Journal of Research in Personality 52 (2014) 20–28

Contents lists available at ScienceDirect

Journal of Research in Personality

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

A meta-analysis of dependability coefficients (test–retest reliabilities) for measures of the Big Five

Timo Gnambs *

Institute of Psychology, Osnabrück University, Germany

ARTICLE INFO

Article history:
Available online 17 June 2014

Keywords:
Big Five
Retest reliability
Transient error
Measurement error
Meta-analysis
Reliability generalization
Dependability

ABSTRACT

Dependability coefficients such as test–retest correlations quantify transient error in test scores due to occasion-specific variations in, for example, current mood or feelings. The meta-analysis summarizes 682 test–retest correlations collected within an interval of up to two months from 74 samples (total $N = 14{,}923$) across different measures of the Big Five. The median aggregated dependability estimate for the five traits was $\rho_{tt} = .816$. Extraversion scales resulted in the most dependable scores, whereas agreeableness scales exhibited slightly larger measurement error. Transient error accounted for about 10% of the observed variance in scores of the Big Five. Meta-regression analyses indicated small moderation effects of the chosen retest interval for three traits, with shorter intervals resulting in higher retest correlations.

1. Introduction

Although the basic traits of personality such as the Big Five (Goldberg, 1981) have a rather stable core they are subject to pronounced developmental changes. While the preponderance of change occurs during childhood and adolescence (e.g., Hopwood et al., 2011; Klimstra, Hale, Raiijmakers, Brjanje, & Meeus, 2009; Robins, Fraley, Roberts, & Trzesniewski, 2001) personality also develops across the entire life course from infancy to old age (e.g., Ferguson, 2010; Lucas & Donnellan, 2011; Mõttus, Johnson, & Deary, 2012; Roberts & DelVecchio, 2000; Wortman, Lucas, & Donnellan, 2012).

One challenge in the study of personality change are psychological measures with less than perfect reliability. Measurement error typically attenuates observed trait scores and, consequently, distorts longitudinal relationships. For the study of developmental change in personality the appropriate indicators of measurement error are dependability coefficients (i.e. test–retest reliabilities) which indicate the similarity of scores when a scale is administered twice within a short period of time (e.g., Anusic, Lucas, & Donnellan, 2012; Becker, 2000; Chmielewski & Watson, 2009; McCrae, Kurtz, Yamagata, & Terracciano, 2011; Schmidt, Le, & Ilies, 2003; Watson, 2004). Unfortunately, dependability coefficients are frequently not available for study measures because a second assessment might be difficult to implement in a given situation. Therefore, researchers have to resort to meta-analyses that summarize dependability estimates for their scales. However, available meta-analyses of dependability coefficients for the Big Five (Caruso, 2000; Viswesvaran & Ones, 2000) are afflicted by a serious limitation: they did not take into account the interval between test and retest. As a consequence, these dependability estimates assign variance associated with true trait changes to error variance. Studies using these estimates to correct for error in their measures would result in an overestimation of their true effects.

Therefore, this study answers the repeated call for a greater emphasis of dependability in personality research (McCrae et al., 2011; Schmidt et al., 2003; Watson, 2004) and presents a comprehensive meta-analysis of dependability coefficients for measures of the Big Five that also acknowledges the chosen interval between test and retest.

2. Personality stability and measurement esrror

Several longitudinal studies examined the stability of the five basic traits of personality across the life course. Meta-analytic summaries (Roberts & DelVecchio, 2000) showed that stability coefficients increase during transition to adulthood, start to slow down at the ages between 30 and 40 years, and reach a peak in old age. Recently, this pattern has also been replicated in two national samples of the general public (Lucas & Donnellan, 2011; Wortman et al., 2012). Moreover, these analyses also highlighted

* Address: Institute of Psychology, Osnabrück University, Seminarstr. 20, 49069 Osnabrück, Germany.
E-mail address: timo.gnambs@uni-osnabrueck.de

http://dx.doi.org/10.1016/j.jrp.2014.06.003
0092-6566/

that personality stability follows an inverted U-shaped curve; that is, between 70 and 80 years of age stability coefficients start to decline again. Thus, there is considerable evidence of personality change from infancy to old age. Unfortunately, many studies neglected to incorporate measurement error of their trait scales in their analyses. This seems rather peculiar since Ferguson (2010) reported that measurement error reduced stability coefficients by up to 26%. As a consequence, even if internal consistent measures were administered at two separate occasions and no true changes in personality took place, empirically observed stability coefficients would rarely reach 1. Rather, transient error that is specific to a single measurement occasion would distort the observed effect. For this reason, longitudinal analyses of personality development are well advised to acknowledge the dependability of their measures (cf. McCrae et al., 2011; Schmidt et al., 2003; Watson, 2004).

2.1. Transient error in personality scales

Correlations of test scores between two measurement occasions obtained from the same scale are typically used as indicators of dependability. These reflect two forms of measurement error: random error that is a consequence of individual fluctuations in attention or distractions and transient error that results from variations in, for example, current levels of mood or feelings (Watson, 2004). While transient error affects responses in a single measurement occasion, it is typically cancelled out across different occasions. For example, when respondents are in a good mood, they tend to provide more favorable self-descriptions to themselves and others, whereas negative moods result in less positive self-attributions (Mayer, Gaschke, Braverman, & Evans, 1992; Sedikides, 1994). Thus, even ratings of rather stable traits partly reflect the current emotional state of the respondent. Because affective states are rather unstable (Leue & Lange, 2011), they are unlikely to replicate across different measurement occasions that are separated by a reasonably long time interval (e.g., several days or even weeks). Although transient error is more severe for measures of affective states (Chmielewski & Watson, 2009), stable traits such as the Big Five also display non-ignorable short-term fluctuations: over an interval of eight weeks, up to 16% of the observed score variance can be attributed to random and transient measurement error (Anusic et al., 2012).

Two meta-analyses of test–retest correlations have been previously presented for the Big Five: Caruso (2000) reported a mean test–retest correlation for the NEO personality scales (Costa & McCrae, 1992) collected from four studies of $\rho_{tt} = .75$, whereas Visveswaran and Ones (2000) summarized correlations from several work-related personality inventories, resulting in mean test–retest correlations from $\rho_{tt} = .73$ to .78 for the five traits. However, both meta-analyses are rather inconclusive because they neglected to take the length of the retest interval between measurement occasions into account. They included all test–retest correlations, independent of the time interval between the two assessments. The mean test–retest interval in Viswesvaran and Ones (2000), for example, exceeded a year. As a consequence, these meta-analyses confounded measurement error variance with variance associated with developmental changes in the trait. These test–retest correlations are likely to be an overestimation of error in measures of the Big Five.

2.2. Length of test–retest interval

Transient error can be examined in-depth using various complex, latent variable modeling techniques (cf. Anusic et al., 2012; Gnambs & Batinic, 2011; Steyer, Schmitt, & Eid, 1999). However, in practice it is typically estimated by correlating two measures of the same trait assessed twice within a short period of time. The accuracy of these estimates is strongly influenced by the length of the chosen test–retest interval. An increase of the interval between two measurements typically leads to a decrease in the resulting test–retest correlations (Roberts & DelVecchio, 2000; Schuerger, Zarrella, & Hotz, 1989). This effect is stronger when more true changes take place between test and retest. It is well established that the Big Five of personality show pronounced developmental changes in childhood and adolescence but also during adulthood (Hopwood et al., 2011; Lucas & Donnellan, 2011; Roberts, Caspi, & Moffitt, 2001; Robins et al., 2001; Wortman et al., 2012). Thus, the longer the retest interval, the more variance associated with these developmental changes is assigned to error variance. For example, test–retest correlations for scores of the Big Five Inventory (BFI; John, Naumann, & Soto, 2008) range from .81 to .84 over an interval of two weeks and hardly change for a two months interval, $r_{tt} = .79–.89$ (Chmielewski & Watson, 2009). In contrast, the respective correlations over a period of three years fall between .62 and .70 (Vaidya, Gray, Haig, Mroczek, & Watson, 2008). Because transient error is assumed to be stable over time, the observed differences in correlations are typically attributed to developmental changes. Comparably, meta-analyses of stability coefficients for neuroticism scores in young adults show a marked decline of retest correlations from 1 year ($\rho_{tt} = .66$) to 2 year ($\rho_{tt} = .58$) retest intervals (Fraley & Roberts, 2005). Although longer timer intervals tend to decrease retest correlations, they do not reach zero but gradually approach a nonzero asymptote. Even within one year, extraversion scores show a gradual decline for longer test–retest intervals (Schuerger et al., 1989): an increase of one week translated to a decrease in test–retest correlations of about $\Delta r = -.06$. However, this result has to be interpreted with caution because the study included rather heterogeneous samples that also comprised of children and psychiatric patients.

3. The present study

The available empirical evidence highlights the importance of the retest interval for dependability coefficients to reflect measurement error, rather than true personality changes: Retest intervals should be short enough to rule out developmental change and, at the same time, should be long enough to minimize the risk of carry-over effects when, for example, participants simply recall previous answers from memory and repeat them without properly rereading the items (Cronbach & Furby, 1970). So far, no universally established bounds for appropriate test–retest intervals have been put forward. However, most researchers (explicitly or implicitly) adhere to Catell's recommendation (Catell, Eber, & Tatsuoka, 1976; Cattell's, 1986) and adopt retest intervals of up to eight weeks. Because empirical studies found essentially no difference in dependability between retest intervals of two weeks and two months (e.g., Anusic et al., 2012; Chmielewski & Watson, 2009) test–retest correlations between measurements collected within two months are unlikely to reflect developmental changes in personality, but rather represent indicators of measurement error. Therefore, the present meta-analysis will be limited to studies that assessed the Big Five twice, no longer than two months apart. Moreover, the analyses will also demonstrate that, even within this short period of time, the length between test and retest yields non-negligible effects on the estimated dependability coefficients.

4. Method

4.1. Literature search

Primary studies reporting relevant test–retest correlations for measures of the Big Five were located using a two-step strategy.