Published in final edited form as:
Pers Individ Dif. 2014 October ; 69: 98–103. doi:10.1016/j.paid.2014.05.015.

Application of Generalizability Theory to the Big Five Inventory

Brooke J. Arterberry, M.S., **Matthew P. Martens, Ph.D.**, **Jennifer M. Cadigan, M.A.**, and **David Rohrer, M.A.**
Department of Educational, School, and Counseling Psychology, University of Missouri

Abstract

The purpose of the present study was to examine the Big Five Personality Inventory score reliability (BFI: John, Donahue, & Kentle, 1991) utilizing Generalizability Theory analyses. Participants were recruited from a large public Midwestern university and provided complete data for the BFI on three measurement occasions (n = 264). Results suggested score reliability for scales with 7-10 items were adequate. However, score reliability for two item scales did not reach a .80 threshold. These findings have indicated BFI score reliability was, in general, acceptable and demonstrated the advantages of using Generalizability Theory analyses to examine score reliability.

Keywords

Big Five Personality; Generalizability Theory; Reliability

Through the decades, researchers have developed a theoretical framework of personality to better understand human behavior. The trait taxonomy of personality has been studied using lexical approaches, self-report measures, and observer ratings, where findings have supported evidence for a five-factor model (FFM: extraversion, neuroticism (vs. emotional stability), conscientiousness, agreeableness, and openness to experience) of personality (see Costa & McCrae, 1992; John, Angleitner, & Ostendorf, 1988; John, Naumann, & Soto, 2008). Furthermore, the FFM has been studied across clinical, organizational, and research settings to identify adaptive and maladaptive personality types (e.g. Judge, Klinger, Simon, & Yang, 2008; Littlefield, Vergés, Wood, & Sher, 2012; Samuel & Widiger, 2008).

One commonly used assessment examining the FFM is the Big Five Inventory (John, Donahue, & Kentle, 1991). To create a novel, brief measure that differentiated the BFI from other personality assessments, John and colleagues (1991) developed 44 prototypic items formed into short phrases (e.g., “I am someone who tends to be lazy.”). In addition, Rammstedt and John (2007) created a shorter version of the BFI with 10 items, two items per scale, to provide a brief measure for settings with time-limited assessment protocols. Although shorter personality assessments may be appealing to both researchers and clinicians, there are limitations regarding the validity of score interpretation. The limited items may under-represent the construct being measured, narrowing the operational

Correspondence concerning this article should be addressed to Brooke J. Arterberry; Department of Educational, School, and Counseling Psychology, 16 Hill Hall; University of Missouri, Columbia, MO 65211; (573) 882-3382; bjaww8@mail.missouri.edu.

definition of the construct resulting in the unintended assessment of a theoretically variant construct (Kane, 2013). In essence, unconditionally restricting the number of items used to assess complex constructs like the FFM could result in diminished measurement of the full range of personality processes and associations present within each construct. Considering the extensive use and importance of constructs the BFI measures, using generalizability theory (GT: Brennan, 2001), which overcomes limitations associated with CTT, to assess BFI score reliability is warranted.

Generalizability Theory

GT-based analyses allow the researcher to examine score reliability by simultaneously identifying multiple sources of systematic and unsystematic measurement error (Brennan, 2001; Shavelson & Webb, 2006). In classical test theory (CTT), the coefficient of reliability estimates true score variance with remaining variance attributed to error (Hoyt & Melby, 1999). For example, internal consistency analyses examine error associated with differences in items while test-retest reliability examines error associated with differences across time; however, in both cases other sources of error are subsumed under the "true" score. This variance could be due to systematic error, the object of measurement, or multiple testing occasions, but CTT cannot disentangle these differing sources of error.

GT methods, though, can assess multiple sources of measurement error (Hoyt & Melby, 1999; Webb et al., 2006). A G-study estimates variance due to the object of measurement and facets (e.g., occasions or raters). Observed scores are drawn from the universe of admissible observations (i.e., all hypothetical observations that could be substituted for actual observations) and can then be used to estimate variance components. The D-study uses G-study estimates to test designs (e.g., nested, random, fixed) that may reduce measurement error (Brennan, 2001; Webb et al., 2006). For example, a researcher could design a D-study that increases/decreases the number of items on a measure or increases/decreases the number of measurement occasions to examine possible avenues to reduce measurement error.

There is a dearth of research using GT methods to assess FFM personality constructs. Given the widespread use of the BFI and advantages associated with assessing score reliability via GT methods, the purpose of the present study was to use GT-based analyses to examine the BFI's score reliability. We were particularly interested in D-study tests involving two items on each scale, considering at least two measures exist that attempt to assess the FFM in this manner (Gosling, Rentfrow, & Swann, 2003; Rammstedt & John, 2007). Such D-study results provide insight into score reliability of shorter FFM assessment protocols.

Method

Participants and Procedure

Participants (N = 365) were recruited as part of a larger clinical trial from a public Midwestern university examining brief interventions aimed at reducing alcohol use among college students (Martens, Smith, & Murphy, 2013). In the present study, analyses were restricted to participants who provided complete data for the BFI on three measurement

occasions (n = 264; 72.3%). The majority of the sample was female (64%) and Caucasian (89.7%), with other ethnic representations: Asian/Asian-American (3.0%), Black/African-American (2.7%), Hispanic (2.7%), Native American (0.4%), and all other ethnicities (1.5%). The mean age of the participants was 20.10 years (SD = 1.38).

Participants were recruited through the university mass communication system via an email announcement with a link for participants to complete a screening questionnaire including demographic information, contact information, and frequency of binge drinking episodes. Eligible individuals were called and asked to participate. Interested participants were asked to attend an enrollment meeting and completed informed consent, baseline questionnaires, and participated in a brief intervention. Participants returned to complete one- and six-month follow-up surveys and were compensated with a $25 gift card after completing questionnaires. The university Institutional Review Board approved these procedures.

Measures

Big Five Inventory (BFI)—Personality traits associated with the FFM were assessed using the BFI (John et al., 1991), a 44-item measure with five scales: Extraversion (8 items), Agreeableness (9 items), Conscientiousness (9 items), Neuroticism (8 items), and Openness (10 items). Participants were instructed to read the phrase “I am someone who…” followed by the item statement (e.g., “Can be moody”). Respondents indicated to what degree they agreed with the statement using a 5-point Likert scale ranging from 1 (*Disagree Strongly*) to 5 (*Agree Strongly*). The score reliability and validity of score interpretation have been examined across age, gender, and culture (e.g., Soto & John, 2009; Worrell & Cross, 2004), where factor analytic studies have supported a five-factor solution (e.g., Fossati et al., 2011). Coefficient alphas (e.g., α from .70 to .80) and test-retest reliabilities (e.g., r from .75 to .90) across scale scores have been considered satisfactory (e.g., Benet-Martínz & John, 1998; Worrell & Cross, 2004) in cross-cultural samples using multiple translations of the measure. Test-retest reliability and internal consistency estimates for the sample are reported in Table 1.

Demographics—Participants completed relevant demographic information including age, gender, race, and ethnicity.

Data Analysis

GT analyses were conducted using SPSS with syntax developed by Mushquash and O’Connor (2006). We employed a random effects design for both the G-study and D-study using a two-facet design: persons (p) by items (i) by occasions (o), represented as $p \times i \times o$, where persons is the object of measurement and not a source of error and not considered a facet. Additionally, we included occasions as a facet, as personality traits should remain stable across items as well as occasions. Main and interaction effects for all facets of an observed score were calculated for the G-study, where X is the observed-score (Shavelson, Webb, & Rowley, 1989):

$X_{pio} =$	
μ	grand mean
$+ \mu_p - \mu$	person effect
$+ \mu_i - \mu$	item effect
$+ \mu_o - \mu$	occasion effect
$+ \mu_{pi} - \mu_p - \mu_i + \mu$	person x item effect
$+ \mu_{po} - \mu_p - \mu_o + \mu$	person x occasion effect
$+ \mu_{io} - \mu_i - \mu_o + \mu$	item x occasion effect
$+ X_{pio} - \mu_{pi} - \mu_{po} - \mu_{io} + \mu_p + \mu_i + \mu_o - \mu$	residual

Each of the effects has a mean (i.e., all means are zero except the grand mean) as shown above and estimated variance components, which identify possible sources of error that may influence measurement, where *MS* is the mean square and *n* represents facet sample size:

$\sigma^2_p = \left(MS_p - MS_{pi} - MS_{po} + MS_{pio}\right) / n_i n_o$	person variance component
$\sigma^2_i = \left(MS_i - MS_{pi} - MS_{io} + MS_{pio}\right) / n_p n_o$	item variance component
$\sigma^2_{pi} = \left(MS_{pi} - MS_{pio}\right) / n_o$	occasion variance component
$\sigma^2_{pi} = \left(MS_{pi} - MS_{pio}\right) / n_o$	person x item variance component
$\sigma^2_{po} = \left(MS_{po} - MS_{pio}\right) / n_i$	person x occasion variance component
$\sigma^2_{io} = \left(MS_{io} - MS_{pio}\right) / n_p$	item x occasion variance component
$\sigma^2_{pio} = MS_{pio}$	residual variance component

In the D-study, coefficients (G-coefficients and Phi-coefficients) were calculated from the object of measurement (*p*), items (*i*), and occasions (*o*). G-coefficients were used to determine the ratio of universe-score variance to expected observed-score variance (Shavelson et al., 1989), while Phi-coefficients were calculated for absolute decisions (e.g., criterion-referenced measures: Brennan, 2001). Considering BFI scores are generally interpreted in terms of a relative versus absolute standard (e.g., relative scores are examined within a sample rather than being compared to a population norm), we focused on G-coefficients utilizing estimated D-study variance components to obtain relative and absolute error variances:

$\sigma^2_I = \frac{\sigma 2(i)}{n_i}$	item variance component
$\sigma^2_O = \frac{\sigma 2(o)}{n_o}$	occasion variance component

$\sigma^2_{pI} = \frac{\sigma 2(pi)}{n_i}$	person x item variance component
$\sigma^2_{pO} = \frac{\sigma 2(po)}{n_o}$	person x occasion variance component
$\sigma^2_{IO} = \frac{\sigma 2(io)}{n_i n_o}$	item x occasion variance component
$\sigma^2_{pIO} = \frac{\sigma 2(pio)}{n_i n_o}$	residual variance component

And it follows that,

$\sigma^2_\delta = \sigma^2_{pI} + \sigma^2_{pO} + \sigma^2_{pIO}$	relative error variance
$\sigma^2_\Delta = \sigma^2_I + \sigma^2_O + \sigma^2_{IO} + \sigma^2_{pI} + \sigma^2_{pO} + \sigma^2_{pIO}$	absolute error variance

We then calculated both G-coefficients and Phi-coefficients to provide global indices of score reliability, where *E* means expected:

$$\text{G} - \text{coefficient}: E\rho^2_{X_{pIO},\mu_p} = E\rho^2 = \frac{E_P(\mu_p - \mu)^2}{E_P E_I E_O \left(X_{pIO} - \mu_{IO}\right)^2} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\delta}$$

$$\text{Phi} - \text{coefficient}: \Phi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\Delta}$$

For the D-study, different designs were created through changing measurement facets for items and occasions (*p x I x O*). For the D-study, we estimated score reliability using 1, 2, 3, 5, 7, 9, 12, and 15 items, and 1 through 5 measurement occasions.

Results

Internal Consistency and Test-retest Reliability

Internal consistency of each scale was assessed via Cronbach's alpha and test-retest reliability with an intraclass correlation coefficient (ICC), with results presented in Table 1. Cronbach's alpha is an ICC calculation analogous to the GT estimation of the squared correlation of average scores and universe scores, while the ICC used to examine test-retest reliability is related to the GT calculation for the squared correlation of individual measurements and universe scores (McGraw & Wong, 1996). Findings showed, with the exception of the Openness scale, all alpha values were ≥ .80. Additionally, BFI test-retest reliability of scores ranged from .93 (Openness/Neuroticism) to .96 (Extraversion).

G-Study Results

G-study variance components for each scale are presented in Table 2. The variance components attributable to persons ranged from 19.9% (Openness) to 37.7% (Extraversion) and were the first or second largest percentage for each scale. The main effect variance

component for persons is analogous to true-score variance in CTT and is not considered error. Therefore, a larger proportion of variance represents systematic individual differences in personality scores (Shavelson et al., 1989). The variance attributable to the main effect for measurement occasions (i.e., variability across time) and the interaction for occasions x items (i.e., inconsistency in item responses across time) accounted for less than 1% of variance across scales. The persons x occasions interaction accounted for 1.5% (Conscientiousness) to 3.7% (Neuroticism) of variance, indicating individual response patterns were relatively consistent across time. In contrast, percent of variance accounted for by items ranged from 14% to 33%, and percent of variance accounted for by the persons x items interaction ranged from 18% to 30%. These findings indicate a relatively large discrepancy in item difficulty and how individuals responded to specific items. When examining this discrepancy among scales, the Conscientiousness scale had the highest amount of variance accounted for by items: 33.2%. In comparison to other scales, the Conscientiousness scale variance attributable to items differed by 14.8 to 19.5 percentage points. Although the three-way persons x items x occasions interaction accounted for a large percentage of variance, this variance represents random error associated with the full study design (i.e., it represents the residual term). This residual term is confounded by random error, making it impossible to interpret in GT studies (Brennan, 2001).

G-Coefficients/Phi-Coefficients

The G- and Phi-coefficients were calculated for each scale (see Table 3). G-coefficients ranged from .81 (Openness) to .89 (Extraversion), while Phi-coefficients ranged from .73 (Conscientiousness) to .85 (Extraversion). A G-coefficient > .80 has often been used as a benchmark for acceptable overall score reliability in GT analyses (e.g., Mushquash & O'Connor, 2006), although this criterion has never been validated against independent benchmarks.

D-Study Results

D-study results suggested increasing measurement occasions beyond two and increasing items from seven or ten resulted in modest increases in reliability scores (see Tables 4 and 5). For measurement occasions, increasing the number of occasions beyond two would have little impact on score reliability. For example, assuming there are seven items per scale, the G-coefficient increased by .05-.08 percentage points between one and two measurement occasions, but only increased an additional .03-.06 percentage points between two and five occasions. These results were consistent with G-study findings indicating that little error variance in items was associated with measurement occasions.

In contrast, the number of items included on each scale was more consistently associated with estimated G-coefficients. There was a substantial increase in the estimated G-coefficient when the number of items increased from one to two, and relatively large increases when items were increased from two to seven or 10. Increases in the G-coefficient were relatively small above 10 items. Across all scales, the D-study indicated that 7-10 items per scale, measured on at least two occasions, yielded G-coefficients close to or in excess of the .80 threshold.

Although two-item scales yielded estimated G-coefficients higher than one-item scales, the magnitude of the coefficients was relatively low (with the exception of the Extraversion scale). G-coefficients for two-item scales ranged from .38 (Openness) to .61 (Extraversion) on one measurement occasion. Between one and two measurement occasions, G-coefficients for two items increased between .07 and .09 percentage points. Modest increases of the G-coefficient were observed between two and five occasions ranging from .01 to .04 percentage points.

Post hoc Analyses

Due to the high level of item variance within the Conscientiousness scale, we performed post hoc analyses to determine possible item/s contributing to this higher variance. Since high variance associated with items indicates inconsistency of item difficulty (Webb et al., 2006), we examined the mean scores within the scale. First, we examined mean scores of all the items across each time point. Mean scores on the “Is easily distracted.” item were 2.51 (baseline), 2.52 (one month), and 2.59 (six month). The closest item mean scores “Can be somewhat careless.” were 3.22 (baseline), 3.12 (one month), and 3.12 (six months). Considering the mean scores on the “Is easily distracted.” item were relatively low, we concluded this item might be contributing most to the variance. We conducted another GT analysis removing the “Is easily distracted.” item. The variance associated with items reduced from 33.2% to 25.4%. Furthermore, the variance associated with persons increased from 21% to 24%. To ensure this item was contributing to the most variance in item inconsistency, we conducted eight more GT analyses removing a different item on the scale each time. Variance attributable to items reduced at most by 4% depending on the item and increased by 3% in some instances

Discussion

The current study was conducted to examine the reliability of scores on the Big Five Inventory using GT-based analyses, which provide a more holistic examination of measurement error. We also examined D-study estimates for two item scales, as two instruments attempt to examine the FFM using this structure. Results from the G-study indicated variance in scores was primarily attributable to differences among persons, items, and their two-way interaction. Additionally, findings suggested reliability of BFI scores was generally acceptable, as evidenced by G-coefficients above .80. Results from the D-study suggested that 7-10 items per scale yielded considerably higher estimated reliability coefficients than estimates with fewer items.

One goal of the study was to examine variance components associated with possible sources of measurement error. When examining variance, a perfectly reliable measurement instrument would have 100% of variance attributable to persons. As the object of measurement, the variance is interpreted as individual differences in personality. Ideally, scales would have the highest amount of variance attributable to persons. In this study, the variance attributable to persons had either the highest or second highest amount of variance, providing some support for the reliability of scores on each of the BFI scales.

When examining the interaction of *p* x *i*, it was associated with larger proportions of variance. Across scales, there was inconsistency in how persons responded to items. For example, Openness had the largest proportion of variance indicating some items may be more endorsed and indicative of the construct being measured. In contrast, facets associated with *o*, *p* x *o*, and *i* x *o* were associated with the least proportion of variance. This means there was little inconsistency in how items were endorsed across testing occasions even when accounting for persons and items.

Another interesting finding was the level of item difficulty associated with the Conscientiousness scale scores. In comparison, Conscientiousness items were associated with the largest proportion of variance across scales and variance components. The item facet indicates possible inconsistency of endorsement or difficulty. Thus, we performed post hoc tests to identify items contributing to this variance. The “Is easily distracted.” item had the most inconsistent mean scores within the scale. When removing this item, the proportion of variance associated with this facet lowered. The item may be indicative of behavioral undercontrol, as lower conscientiousness has been shown to be associated with high-risk drinking among college students (e.g., Ruiz, Pincus, & Dickenson, 2003). Therefore, this item may be under endorsed by this sample of high-risk college student drinkers.

Within GT, variance due to items cannot be interpreted as “acceptable” or “unacceptable”, as we can only identify inconsistency within items. Inconsistency of item endorsement can show level of item difficulty among individuals measuring high/low on the construct or that items are not assessing the intended construct. In the present case, one could argue variance in items on the BFI is desirable, assuming researchers are interested in using items that assess each trait across a high/low continuum. Future researchers should consider linking GT and Item Response Theory (IRT) analyses to more precisely examine differences in item difficulty and item discrimination within each scale. By combining these two analyses, a sequential examination can provide more robust interpretations regarding item functioning (see Bachman, Lynch, & Mason, 1995; Bock, Brennan, & Muraki, 2002, Brennan, 2011).

Another goal of this study was to examine the reliability of scores when using two items per scale. Our results indicated reliability of scale scores reduced when assessing personality constructs with two items; however, additional testing occasions may increase the reliability of scores. Assessing personality constructs across three to five occasions to gain score reliability and validity of scale score interpretation may be impractical in time-limited settings. Increasing the number of items to adequately represent and assess constructs in a cross-sectional study would increase both score reliability and validity of score interpretation. In this study, none of the scales reached the .80 threshold when only using two items. However, including a domain definition or a universe of admissible observations is important as reliability estimates become unclear if selection of items does not reasonably satisfy a domain definition of the actual construct. Findings suggested the Extraversion scale required the least amount of items (seven items) to obtain an adequate representation of the construct’s score interpretation and score reliability, where the Neuroticism and Openness scales required +15 items per single occasion to adequately assess the constructs. Despite the attraction of shortening comprehensive scales, research has suggested unconditionally abbreviating constructs raises concerns regarding validity of score interpretation (Smith,

McCarthy, & Anderson, 2000). Researchers and clinicians may benefit from using a shorter version of the BFI, however, caution may be warranted when interpreting results of such a short personality assessment, as the constructs may not be fully represented when utilizing so few items and reduce validity of score interpretation.

There were limitations to the current study. Although GT has many benefits, the analysis does not provide context specific information of facets such as items and persons. Although these analyses do not provide item level or person level information, one can perform post hoc procedures to identify possible item level contextual information. Furthermore, the current sample was composed of primarily Caucasian female college students, which limits the generalizability of the results across different ethnicities and populations. The sample also included a sample of high-risk drinkers that limits generalizability, as this population may be more likely to endorse specific personality traits. Another limitation of this study was the data being obtained through self-report measures.

Future researchers could examine the BFI's score reliability using GT-based analyses across diverse samples to further establish the measure's score reliability above and beyond that of CTT methods. Those wanting to extend these findings can access several Fortran/C program software programs developed to conduct GT-based analyses such as GENOVA (GENeralized analysis Of Variance: Crick & Brennan, 1983), urGENOVA, and mGENOVA, available from the University of Iowa at no cost. Additionally, GT-based analyses are becoming more frequently used across disciplines (e.g., Arterberry, Martens, Smith, & Cadigan, 2012; Xu & Shrout, 2013) due to the accessibility of programs developed for more commonly used statistical packages like SPSS, SAS, and MATLAB (see Mushquash & O'Connor, 2006). We also encourage researchers to continue conducting psychometric analyses such as IRT and/or equipercentile, linking methods on both the long- and short-version of the BFI.

In conclusion, we believe this study achieved our aim in providing supporting evidence of the BFI's score reliability. Moreover, researchers would also benefit from using the BFI to provide more understanding of the personality traits that lead to both adaptive and maladaptive behaviors. Most importantly, clinicians and researchers must use reliable measures such as the BFI, across disciplines to ensure we provide the most accurate information regarding measurement of personality traits.

Acknowledgments

The project was in part supported by National Institute of Health Grant #R21AA016779.

References

Arterberry BJ, Martens MP, Cadigan JM, Smith AE. Assessing the dependability of drinking motives via generalizability theory. Measurement and Evaluation in Counseling and Development. 2012; 45:292–302. doi: 10.1177/0748175612449744.

Bachman LF, Lynch BK, Mason M. Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. Language Testing. 1995; 12:238–257.

Benet-Martínez V, John OP. Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. Journal of Personality and Social Psychology. 1998; 75:729–750. doi: 10.1037/0022-3514.75.3.729. [PubMed: 9781409]

Bock RD, Brennan RL, Muraki E. The information in multiple ratings. Applied Psychological Measurement. 2002; 26:364–375.

Brennan, RL. Generalizability theory. Springer-Verlag Publishing; New York, NY US: 2001.

Brennan, Robert L. Generalizability theory and classical test theory. AppliedMeasurement in Education. 2011; 24:1–21.

Crick, JE.; Brennan, RL. Manual for GENOVA: A generalized analysis of variance system. Research and Development Division, American College Testing Program; 1983.

Costa PT, McCrae RR. Four ways five factors are basic. Personality andIndividual Differences. 1992; 13:653–665. doi: 10.1016/0191-8869(92)90236-I.

Fossati A, Borroni S, Marchione D, Maffei C. The Big Five Inventory (BFI): Reliability and validity of its Italian translation in three independent nonclinical samples. European Journal of Psychological Assessment. 2011; 27:50–58. doi: 10.1027/1015-5759/a000043.

Gosling SD, Rentfrow PJ, Swann WB Jr. A very brief measure of the Big-Five personality domains. Journal of Research in Personality. 2003; 37:504–528. doi: 10.1016/S0092-6566(03)00046-1.

Hoyt WT, Melby JN. Dependability of measurement in counseling psychology: An introduction to generalizability theory. The Counseling Psychologist. 1999; 27:325–352. doi: 10.1177/0011000099273003.

John OP, Angleitner A, Ostendorf F. The lexical approach to personality: A historical review of trait taxonomic research. European Journal of Personality. 1988; 2:171–203. doi: 10.1002/per.2410020302.

John, OP.; Donahue, EM.; Kentle, RL. The Big Five Inventory—Versions 4a and54. University of California, Berkeley, Institute of Personality and Social Research; Berkeley, CA: 1991.

John, OP.; Naumann, LP.; Soto, CJ. Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In: John, OP.; Robins, RW.; Pervin, LA., editors. Handbook of personality: Theory and research. 3rd ed. Guilford Press; New York, NY US: 2008. p. 114-158.

Judge TA, Klinger R, Simon LS, Yang IWF. The contributions of personality to organizational behavior and psychology: Findings, criticisms, and future research directions. Social and Personality Psychology Compass. 2008; 2:1982–2000. doi: 10.1111/j.1751-9004.2008.00136.x.

Kane MT. Validating the interpretations and uses of test scores. Journal of EducationalMeasurement. 2013; 50:1–73.

Littlefield AK, Vergés A, Wood PK, Sher KJ. Transactional models between personality and alcohol involvement: A further examination. Journal of Abnormal Psychology. 2012; 121:778–783. doi: 10.1037/a0026912. [PubMed: 22288908]

Martens MP, Smith AE, Murphy JG. The efficacy of single-component brief motivational interventions among at-risk college drinkers. Journal of Consulting and Clinical Psychology. 2013 Advance online publication.

McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996; 1:30.

Mushquash C, O'Connor BP. SPSS and SAS programs for generalizability theory analyses. Behavior Research Methods. 2006; 38:542–547. doi: 10.3758/BF03192810. [PubMed: 17186766]

Rammstedt B, John OP. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality. 2007; 41:203–212. doi: 10.1016/j.jrp.2006.02.001.

Ruiz MA, Pincus AL, Dickinson KA. NEO PI-R predictors of alcohol use and alcohol-related problems. Journal of Personality Assessment. 2003; 81:226–236. doi: 10.1207/S15327752JPA8103_05. [PubMed: 14638447]

Samuel DB, Widiger TA. A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: A facet level analysis. Clinical Psychology Review. 2008; 28:1326–1342. doi: 10.1016/j.cpr.2008.07.002. [PubMed: 18708274]

Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. AmericanPsychologist. 1989; 44:599–612.

Shavelson, RJ.; Webb, NM. Generalizability Theory. In: Green, JL.; Camilli, G.; Elmore, PB., editors. Handbook of complementary methods in education research. Lawrence Erlbaum Associates Publishers; Mahwah, NJ US: 2006. p. 309-322.

Smith GT, McCarthy DM, Anderson KG. On the sins of short-form development. Psychological Assessment. 2000; 12:102–111. [PubMed: 10752369]

Soto CJ, John OP. 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. 2009; 43:84–90. doi: 10.1016/j.jrp.2008.10.002.

Webb NM, Shavelson RJ, Haertel EH. Reliability coefficients and generalizability theory. Handbook of statistics. 2006; 26:81–124. doi: 10.1016/S0169-7161(06)26004-8.

Worrell FC, Cross WE Jr. The reliability and validity of Big Five Inventory Scores with african american college students. Journal of Multicultural Counseling and Development. 2004; 32:18–32. doi: 10.1002/j.2161-1912.2004.tb00358.x.

Xu JH, Shrout PE. Assessing the reliability of change: A comparison of two measures of adult attachment. Journal of Research in Personality. 2013; 47:202–208.

Table 1

Internal Consistency and Test-Retest Estimates

Scale	α[a]-Baseline	α-1 Month	α-6 Month	ICC[b]
Extraversion	.87	.88	.88	.96
Agreeableness	.81	.83	.83	.94
Conscientiousness	.81	.81	.83	.95
Neuroticism	.82	.83	.84	.93
Openness	.78	.80	.79	.93

[a] *Note*. α = Cronbach's Alpha for internal consistency;

[b] ICC = Intraclass Correlation Coefficient for test-retest reliability of scores

Table 2

Estimated variance components for the G-study p x i x o design

	Extraversion		Agreeableness		Conscientiousness		Neuroticism		Openness	
Effect	**Variance**	**Percent**	**Variance**	**Percent**	**Variance**	**Percent**	**Variance**	**Percent**	**Variance**	**Percent**
a_{p}	.55	37.7	.32	27.2	.33	20.7	.45	27.2	.25	19.9
b_{i}	.23	15.8	.16	13.7	.52	33.2	.31	18.4	.23	18.1
c_{o}	.00	0.0	.00	0.0	.00	0.0	.00	0.1	.00	0.0
$d_{p\ x\ i}$	.34	23.2	.32	27.2	.37	23.3	.40	24.2	.38	30.0
$e_{p\ x\ o}$	.03	2.3	.03	2.2	.02	1.5	.06	3.7	.03	2.1
$f_{i\ x\ o}$	.00	0.1	.00	0.0	.00	0.0	.00	0.1	.00	0.0
$g_{p\ x\ i\ x\ o,e}$	.30	20.8	.35	29.7	.34	21.3	.44	26.3	.38	29.9

[a]*Note*. *p* person effect,

[b]*i* item effect,

[c]*o* occasion effect,

[d]*p x i* person by item effect,

[e]*p x o* person by occasion effect,

[f]*i x o* item by occasion effect,

[g]*p x i x o,e* person by item by occasion effect, plus error

Pers Individ Dif. Author manuscript; available in PMC 2015 October 01.

Table 3

G-coefficients and Phi-coefficients for p x I x O design

Scale	G-Coefficient	Phi-Coefficient
Extraversion	.89	.85
Agreeableness	.85	.81
Conscientiousness	.84	.73
Neuroticism	.84	.78
Openness	.81	.75

Table 4

G-Coefficients for p x I x O design

	Extraversion				Agreeableness				Conscientiousness			
	Measurement Occasions				Measurement Occasions				Measurement Occasions			
Items	**1**	**2**	**3**	**5**	**1**	**2**	**3**	**5**	**1**	**2**	**3**	**5**
1	.45	.52	.55	.58	.32	.39	.42	.45	.31	.37	.40	.43
2	.61	.68	.70	.73	.47	.55	.59	.62	.47	.54	.57	.60
3	.69	.75	.78	.80	.56	.64	.68	.70	.56	.63	.66	.69
5	.77	.83	.85	.86	.67	.74	.77	.79	.67	.73	.76	.78
7	.81	.86	.88	.90	.73	.79	.82	.84	.73	.79	.81	.83
10	.85	.89	.91	.92	.78	.84	.86	.88	.78	.83	.85	.87
12	.86	.91	.92	.93	.80	.86	.88	.90	.80	.85	.87	.89
15	.88	.92	.93	.94	.82	.87	.90	.91	.83	.87	.89	.91

Pers Individ Dif. Author manuscript; available in PMC 2015 October 01.

Table 5

G-Coefficients for p x I x O design

	Neuroticism				Openness			
	Measurement Occasions				Measurement Occasions			
Items	**1**	**2**	**3**	**5**	**1**	**2**	**3**	**5**
1	.33	.41	.44	.47	.24	.30	.33	.35
2	.48	.57	.61	.64	.38	.46	.49	.52
3	.57	.66	.69	.72	.48	.56	.59	.62
5	.66	.74	.78	.80	.59	.67	.70	.72
7	.71	.79	.82	.85	.65	.73	.76	.78
10	.76	.83	.86	.88	.71	.78	.81	.83
12	.77	.85	.87	.90	.74	.81	.83	.85
15	.79	.86	.89	.91	.77	.83	.86	.88