–retest reliability ranged from .55 to .85, with a mean reliability of .72. These estimates were the smallest of those using the Likert-type scales, as was expected given the impact of the many sources of measurement error, and they likely reflect the same sources of error impacting the FC personality CAT estimates. Consequently, these reliability estimates are also relatively close to those found for the dimension-level

**Table 6** Gender Subgroup Reliability Estimates for the Forced-Choice Personality Measure

| Big Five construct | Dimension | Male $N = 347$ | Female $N = 391$ | Diff. $z$ | $p$ |
|---|---|---|---|---|---|
| Agreeableness | Cooperation | .53 | .59 | −1.19 | .12 |
| | Generosity | .54 | .50 | .79 | .21 |
| Conscientiousness | Achievement | .63 | .57 | 1.17 | .12 |
| | Order | .72 | .63 | 2.22 | **.01** |
| | Responsibility | .54 | .52 | .33 | .37 |
| | Self-control | .50 | .39 | 1.83 | **.03** |
| Emotional stability | Adjustment | .60 | .65 | −1.10 | .13 |
| | Well-being | .73 | .73 | .08 | .47 |
| Extraversion | Dominance | .78 | .75 | 1.24 | .11 |
| | Sociability | .72 | .71 | .16 | .43 |
| Openness | Curiosity | .61 | .54 | 1.36 | .09 |
| | Ingenuity | .72 | .69 | .89 | .19 |
| | Intellectual efficiency | .69 | .58 | 2.37 | **.01** |
| *Mean* | | *.64* | *.60* | | |
| Big Five–related composite | Agreeableness | .61 | .64 | −.61 | .27 |
| | Conscientiousness | .70 | .62 | 1.95 | **.03** |
| | Emotional stability | .75 | .79 | −1.38 | .08 |
| | Extraversion | .81 | .79 | .60 | .27 |
| | Openness | .78 | .75 | 1.14 | .13 |

*Note*. Diff. $z = z$ test of the difference between the two correlations; $p = p$ value of significance of the difference $z$; $p < .05$ indicated in bold.

scores for the FC personality CAT. Examining the alternate form test–retest reliability for the Big Five composites, the estimates ranged from .81 to .90, larger than those for the FC personality CAT but similar to the meta-analytic estimates.

## Subgroup Reliability

Table 6 provides the test–retest reliability estimates for the FC personality CAT measure separated by gender. It can be seen that there were significant differences in test–retest reliability between males and females for three of the lower order dimensions (order, self-control, and intellectual efficiency) and the conscientiousness-related composite. In each case, the reliability estimate was lower for females than for males. Table 7 provides the test–retest reliability estimates separated by race. Due to the relatively small sample sizes available for each of the non-White samples, reliability was examined based on a White versus non-White categorization. As can be seen, significant differences in reliability were observed for five of the lower order dimensions (generosity, achievement, well-being, curiosity, and intellectual efficiency) but not for any of the Big Five-related composites.

Table 8 provides the test–retest reliability estimates for the FC personality CAT, separated into education-level clusters. Looking at the mean of the lower order dimension results, it can be seen that the reliability decreases as the level of education increases. At the dimension level, four have significant differences among the education levels (cooperation, generosity, achievement, and agreeableness). Examining these differences with group-level post hoc comparisons of groups with Bonferroni corrections, it can be seen that generosity showed the largest number of significant group differences and also illustrated the general trend of lowering reliability with education. Participants with a high school or equivalent education were the most reliable on generosity ($r = .68$), whereas those with an advanced degree had the lowest level of reliability ($r = .39$).

Examining the Big Five–related composites in Table 8, agreeableness showed significant differences among the groups, with a significant group difference observed between those with a high school education and those with an advanced degree. This result was unsurprising, given that the two dimensions comprised by the composite also showed significant differences. More surprising was that the extraversion and openness-related composites had significant differences across groups despite their individual dimensions each not being significantly different. Extraversion showed a significant difference only between the high school and advanced degree groups, whereas for openness, both those with some college and those with a bachelor's degree significantly differed from those with an advanced degree.

**Table 7** Race Subgroup Reliability Estimates for the Forced-Choice Personality Measure

| Big Five construct | Dimension | White $N = 518$ | Not White $N = 225$ | Diff. $z$ | $p$ |
|---|---|---|---|---|---|
| Agreeableness | Cooperation | .58 | .50 | 1.41 | .08 |
| | Generosity | .56 | .40 | 2.70 | **.00** |
| Conscientiousness | Achievement | .64 | .48 | 2.90 | **.00** |
| | Order | .69 | .65 | .86 | .19 |
| | Responsibility | .54 | .49 | .88 | .19 |
| | Self-control | .46 | .40 | .87 | .19 |
| Emotional stability | Adjustment | .66 | .63 | .54 | .29 |
| | Well-being | .71 | .78 | −2.04 | **.02** |
| Extraversion | Dominance | .78 | .74 | 1.04 | .15 |
| | Sociability | .72 | .71 | .18 | .43 |
| Openness | Curiosity | .61 | .47 | 2.60 | **.00** |
| | Ingenuity | .72 | .71 | .15 | .44 |
| | Intellectual efficiency | .66 | .57 | 1.73 | **.04** |
| *Mean* | | *.64* | *.58* | | |
| Big Five – related composite | Agreeableness | .65 | .57 | 1.46 | .07 |
| | Conscientiousness | .68 | .62 | 1.25 | .11 |
| | Emotional stability | .77 | .80 | −.83 | .20 |
| | Extraversion | .80 | .80 | .08 | .47 |
| | Openness | .78 | .74 | 1.23 | .11 |

*Note*. Diff. $z = z$ test of the difference between the two correlations; $p = p$ value of significance of the difference $z$; $p < .05$ indicated in bold.

**Table 8** Education Subgroup Reliability Estimates for the Forced-Choice Personality Measure

| Big Five construct | Dimension | High school $N = 91$ | College/Associate's $N = 262$ | Bachelor's $N = 292$ | Advanced $N = 97$ | Diff. $\chi^2$ | $p$ |
|---|---|---|---|---|---|---|---|
| Agreeableness | Cooperation | .55 | .59[a] | .56[b] | .44[a,b] | 9.43 | **.02** |
| | Generosity | .68[a,b] | .56[c] | .48[a] | .39[b,c] | 18.02 | **.00** |
| Conscientiousness | Achievement | .56 | .63[a] | .60 | .49[a] | 8.63 | **.03** |
| | Order | .75 | .64 | .70 | .72 | 5.07 | .17 |
| | Responsibility | .60 | .57 | .49 | .48 | 4.44 | .22 |
| | Self-control | .43 | .46 | .46 | .41 | .85 | .84 |
| Emotional stability | Adjustment | .66 | .62 | .69[a] | .55[a] | 9.88 | **.02** |
| | Well-being | .80 | .75 | .71 | .70 | 4.97 | .17 |
| Extraversion | Dominance | .77 | .76 | .77 | .76 | .08 | .99 |
| | Sociability | .73 | .73 | .71 | .64 | 7.17 | .07 |
| Openness | Curiosity | .55 | .60 | .57 | .52 | 2.52 | .47 |
| | Ingenuity | .71 | .70 | .73 | .69 | 1.57 | .67 |
| | Intellectual efficiency | .65 | .64 | .65 | .60 | 1.59 | .66 |
| *Mean* | | *.65* | *.65* | *.64* | *.61* | | |
| Big Five – related composite | Agreeableness | .72[a] | .63 | .62 | .54[a] | 9.10 | **.03** |
| | Conscientiousness | .66 | .69 | .65 | .67 | .74 | .86 |
| | Emotional stability | .80 | .79 | .78 | .74 | 3.83 | .28 |
| | Extraversion | .85[a] | .81 | .79 | .74[a] | 9.50 | **.02** |
| | Openness | .76 | .78[a] | .78[b] | .68[a,b] | 11.75 | **.01** |

*Note*. Diff. $\chi^2$ = chi-square test of the differences among the four correlations; $p = p$ value of significance of the differences; $p < .05$ indicated in bold.

[a,b,c]This information indicates group comparisons within a dimension or composite that showed significant difference.

**Table 9** Correlation Between Forced-Choice (FC) Personality Scores at Time 1 and Likert-Type Big Five Composites

| | | BFI2 Time 1 | | | | | BFI2 Time 2 | | | | | IPIP Time 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Big Five construct | Dimension | A | C | ES | Ex | O | A | C | ES | Ex | O | A | C | ES | Ex | O |
| Agreeableness | Cooperation | **.48** | **.12** | **.17** | **.13** | −.01 | **.48** | **.14** | **.16** | **.14** | .00 | **.46** | **.11** | **.18** | **.19** | .02 |
| | Generosity | **.34** | .00 | −.05 | .03 | **.12** | **.33** | .01 | −.02 | .06 | **.13** | **.32** | .04 | −.02 | **.09** | **.21** |
| Conscientiousness | Achievement | .07 | **.38** | **.15** | **.32** | **.23** | **.09** | **.35** | **.15** | **.31** | **.22** | .02 | **.40** | **.15** | **.24** | **.14** |
| | Order | −.01 | **.46** | .00 | −.03 | **−.12** | −.02 | **.42** | .02 | −.02 | **−.11** | −.05 | **.26** | .01 | −.06 | **−.15** |
| | Responsibility | **.21** | **.39** | **.18** | **.16** | **.12** | **.22** | **.35** | **.17** | **.15** | **.09** | **.21** | **.38** | **.18** | **.13** | .06 |
| | Self-control | **.13** | **.39** | **.18** | −.01 | **.09** | **.14** | **.39** | **.17** | −.01 | .07 | **.13** | **.34** | **.16** | −.02 | .05 |
| Emotional stability | Adjustment | **.09** | **.17** | **.67** | **.38** | .06 | **.11** | **.19** | **.64** | **.37** | **.08** | **.11** | **.21** | **.60** | **.37** | −.01 |
| | Well-being | **.24** | **.27** | **.64** | **.49** | .04 | **.26** | **.28** | **.65** | **.48** | .05 | **.22** | **.33** | **.66** | **.48** | −.03 |
| Extraversion | Dominance | **−.22** | **.10** | **.16** | **.54** | **.17** | **−.19** | **.09** | **.18** | **.53** | **.20** | **−.24** | **.10** | **.15** | **.44** | **.11** |
| | Sociability | **.14** | .06 | **.28** | **.62** | .06 | **.16** | .07 | **.29** | **.62** | .07 | **.10** | .06 | **.27** | **.67** | .02 |
| Openness | Curiosity | .00 | **.08** | .04 | .03 | **.49** | −.03 | **.08** | .03 | −.01 | **.46** | .02 | **.13** | .02 | −.01 | **.46** |
| | Ingenuity | .00 | .00 | .04 | **.13** | **.56** | .00 | .02 | .04 | **.12** | **.57** | −.02 | .04 | .03 | **.12** | **.46** |
| | Intellectual efficiency | −.06 | **.15** | **.13** | **.09** | **.38** | −.04 | **.16** | **.12** | **.08** | **.36** | −.05 | **.19** | **.10** | .04 | **.31** |
| Big Five–related composite | Agreeableness | **.51** | **.08** | **.08** | **.10** | .07 | **.50** | **.09** | **.09** | **.12** | **.09** | **.49** | **.09** | **.10** | **.18** | **.14** |
| | Conscientiousness | **.15** | **.62** | **.19** | **.17** | **.12** | **.16** | **.57** | **.19** | **.16** | **.10** | **.12** | **.53** | **.19** | **.11** | .04 |
| | Emotional stability | **.19** | **.25** | **.75** | **.50** | .05 | **.21** | **.27** | **.73** | **.49** | **.07** | **.19** | **.31** | **.72** | **.48** | −.02 |
| | Extraversion | −.05 | **.10** | **.27** | **.71** | **.14** | −.02 | **.09** | **.29** | **.70** | **.16** | **−.09** | **.10** | **.26** | **.68** | **.08** |
| | Openness | −.03 | **.10** | **.09** | **.11** | **.62** | −.03 | **.11** | **.08** | **.08** | **.60** | −.02 | **.16** | .07 | .06 | **.53** |

*Note*. $N = 743$; A = agreeableness; C = conscientiousness; ES = emotional stability; Ex = extraversion; O = openness; BFI2 = Big Five Inventory Two; IPIP = International Personality Item Pool. Significant correlations of $p < .05$ are indicated in bold. Expected convergent relationships between FC scores and Likert-type scales are highlighted in orange.

## Construct Validity

The construct validity of the FC personality CAT measure was addressed through an examination of the correlation of each dimension and composite score with the Big Five constructs from the BFI and IPIP scales. Table 9 provides the correlation between the FC personality scores at Time 1 and the Big Five constructs at both time points. The highlighted cells in the table indicate the convergence between related dimensions and constructs. For example, the agreeableness constructs of cooperation and generosity for the FC personality scale were expected to demonstrate the largest correlations with the agreeableness composite of the BFI2 and IPIP scales as compared to the other Big Five scores. Examining Table 9, it can be seen that across every lower order dimension, the FC scores had the largest correlations with the expected Big Five construct. This held true for every score, both at the same time point (Time 1) and the two scales at Time 2. Only achievement showed correlations with unrelated Big Five constructs that approached those for the expected convergent construct. The largest relationships observed were those for the Big Five-related composites. The emotional stability and extraversion composites, in particular, showed consistently high relationships with those same constructs from the BFI2 and IPIP scales.

Table 10 provides the same correlations between the FC personality scores as Table 9, but at Time 2 for the FC personality scores. As with the observed relationships at Time 1, the lower order personality dimensions showed the largest correlations with convergent Big Five constructs at both Time 1 and Time 2. Similarly, the largest correlations were for the relationship between the Big Five-related composites for the FC personality measure and the convergent construct from the BFI2 and IPIP, regardless of the time point at which the data were collected. Indeed, there was on average a less than .02 difference in the correlation of FC personality scores at Time 1 as compared to Time 2. Considering Tables 9 and 10, the expected relationships between the FC and Likert-type scales were observed, providing some evidence that each of the dimensions of the FC measure was targeting an aspect of the intended Big Five construct. That the Big Five-related composites correlated .49 to 75 with those from the BFI2 and IPIP was notable, as well-developed scales of the Big Five often correlate less than those observed here. Using meta-analysis, Pace and Brannick (2010) found that Big Five constructs correlate less than .50 on average. The stability of the observed relationships across each time point was also of

**Table 10** Correlation Between Forced-Choice (FC) Personality Scores at Time 2 and Likert-Type Big Five Composites

| | | BFI2 Time 1 | | | | | BFI2 Time 2 | | | | | IPIP Time 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Big Five construct | Dimension | A | C | ES | Ex | O | A | C | ES | Ex | O | A | C | ES | Ex | O |
| Agreeableness | Cooperation | **.47** | .05 | **.14** | .07 | .00 | **.49** | .06 | **.15** | **.09** | .00 | **.47** | **.08** | **.16** | **.17** | .02 |
| | Generosity | **.34** | .01 | **−.02** | .03 | **.12** | **.35** | .02 | .00 | .06 | **.13** | **.33** | .05 | .00 | **.07** | **.20** |
| Conscientiousness | Achievement | .06 | **.35** | **.12** | **.29** | **.18** | .07 | **.36** | **.15** | **.31** | **.18** | .03 | **.41** | **.15** | **.23** | **.10** |
| | Order | **−.01** | **.47** | **−.02** | **−.04** | **−.07** | **−.04** | **.47** | **−.01** | **−.02** | **−.08** | **−.03** | **.32** | .00 | **−.06** | **−.12** |
| | Responsibility | **.22** | **.37** | **.13** | **.12** | **.07** | **.23** | **.38** | **.13** | **.12** | **.08** | **.20** | **.39** | **.14** | .06 | .03 |
| | Self-control | **.09** | **.31** | **.15** | **−.08** | .06 | **.09** | **.34** | **.15** | **−.09** | .04 | **.13** | **.29** | **.13** | **−.09** | .04 |
| Emotional stability | Adjustment | .06 | **.14** | **.61** | **.36** | .06 | **.08** | **.15** | **.64** | **.38** | **.08** | **.09** | **.18** | **.61** | **.37** | .00 |
| | Well-being | **.24** | **.31** | **.65** | **.49** | .03 | **.25** | **.31** | **.67** | **.52** | .05 | **.24** | **.36** | **.70** | **.49** | **−.03** |
| Extraversion | Dominance | **−.22** | .07 | **.13** | **.50** | **.11** | **−.20** | **.08** | **.17** | **.52** | **.15** | **−.25** | **.08** | **.13** | **.44** | .06 |
| | Sociability | **.12** | .07 | **.27** | **.62** | .04 | **.14** | .07 | **.29** | **.65** | .05 | **.10** | .07 | **.27** | **.70** | .03 |
| Openness | Curiosity | **−.01** | .03 | .00 | .00 | **.44** | **−.02** | .02 | **−.01** | **−.01** | **.45** | .01 | **.10** | **−.01** | .00 | **.45** |
| | Ingenuity | **−.01** | **−.02** | .05 | **.16** | **.54** | **−.03** | .02 | .06 | **.15** | **.58** | **−.04** | .03 | .03 | **.15** | **.46** |
| | Intellectual efficiency | **−.03** | **.10** | **.11** | **.12** | **.38** | **−.03** | **.12** | **.11** | **.12** | **.37** | **−.03** | **.17** | **.10** | **.09** | **.34** |
| Big Five–related composite | Agreeableness | **.51** | .04 | .07 | .06 | **.08** | **.52** | .05 | **.09** | **.09** | **.08** | **.50** | **.08** | **.10** | **.15** | **.14** |
| | Conscientiousness | **.14** | **.57** | **.15** | **.11** | **.10** | **.14** | **.59** | **.16** | **.13** | **.09** | **.12** | **.54** | **.16** | .06 | .02 |
| | Emotional stability | **.17** | **.26** | **.71** | **.49** | .05 | **.19** | **.26** | **.75** | **.51** | **.08** | **.19** | **.30** | **.74** | **.49** | **−.02** |
| | Extraversion | **−.07** | **.09** | **.24** | **.68** | **.09** | **−.04** | **.09** | **.28** | **.71** | **.12** | **−.09** | **.09** | **.24** | **.69** | .05 |
| | Openness | **−.03** | .05 | .07 | **.12** | **.59** | **−.03** | .07 | .07 | **.11** | **.61** | **−.03** | **.13** | .05 | **.10** | **.55** |

*Note.* $N = 743$; A = agreeableness; C = conscientiousness; ES = emotional stability; Ex = extraversion; O = openness; BFI2 = Big Five Inventory Two; IPIP = International Personality Item Pool. Significant correlations of $p < .05$ are indicated in bold. Expected convergent relationships between FC scores and Likert-type scales are highlighted in orange.

note, as this stability suggested that scores from the FC personality CAT may demonstrate consistent validity over time, despite the observed reliabilities.

## Discussion

The purpose of this study is to investigate the test–retest reliability of an FC personality CAT, given the continued growth in the use of this approach for high-stakes testing purposes. The observed test–retest reliability estimates for the FC scores ranged from .44 to .77, which represents a relatively low level of reliability if used for consequential decisions. These estimates, however, are not substantially lower than comparable dimension-level alternate form test–retest reliabilities for the Likert-type scales. This suggests, in context, that although the reliabilities may be low, they may not be substantially lower than those of more established measures in the field (Gnambs, 2014; Viswesvaran & Ones, 2000). This is especially true if the Big Five-related composites are examined, which for the FC scores ranged from .63 to .81 and are well within the range of similar measures of personality and are within the range of meta-analytic estimates reported by Gnambs (2014) and Viswesvaran and Ones (2000).

Examining the reliability of the FC personality CAT separated by subgroups, there are some concerning differences between certain groups. For example, self-control had a low reliability of .50 for males but dropped even lower, to .39, for females. That all differences between males and females were in the direction of reliability being worse for females is of particular concern, as that may indicate group-level differences in responding that result in less consistent scores across administrations. The same trend is almost true when comparing Whites to non-Whites at the lower order dimension level (with the exception of well-being), but these differences become negligible when considered at the composite level. What is particularly interesting when examining the subgroups is that reliability tended to decrease as the educational level of the subgroup increased. For example, the agreeableness-related composite had a reliability of .72 for those with a high school education but dropped to .54 for those with an advanced degree. Given the number of comparisons examined in this study and the sample sizes available here, this finding may not hold in other samples. Consequently, this consistent trend in reduction of reliability merits further exploration.

Despite the relatively low reliability observed at the lower order dimension level, and arguably at the Big Five–composite level, there were remarkably consistent relationships with the related Big Five constructs for the BFI2 and IPIP. FC personality scores showed very similar correlations with Big Five constructs, both within and between the same time points.

The Big Five-related composites, in particular, showed a very consistent relationship and notably small relationships with dissimilar Big Five constructs. The exception to this was the approximately .50 correlation between the Big Five-related composite for emotional stability and the BFI2 and IPIP extraversion constructs. This same relationship held true for the BFI2 itself, with the correlation between emotional stability and extraversion being .52 and .54 at Time 1 and Time 2, respectively. This consistent convergent validity evidence across time points suggests that the scores from the FC personality CAT may be sufficiently reliable for use in many purposes, particularly if considered at the composite level, which showed both high levels of convergent validity and moderately high levels of reliability. When this evidence is combined with the findings from the research literature on the gains in criterion-related validity, it indicates that the FC personality CAT provides sufficient evidence for decision-making purposes.

## Directions for Future Research

There has been little to no research on the stability of scores produced using FC measures and even less in regard to the delivery of these items via CAT. The test–retest reliabilities found in this study suggest that there is still much work to be done in this area. It would be useful for researchers to examine not only the item type but also the influence of various decisions in scale construction, such as dimensionality and number of items per dimension. The current FC measure studied used relatively few items to measure 13 individual dimensions. Despite observing16 statements for each dimension, the items themselves were dichotomous and intentionally multidimensional, resulting in relatively low information from each response. Future research may explore the impact of test length on reliability, as there may be a minimum threshold of items needed to reach a target level of reliability.

The differences across subgroups found here also warrant further exploration. The results showed that reliability estimates were highest for males and for Whites, which suggests that scores on the measure may interact with demographic variables in their performance. Given that the use of CAT requires the precalibration of parameters for administered items, it may be that the sample from which these parameter estimates were obtained influenced the performance of the items across subgroups. Further research may examine how the fit of subgroups across parameter estimates used in FC CAT may impact the reliability and validity of scores.

The promise of FC personality measures is that they may reduce the influence of response biases and distortions that can reduce the accuracy of scores. Conversely, those same biases and distortions may contribute to the consistency of scores for the Likert-type items, as the individual is systematically influencing scores in the same way across both time periods. Future research may explore whether the lower observed reliabilities for the FC personality CAT dimensions may, in part, be explained by the reduction in the response biases.

Additionally, some personality dimensions showed quite low reliability despite their content being intertwined with content from dimensions with relatively high reliability. Most notable was self-control, which not only was low overall but was remarkably low when examined for the female subgroup. There may be content issues with the personality dimension that led to this result. As the FC measure pairs statements from different dimensions, it may be that eliminating particularly problematic personality dimensions results in an overall improvement of reliability for the remaining dimensions. As the research on FC measurement is relatively undeveloped, addressing such issues may have a significant impact on the field.

## References

Barker, T. (2008). Computer adaptive testing in higher education: The validity and reliability of the approach. In F. Khandia (Ed.), *12th CAA international computer assisted assessment conference: Proceedings of the conference on 8th and 9th July 2008 at Loughborough University* (pp. 25–40). Loughborough, England: Loughborough University.

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

Bartram, D. (2013). Scalar equivalence of OPQ32: Big Five profiles of 31 countries. *Journal of Cross-Cultural Psychology*, *44*(1), 61–83. https://doi.org/10.1177/0022022111430258

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods*, *9*, 453–465. https://doi.org/10.1037/1082-989X.9.4.453

Borman, W. C., Buck, D., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology*, *86*, 965–973. https://doi.org/10.1037/0021-9010.86.5.965

Boyce, A. S., Conway, J. S., & Caputo, P. (2014). *Development and validation of Aon Hewitt's personality model and Adaptive Employee Personality Test (ADEPT-15)*. New York, NY: Aon Hewitt.

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

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement*, *60*, 236 – 254. https://doi.org/10.1177/00131640021970484

CEB. (2014). *Global personality inventory – adaptive* [Technical manual]. Surrey, England: Author.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297 – 334. https://doi.org/10.1007/BF02310555

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement*, *37*, 201 – 225. https://doi.org/10.1177/0146621612470210

Drasgow, F., Stark, S, Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support army selection and classification decisions* (Technical Report 1311). Fort Belvoir, VA: Army Research Institute for the Behavioral and Social Sciences.

Gnambs, T. (2014). A meta-analysis of dependability coefficients (test – retest reliabilities) for measures of the Big Five. *Journal of Research in Personality*, *52*, 20 – 28. https://doi.org/10.1016/j.jrp.2014.06.003

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*, 84 – 96. https://doi.org/10.1016/j.jrp.2005.08.007

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*, 167 – 184. https://doi.org/10.1037/h0029780

Hough, L., & Dilchert, S. (2010). Personality: Its measurement and validity for employee selection. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 299 – 319). New York, NY: Routledge Taylor & Francis Group.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2014). *Personality, motivation, and college readiness: A prospectus for assessment and development* (Research Report No. RR-14-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12004.

Lee, P., Joo, S. H., Stark, S., & Chernyshenko, O. S. (2018). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement*, *42*, 1 – 15. https://doi.org/10.1177/0146621618768294

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28 – 50. https://doi.org/10.1177/1088868310366253

Naemi, B., Seybert, J., Robbins, S., & Kyllonen, P. (2014). *Examining the WorkFORCE Assessment for Job Fit and core capabilities of FACETS* (Research Report No. RR-14-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12040.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017). Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality and Individual Differences*, *106*, 183 – 189. https://doi.org/10.1016/j.paid.2016.11.014

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

Pace, V. L., & Brannick, M. T. (2010). How similar are personality scales of the "same" construct? A meta-analytic investigation. *Personality and Individual Differences*, *49*(7), 669 – 676. https://doi.org/10.1016/j.paid.2010.06.014

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*(3), 270 – 290. https://doi.org/10.1080/08959285.2011.580808

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(4), 311 – 327. https://doi.org/10.1111/j.1745-3984.1998.tb00541.x

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353. https://doi.org/10.1037/a0026838, 387

Ryan, A. M., Inceoglu, I., Bartram, D., Golubovich, Y., Grand, J., Reeder, M., … Yao, X. (2015). Trends in testing: Highlights of a global survey. In I. Nikolaou & J. K. Oostrom (Eds.), *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice* (pp. 136 – 153). New York, NY: Psychology Press/Taylor & Francis.

Salgado, J. F., & Tauriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, *23*(1), 3 – 30. https://doi.org/10.1080/1359432X.2012.716198

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, *8*(2), 206. https://doi.org/10.1037/1082-989X.8.2.206, 224

SHL Group. (2006). *OPQ32 technical manual*. Thames Ditton, England: Author.

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*(1), 84–90. https://doi.org/10.1016/j.jrp.2008.10.002

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117. https://doi.org/10.1037/pspp0000096, 143

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, *29*(3), 184–203. https://doi.org/10.1177/0146621604273988

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*, *15*, 463–487. https://doi.org/10.1177/1094428112444611

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications* (*3*). Thousand Oaks, CA: SAGE.

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346–360. https://doi.org/10.1177/0022022104264126

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement*, *60*, 224–235. https://doi.org/10.1177/00131640021970475

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*, 319–350. https://doi.org/10.1016/j.jrp.2004.03.001

White, L. A., & Young, M. C. (1998, August). *Development and validation of the Assessment of Individual Motivation (AIM)*. Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Yang, W. L., Bontya, A. M., & Moses, T. P. (2011). *Repeater effects on score equating for a graduate admissions exam* (Research Report No. RR-11-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02253.x.

## Suggested Citation

Seybert, J. W., & Becker, D. (2019). *Examination of the test–retest reliability of a forced-choice personality measure* (Research Report No. RR-19-37). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12273

**Action Editor:** John Sabatini

**Reviewers:** Margarita Olivera Aguilar and Steven Robbins