Personality and Individual Differences xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Facets of measurement error for scores of the Big Five: Three reliability generalizations

Timo Gnambs *

Institute of Psychology, Osnabrück University, Germany

ARTICLE INFO

Article history:
Received 9 July 2014
Received in revised form 14 August 2014
Accepted 17 August 2014
Available online xxxx

Keywords:
Big Five
Reliability
Measurement error
Meta-analysis
Generalizability theory

ABSTRACT

Measurement error in self-reports of personality consists of multiple facets that include random, transient, item- and scale-specific error components. Different reliability coefficients reflect different facets of measurement error. This study presents three reliability generalizations for measures of the Big Five based on 71 independent samples (total $N = 38{,}944$) that derived estimates for five types of reliability. The median aggregated coefficient of equivalence for the five traits was .82, the median coefficient of stability fell at .84, and the respective value for the generalized coefficient of equivalence was .74. The four facets of measurement error accounted for up to a half of the variance in observed scores. Estimates of different reliability coefficients are presented that can be used in future artifact corrections to derive construct-level relationships for the Big Five of personality.

1. Introduction

Observed statistics are always distorted to some degree by measurement error. Therefore, construct-level relationships are derived by correcting observed effects and taking the instruments' unreliabilities into account (Ree & Carretta, 2006). For example, in recent years, several meta-analyses linked the Big Five personality dimensions, namely openness to experiences, conscientiousness, extraversion, agreeableness and neuroticism (or emotional stability), to various important outcomes such as psychopathological disorders (Kotov, Gamez, Schmidt, & Watson, 2010), general psychological functioning (Steel, Schmidt, & Schultz, 2008), and even academic performance (Richardson, Abraham, & Bond, 2012) or political orientation (Sibley, Osborne, & Duckitt, 2012). The prevalent indicator of reliability used for artifact corrections in these studies is coefficient alpha (Cronbach, 1947) that quantifies measurement error in terms of consistency between item responses within a specific measurement occasion. However, coefficient alpha can lead to an overestimation of a measure's reliability, if systematic measurement error specific to the current measurement occasion or the administered instrument is present. Therefore, a variety of more general reliability indices have been suggested in recent years that acknowledge different sources of error in observed scores (e.g., Le, Schmidt, & Putka, 2009; McCrae, Kurtz, Yamagata, & Terracciano, 2011; Schmidt, 2010; Schmidt, Le, & Ilies, 2003; Watson, 2004). Unfortunately, these are seldom reported in primary studies. Therefore, this study presents a series of meta-analyses on measures of the Big Five and derives estimates of five types of reliability that can be used in future research to correct observed statistics for measurement error.

2. Measurement error in self-reports

In classical test theory, the observed test score variance is assumed to represent an additive combination of two variance components: true score variance and measurement error variance (Lord & Novick, 1968). For most research questions, the true score component is of focal interest, whereas the error variance represents a nuisance factor that distorts observed relationships and results in a downward bias between the scores on two measures (Ree & Carretta, 2006). Therefore, it is crucial to obtain precise estimates of the error component in test scores to adjust observed statistics and derive true score relationships between constructs. The size and structure of the error variance is the focus of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), which examines different sources (or "facets") of measurement error that contribute to the observed test score variance. In self-reports, the most important sources of error are random response errors, transient errors and factor errors (Le et al., 2009; Schmidt et al., 2003).

* Address: Institute of Psychology, Osnabrück University, Seminarstr. 20, 49069 Osnabrück, Germany. Tel.: +49 (0)541 969 4417; fax: +49 (0)541 969 14200.
E-mail address: timo.gnambs@uni-osnabrueck.de

http://dx.doi.org/10.1016/j.paid.2014.08.019
0191-8869/

Table 1
Sources of measurement error and reliability indices.

	Coefficient of equivalence (CE)	Coefficient of stability (CS)	Coefficient of equivalence and stability (CES)	Generalized coefficient of equivalence (GCE)	Generalized coefficient of equivalence and stability (GCES)
Random error	x	x	x	x	x
Transient error		x	x		x
Item-specific factor error	x		x	x	x
Scale-specific factor error				x	x

2.1. Sources of measurement error

Random measurement error is a consequence of individual fluctuations in attention or distractions. It results in different responses to the same item within the same measurement occasion. Random error variance can be reduced by increasing the length of the scale and including more items. Transient error represents measurement error specific to a certain measurement occasion and is a result of situational variations in, for example, current levels of mood (Watson, 2004). It affects responses in a single measurement occasion, but gets canceled out across different occasions. Item-specific factor error results from inter-individual differences in the interpretation of an item or from inter-individual differences in constructs that are specific to an item (i.e. reliable item variance not shared with other items). Because it does not capture the theoretical construct of interest, item-specific error is canceled out across different items, while it reproduces for the same item across different measurement occasions (Schmidt et al., 2003). When generalized to the scale level (cf. Le et al., 2009), factor error also results from specific, idiosyncratic ways entire scales operationalize the theoretical construct of interest. Scale-specific differences in, for example, the construction process (e.g., sampling items from a specific content domain) or the choice of specific response formats (e.g., rating vs. forced-choice scales) result in variance components that are not relevant to the construct to be measured but are specific to a given scale. As a consequence, a scale-specific factor error reproduces across different measurement occasions for a specific instrument, but is canceled out across different instruments.[1] Together, these four forms of measurement error—that is, random response error, transient error, item-specific and scale-specific factor error—attenuate observed test score variances and bias observed relationships between constructs.

2.2. Indices of measurement error

Although measurement error can be analyzed using various latent variable techniques (cf. Gnambs, Appel, Schreiner, Richter, & Isberner, 2014; Gnambs & Batinic, 2011; Steyer, Mayer, Geiser, & Cole, 2014), it is more commonly quantified by forms of reliability. Reliability is defined as the ratio of true score variability to total score variability in classical test theory (Lord & Novick, 1968). While several methods have been proposed to calculate test score reliabilities, they differ in the way they define and measure the true score variance. As a result, different measures of reliability quantify different sources of measurement error (cf. Schmidt et al., 2003): Coefficients of equivalence (CE) focus on the shared variance between different items at a single measurement occasion. They quantify measurement error in terms of random and item-specific factor error because these cancel each other out across different items. On the other hand, correlations of test scores across two measurement occasions obtained from the same scale are typically used as measures of test–retest reliabilities (coefficient of stability, CS). These assess random measurement error and transient error, but do not reflect item-specific error. All three forms of measurement error are incorporated in the coefficient of equivalence and stability (CES), which results from correlating two parallel forms of a measure that have been administered on separate occasions. Moreover, Le and colleagues (2009) proposed extensions of CE and CES that also acknowledge scale-specific factor errors. The generalized coefficient of equivalence (GCE) and the generalized coefficient of equivalence and stability (GCES) represent the correlations of test scores from different scales measuring the same construct, each either administered on the same (GCE) or on separate occasions (GCES). Of these coefficients, the GCES represents the most general indicator of reliability that accounts for all four sources of measurement error (see Table 1).

3. The present study

In response to repeated calls for a stronger focus on more appropriate indicators of reliability beyond CE (McCrae et al., 2011; Schmidt, 2010; Schmidt et al., 2003) three reliability generalizations are presented that derive five types of reliability estimates (CE, CS, CES, GCE, and GCES) for the Big Five of personality. Although measurement error across different measures of the Big Five has been examined in previous meta-analyses (e.g., Gnambs, 2014; Pace & Brannick, 2010; Viswesvaran & Ones, 2000), the present study extends these results in several important ways: First, previous reliability generalizations on CE (e.g., Viswesvaran & Ones, 2000) exclusively focused on coefficient alpha. However, coefficient alpha is frequently criticized as being a lower bound of CE and, thus, underestimates the true reliability (Sijtsma, 2009). Therefore, this study focuses on ω_h that represents a more precise indicator of CE (Dunn, Baguley, & Brunsden, 2014; Gignac, 2014). Second, previous reliability generalizations typically included a broad array of instruments that were grouped *posthoc* within the Big Five framework. Because imperfect construct validities might also compromise reliability (see Salgado, 2003, for a respective effect on criterion validity), particularly GCE and GCES, the analyses exclusively focus on instruments that were explicitly constructed according to the Big Five model. Finally, this study is the first to also derive more general types of reliability such as CES or GCES that have not yet been examined for the Big Five from a meta-analytically perspective.

4. Method

4.1. Meta-analytic procedure

4.1.1. Effect sizes

In order to quantify different facets of measurement error the meta-analyses focused on three indices of reliability that are frequently reported in research articles: (a) CE in the form of coefficient ω_h, (b) CS in the form of test–retest correlations, and (c) GCE in the form of correlations between different measures of the Big Five.

[1] It is important to note that the concept of scale-specific error does not apply when scales conceptualize constructs differently—even if the constructs have the same name as, for example, the agreeableness traits in the Big Five and HEXACO models (Ashton, Lee, & de Vries, 2014). In this case the concept of error is not meaningful because different constructs are being measured.

Please cite this article in press as: Gnambs, T. Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personality and Individual Differences* (2014), http://dx.doi.org/10.1016/j.paid.2014.08.019