Personality and Individual Differences 69 (2014) 98–103

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Application of Generalizability Theory to the Big Five Inventory

Brooke J. Arterberry [*], Matthew P. Martens, Jennifer M. Cadigan, David Rohrer

Department of Educational, School, and Counseling Psychology, University of Missouri, United States

ARTICLE INFO

Article history:
Received 4 February 2014
Received in revised form 28 April 2014
Accepted 2 May 2014
Available online 9 June 2014

Keywords:
Big Five Personality
Generalizability Theory
Reliability

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.

1. Introduction

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 et al., 1991). To create a novel, brief measure that differentiated the BFI from other personality assessments, John et al., 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 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.

2. 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, Shavelson, & Haertel, 2006). A G-study estimates variance due to the object of measure-

[*] Corresponding author. Address: Department of Educational, School, and Counseling Psychology, 16 Hill Hall, University of Missouri, Columbia, MO 65211, United States. Tel.: +1 573 882 3382.

E-mail address: bjaww8@mail.missouri.edu (B.J. Arterberry).

http://dx.doi.org/10.1016/j.paid.2014.05.015
0191-8869/© 2014 Elsevier Ltd. All rights reserved.

ment 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.

3. Method

3.1. 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.

3.2. 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, Borroni, Marchione, & Maffei, 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-Martínez & 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.

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] α = Cronbach's Alpha for internal consistency;
[b] ICC = Intraclass Correlation Coefficient for test–retest reliability of scores.

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

3.3. 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):

$$\begin{aligned}
X_{pio} = & \\
& \mu && \text{grand mean}\\
& + \mu_p - \mu && \text{person effect}\\
& + \mu_i - \mu && \text{item effect}\\
& + \mu_o - \mu && \text{occasion effect}\\
& + \mu_{pi} - \mu_p - \mu_i + \mu && \text{person} \times \text{item effect}\\
& + \mu_{po} - \mu_p - \mu_o + \mu && \text{person} \times \text{occasion effect}\\
& + \mu_{io} - \mu_i - \mu_o + \mu && \text{item} \times \text{occasion effect}\\
& + X_{pio} - \mu_{pi} - \mu_{po} && \text{residual}\\
& \quad - \mu_{io} + \mu_p + \mu_i + \mu_o && \\
& \quad - \mu &&
\end{aligned}$$

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 (Shavelson et al., 1989):