Learning Accelerator Research Paper

ON THE VALIDITY OF FORCED CHOICE SCORES DERIVED FROM THE THURSTONIAN ITEM RESPONSE THEORY MODEL

Kate E. Walton

Lina Cherkasova

Richard D. Roberts

2019

Walton, K. E., Cherkasova, L., & Roberts, R. D. (2019). On the validity of forced choice scores derived from the Thurstonian Item Response Theory model. *Assessment*.

This is a draft of "On the validity of forced choice scores derived from the Thurstonian Item Response Theory model," and the copy of record is with SAGE. (DOI: 10.1177/1073191119843585)

On the Validity of Forced Choice Scores Derived from the Thurstonian Item Response Theory Model

Kate E. Walton[1], Lina Cherkasova[2], Richard D. Roberts[3]

[1]ACT, Inc.

[2]St. John's University

[3]Research and Assessment Design (RAD): Science Solutions

Abstract

Forced choice (FC) measures may be a desirable alternative to single stimulus (SS) Likert items, which are easier to fake and can have associated response biases. However, classical methods of scoring FC measures lead to ipsative data, which have a number of psychometric problems. A Thurstonian item response model (TIRT) has been introduced as a way to overcome these issues, but few empirical validity studies have been conducted to ensure its effectiveness. This was the goal of the current three studies, which used FC measures of domains from popular personality frameworks including the Big Five and HEXACO, and both statement and adjective item stems. We computed TIRT and ipsative scores and compared their validity estimates. Convergent and discriminant validity of the scores were evaluated by correlating them with SS scores, and test-criterion validity evidence was evaluated by examining their relationships with meaningful outcomes. In all three studies, there was evidence for the convergent and test-criterion validity of the TIRT scores, though at times this was on par with the validity of the ipsative scores. The discriminant validity of the TIRT scores was problematic and was often worse than the ipsative scores'.

The most commonly used item type in personality assessment is a single stimulus (SS) item, such as the Likert-type item, even though its shortcomings, such as susceptibility to responses biases, have been widely discussed (Cheung & Chan, 2002; Danner et al., 2016). One alternative is the forced choice (FC) method. Respondents are presented with items representing opposite poles of the same trait continuum (e.g., introversion and extraversion), or respondents are presented with multidimensional FC blocks containing at least two items, each one an indicator of a different latent trait. Respondents are instructed to either partially or fully rank the items according to which is most-least descriptive of them. Items within multidimensional blocks are often matched in terms of their level of social desirability or perceived relevance, which reduces the potential for faking. That is, respondents cannot simultaneously make themselves look good on all traits represented within that block so the tendency to respond in a desirable manner is diminished. As a result, the FC method has been conceptualized as a solution to the problem of faking, particularly on high stakes assessments such as those used for personnel selection. Indeed, several empirical studies have established that FC scales are less susceptible to faking (Christiansen, Burns, & Montgomery, 2005; Jackson, Wroblewski, & Ashton, 2000; cf. Heggestad, Morrison, Reeve, & McCloy, 2006). Moreover, given that no scales are used (e.g., multi-point Likert scales), response biases such as acquiescent responding or halo effect should be eliminated (Cheung & Chan, 2002). Finally, meta-analytic findings suggest that FC inventories have greater predictive validity than normative data (Bartram, 2007; Salgado & Táuriz, 2014).

Despite some noted advantages of FC assessments, they have also been met with criticism. One commonly cited problem with the multidimensional FC design is the scoring procedure. Typically with a classical ipsative approach, the item selected to be most like the

respondent is given a score of 2, the item(s) not selected is given a score of 1, and the item selected to be least like the respondent is given a score of 0 (or some mathematical equivalent of these scores). Alternatively, in the case of fully ranking items, scores of 4, 3, 2, and 1 are given to items ranked 1, 2, 3, and 4, respectively (in blocks consisting of four items; Hontangas et al., 2015). This yields ipsative scores; that is, the total score (i.e., the sum across dimensions) is constant for all individuals. Moreover, given that respondents must select among items representing different traits, there is a degree of dependability among trait scores; being high on one trait necessitates being lower on other traits. Therefore, an individual's trait scores are relative to one another and cannot be considered absolute, rendering meaningful inter-individual comparisons difficult. Meade (2004) provided an in-depth look at several problems associated with FC ipsative data and noted that the psychometric properties of such data render reliability and factor analyses impossible. Furthermore, he showed that normative vs. ipsative measures can lead to different selections in hypothetical hiring decision processes, particularly when using moderate or high cut-scores, and cautioned against using FC ipsative data for such high stakes purposes.

Several authors have argued that such items can be more appropriately modeled with item response models (IRT), which would allow for normative comparisons (Maydeu-Olivares & Brown, 2010; Wang, Lee, Joo, Stark, & Louden, 2016; Wang, Qiu, Chen, Ro, & Jin, 2017). There are a few such models. For example, Stark, Chernyshenko, and Drasgow (2005) proposed the multi-unidimensional pairwise-preference (MUPP) model for two-item blocks, which can be extended to handle different response formats with more than two items per block (Hontangas et al., 2015), and McCloy, Heggestad, and Reeve (2005) proposed an IRT model for scoring data that are partially ipsative (i.e., being high on one trait necessitates being lower on other traits, but

total score variability is allowed). We will focus on the Thurstonian IRT model (TIRT; Maydeu-Olivares & Brown, 2010), a normal ogive model with structured loadings, uniquenesses, and local dependencies. Binary outcomes, which are comparisons of different items within the same block, are modeled (see Brown & Maydeu-Olivares, 2011, or Dueber, Love, Toland, & Turner, 2018, for a more "elementary level" introduction to the model). This model is increasingly popular in part due to the easily accessible Mplus scoring syntax (Brown & Maydeu-Olivares, 2012) and its ability to yield accurate latent trait estimates. To fully evaluate the appropriateness of this model, Brown and Maydeu-Olivares (2011) carried out a number of simulation studies to assess the recovery of item parameters and latent trait scores under various conditions (see also Xiao, Liu, & Li, 2017). They concluded that when design recommendations are met for parameters such as the number of traits assessed, block size, number of items per trait, correlations among traits, and the keyed direction of the items, the method yields accurate latent trait estimates.

The introduction of such models seems to have spurred greater interest in the use of FC inventories for both research and practice. It is critical, however, to ensure that such scoring approaches produce valid scores, relative to classical ipsative approaches. Given that these approaches are in their infancy, there is limited validity evidence to support them. In one study, researchers examined the validity evidence for an emotional intelligence FC measure and concluded that TIRT-derived scores had greater convergent and test-criterion validity evidence than scores derived from the classical ipsative approach (Anguiano-Carrasco, MacCann, Geiger, Seybert, & Roberts, 2015). In contrast, findings from a study on narcissism indicated that the scores derived from the two approaches had similar correlations with criterion-related variables (Wetzel, Roberts, Fraley, & Brown, 2016). The current study focuses on two popular and well-

validated personality frameworks, the Big Five and HEXACO models, thus a study examining validity evidence for Big Five personality scores derived from the TIRT model is particularly relevant. In a study of Korean university students, Lee, Lee, and Stark (2018) reported acceptable convergent, discriminant, and test-criterion validity estimates for TIRT Big Five scores, though these were similar to or only slightly better than ipsative scores.

Rationale for the Current Studies

Alone or in concert with other item types, FC inventories may be desirable when faking is a concern or when one wants to mitigate response biases. They have the added advantage of being as or more valid than SS items, according to meta-analytic data. When using FC inventories, one must consider various scoring options. Until recently, scores derived from FC items have been ipsative, not allowing for inter-individual comparison. Newer IRT-based approaches are necessary for creating normative scores that allow for inter-individual comparisons, and the TIRT approach is experiencing an upswing in usage. Particularly when the goal is to generate broad personality factor scores often used in high stakes settings like personnel selection (e.g., Barrick & Mount, 1991; Rothstein & Goffin, 2006), it is necessary to establish the validity of this approach, specifically with a head-to-head comparison of scores derived from the TIRT approach versus the ipsative approach. Lee and colleagues (2018) made a solid first step in this endeavor. We extend this work by exploring multiple item types and personality taxonomies and use three different samples ranging from high school students to

community adults. Specifically, across three studies, our goal was to compare the convergent, discriminant, and test-criterion validity evidence of TIRT scores to classical ipsative FC scores.[1]

Study 1

Rationale

To our knowledge, the only empirical study on the validity of Big Five TIRT scores examined statement item stems. We were interested in extending this work to examine the robustness of the TIRT model fit to responses to adjective ratings rather than statements.

Method

Participants

Prospective participants were all registered ACT test takers for the June 2017 national test date who were in grades 9-12 and were in the United States (N = 363,797). Of those who

[1] Here we focus on comparing TIRT FC scores with ipsative FC scores rather than with SS scores because we are considering a scenario in which one has opted to use a FC measure to capitalize on their known advantages and must consider which scoring approach to use. Aside from it not being our goal, comparisons between TIRT FC and SS scores were not possible in our studies given that the item content across the FC and SS inventories were not equivalent. If interested in how TIRT scores compare with SS scores in terms of validity, one can consult our findings and meta-analytic data (Bartram, 2007; Salgado & Táuriz, 2014) showing that the validity of ipsative scores is better than or equal to the validity of SS scores and apply the transitive property to draw loose conclusions.

received an invitation to take part in the study, 7,373 initiated the survey. Cases were removed from the dataset if they skipped more than 20% of the FC items. The final dataset included 5,268 students in grades 10-12. Three thousand eight hundred seventy-one (73.5%) were female. Three percent of the participants chose not to report their race/ethnicity, but of the remaining participants, the percentage of each category represented were: White (69.1%), Hispanic or Latino (12.9%), Black or African American (6.7%), Asian (5.2%), and two or more races (3.1%).

Procedure and Measures

Participants' contact information was obtained from ACT's national database of registered test-takers. After completing the ACT, students received an email inviting them to participate in a short survey. It described the purpose of the study, indicated that participation was voluntary and would in no way affect students' ACT scores, and stated that survey responses would not be provided to students' chosen universities. The message included a Qualtrics survey link unique to each participant where they were directed to the measures described below. These survey responses were then matched to the ACT database that includes students' ACT scores (i.e., composite score and subject specific scores), GPA, school involvement information, and self-reported demographic information, among other things.

In addition to some of the measures discussed below, school involvement, GPA, and ACT scores were included to evaluate test-criterion validity. Students were asked whether they participated in 12 school activities (e.g., band, student government, athletics, and theater), and participation was quantified by summing these 12 binary variables. One would expect school participation to have the greatest correlation with extraversion, given that individuals high in extraversion tend to be more social and active. Participants indicated their GPA on a seven-point scale in .5 increments ranging from ".5-.9 or lower" to "3.5-4.0 or higher". GPA was expected to

have the strongest relationship with conscientiousness (Poropat, 2009). Specific ACT-personality relationships were expected based on prior research with standardized test scores. In one study (Noftle & Robins, 2007), SAT verbal scores were positively related to openness, but no other Big Five traits were consistently related to SAT verbal scores, and no consistent associations with SAT math scores were observed.

It is common for students to quit such surveys before finishing so the amount of available data decreases with each measure in the survey. The amount of missing data for each measure is described below. Missing data were handled by excluding cases pairwise. Distributions of scores on all measures were close to normal.

Big Five Forced Choice – Adjectives. The questionnaire consists of ten blocks containing three adjectives each, one of which was negatively keyed. An example triad might include the adjectives: *sympathetic*, *organized*, and *shy*. The blocks were balanced so all permutations of three out of five traits were represented and each trait was assessed with the same number of items. In response to each block, participants were instructed to select the adjective that is "most like" them, select the adjective that is "least like" them, and do nothing with the third adjective. To score the FC measure, two approaches were taken. First, we computed classical ipsative scores (below referred to as "ipsative"). Each item received a score of 1 ("least like me"), 2 (not selected), or 3 ("most like me"). Within each trait, items were averaged to derive domain scores. Next we estimated the TIRT model using Mplus (below scores derived from this model are referred to as "TIRT"). The TIRT model converged, had a reasonable fit to the data (CFI = .92; RMSEA = .03), and produced reasonable parameter estimates with small standard errors. Factor scores from this model estimation were saved and

used in subsequent analyses. One hundred nineteen participants were missing data for one triad but were still scored (Mplus uses all available data to estimate the model).

Big Five Inventory-10 (BFI-10). The BFI-10 (Rammstedt & John, 2007) was included to evaluate the convergent and discriminant validity of the FC scores. The BFI-10 measures the broad Big Five domains with ten SS items. Respondents indicated their level of agreement with each item on a scale from "strongly disagree" (1) to "strongly agree" (5). Given that these are two-item scales, we are not reporting Cronbach's alpha. One thousand sixty-two participants did not complete the BFI-10.

Students' Life Satisfaction Scale (SLSS). We used the seven-item SLSS (Huebner, 1991) to assess students' global well-being. Previous research suggests that life satisfaction should have the highest correlations with extraversion and emotional stability (Diener & Lucas, 1999). Respondents indicated their level of agreement with each item from "strongly disagree" (1) to "strongly agree" (5). The scale was reliable in our sample with Cronbach's alpha reaching .85. One thousand one hundred thirty-six participants did not complete the SLSS.

Missing Data Analysis

Given the high number of students who did not complete the BFI-10 or the SLSS, we created a binary variable of "completers" vs. "non-completers" where non-completers were students who were missing data for the BFI-10 or the SLSS. We compared the two groups on TIRT FC scores, school participation, GPA, and composite ACT scores. They differed significantly on extraversion (d = -.08), emotional stability (d = .08), openness (d = .07), GPA (d = .09), and ACT scores (d = .17), though the effect sizes were small (positive effect sizes indicate the completers had higher scores).

Results

Means and standard deviations for variables can be found in Tables 1 and 2.

Evidence for Convergent and Discriminant Validity

To evaluate evidence for their convergent and discriminant validity, we compared the two sets of FC scores' associations with the BFI-10 scales (see Table 1). The two sets of FC scores were highly similar in terms of their correlations with their respective BFI-10 scales. These ranged from .36 to .72 (M = .48) for the TIRT scores and from .34 to .70 (M = .45) for the ipsative scores. The two sets of scores were also highly similar in terms of discriminant validity. The TIRT scores' off-trait correlations ranged (in absolute value) from .01 to .30 (M = .16), and the ipsative scores' correlations ranged from .02 to .29 (M = .15).

Evidence for Test-Criterion Validity

We next evaluated the test-criterion validity of the FC scores by correlating them with the various outcome measures (see Table 2). The two sets of FC scores had highly similar correlations with SLSS, school participation, and GPA. SLSS's correlations with extraversion and emotional stability were high, though the correlations with agreeableness and conscientiousness were approximately the same magnitude. School participation had a modest yet significant correlation with extraversion, which is in line with what one would expect. Also as expected, GPA had the highest correlation with conscientiousness. We correlated the two sets of Big Five scores with the ACT composite score and the scores from the four sections. The TIRT and ipsative scores had a similar pattern of associations with the ACT scores, but the ipsative openness scores' correlations with the English and Reading scores were a bit higher than the TIRT openness scores'.

Finally, we fit separate multiple linear regression models with the five ipsative FC scores or the five TIRT scores as the predictors. Each outcome variable was a separate criterion variable. Note that we included only ACT composite scores here to conserve space and because there were not many notable subtest differences in terms of correlations. We sought to determine which set of FC scores could account for the greatest amount of outcomes' variance. The two sets of scores were highly similar in terms of variance accounted for in all outcome variables (see Table 3).

Discussion

In Study 1, we were able to evaluate evidence for convergent, discriminant, and test-criterion validity of FC scores derived from the TIRT model. On the whole, we conclude that there is evidence for the validity of these scores. The TIRT and ipsative scores were fairly similar in terms of the three types of validity. For the most part, all exhibited the pattern of associations one would expect to see based on theory and previous findings.

Study 2

Rationale

Study 1 had the advantage of being the first time to apply the TIRT model to a FC adjective measure. In Study 2, we examined a slight deviation from the Big Five model, one in which one of the five domains is split into two related, yet distinct factors. This will enable us to further evaluate the robustness of the TIRT model in an instance when two dimensions are more highly correlated than the others (and possibly more highly correlated than factors in Brown & Maydeu-Olivares's (2011) simulation studies), which could arise if one were interested in measuring personality trait facets instead of, or in addition to, broad factors.

Method

Participants

Participants were undergraduate and graduate students at a large private university in the Northeast U.S. who participated for extra course credit or to fulfill a course requirement. Three hundred ninety-five individuals initiated the survey, but a number did not complete it. Cases were removed from the dataset if they were missing five or more of the seven questionnaires (described below) and had completed only one of the two FC questionnaires (n = 55). One additional case was removed for missing both data check items (i.e., items designed to flag individuals responding carelessly). The final dataset included 339 individuals ranging in age from 17 to 37 years (M = 20.0, SD = 2.3; age missing for one participant). Two hundred fifty-five (75.2%) were female. The largest race/ethnicity categories represented were: White (38.3%), Hispanic or Latino (18.3%), Asian (17.4%), and Black or African American (17.4%). The remaining participants selected American Indian or Alaska Native, or "other", or did not report this information.

Procedure and Measures

Participants completed the assessment online in Qualtrics. In addition providing demographic information, participants were asked to complete the measures described below. We asked participants to report their cumulative GPAs because it was expected to correlate positively with conscientiousness-related scores (Poropat, 2009). Unless otherwise stated, there were no missing data and distributions of scores on these measures were close to normal. Missing data were handled by excluding cases pairwise.

Big Five Variation Forced Choice – Statements. This measure, which was designed for use as part of a different assessment (ACT, Inc., 2018), assesses six personality constructs stemming from the Big Five personality model (McCrae & Costa, 2003). The traits of extraversion, agreeableness, emotional stability, and openness to experience were assessed, as well as two components of conscientiousness, namely grit and responsibility. The measure is typically used operationally in educational settings where conscientiousness is the best predictor of academic success (Poropat, 2009), which is why those who developed the measure included a more fine-grained assessment of conscientiousness. The measure consists of 20 blocks containing three items each, one of which was negatively keyed. The blocks were balanced so all permutations of three out of six traits were represented and each trait was assessed with the same number of items. Below is an example.

I complete my assignments on time.
I enjoy creating art projects.
I avoid challenging tasks.

In response to each block, participants were instructed to select the one item that is "most like" them, select the one that is "least like" them, and do nothing with the third item. Two participants did not complete this measure. The same two scoring approaches used in Study 1 were used here. The TIRT model converged, had a reasonable fit to the data (CFI = .89; RMSEA = .02), and produced reasonable parameter estimates with small standard errors. Factor scores from this model estimation were saved and used in subsequent analyses.

Big Five Inventory-2 (BFI-2). The BFI-2 (Soto & John, 2017) was included to evaluate the convergent and discriminant validity of the FC scores. The BFI-2 measures the broad Big Five domains and facets with 60 SS items. Respondents indicated their level of agreement with

each item on a scale from “disagree strongly” (1) to “agree strongly” (4). We calculated mean scores for the domains only. All scales were reliable with the following Cronbach’s alpha values in the current sample: .85 (extraversion), .74 (agreeableness), .85 (conscientiousness), .89 (emotional stability), and .83 (openness to experience).

Satisfaction with Life Scale (SWLS). The widely used, reliable, and valid five-item SWLS (Diener, Emmons, Larson, & Griffin, 1985) assesses individuals’ satisfaction with life as a whole. SWLS was expected to have the highest correlations with extraversion and emotional stability (Diener & Lucas, 1999). Respondents indicated their level of agreement with each item from “disagree strongly” (1) to “agree strongly” (4). The scale was reliable in our sample with Cronbach’s alpha reaching .87. Two participants did not complete this measure.

Computerized Adaptive Assessment of Personality Disorder (CAT-PD). The CAT-PD (Simms et al., 2011) is a reliable and valid self-report measure of personality disorder trait dimensions. We selected and administered four scales with high face validity for internalizing (anxiousness and depressiveness) and externalizing (hostile aggression and norm violation) psychopathology. The CAT-PD internalizing scores were expected to have the strongest (negative) associations with emotional stability, extraversion, and conscientiousness-related scales (Kotov, Gamez, Schmidt, & Watson, 2010), and the CAT-PD externalizing scores were expected to have the strongest (negative) associations with agreeableness and the conscientiousness-related scales (Ruiz, Pincus, & Schinka, 2008). Participants indicated how well each of the 28 statements described them on a scale from “very untrue of me” (1) to “very true of me” (5). The two internalizing scales and the two externalizing scales were combined, and the two resulting scales were reliable with Cronbach’s alpha values reaching .91 (internalizing) and .88 (externalizing). Two participants did not complete this measure.

Counterproductive School Behaviors (CSB). The CSB was designed to assess frequency of students' counterproductive behavior in the school setting in the past year. The items were adapted from a measure of counterproductive behaviors in the workforce (Bennett & Robinson, 2000). CSB was expected to be most highly (negatively) associated with agreeableness and the conscientiousness scales (Berry, Ones, & Sackett, 2007). It consists of seven items to which respondents indicate their level of behavioral frequency from "never" (1) to "daily" (7). The scale was reliable in our sample with Cronbach's alpha reaching .83. More than 45% of the sample claimed to have never engaged in any of the behaviors so scores on this measure were positively skewed (2.67), and the distribution had a high kurtosis value (8.42). To correct for this, we used an inverse transformation, which made the variable's distribution more closely resemble the normal distribution (skew = -.77, kurtosis = -.55). However, this also reversed the signs for all CSB-Big Five correlations. The absolute value of the correlations with the Big Five changed very little from raw to transformed CSB scores (a maximum of .04), though, so we used the raw CSB scores in all analyses reported below.

Results

Means and standard deviations for all SS scales and GPA can be found in Tables 4 and 5.

Evidence for Convergent and Discriminant Validity

To evaluate evidence for their convergent and discriminant validity, we compared the two sets of FC scores' associations with the BFI-2 scales (see Table 4). The TIRT scores on average proved to have slightly better convergent validity, exhibiting the highest correlations with their respective BFI-2 scales. These ranged from .44 to .73 (M = .60). The ipsative scores' correlations ranged from .45 to .63 (M = .56). Differences were slight, though, except for the

conscientiousness scales. On average, the scores derived from the ipsative approach had better discriminant validity than the scores derived from the TIRT approach. The ipsative scores' off-trait correlations ranged from .08 to .37 (M = .23), and the TIRT scores' correlations ranged from .12 to .46 (M = .29). There were some instances of the discriminant validity of the TIRT scores being considerably worse than the ipsative scores'. As one example, the correlation between BFI-2 conscientiousness reached .46 with TIRT agreeableness vs. .21 with the ipsative agreeableness score.

Evidence for Test-Criterion Validity

We next evaluated the test-criterion validity of the FC scores by correlating them with the various outcome measures (see Table 5). In general, the TIRT scores slightly outperformed the ipsative scores. As expected, SWLS correlated highly with extraversion and emotional stability (Diener & Lucas, 1999), but it correlated with other domains to a relatively large degree. Both sets of FC scores correlated with the CAT-PD internalizing and externalizing scores, CSB, and GPA as expected.

We next regressed the outcome variables on the two sets of FC scores to compare their criterion validity (see Table 6). Both sets of FC scores were statistically significant and performed very similarly in terms of variance accounted for (see Table 4).

Discussion

In this study, we were able to evaluate evidence for the convergent, discriminant, and test-criterion validity of Big Five FC scores derived from the TIRT model in a student sample. The model was somewhat different than the pure Big Five model used in Study 1; here two subcomponents of conscientiousness were assessed. The convergent and test-criterion validity

evidence for the TIRT scores was slightly better than the evidence for the ipsative scores (with the exception of convergent validity for conscientiousness scores where the TIRT scores performed considerably better than the ipsative scores). However, when evaluating the discriminant validity, the TIRT scores often had high correlations with off-target traits, and this was often significantly worse than for the ipsative scores.

Study 3

Rationale

As discussed above, the literature on the use of the TIRT model for FC scales is relatively limited; therefore, additional research on the applicability of this model for various item types and personality models, for example, is needed. In Study 1, we extended previous work to examine the validity of the model as applied to adjective ratings of the Big Five in a large national sample of prospective college students. In Study 2, we altered the Big Five model to score two components of one of the broad domains. In our final study, we extend this to examine the validity of the model as applied to adjective ratings of the HEXACO domains in an adult online community sample. Study 3 allows us to further assess the TIRT model robustness as the FC measure used here is fairly dissimilar to Big Five statement measures used in practice and prior research.

Method

Participants

Participants were Amazon Mechanical Turk workers who were paid for their participation. Cases were removed from the dataset if they did not complete the FC questionnaire or if they missed both data check items (i.e., items designed to flag individuals responding

carelessly). The final dataset included 721 individuals ranging in age from 18 to 70 years (M = 34.5, SD = 9.9). Three hundred sixty-two (50.2%) were female. The percentage of each race/ethnicity category represented were: White (72.4%), Black or African American (14.7%), and Hispanic or Latino (12.9%).

Procedure and Measures

Participants completed the assessment online in Qualtrics. Participants were asked to complete the measures described below in addition to giving informed consent and providing demographic information, including highest level of education, which was on a nine-point scale ranging from "up to grade 8" to "graduate degree". The mean education level was 5.96, where 6 indicates a 2-year associate's degree. Prior research on the relationship between the Big Five and educational attainment shows that (positive) openness and emotional stability are the best predictors of educational attainment followed by (negative) extraversion (Rammstedt, Danner, & Lechner, 2017). Distributions of scores on these measures were close to normal. Missing data, which is detailed below when relevant, were handled by excluding cases pairwise.

HEXACO Forced Choice. This questionnaire assesses the six personality constructs of the HEXACO model (Ashton et al., 2004), namely honesty/humility, emotionality, extraversion, agreeableness, conscientiousness, and openness to experience. The questionnaire consists of 20 blocks containing three adjectives each, one of which was negatively keyed. The blocks were balanced so all permutations of three out of six traits were represented and each trait was assessed with the same number of items. In response to each block, participants were instructed to select the adjective that is "most like" them, select the adjective that is "least like" them, and do nothing with the third adjective. The same two scoring approaches used in Studies 1 and 2 were used here. The TIRT model converged, had a reasonable fit to the data (CFI = .87; RMSEA

= .03), and produced reasonable parameter estimates with small standard errors. Factor scores from this model estimation were saved and used in subsequent analyses.

HEXACO Single Stimulus. This questionnaire was included to evaluate the convergent and discriminant validity of the FC scores. It assesses the six HEXACO domains and facets with 72 SS items. Respondents indicated their level of agreement with each item on a scale from "strongly disagree" (1) to "strongly agree" (6). We calculated mean scores for the domains only. In the current sample, the scales had the following Cronbach's alpha values: .67 (honesty/humility), .65 (emotionality), .85 (extraversion), .83 (agreeableness), .78 (conscientiousness), and .86 (openness to experience). Seven individuals were not scored due to complete or excessive missing data. Due to a technical error, 505 participants did not complete a total of 15 items which covered all construct domains.

Big Five Inventory-10 (BFI-10). The BFI-10 (Rammstedt & John, 2007; described above in Study 1) was included to further assess convergent and discriminant validity. According to prior research reporting correlations between the two constructs reaching .30 or higher (Ashton, Lee, & de Vries, 2014; Gaughn, Miller, & Lynam, 2012), the following correlations were expected to be the highest: HEXACO honesty – BFI-10 agreeableness, HEXACO emotionality – BFI-10 emotional stability, HEXACO emotionality – BFI-10 agreeableness, HEXACO extraversion – BFI-10 extraversion, HEXACO extraversion – BFI-10 emotional stability, HEXACO agreeableness – BFI-10 agreeableness, HEXACO agreeableness – BFI-10 emotional stability, HEXACO conscientiousness – BFI-10 conscientiousness, and HEXACO openness – BFI-10 openness. Eighteen participants did not complete the BFI-10.

Results

Means and standard deviations for variables can be found in Tables 7 and 8.

Evidence for Convergent and Discriminant Validity

To evaluate evidence for their convergent and discriminant validity, we compared the two sets of FC scores' associations with the HEXACO SS scales (see Table 7). On the whole, the ipsative scores proved to have slightly better convergent validity. The ipsative scores' correlations ranged from .32 to .72 (M = .51). The TIRT scores' correlations with their respective HEXACO SS scales ranged from .17 to .69 (M = .48). The scores derived from the ipsative approach had better discriminant validity than the scores derived from the TIRT approach. The absolute values of the ipsative scores' off-trait correlations ranged from .01 to .46 (M = .22), and the absolute values of the TIRT scores' correlations ranged from .01 to .67 (M = .27). For both sets of FC scores, in some instances, the off-trait correlations were higher than the target trait correlations (e.g., the FC emotionality scores), though this was particularly problematic for the TIRT scores. As one example, the correlation between the SS conscientiousness scores reached .40 with TIRT openness vs. .09 with the ipsative openness score.

We also evaluated the convergent and discriminant validity of the FC scores by correlating them with the BFI-10 (see Table 8). In general, both sets of FC scores had high correlations with the expected BFI-10 scales; however, there were some high correlations that were not expected. It was difficult to distinguish the two sets of FC scores in terms of which had stronger validity evidence. Relative to the ipsative scores' correlations, in some cases the TIRT scores' correlations were greater in magnitude (e.g., HEXACO agreeableness – BFI-10 agreeableness), in some cases they were very similar (e.g., HEXACO extraversion – BFI-10 extraversion), and in some cases, they were smaller in magnitude (e.g., HEXACO openness – BFI-10 openness).

Evidence for Test-Criterion Validity

The two sets of FC scores had a similar pattern of correlations with educational attainment (see Table 8). We fit two regression models predicting educational attainment with the two sets of scores. Both models were statistically significant, though the amount of variance accounted for was slight (TIRT: $R^2_{(adjusted)} = .02$, $F_{6,714} = 2.85$, $p < .05$; ipsative: $R^2_{(adjusted)} = .01$, $F_{6,714} = 2.64$, $p < .05$).

Discussion

In this study, we were able to evaluate evidence for convergent, discriminant, and test-criterion validity of FC scores of an adjective-based HEXACO measure derived from the TIRT model in a community-based online sample. In terms of convergent and discriminant validity, the ipsative scores generally fared better than the TIRT scores. Discriminant validity for the TIRT scores was particularly problematic. The TIRT and ipsative scores were fairly similar in terms of evidence for test-criterion validity.

The validity evidence for the TIRT scores was arguably worse in Study 3 than in Studies 1 and 2. It is possible that there could be something inherent to HEXACO model that poses a problem for the TIRT model. One possibility is that the latent factor correlations are too high. In this model, nine of the 15 factor correlations exceeded .7. In Study1, zero of the ten exceeded .7, and only one of the 15 did in Study 2 (the two conscientiousness-related factors).

General Discussion

IRT models for scoring FC data, including the TIRT, are relatively new and are likely to increase in popularity given that they are a solution to what has been thought to be a key problem with FC data, namely its ipsative nature (Meade, 2004). Prior to embracing the use of TIRT

scoring, the validity of the scores needs to be fully evaluated. Our objective was to do this by comparing the validity of TIRT scores to ipsative scores and to focus on domain scores from two major personality frameworks. To accomplish this, across three samples with varied demographic characteristics, we examined both Big Five and HEXACO scores assessed with different item types, including both statements and adjectives. In general, we found convergent and test-criterion validity evidence for the TIRT scores, although the evidence was a bit stronger in Studies 1 and 2 in which the Big Five personality model (or a slight variant of it) was used rather than the HEXACO model. There was not strong evidence for the discriminant validity of the TIRT scores.

It is worth pointing out that there was evidence for the validity of the ipsative scores as well, and the ipsative scores had better discriminant validity evidence than the TIRT scores. TIRT score discriminant validity evidence was stronger in previous research (e.g., Lee et al., 2018) so additional data are needed to more fully evaluate whether ipsative scores are generally superior to TIRT scores in terms of discriminant validity. Problems with ipsative scores' test-criterion validity coefficients have been discussed in prior literature (e.g., Baron, 1996). In their study of emotional intelligence, Anguiano-Carrasco and colleagues (2015) reported that the validity evidence for the IRT-derived scores was much stronger than scores derived from the ipsative approach and noted that many of those correlations were actually in the opposite direction than one would expect based on theory. In contrast, others have reported that their ipsative scoring approach did not distort test-criterion validity estimates relative to TIRT scores (Lee et al., 2018; Wetzel et al., 2016). Likewise, in the current studies, the ipsative scores' correlations with external criteria did not appear to be distorted. All were in the expected direction and were generally on par with the TIRT scores. The ipsative and TIRT scores were

highly correlated ranging from .91 to .98 (M = .95) in Study 1, from .77 to .93 (M = .86) in Study 2, and from .74 to .95 (M = .79) in Study 3. This is consistent with prior work reporting correlations exceeding .90 between such scores (Lee et al., 2018).

Although there is some mixed evidence regarding the validity of TIRT scores relative to ipsative scores, IRT-based FC scores can alleviate other problems with ipsative data such as the inability to carry out factor analyses or make inter-individual comparisons. Moreover, studies to date, including this one, provide evidence of their convergent and test-criterion validity. As a result, we conclude that TIRT scoring may be a solution to the problem of ipsative data for those who wish to use FC inventories to alleviate concerns about faking or responses biases associated with Likert items. However, before using TIRT scores in place of ipsative scores, one must carefully consider whether there is sufficient discriminant validity evidence and in the event that any validity estimates are poor, one should use ipsative scores instead. In addition, certain features of the FC design may preclude the use of the TIRT. These include the keyed direction of items, the number of traits measured, the correlations between the traits, and the number of items per block (Brown & Maydeu-Olivares, 2011). Brown and Maydeu-Olivares concluded that the TIRT model can reproduce latent trait estimates well when recommended design guidelines are followed. However, most often researchers use preexisting measures and do not have control over the design. When the parameters of the design fail to conform to the TIRT model specifications or when validity estimates are poor, one can use ipsative scores and have confidence in their validity based on the findings presented here and elsewhere (Lee et al., 2018; Wetzel et al., 2016).

Some additional considerations should be made in future research. We considered only triads and fully ipsative data. Although triads seem to be a popular choice in recent research

using the TIRT model (Anguiano-Carrasco et al., 2014; Guenole, Brown, & Cooper, 2016; Lee et al., 2018), the validity of scores garnered from other designs and partially ipsative data should be evaluated as well. In addition, the outcomes used to evaluate test-criterion validity evidence were included in these studies simply out of convenience. That is, in some cases, the data used here were collected for purposes for a different study. Many of the outcome measures were self-report Likert scales, and more varied outcomes would be beneficial.

References

ACT, Inc. (2018). *ACT® Tessera® technical bulletin*. Iowa City, IA: ACT. Inc.

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

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*, 139-152.

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., …De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology*, *86*, 356-366.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*, 49-56.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1-26.

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

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology*, *85*, 349-360.

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology*, *92*, 410-424.

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

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behaivoral Research*, *44*, 1135-1147.

Cheung, M. W., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling*, *9*, 55-77.

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

Danner, D., Blasius, J., Breyer, B., Eifler, S., Menold, N., Paulhus, D. L., Rammstedt, B., Roberts, R. D., Schmitt, M., & Ziegler, M. (2016). Current challenges, new developments, and future directions in scale construction. *European Journal of Psychological Assessment*, *32*, 175-180.

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

Diener, E., & Lucas, R. E. (1999). Personality and subjective well-being. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 213-229). New York: Russell Sage Foundation.

Dueber, D. M., Love, A. M. A., Toland, M. D., & Turner, T. A. (2018). Comparison of single-response format and forced-choice format instruments using Thurstonian Item Response Theory. *Educational and Psychological Measurement*. Advance online publication.

Gaughn, E. T., Miller, J. D., & Lynam, D. R. (2012). Examining the utility of general models of personality in the study of psychopathy: A comparison of the HEXACO-PI-R and NEO PI-R. *Journal of Personality Disorders*, *26*, 513-523.

Guenole, N., Brown, A. A., & Cooper, A. J. (2016). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of Thurstonian Item Response Modeling. *Assessment*, *23*, 1-14.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology*, *91*, 9-24.

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, *39*, 598-612.

Huebner, E. S. (1991). Initial development of the student's life satisfaction scale. *School Psychology International*, 12, 231-240.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance*, *13*, 371-388.

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin, 136,* 768-821.

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences, 123,* 229-235.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45,* 935-974.

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

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77,* 531-552.

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology, 93,* 116-130.

Poropat, A. (2009). A meta-analysis of the Five-Factor Model of personality and academic performance. *Psychological Bulletin, 135,* 322-338.

Rammstedt, B., Danner, D., & Lechner, C. (2017). Personality, competencies, and life outcomes: results from the German PIAAC study. *Large-scale Assessments in Education, 5,* 1-19.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155-180.

Ruiz, M. A., Pincus, A. L., & Schinka, J. A. (2008). Externalizing psychopathology and the five-factor model: A meta-analysis of personality traits associated with antisocial personality disorder, substance use disorder, and their co-occurrence. *Journal of Personality Disorders*, *22*, 365-388.

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, *23*, 3-30.

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: Introducing the CAT-PD project. *Journal of Personality Assessment, 93*, 380-389.

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,* 117-143.

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

Wang, W., Lee, P., Joo, S., Stark, S., & Louden, R. (2016). MCMC Z-G: An IRT computer program for forced-choice noncognitive measurement. *Applied Psychological Measurement*, *40*, 551-553.

Wang, W., Qiu, X., Chen, C., Ro, S., & Jin, K. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement*, 1-14.

Wetzel, E., Roberts, B. W., Fraley, R. C., & Brown, A. (2016). Equivalence of Narcissistic Personality Inventory constructs and correlates across scoring approaches and response formats. *Journal of Research in Personality*, *61*, 87-98.

Xiao, Y., Liu, H., & Li, H. (2017). Integration of the forced-choice questionnaire and the Likert scale: A simulation study. *Frontiers in Psychology: Methods*, *8*, 1-13.

Table 1

Correlations between forced choice and single stimulus scores: Study 1

	BFI-10				
	E	A	C	ES	O
Extraversion (E)					
TIRT	.72*	.12*	.16*	.30*	.12*
Ipsative	.70*	.13*	.20*	.29*	.16*
Agreeableness (A)					
TIRT	.26*	.36*	.15*	.15*	.16*
Ipsative	.22*	.34*	.10*	.09*	.15*
Conscientiousness (C)					
TIRT	.04*	.13*	.43*	.27*	-.10*
Ipsative	.02	.06*	.40*	.25*	-.15*
Emotional Stability (ES)					
TIRT	.01	.29*	.27*	.38*	-.04*
Ipsative	-.05*	.27*	.19*	.34*	-.09*
Openness (O)					
TIRT	.24*	.11*	.10*	.08*	.50*
Ipsative	.28*	.06*	.13*	.13*	.46*
M	3.08	3.68	3.78	2.93	3.61
(*SD*)	(1.02)	(.85)	(.81)	(1.06)	(.95)

Note. * = $p < .05$.

Table 2

Correlations between forced choice scores and outcome variables: Study 1

	SLSS	Part	GPA	ACT-Comp	ACT-E	ACT-R	ACT-M	ACT-S
Extraversion								
TIRT	.17*	.05*	.02	-.02	-.01	-.01	-.01	-.02
Ipsative	.17*	.05*	.03*	-.02	-.01	-.01	-.01	-.02
Agreeableness								
TIRT	.22*	.01	.07*	.00	.01	.00	.00	.01
Ipsative	.20*	.01	.05*	-.02	-.01	-.02	-.02	-.01
Conscientiousness								
TIRT	.21*	.00	.17*	.07*	.04*	.03*	.10*	.08*
Ipsative	.18*	.01	.16*	.10*	.07*	.06*	.13*	.11*
Emotional Stability								
TIRT	.26*	-.02	.07*	-.01	-.03	-.05*	.04*	.02
Ipsative	.24*	-.02	.04*	-.05*	-.07*	-.08*	.00	-.02
Openness								
TIRT	.02	-.01	.00	.02	.05*	.04*	-.02	.01
Ipsative	.03	.00	.05*	.09*	.10*	.10*	.04*	.07*
M	3.75	4.48	6.73	25.19	25.42	25.91	24.36	24.60
(*SD*)	(.78)	(5.33)	(.61)	(5.04)	(6.02)	(6.10)	(5.23)	(5.00)

Note. * = $p < .05$. Part = participation in school activities, ACT-Comp = ACT composite score, ACT-E = ACT English score, ACT-R = ACT reading score, ACT-M = ACT math score, ACT-S = ACT science score.

Table 3

Regressing outcomes on forced choice scores: Study 1

	TIRT		Ipsative	
	$R^2_{adjusted}$	F	$R^2_{adjusted}$	F
SLSS	.11	102.16*	.10	98.73*
Participation	.00	2.72	.00	2.90
GPA	.04	40.43*	.04	44.30*
ACT	.01	13.04*	.04	44.17*

Note. * = $p < .05$. $df_{regression} = 5$, and $df_{residual}$ ranged from 4,161 to 5,205, depending on missing data.

Table 4

Correlations between forced choice and single stimulus scores: Study 2

	E	A	C	ES	O
Extraversion (E)					
TIRT	.62*	.16*	.42*	.33*	.31*
Ipsative	.63*	.10	.31*	.26*	.35*
Agreeableness (A)					
TIRT	.44*	.44*	.46*	.25*	.35*
Ipsative	.27*	.45*	.21*	.17*	.29*
Grit					
TIRT	.30*	.33*	.66*	.38*	.30*
Ipsative	.35*	.21*	.55*	.33*	.32*
Responsibility					
TIRT	.35*	.24*	.73*	.31*	.26*
Ipsative	.12*	.19*	.59*	.18*	.08
Emotional Stability (ES)					
TIRT	.43*	.14*	.42*	.51*	.20*
Ipsative	.22*	.21*	.37*	.53*	.16*
Openness (O)					
TIRT	.21*	.19*	.12*	.16*	.65*
Ipsative	.25*	.18*	.10	.20*	.61*
M	2.68	3.06	2.91	2.48	3.01
(*SD*)	(.56)	(.43)	(.54)	(.64)	(.52)

Note. * = $p < .05$. C = conscientiousness.

Table 5

Correlations between forced choice scores and outcome variables: Study 2

	SWLS	Int	Ext	CSB	GPA
Extraversion					
TIRT	.36*	-.42*	-.11*	-.04	.09
Ipsative	.31*	-.37*	.01	.01	.01
Agreeableness					
TIRT	.30*	-.30*	-.35*	-.26*	.12*
Ipsative	.16*	-.20*	-.33*	-.24*	-.02
Grit					
TIRT	.33*	-.38*	-.33*	-.28*	.22*
Ipsative	.37*	-.36*	-.18*	-.18*	.22*
Responsibility					
TIRT	.34*	-.34*	-.30*	-.25*	.25*
Ipsative	.26*	-.21*	-.33*	-.26*	.21*
Emotional Stability					
TIRT	.34*	-.51*	-.09	-.03	.11
Ipsative	.26*	-.46*	-.17*	-.08	.11*
Openness					
TIRT	.11	-.15*	-.10	-.14*	-.04
Ipsative	.09	-.21*	-.06	-.12*	-.03
M	2.61	2.70	1.69	1.45	3.42
(*SD*)	(.78)	(.95)	(.64)	(.73)	(.41)

Note. * = $p < .05$. Int = CAT-PD internalizing, Ext = CAT-PD externalizing.

Table 6

Regressing outcomes on forced choice scores: Study 2

	TIRT		Classical	
	$R^2_{adjusted}$	*F*	$R^2_{adjusted}$	*F*
SWLS	.15	10.61*	.16	11.98*
CAT-PD Int	.27	21.17*	.26	20.49*
CAT-PD Ext	.21	15.41*	.21	16.20*
CSB	.14	10.04*	.12	8.79*
GPA	.07	4.77*	.06	4.49*

Note. * = $p < .05$. $df_{regression} = 6$, and $df_{residual}$ ranged from 322 to 331, depending on missing data. Int = internalizing, Ext = externalizing.

Table 7

Correlations between forced choice and single stimulus scores: Study 3

	H	E	eX	A	C	O
Honesty (H)						
TIRT	.44*	-.06	.05	.43*	.49*	.18*
Ipsative	.42*	-.03	.24*	.40*	.40*	.14*
Emotionality (E)						
TIRT	.18*	.17*	.67*	.37*	.35*	.16*
Ipsative	.18*	.32*	.46*	.30*	.30*	.14*
Extraversion (eX)						
TIRT	-.11*	.18*	.69*	.13*	.02	.21*
Ipsative	-.01	.14*	.72*	.23*	.15*	.28*
Agreeableness (A)						
TIRT	.37*	.01	.45*	.53*	.47*	.30*
Ipsative	.35*	-.19*	.20*	.45*	.34*	.18*
Conscientiousness (C)						
TIRT	.33*	.06	.43*	.47*	.56*	.24*
Ipsative	.30*	-.02	.16*	.34*	.57*	.16*
Openness (O)						
TIRT	.31*	.00	.34*	.45*	.40*	.48*
Ipsative	.08*	.13*	.34*	.21*	.09*	.55*
M	4.23	3.24	3.50	4.53	4.46	4.42
(*SD*)	(.65)	(.63)	(.93)	(.67)	(.67)	(.73)

Note. * = $p < .05$.

Table 8

Correlations between forced choice scores and outcome variables: Study 3

	BFI-10					
	E	A	C	ES	O	Edu
Honesty						
TIRT	-.10*	.37*	.51*	.26*	.13*	-.01
Ipsative	.19*	.36*	.47*	.35*	.09*	.02
Emotionality						
TIRT	.63*	.40*	.48*	.59*	.04	.10*
Ipsative	.36*	.31*	.45*	.64*	.01	.09*
Extraversion						
TIRT	.79*	.15*	.12*	.39*	.13*	.13*
Ipsative	.80*	.23*	.26*	.41*	.20*	.12*
Agreeableness						
TIRT	.33*	.51*	.53*	.46*	.21*	.06
Ipsative	.06	.46*	.31*	.14*	.13*	.01
Conscientiousness						
TIRT	.31*	.42*	.67*	.47*	.12*	.07
Ipsative	.05	.25*	.65*	.26*	.06	.01
Openness						
TIRT	.27*	.36*	.43*	.30*	.41*	.08*
Ipsative	.39*	.13*	.12*	.21*	.48*	.11*
M	2.78	3.64	4.06	3.44	3.79	5.96
(*SD*)	(1.21)	(1.03)	(.88)	(1.17)	(.97)	(1.73)

Note. * = $p < .05$. E = extraversion, A = agreeableness, C = conscientiousness, ES = emotional stability, O = openness, Edu = educational level.