A Meta-Analysis of Dependability Coefficients (Test-Retest Reliabilities) for Measures of the Big Five

Timo Gnambs

Osnabrück University

Author Note

Timo Gnambs, Department of Psychology, University of Osnabrück, Germany

Correspondence concerning this article should be addressed to Timo Gnambs, Institute of Psychology, Osnabrück University, Seminarstr. 20, 49069 Osnabrück, Germany, E-mail: timo.gnambs@uni-osnabrueck.de

Accepted for publication in the *Journal of Research in Personality*

Highlights

- A meta-analysis summarized short-term test-retest correlations for the Big Five.
- The median aggregated dependability estimate for the five traits was $\rho_{tt} = .816$.
- Transient error accounted for about 10% of the observed variance in trait scores.
- Shorter retest intervals resulted in more dependable scores for three traits.

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 ρ_{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.

Keywords: Big Five, retest reliability, transient error, measurement error, meta-analysis, reliability generalization, dependability

A Meta-Analysis of Dependability Coefficients (Test-Retest Reliabilities) for Measures of the Big Five

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, Raajmakers, Branje, & 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.

Personality Stability and Measurement Error

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 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 (e.g., Hopwood et al., 2011). 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).

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.

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 & DelVeccio, 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 to .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 (ρ_{tt} = .66) to 2 year (ρ_{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 Δr = -.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.

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 Cattell's (1986; Catell, Eber, & Tatsuoka, 1976) recommendation 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.

Method

Literature Search

Primary studies reporting relevant test-retest correlations for measures of the Big Five were located using a two-step strategy. First, a list of 43 personality inventories that either explicitly operationalized the Big Five (e.g., NEO-PI-R; Costa & McCrae, 1992) or, following a different theoretical model, measured traits that could potentially represent one or more

traits of the Big Five (e.g., Occupational Personality Inventory; Saville, Holdsworth, Nyfiled, Cramp, & Mabey, 1996) were compiled from previous meta-analyses (Birkland, Manson, Kisamore, Brannick, & Smith, 2006; Hough & Ones, 2001; Salgado, 2003), products of several commercial test publishers, and a web-based search for contemporary personality instruments used in psychological practice. Only validated, multi-item instruments were considered; thus, ad-hoc constructed scales or single item measures were excluded. In the second step, relevant studies that administered one of these instruments were located by searching several computerized databases (PsycINFO, Psyndex, EconLit, Psychology & Behavioral Sciences Collection, and Google Scholar) using the search terms "*retest reliability*", "*dependability*", "*stability*" and "*transient error*". Additional studies were identified from previous reliability generalizations (Caruso, 2000; Viswesvaran & Ones, 2000) and the manuals of published personality inventories.

Studies were included in the meta-analysis when they met the following criteria: (a) The study was written in English or German, (b) was published in1990 or later[1], and (c) the interval between test and retest did not exceed two months. Initially, the search also extended to studies with retest intervals up to six months. However, due to the small number of studies with intervals of more than two months, they were not included in the present analyses. Two month retest intervals also fall in line with Catell's (1986; Catell et al., 1976) recommendation which seems to represent an implicit convention most test authors adhere to in practice (e.g., Anusic et al., 2012; Chmielewski & Watson, 2009). (d) Participants were at least 18 years of

[1] After Goldberg (1981) coined the term "Big Five" and wide-spread acceptance of the five factor model as a broad taxonomy of human personality began to emerge during the eighties (cf. John et al., 2008), the first validated Big Five instruments were introduced in the late eighties / early nineties (e.g., Goldberg's, 1992, Big Five Markers). To examine if instruments that were constructed within the Big Five framework displayed divergent dependability coefficients as compared to instruments using a different theoretical model (see section on moderators), studies published prior to 1990 were excluded. Otherwise, the adopted theoretical model of the instruments would have been severely confounded with the publication date of the respective studies.

age and (e) of sound physical and psychological health. Studies on children or patients with severe physical traumata or mental illnesses were not considered to exclude individuals with unstable personalities for whom temporary personality changes seemed likely.

This search identified 68 sources reporting on 75 independent samples. To prevent carry-over effects due to memory (Cronbach & Furby, 1970), one study with an extremely short test-retest interval of 1 day, as compared to the remaining studies (*Min* = 1 week), was excluded. This had no noticeable effect on the results of the meta-analysis. A sensitivity analysis that included the two dependability coefficients from the respective study did not lead to different conclusions.

Coding Process

Classification of scales. The scales from the identified personality inventories were grouped into the Big Five dimensions using a multi-step strategy. Scales from inventories that were developed within the Big Five framework were directly assigned to the corresponding Big Five dimension as indicated by the test authors, whereas scales from the remaining inventories were classified into the respective trait dimensions: First, a description of the Big Five was generated using a short summary adapted from John et al. (2008) and a list of typical attributes of high- and low-scorers for each trait taken from Goldberg (1992). Second, a list of all scales and their respective descriptions was compiled. Based on these descriptions the scales were classified into the Big Five taxonomy by two independent raters. The mean percentage of agreement between the ratings was .80. Inconsistencies between raters were resolved by discussion.

Moderators. Several study characteristics were extracted from the primary studies to test the extent different cross-study differences moderate the size of the test-retest correlations. (a) The length of the test-retest interval (coded in weeks) was included as a continuous covariate to demonstrate the stability of retest correlations across the two month period. Because retest correlations measure random and transient error, two variables were

included to separate these two error components: (b) the number of items in the scale and (c) the coefficient alpha. For studies that did not report coefficient alpha, the respective values from norm samples were substituted. (d) Instruments that were constructed according to the Big Five model (coded 1) were compared to instruments using a different theoretical basis (coded as -1) as prior research identified divergent validities for the latter (Salgado, 2003). (e) Because test score reliabilities depend on the current sample, that is, the same measure will yield higher reliabilities in more heterogeneous samples (Vacha-Haase, 1998), students (coded 1) were compared to samples from the general, adult population (coded -1). Initially, a variety of additional characteristics such as age, sex or score variance were also extracted. But due to an excessive amount of missing data (over 50%) these variables were not included in the analyses. However, dependability coefficients are typically far less affected by range restrictions in test scores than coefficient alpha (Fife, Mendoza, & Terry, 2012). (f) Possible cross-cultural differences were examined by contrasting samples from the United States (coded as -1) with samples from other countries (coded as 1). Finally, two indicators of publication bias were extracted: (g) the publication year was included as a continuous covariate and (h) correlations published in research reports (coded -1) were compared to those taken from test manuals (coded 1).

Meta-Analytic Procedure

Test-retest correlations were aggregated separately for each of the five dimensions of personality using a random effects meta-analysis (Cheung, 2013). To account for sampling error the correlations were weighted by the inverse of their variances. Unbiased estimates of the sampling variances were calculated using the formulas in Hedges (1989). Two web-based studies with extraordinarily large samples ($N > 1{,}000$) were truncated to the maximum sample size of the remaining studies before calculating the variances (cf. Gnambs, 2013). Otherwise the aggregated, variance-weighted correlations would predominantly reflect these two

samples and give hardly any weight to the other studies. The accuracy and significance of the synthesized mean effect is gauged by means of a 90% credibility interval.

In order to cope with dependencies between effects that resulted from studies reporting multiple correlations (e.g., obtained with different instruments or varying retest intervals), the meta-analytic model was formulated as a multilevel model where individual effects are nested within studies (cf. Cheung, 2013). This approach models the data on three hierarchical levels: (a) Level 1 refers to the individual effect sizes, (b) level 2 refers to differences between effect sizes within a sample, and (c) level 3 refers to difference between samples. Thus, the random level 2 variance $\tau^2_{(2)}$ reflects the heterogeneity of effects due to differences between Big Five measures, whereas the random level 3 variance $\tau^2_{(3)}$ indicates the heterogeneity of effect sizes across samples after controlling for the different types of measures at level 2.

In addition to the dependability coefficients for each trait, the aggregated square roots of these indices are also reported to be used in future meta-analytic artifact corrections (Hunter & Schmidt, 2001). All meta-analytical models were estimated with the *metaSEM* software (Cheung, 2014).

Results

The meta-analysis included 67 studies reporting $k = 123$ test-retest correlations for openness, $k = 136$ for conscientiousness, $k = 152$ for extraversion, $k = 107$ for agreeableness and $k = 164$ for neuroticism. The 74 independent samples included a total of $N = 14{,}923$ individuals ($Mdn = 92$, range: 17 to 5,759), predominantly from the United States (47%). About 63 percent of the participants were female. The samples' mean age fell at $M = 25.30$ ($SD = 7.26$). Most dependability coefficients were available for variants of Costa and McCrae's (1992) NEO scales (24%), Goldberg's (1999) statements from the International Personality Item Pool (11%), and various adjective lists such as Goldberg's (1992) Big Five markers (10%). The test-retest intervals varied between one and eight weeks ($M = 3.68$, $SD = 2.25$). The correlations between the focal study variables are summarized in Table 1. It is

noteworthy that the two different indicators of measurement error, dependability and coefficient alpha, were only moderately correlated, *Mdn*(*r*) = .46—similar to correlations in previous single sample studies (cf. McCrae et al., 2011). This indicates that the two reliability estimates capture, albeit related, by no means identical error components in measures of the Big Five.

Overall Dependability

For each of the five traits, the mean inverse-variance weighted test-retest correlations are reported in Table 2. As noted previously (Chmielewski & Watson, 2009; Viswesvaran & Ones, 2000; Wood, Nye, & Saucier, 2010), extraversion yielded the most dependable test scores, $\rho_{tt} = .851$, whereas agreeableness scales had the lowest dependability, $\rho_{tt} = .778$. The respective coefficients were $\rho_{tt} = .810$ for openness to experiences, $\rho_{tt} = .817$ for conscientiousness, and $\rho_{tt} = .816$ for neuroticism. To some degree these test-retest correlations were instrument-specific (see Table 3). The NEO-PI-R scales (Costa & McCrae, 1992) generally showed the highest dependability coefficients, reaching up to .918, whereas the short TIPI (Gosling et al., 2003) was less dependable with test-retest correlations falling between .664 and .807. For all traits the random variance components τ^2 were significant at $p < .05$, indicating unaccounted heterogeneity which could be attributed to the effects of one or more moderators.

Sensitivity analyses. Potential outliers were identified using the studentized deleted residual (Viechtbauer & Cheung, 2010), a standardized difference between an individual effect size and the aggregated mean effect. Using a significance level of $\alpha = .05$, these marked between 0% and 3% of all correlations as extreme (see Table 2). To study their influence on the aggregated correlations, the identified outliers were substituted with the upper or lower bounds of the 90% credibility interval obtained from a truncated data set where the extreme correlations had been excluded (cf. Gnambs. 2013). In this way, the effects of the outliers were controlled for, while keeping the same number of effect sizes as in the original analyses.

The identified outliers had a negligible effect on the aggregated correlations (see Table 2). Dependability estimates from the original and modified data set resulted in a maximum difference of $Max(\Delta\rho_{tt}) = .002$; albeit, controlling for the outliers slightly reduced the unaccounted heterogeneity.

Publication bias. The fail-safe number of missing studies with unreliable test scores that would be needed to alter the conclusions from the meta-analysis was calculated using the formulas in Howell and Shields (2008). Because reliability coefficients below .70 are frequently viewed as problematic for most assessment purposes, the number of file drawer studies needed to lower the population reliability below this threshold was determined. Following Howell and Shields (2008), the worst-case average reliabilities of unpublished studies was assumed to be .80 standard deviations below the chosen threshold. The number of file drawer studies required to lower the population reliability below .70 was estimated to be about 1.2 (agreeableness) to 3.0 (extraversion) times larger than the number of studies included in the meta-analysis (see Table 2). Thus, it seems safe to conclude that the measures of the Big Five generate reliable test scores when contend with a dependability coefficient of .70.

Moderator Analyses

For each trait a separate inverse-variance weighted, mixed-effects regression was specified to examine the effects of the selected moderators on the retest correlations. Because categorical moderators were contrast- (-1 and 1) instead of dummy-coded (0 and 1) the intercept in the regression models can be interpreted as the mean population correlation when controlling for the specified cross-study differences. Moreover, the scale length (as 10 minus number of items), coefficient alpha (as 1 minus coefficient alpha), retest interval (as deviation from 4), and publication year (as 2013 minus year) were recoded in such a way that the intercept reflects the retest correlation in the year 2013 at a retest interval of four weeks for instruments containing ten items and no random error.

The results of the five regressions are summarized in Table 4. Overall, inclusion of the moderators reduced the random level 2 variances that reflect instrument-specific differences by 49% (extraversion) to 87% (conscientiousness). Between-sample heterogeneity as quantified by the random level 3 variance was reduced up to 53%. The test-retest correlations corrected for the included moderators were about $M(\Delta\rho_{tt}) = .095$ ($SD_{\Delta\rho} = .019$, $Max = .116$) larger than the uncorrected correlations and averaged at $Mdn(\rho_{tt}) = .904$; thus, about 10% of the observed score variance in the Big Five can be attributed to transient error. These differences can be explained by three main effects:

Random response error. Random error was quantified twofold, as part of coefficient alpha and also indirectly in form of the scales' lengths. Instruments with higher coefficient alpha consistently resulted in significantly, $p < .05$, larger retest correlations for all five traits. Thus, random error variance partly attenuates retest correlations as indicators of transient error. The number of items yielded no significant effects. Although both variables were only moderately correlated, $Mdn(r) = .48$, coefficient alpha captured most of the random error component in test scores.

Test-retest interval. Despite the short retest intervals of the included studies (eight weeks at the most), studies with longer retest intervals resulted in significantly, $p < .05$, lower dependability coefficients for three of the five traits: openness, extraversion, and neuroticism. The respective effects were rather small, an increase of one week corresponded to a decrease in transient error of about $\Delta r = .006$ to .008 (see Table 4). This effect was not conditional on the included covariates (cf. Spector & Brannick, 2011) but also reproduced when the other moderators were excluded, $B = -.007$, $SE = .003$, $p = .03$ for openness, $B = -.006$, $SE = .002$, $p = .02$ for extraversion and $B = -.006$, $SE = .003$, $p = .04$ for neuroticism.

The mean test-retest correlations for the three traits at two week intervals are plotted in Figure 1. For comparison, the respective correlations for three to six months retest intervals from the studies discarded during the literature search and the respective values from

Viswesvaran and Ones (2000, Table 1) are included as well. The test-retest correlations show a continuous decline over the plotted intervals. The decline seems somewhat stronger during the first two to four weeks and slows down afterwards. In contrast, the long-term retest correlations reported in Viswesvaran and Ones (2000) assign variance associated with true trait changes to measurement error and, as a consequence, are markedly lower than the respective short-term correlations from the present meta-analysis.

Further moderators. Studies conducted outside the United States that frequently administered instruments adapted from another language resulted in significantly, $p < .05$, lower dependability coefficients. Thus, it might be speculated that the adaption of personality instruments to other languages impaired their measurement precision to some degree. Instruments that were constructed according to the Big Five taxonomy did not, $p > .05$, result in higher dependability coefficients than instruments developed within another theoretical framework. Thus, the impaired criterion validities for non-Big Five measures reported previously (Salgado, 2003) do not seem to be a consequence of larger transient error. The publication year yielded a significant effect for only one trait; dependability was slightly lower for extraversion scores in the 1990ies as compared to recent years. Neither the publication type nor the use of student samples showed significant effects.

Discussion

Most measures in psychological research are biased to some degree by less than perfect reliability (cf. McCrae et al., 2011; Schmidt et al., 2003; Watson, 2004). As a consequence, studies on personality development require information on the dependability of their measures to derive undistorted estimates of longitudinal relationships. Unfortunately, it is frequently not possible to obtain dependability information for a given sample. In these cases, researchers depend on precise meta-analytical estimates for the construct in question. The present study reported the respective dependability coefficients for the five broad dimensions of personality across different measures of the Big Five. On average, about ten

percent of the observed scores' variances could be attributed to occasion-specific variations in, for example, transient moods or feelings. Although transient error for measures of the Big Five is not negligible, it is not as serious as previous research has suggested. Compared to the dependability coefficients reported in Viswesvaran and Ones (2000, Table 1) that did not take into account the length of the test-retest interval, the dependabilities presented here are about $\Delta\rho_{tt} = .06$ higher on average. The size of the dependability estimates exhibited some variations within the five factor space and also between different instruments. For example, scores from the extraversion and the NEO-PI-R scales (Costa & McCrae, 1992) tended to exhibit slightly higher test-retest correlations, whereas the agreeableness and TIPI scales (Gosling et al., 2003) showed somewhat less dependable scores. In addition, minor differences in the chosen retest interval affected the dependability coefficients to some degree. What might account for these differences?

Scale-Specific Effects on Dependability

Some differences between instruments can be explained by variations in random error. Because the number of items per scale influences the degree of random error (Schmidt et al., 2003), longer instruments such as the NEO-PI-R (Costa & McCrae, 1992) that includes 48 items per scale tend to exhibit larger retest correlations than instruments with fewer items such as the 2 item scales of the TIPI (Gosling et al., 2003). In addition, scale-specific features (e.g., the choice of specific item formats or item wordings) might have contributed to the observed differences. Reliabilities are known to be affected by, for example, varying numbers of response options (Lozano, García-Cueto, & Muñiz, 2008) and reversed items (Swain, Weathers, & Niedrich, 2008), the degree of social desirability in items (Kuncel & Tellegen, 2009), and even differences in item contexts (Rivers, Meade, & Fuller, 2009). Recently, linguistic analyses of self-report scales also highlighted that various forms of miscomprehension attenuate observed test scores (Hardy & Ford, 2014). A significant proportion of respondents interprets items and response instructions differently and, thus,

produces error variance. If linguistic properties vary considerably between instruments (see, for example, Mõttus, Pullman, & Alik, 2006) these differences might contribute to the identified differences in dependability coefficients between scales.

Trait-Specific Effects on Dependability

Differences in retest correlations between the five traits of personality have been observed repeatedly (e.g., Anusic et al., 2012; Chmielewski & Watson, 2009; Viswesvaran & Ones, 1999; Wood et al., 2010): Typically, extraversion scales yield the most dependable and agreeableness scales the least dependable scores. Linguistic differences in trait scales are unlikely to account for these differences. There are rather modest differences in item comprehension, for example, between the five scales of the BFI (Soto, John, Gosling, & Potter, 2008) or the NEO-PI-R (De Fruyt, Mervielde, Hoekstra, & Rolland, 2000). In contrast, differences in trait contents might render more promising explanations. Generally, people rate observable behaviors more consistently than emotional reactions or mental states (Johnson, 2004). For example, extraversion can be readily inferred from thin slices of behavior (Canrey, Colvin, & Hall, 2007) and, consequently, also exhibits high-levels of agreement across self and peer reports (Gnambs, 2013). In contrast, the others traits in the five-factor space are less clearly manifested in observable behaviors (Simms, Zelazny, Yam, & Gros, 2010). Therefore, some traits might be easier to infer than others and, thus, result in more dependable scores. Furthermore, the lower retest correlation of agreeableness could be a consequence of its interpersonal nature (Jensen-Campbell & Graziano, 2001) and its susceptibility to situational influences. Experience sampling studies revealed that individuals report increased agreeableness in the presence of unfriendly people and more disagreeableness the friendlier others are (Fleeson, 2007). Thus, the lower retest correlations might be a consequence of situational differences that affect agreeableness ratings stronger than other trait ratings.

Length of Test-Retest Interval

Although this study cannot provide a definitive cut-off for appropriate retest intervals in dependability studies because the present meta-analysis was limited to studies with two month retest periods, it seems clear that retest intervals of more than a year that have been included in previous meta-analyses (Caruso, 2000, Viswesvaran & Ones, 2000) are not appropriate to quantify transient error in Big Five scores. The one year retest correlation for neuroticism (Fraley & Roberts, 2005), for example, is about $\Delta\rho_{tt} = .15$ lower than the actual dependability coefficient found in the present study. In line with prevalent recommendations and contemporary practice (e.g., Anusic et al., 2012; Cattell, 1986; Chmielewski & Watson, 2009) retest intervals of about two months revealed to be more appropriate. However, even within this interval different retest periods still influence the dependability estimates for three of the five traits (openness, extraversion, and neuroticism). Studies with shorter retest intervals of about one to two weeks had higher retest correlations as compared to studies that adopted intervals closer to eight weeks (see Figure 1). Although the size of these effects might be considered small—that is, increasing the retest interval by one week decreased transient error in trait scores by less than one percent—authors of dependability studies should be aware that extremely short retest intervals might overestimate the dependability of trait scores to some degree. Memory and mood effects might serve as potential explanations for these differences.

Memory effects. Larger dependability coefficients might arise when participants simply recall previous answers from memory without properly rereading the items (Cronbach & Furby, 1970). Thus, if memory effects present themselves they are more likely for extremely short retest intervals whereas they should gradually diminish over time. Although in the present study memory effects might have contributed to a *general* decline in retest correlations across the two months period, they are unlikely to explain the *differential* effects for the five traits. It is conceivable that memory effects are more pronounced for easy to recall items as compared to more complicated items. However, differences in item comprehension

between the five traits are typically rather modest (De Fruyt et al., 2000; Soto et al., 2008). Moreover, McCrae et al. (2011) found no relationship between various indicators of item ambiguity and one-week dependability coefficients for the facets of the NEO-PI-R. Thus, memory effects seem to be less likely explanations for the differential decline in dependability across different retest intervals.

Mood effects. The decline in retest correlations was strongest for traits with a pronounced affective component. Negative affect is a central component of neuroticism (McCrae & Costa, 1987) whereas extraversion has been attributed with a core of positive affect (Hermes, Hagemann, Naumann, & Walter, 2011; Watson & Clark, 1997). Indeed, analyses that divided that item content of various Big Five instruments into behavioral, cognitive and affective components found that about 70% of the item content in neuroticism scales and 38% in extraversion scales are affective in nature (Pytlik Zillig, Hemenover, & Dienstbier, 2002). Even the openness scale of the NEO-PI-R (Costa & McCrae, 1992) that provided the majority of retest correlations for this meta-analysis is heavily loaded with affective content (35%). In contrast, conscientiousness and agreeableness scales are significantly less affect-laden (6% to 26%). Similarly, responding to items indicating extraversion, openness and low neuroticism (but also conscientiousness) evoke more positive feelings than agreeableness items (Johnson, 2006). Thus, for test-retest correlations to properly reflect the transient error of these scales, it is necessary not only to adopt retest intervals that are short enough to preclude true trait changes but, at the same time, also long enough to separate momentary from dispositional affective components. Retest intervals of only one to two weeks seem to be too short to achieve the latter objective (see Figure 1).

Limitations and Recommendations for Future Studies

There are some limitations of this study that should be addressed in future research. The meta-analysis identified significant between-sample heterogeneity in the aggregated dependability coefficients. Thus, characteristics of the studied sample affected dependability

coefficients beyond the moderators included in the analyses. For example, differences in respondents' literacy and educational attainment are known to affect the factor structure of Big Five instruments (cf. Rammstedt, Goldberg, & Borg, 2010; Sutin, Costa, Evans, & Zonderman, 2013). Whether they also distort other psychometric properties of measurement instruments is of yet not well known. In particular, authors are encouraged to explore the stability of dependability coefficients across the life course. Whereas internal consistency seems to be age invariant (Fraley & Roberts, 2005), it is not known whether dependability remains stable from childhood to old age. Unfortunately, the presented results are limited to self-reports of personality. To date, there are few studies investigating the transient error of peer reports. Preliminary evidence for the NEO-PI-R (Costa & McCrae, 1992) suggests that transient error—assessed over a test-retest interval of six months—is not considerably different for observer reports of the five traits (Kurtz, Lee, & Sherker, 1999). Similarly, long-term stabilities across two years are nearly identical for self- and other-reports of personality(Roberts & DelVecchio, 2000; Watson & Humrichouse, 2006). However, future studies should address this aspect in more detail and also ascertain the transient error in observer reports for different instruments. For the time being, there are no compelling reasons why the presented dependability coefficients should not be used as substitutes for peer reports of the Big Five as well.

Conclusion

In response to recent pleas for a stronger focus on dependability in personality research (McCrae et al., 2011; Schmidt et al. 2003; Watson, 2004) this meta-analysis reported estimates of dependability coefficients for the Big Five of personality. This information should represent an important resource for authors examining longitudinal relationships of the Big Five. Moreover, the analyses also highlighted the importance of carefully choosing an appropriate retest interval in dependability studies that are able to separate transient error not

only from developmental changes but also from situational carry-over effects. Based on the presented results, retest intervals of about four weeks are recommended.

References

Anusic, I., Lucas, R. E., & Donnellan, M. B. (2012). Dependability of personality, life satisfaction, and affect in short-term longitudinal data. *Journal of Personality, 80*, 33-58. doi:10.1111/j.1467-6494.2011.00714.x

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5*, 370-379. doi:10.1037/1082-989X.5.3.370

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*, 317-335. doi:10.1111/j.1468-2389.2006.00354.x

Carney, D. A., Colvin, C. R., & Hall, J. A. (2007).A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*, 1054-1072. doi:10.1016/j.jrp.2007.01.004

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement, 60*, 236-254. doi:10.1177/00131640021970484

Cattell, R. B. (1986). The psychometric properties of tests: Consistency, validity, and efficienc. In R. B. Cattell & R. C. Johnson (Eds.), *Functional psychological testing* (pp. 54-78). New York, NY: Brunner/Mazel.

Catell, R. B., Eber, H. W., & Tatsuoka, M. M. (1976). *Handbook for the Sixteen Personality Factors Questionnaire (16PF)*. Champaign, IL: Institute for Personality and Ability Testing.

Cheung, M. W.-L. (2013). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*. Advance online publication. doi:10.1037/a0032968

Cheung, M. W.-L. (2014). Fixed- and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavior Research Methods, 46*, 29-40. doi:10.3758/s13428-013-0361-y

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology, 97*, 186-202. doi:10.1037/a0015618

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R: Professional manual*. Odessa, TX: Psychological Assessment Resources.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"- or should we? *Psychological Bulletin, 74*, 68-80. doi:10.1037/h0029382

De Fruyt, F., Mervielde, I., Hoekstra, H. A., & Rolland, J. P. (2000). Assessing adolescents' personality with the NEO PI-R. *Assessment*, *7*, 329-345. doi:10.1177/107319110000700403

Fife, D. A., Mendoza, J. L., & Terry, R. (2012). The assessment of reliability under range restriction. *Educational and Psychological Measurement, 72*, 862-888. doi:10.1177/0013164411430225 19

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality, 75*, 825-862. doi:10.1111/j.1467-6494.2007.00458.x

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review, 112*, 60-74. doi:10.1037/0033-295X.112.1.60

Gnambs, T. (2013). The elusive general factor of personality: The acquaintance effect. *European Journal of Personality, 27*, 507-520. doi:10.1002/per.1933

Gnambs, T., & Batinic, B. (2011). Evaluation of measurement precision with Rasch-type models. *Personality and Individual Differences, 50*, 53-58. doi:10.1016/j.paid.2010.08.021

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141.165). Beverly Hills, CA: Sage.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26-42. doi:10.1037/1040-3590.4.1.26

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504-528. doi:10.1016/S0092-6566(03)00046-1

Hardy, B., & Ford, L. R. (2014). It's not me, it's you: Miscomprehension in surveys. *Organizational Research Methods, 17*, 138-162. doi:1094428113520185.

Hedges, L. V. (1989). An unbiased correction for sampling error in validity generalization studies. *Journal of Applied Psychology, 74*, 469-477. doi:10.1037/0021-9010.74.3.469

Hermes, M., Hagemann, D., Naumann, E., & Walter, C. (2011). Extraversion and its positive emotional core - Further evidence from neuroscience. *Emotion, 11*, 367-378. doi:10.1037/a0021550

Hopwood, C. J., Donnellan, M. B., Blonigen, D. M., Krueger, R. F., McGue, M., Iacono, W. G., & Burt, S. A. (2011). Genetic and environmental influences on personality trait stability and growth during the transition to adulthood: A three-wave longitudinal

study. *Journal of Personality and Social Psychology, 100*, 545-556. doi:10.1037/a0022409

Hough, L. M., & Ones, D. S. (2001).The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology* (Vol. 1: Personnel psychology, pp. 233-277). London, England: Sage.

Howell, R. T., & Shields, A. L. (2008). The file drawer problem in reliability generalization: A strategy to compute a fail-safe N with reliability coefficients. *Educational and Psychological Measurement*, *68*, 120-128. doi:10.1177/0013164407301528

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis*. Thousand Oaks, CA: Sage.

Jensen-Campbell, L. A., & Graziano, W. G. (2001). Agreeableness as a moderator of interpersonal conflict. *Journal of Personality, 69*, 323-361. doi:10.1111/1467-6494.00148

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.

Johnson, J. A. (2004). The impact of item characteristics on item and scale validity. *Multivariate Behavioral Research, 39*, 273-302. doi:10.1207/s15327906mbr3902_6

Johnson, J. A. (2006). Ego-syntonicity in responses to items in the California Psychological Inventory. *Journal of Research in Personality, 40*, 73-83. doi:10.1016/j.jrp.2005.08.008

Klimstra, T. A., Hale, W. W., Raiijmakers, Q. A., Brjanje, S. J., & Meeus, W. H. (2009). Maturation of personality in adolescence. *Journal of Personality and Social Psychology, 96*, 898-912. doi:10.1037/a0014746

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology, 62*, 201-228. doi:10.1111/j.1744-6570.2009.01136.x

Kurtz, J. E., Lee, P. A., & Sherker, J. L. (1999). Internal and temporal reliability estimates for informant ratings of personality using the NEO PI-R and IAS. *Assessment, 6*, 103-113. doi:10.1177/107319119900600201

Leue, A., & Lange, S. (2011). Reliability generalization: An examination of the Positive Affect and Negative Affect Schedule. *Assessment, 18*, 487-501. doi:10.1177/1073191110374917

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*, 73-79. doi:10.1027/1614-2241.4.2.73

Lucas, R. E., & Donnellan, M. B. (2011). Personality development across the life span: Longitudinal analyses with a national sample from Germany. *Journal of Personality and Social Psychology, 101*, 847-861. doi:10.1037/a0024298

Mayer, J. D., Gaschke, Y. N., Braverman, D. L., & Evans, T. W. (1992). Mood-congruent judgment is a general effect. *Journal of Personality and Social Psychology, 63*, 119-132. doi:10.1037/0022-3514.63.1.119

McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of a five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81-90. doi:10.1037/0022-3514.52.1.81

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*, 28-50. doi:10.1177/1088868310366253

Mõttus, R., Johnson, W., & Deary, I. J. (2012). Personality traits in old age: measurement and rank-order stability and some mean-level change. *Psychology and Aging, ;27*, 243-249. doi:10.1037/a0023690

Mõttus, R., Pullmann, H., & Allik, J. (2006). Toward more readable Big Five personality inventories. *European Journal of Psychological Assessment, 22*, 149-157. doi:10.1027/1015-5759.22.3.149

Pytlik Zillig, L. M., Hemenover, S. H., & Dienstbier, R. A.(2002). What do we assess when we assess a Big 5 trait? A content analysis of the affective, behavioral, and cognitive processes represented in Big 5 personality inventories. *Personality and Social Psychology Bulletin, 28*, 847-858. doi:10.1177/0146167202289013

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of Big-Five factor markers for persons with different levels of education. *Journal of Research in Personality*, *44*, 53-61. doi:10.1016/j.jrp.2009.10.005

Rivers D. C., Meade A. W., & Fuller, L. W. (2009). Examining question and context effects in organization survey data using item response theory. *Organizational Research Methods, 12*, 529-553. doi:10.1177/1094428108315864

Roberts, B. W., Caspi, A., & Moffitt, T. E. (2001). The kids are alright: Growth and stability in personality development from adolescence to adulthood. *Journal of Personality and Social Psychology, 81*, 670-683. doi:10.1037/0022-3514.81.4.670

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*, 3-25. doi:10.1037/0033-2909.126.1.3

Robins, R. W., Fraley, R. C., Roberts, B. W., & Trzesniewski, K. H. (2001). A longitudinal study of personality change in young adulthood. *Journal of Personality, 69*, 617-640. doi:10.1111/1467-6494.694157

Salgado, S. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology*, *76*, 323-346. doi:10.1348/096317903769647201

Saville, P., Holdsworth, R., Nyfiled, G., Cramp, L., & Mabey, W. (1996). *Occupational Personality Questionnaire: Manual and user's guide*. Boston, MA: SHL.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*, 206-224. doi:10.1037/1082-989X.8.2.206

Schuerger, J. M., Zarrella, K. L., & Hotz, A. S. (1989). Factors that influence the temporal stability of personality by questionnaire. *Journal of Personality and Social Psychology, 56*, 777-783. doi:10.1037/0022-3514.56.5.777

Sedikides, C. (1994). Incongruent effects of sad mood on self- conception valence: It's a matter of time. *European Journal of Social Psychology, 24*, 161-172. doi:10.1002/ejsp.2420240112

Simms, L. J., Zelazny, K., Yam, W. H., Gros, D. F. (2010). Self-informant agreement for personality and evaluative person descriptors: Comparing methods for creating informant measures. *Journal of Personality, 24*, 207-221. doi:10.1002/per.763

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology*, *94*, 718-737. doi:10.1037/0022-3514.94.4.718

Spector, P. E., & Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods, 14*, 287-305. doi:10.1177/1094428110369842

Steyer, R., Schmitt, M., & Eid, M. (1999), Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13*, 389-408. doi: 10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A

Sutin, A. R., Costa Jr, P. T., Evans, M. K., & Zonderman, A. B. (2013). Personality assessment in a diverse urban sample. *Psychological Assessment*, *25*, 1007-1012. doi:10.1037/a0032396

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*, 116-131. doi:10.1509/jmkr.45.1.116

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6-20. doi:10.1177/0013164498058001002

Vaidya, J. G., Gray, E. K., Haig, J., Mroczek, D. K., & Watson, D. (2008). Differential stability and individual growth trajectories of Big Five and affective traits during young adulthood. *Journal of Personality, 76*, 267-303. doi:10.1111/j.1467-6494.2007.00486.x

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*, 112-125. doi:10.1002/jrsm.11

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in “Big Five Factors” personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement*, *60*, 224-235. doi:10.1177/00131640021970475

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality, 38,* 319-350. doi:10.1016/j.jrp.2004.03.001

Watson, D., & Clark, L. A. (1997). Extraversion and its positive emotional core. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 767-793). San Diego, CA: Academic Press.

Watson, D., & Humrichouse, J. (2006). Personality development in emerging adulthood: Integrating evidence from self-ratings and spouse ratings. *Journal of Personality and Social Psychology, 91*, 959 –974. doi:10.1037/0022-3514.91.5.959

Wood, D., Nye, C., & Saucier, G. (2010). Identification and measurement of a more comprehensive set of person-descriptive trait markers from the English lexicon. *Journal of Research in Personality, 44*, 257-272. doi:10.1016/j.jrp.2010.02.003

Wortman, J., Lucas, R. E., & Donnellan, M. B. (2012). Stability and change in the Big Five personality domains: Evidence from a longitudinal study of Australians. *Psychology and Aging, 27*, 867-874. doi:10.1037/a0029322

Articles included in the meta-analysis are listed in the online supplement.

Table 1.

Descriptive Statistics and Correlations between Study Variables

		Mdn / %	Correlations						
			1.	2.	3.	4.	6.	7.	8.
	Within-sample variables [a]								
1.	Dependability coefficient	.82							
2.	Coefficient alpha	.77	.46*						
3.	Number of items	8	.41*	.48*					
4.	Theoretical model								
	non-Big Five = -1	65%	.07	.25*	.07				
	Big Five = 1	35%							
5.	Test-retest interval	4	-.09	-.04	-.26*	.19*			
	Between-sample variables								
6.	Publication year	2005							
7.	Publication type								
	Research article = -1	86%					-.10		
	Test manual = 1	14%							
8.	Geographical region								
	United States = -1	47%					-.22	.19	
	other countries = 1	53%							
9.	Student sample								
	no = -1	30%					-.23	-.06	.13
	yes = 1	70%							

Note. [a] Median across all five traits. * $p < .05$

Table 2.

Meta-Analysis of Dependability Coefficients for Measures of the Big Five

				Interval	Aggregated effects							
	k_1	k_2	N	M_i (SD_i)	r_{tt}	SD_r	ρ_{tt}	90% CRI	$\tau_{(2)}$	$\tau_{(3)}$	$I^2_{(2)}$	$I^2_{(3)}$
Openness	123	51	12,773	4.07 (2.28)	.802	.077	.810*	[.702, .918]	.044*	.048*	.398	.471
Conscientiousness	136	56	13,510	3.86 (2.27)	.815	.067	.817*	[.719, .916]	.040*	.045*	.379	.479
Extraversion	152	72	14,923	3.69 (2.30)	.834	.081	.851*	[.750, .953]	.027*	.056*	.167	.734
Agreeableness	107	63	14,277	3.90 (2.28)	.766	.097	.778*	[.649, .908]	.046*	.064*	.308	.588
Neuroticism	164	68	14,708	3.76 (2.27)	.802	.089	.816*	[.697, .936]	.030*	.066*	.155	.743

Note. k_1 = Number of effect sizes; k_2 = Number of independent samples; k_o = Number of identified outliers (based on $\alpha = .01$); N = Total sample size; M_i = Mean time interval between measurement occasions (in weeks); r_{tt} = Unweighted dependability coefficient; ρ_{tt} = Weighted dependability coefficient; τ = Random level 2 and level 3 *SD* of ρ_{tt}; I^2 = Proportion of total variance in ρ_{tt} due to level 2 or level 3 between-study heterogeneity (Cheung, 2013); CRI = 90% credibility interval; $\sqrt{\rho_{tt}}$ = Mean square root of test-retest correlations; $\rho_{tt.o}$ = True dependability coefficient with outliers truncated to the bounds of the 90% CRI with all outliers excluded (Gnambs, 2013); $\rho_{.70}$ = Reliability of file drawer studies estimated as .80 *SD*ρ below the threshold of .70 (Howell & Shields, 2008); $N_{.70}$ = Fail-Safe *N* for a threshold of .70

* $p < .05$

Table 2. (continued)

	Sensitivity analysis			Artifact distribution		Fail-safe *N*	
	k_o	$\rho_{tt.o}$	90% CRI_o	$\sqrt{\rho_{tt}}$	$SD_{\sqrt{\rho_{tt}}}$	$\rho_{.70}$	$N_{.70}$
Openness	0	.810*	[.702, .918]	.907	.027	.648	258
Conscientiousness	1	.818*	[.719, .916]	.910	.026	.652	332
Extraversion	3	.853*	[.758, .948]	.929	.024	.651	466
Agreeableness	3	.780*	[.657, .904]	.889	.036	.637	133
Neuroticism	0	.816*	[.697, .936]	.909	.034	.642	329

Table 3.

Meta-Analysis of Dependability Coefficients by Instrument Type

	TDA	TIPI	BFI	IPIP	NEO-PI-R
k	7	8	14	21	10
N	1,140	1,859	1,813	6,668	738
Items	10 [a]	2	8-10 [a b]	10 [a]	48
			Openness		
ρ_{tt}	.848	.725	.856	.778	.885
90% CRI	[.818, .878]	[.625, .825]	[.772, .941]	[.730, .823]	[.844, .926]
			Conscientiousness		
ρ_{tt}	.821	.730	.831	.798	.916
90% CRI	[.760, .882]	[.679, .781]	[.770, .892]	[.725, .871]	[.882, .950]
			Extraversion		
ρ_{tt}	.868	.807	.876	.859	.918
90% CRI	[.847, .888]	[.765, .849]	[.787, .965]	[.808, .910]	[.918, .918]
			Agreeableness		
ρ_{tt}	.736	.664	.818	.736	.878
90% CRI	[.674, .798]	[.543, .785]	[.748, .888]	[.690, .783]	[.825, .932]
			Neuroticism		
ρ_{tt}	.802	.753	.832	.799	.914
90% CRI	[.769, .835]	[.691, .815]	[.778, .886]	[.736, .862]	[.878, .950]

Note. TDA = Trait descriptive adjectives based on Goldberg (1992); TIPI = Ten Item Personality Inventory (Gosling et al., 2003); BFI = Big Five Inventory (John et al., 2008); IPIP = International Personality Item Pool (Goldberg et al., 1999); NEO-PI-R = NEO Personality Inventory - Revised (Costa & McCrae, 1992); *k* = Number of dependability coefficients; *N* = Total sample size; ρ_{tt} = True dependability coefficient; CRI = 90% credibility interval

[a] Scales were corrected to a common number of items by including the scale length as moderator in the model (for further details see section on moderator analyses).

[b] Scales were corrected to a length of 8 items for extraversion and neuroticism, 9 items for conscientiousness and agreeableness, and 10 items for openness (cf. John et al., 2008).

Table 4.

Meta-Regression Analyses for Dependability Coefficients

	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	γ	(*SE*)	γ	(*SE*)	γ	(*SE*)	γ	(*SE*)	γ	(*SE*)
Intercept	.924*	(.024)	.904*	(.022)	.935*	(.020)	.894*	(.031)	.891*	(.025)
Publication year (as deviation from year 2013)	.000	(.002)	.000	(.001)	.002*	(.001)	.001	(.002)	.001	(.001)
Publication type (Research article = -1, Test manual = 1)	.006	(.014)	.016	(.012)	-.003	(.009)	.021	(.014)	.017	(.014)
Geographical region (United States = -1, other countries = 1)	-.026*	(.009)	-.024*	(.008)	-.026*	(.006)	-.025*	(.010)	-.030*	(.009)
Student sample (no = -1, yes = 1)	-.020	(.011)	.004	(.008)	-.001	(.007)	-.013	(.010)	.005	(.009)
Theoretical model (-1 = non-Big Five,1 = Big Five)	.006	(.010)	-.003	(.008)	-.004	(.006)	.006	(.009)	.001	(.007)
Number of items (as difference to 10)	.000	(.001)	-.001	(.001)	.000	(.000)	-.001	(.001)	-.001	(.000)
1 - Coefficient alpha	-.366*	(.043)	-.326*	(.052)	-.313*	(.061)	-.333*	(.070)	-.256*	(.055)
Test-retest interval (as deviation from 4 weeks)	-.008*	(.004)	-.003	(.003)	-.006*	(.003)	-.004	(.004)	-.007*	(.003)

Table 4. (continued)

	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	γ	(*SE*)	γ	(*SE*)	γ	(*SE*)	γ	(*SE*)	γ	(*SE*)
$\tau_{(2)}$ / $\tau_{(3)}$	.018 / .042*		.016 / .040*		.021* / .037*		.027* / .051*		.021* / .052*	
$R^2_{(2)}$ / $R^2_{(3)}$	.85 / .08		.87 / .06		.49 / .53		.74 / .29		.61 / .38	
k_1 / k_2	109 / 46		122 / 50		135 / 65		91 / 56		150 / 63	

Notes. k_1 = Number of effect sizes; k_2 = Number of independent samples; τ = Random level 2 and level 3 *SD*. R^2 = Explained variance at level 2 and 3 (Cheung, 2013). Due to missing values on some moderators the number of included effect sizes differs from those presented in Table 2.

* $p < .05$

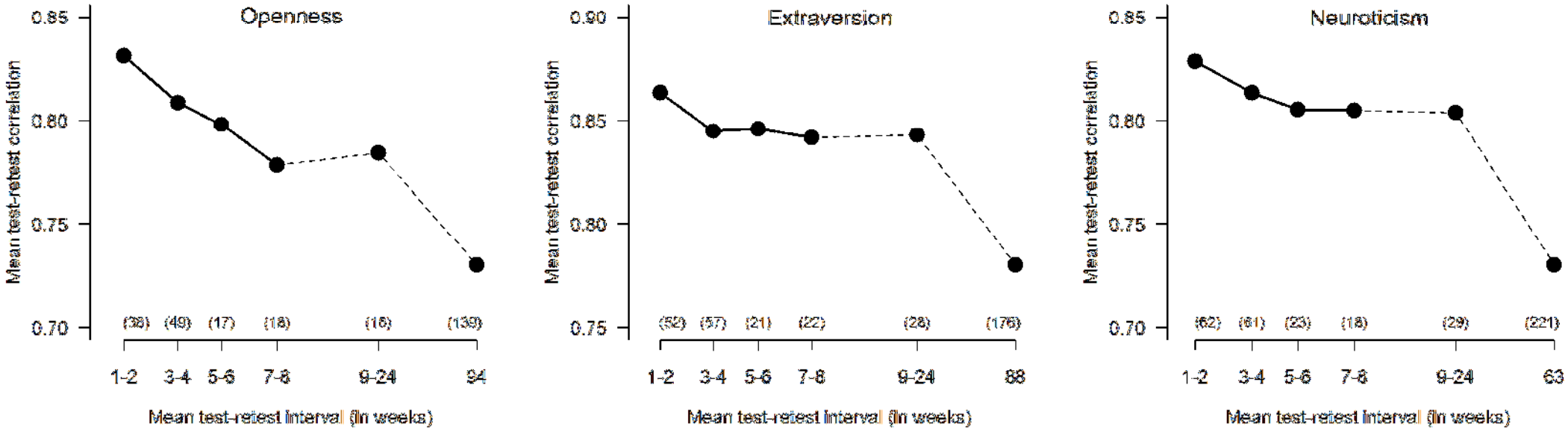

Figure 1. Aggregated dependability coefficients ρ_{tt} at different retest intervals; long-term retest correlations are from Viswesvaran and Ones (2000, Table 1); number of included correlations are in parentheses

Running Head: META-ANALYSIS OF DEPENDABILITY

Online Supplement for

"A Meta-Analysis of Dependability Coefficients for Measures of the Big Five"

Timo Gnambs

Supplemental Tables

Table S1.

Instruments included in the meta-analysis of dependability coefficients

Short	Instrument	Source
ABLE	Assessment of Background and Life Experiences	(Peterson, Hough, Dunnette, Rosse, Houston, Toquam, & Wing, 1990)
BFI	Big Five Inventory	(John, Naumann, & Soto, 2008)
BFQ	Big Five Questionnaire	(Caprara, Barbaranelli, Borgogni, & Perugini, 1993)
BIP	Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung	(Hossiep, Paschen, & Mühlhaus, 2003)
BPI	Basic Personality Inventory	(Jackson, 1997a)
COPAS	Comprehensive personality and affect scales	(Lubin, & Van Whitlock, 2002)
DiSC	DiSC Classic	(Inscape, 2005)
EPQ	Eysenck Personality Questionnaire	(Eysenck, & Eysenck, 1975)
FFMQ	Five-Factor Model Questionnaire	Gill, & Hodgkinson (2007)
FFPI	Five-Factor Personality Inventory	Hendriks, Hofstee, & De Raad (1999)
GPAC	Greek Personality Adjective Checklist	(Tsaousis, & Georgiades, 2009)
HPI	Hogan Personality Inventory	(Hogan & Hogan, 1992)
ICES	ICES Personality Inventory	(Bartram, Lindley, & Coine, 2000)
IPIP	International Personality Item Pool	(Goldberg, 1999)
MMPI	Minnesota Multiphasic Personality Inventory	(Hathaway, & McKinley, 1951)
MMPI-2	Minnesota Multiphasic Personality Inventory - 2	(Hathaway, & McKinley, 1989)
NEO-FFI	NEO Five Factor Inventory	(Costa & McCrae, 1992)
NEO-PI-R	NEO Personality Inventory – Revised	(Costa & McCrae, 1992)
OPQ32	Occupational Personality Questionnaire	(Saville, Holdsworth, Nyfield, Cramp, & Mabey, 1996)
PI	Predictive Index	(Harris, Tracey, & Fisher, 2011)
PM	Profile Match	(Trickey, & Hyde, 2009)
PRF	Personality Research Form	(Jackson, 1997b)
PXT	ProfileXT	(Profiles International, 2010)
TDA	Trait Descriptive Adjectives / Big Five Markers	(Goldberg, 1992)
TIPI	Ten Item Personality Inventory	(Gosling, Rentfrow, & Swann, 2003)
TPQUe	Traits Personality Questionnaire	(Tsaousis, 1999)
ZKPQ	Zuckerman-Kuhlman Personality Questionnaire	(Zuckerman, Kuhlman, Joireman, Teta, & Kraft, 1993)

Supplemental References

Instruments included in the meta-analysis of dependability coefficients

Bartram D., Lindley, P., & Coine, I. J. (2000). *Technical Manual for the ICES Plus*. Vancouver, Canada: ICES Assessment Systems.

Caprara, G. V., Barbaranelli, C. Borgogni, L., & Perugini, M. (1993). The "Big Five Questionnaire": A new questionnaire to assess the five factor model. *Personality and Individual Differences, 15*, 281-288. doi:10.1016/0191-8869(93)90218-R

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R: Professional manual*. Odessa, TX: Psychological Assessment Resources.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. London, England: Hodder and Stoughton.

Gill, C. M., & Hodgkinson, G. P. (2007). Development and validation of the Five-Factor Model Questionnaire (FFMQ): An adjectival-based personality inventory for use in occupational settings. *Personnel Psychology, 60*, 731-766. doi:10.1111/j.1744-6570.2007.00090.x

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26–42. doi:10.1037/1040-3590.4.1.26

Goldberg, L. R. (1999). A broad-bandwith, public-domain, personality inventory measuring the lower-level facets of several Five-Factor models. In I. Mervielde, I. J. Deary, F. de Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, Netherlands: Tilburg University Press.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504–528. doi:10.1016/S0092-6566(03)00046-1

Harris, T. C., Tracey, A. J., & Fisher, G. G. (2011). *Predictive Index Technical Overview*. Wellesley Hills, MA: PI Worldwide.

Hathaway, S. R., & McKinley, J. C. (1951*). The Minnesota Multiphasic Personality Inventory: Manual*. New York, NY: University of Minnesota Press.

Hathaway, S. R., & McKinley, J. C. (1989). *The Minnesota Multiphasic Personality Inventory – 2: Manual*. New York, NY: University of Minnesota Press.

Hendriks, A. A. J, Hofstee, W. K. B., & De Raad, B. (1999). The Five-Factor Personality Inventory (FFPI). *Personality and Individual Differences, 27*, 307-325. doi:10.1016/S0191-8869(98)00245-1

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory: Manual*. Tulsa, TX: Hogan Assessment Systems.

Hossiep, R., Paschen, M., & Mühlhaus, O. (2003). *Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung* [Business-Focused Inventory of Personality]. Göttingen, Germany: Hogrefe.

Inscape. (2005). *DiSC validation* (Research report). Inscape Publishing.

Jackson, D. N. (1997a). *Manual for the Basic Personality Inventory Revised*. Port Huron, MI: Sigma Assessment Systems.

Jackson, D. N. (1997b). *Manual for the Personality Research Form*. Port Huron, MI: Sigma Assessment Systems.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.

Lubin, B., & Van Whitlock, R. (2002). Development of a measure that integrates positive and negative affect and personality: The comprehensive personality and affect scales. *Journal of Clinical Psychology, 58*, 1135–1156. doi:10.1002/jclp.10042

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. L., Houston, J. S., Toquam, J. L., & Wing, H. (1990). Project A: Specification of the predictor domain and development of new selection / classification tests. *Personnel Psychology, 43*, 247-276. doi:10.1111/j.1744-6570.1990.tb01558.x

Profiles International (2010). *ProfileXT: Technical manual*. Waco, TX: Profiles International.

Saville, P., Holdsworth, R., Nyfiled, G., Cramp, L., & Mabey, W. (1996). *Occupational Personality Questionnaire: Manual and user's guide*. Boston, MA: SHL.

Trickey, G., & Hyde, G. (2009). *Profile:Match Manual*. Turnbridge Wells, England: Psychological Consultancy.

Tsaousis, I. (1999). The traits personality questionnaire (TPQue): A Greek measure for the five factor model. *Personality and Individual Differences, 26*, 271-283. doi:10.1016/S0191-8869(98)00131-7

Tsaousis, I., & Georgiades, S. (2009). Development and psychometric properties of the Greek Personality Adjective Checklist (GPAC). *European Journal of Psychological Assessment, 25*, 164-174. doi:10.1027/1015-5759.20.3.180

Zuckerman, M., Kuhlman, D. M., Joireman, J., Teta, P., & Kraft, M. (1993). A comparison of three structural models for personality: The Big Three, the Big Five, and the Alternative Five. *Journal of Personality and Social Psychology, 65*, 757-768. doi:10.1037/0022-3514.65.4.757

Articles included in the meta-analysis of dependability coefficients

Adebayo, S. O., & Arogundade, O. B. (2011). Determinants of significant single best predictor of life satisfaction among Nigerian adults. *Interdisciplinary Review of Economics and Management, 1*, 39-46.

Al-Jurany, K. A. H. (2013). *Personality characteristics, trauma, and symptoms of PTSD: A population study in Iraq*. Unpublished doctoral thesis, Heriot-Watt University, England. http://hdl.handle.net/10399/2641

Anusic, I., Lucas, R. E., & Donnellan, M. B. (2012), Dependability of personality, life satisfaction, and affect in short-term longitudinal data. *Journal of Personality, 80*, 33–58. doi:10.1111/j.1467-6494.2011.00714.x

Bartram D., Lindley, P., & Coine, I. J. (2000). *Technical Manual for the ICES Plus*. Vancouver, Canada: ICES Assessment Systems.

Biesanz, J. C., & West, S. G. (2004). Towards understanding assessments of the Big Five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. *Journal of Personality, 72*, 845-876. doi:10.1111/j.0022-3506.2004.00282.x

Björgvinsson, T., & Thompson, A. P. (1996). Evaluation of an Icelandic translation of the Basic Personality Inventory using a bilingual sample. *Journal of Clinical Psychology, 52*, 431-435. doi:10.1002/(SICI)1097-4679(199607)52:4<431::AID-JCLP7>3.0.CO;2-S

Boals, A., Southard-Dobbs, S., & Blumenthal, H. (2014). Adverse events in emerging adulthood are associated with increases in neuroticism. *Journal of Personality*. Advance online publication. doi:10.1111/jopy.12095

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3-5. doi:10.1177/1745691610393980

Butcher, J. N., Graham, J. R., Dahlstrom, W. G., & Bowman, E. (1990). The MMPI-2 with college students. *Journal of Personality Assessment, 54*, 1-15. doi:10.1080/00223891.1990.9673968

Caldwell-Andrews, A., Baer, R. A., & Berry, D. T. R. (2000). Effects of response sets on NEO-PI-R scores and their relations to external criteria. *Journal of Personality Assessment, 74*, 472-488. doi:10.1207/S15327752JPA7403_10

Caprara, G. V., Barbaranelli, C. Borgogni, L., & Perugini, M. (1993). The "Big Five Questionnaire": A new questionnaire to assess the five factor model. *Personality and Individual Differences, 15*, 281-288. doi:10.1016/0191-8869(93)90218-R

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, *97*, 186–202. doi:10.1037/a0015618

Dennissen, J. A., Geenen, R., Selfhout, M., & Van Aken, M. A. G.(2008). Single-item Big Five ratings in a social network design. *European Journal of Personality, 22*, 37-54. doi:10.1002/per.662

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18*, 192-203. doi:10.1037/1040-3590.18.2.192

Fossati, A., Borroni, S., Marchione, D., & Maffei, C. (2011). The Big Five Inventory (BFI): Reliability and validity of its Italian translation in three independent nonclinical samples.

European Journal of Psychological Assessment, 27, 50-58. doi:10.1027/1015-5759/a000043

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2013). Assessing the stability of psychological and political survey measures. *American Political Research, 41*, 54-75. doi:10.1177/1532673X12446215

Gill, C. M., & Hodgkinson, G. P. (2007). Development and validation of the Five-Factor Model Questionnaire (FFMQ): An adjectival-based personality inventory for use in occupational settings. *Personnel Psychology, 60*, 731-766. doi:10.1111/j.1744-6570.2007.00090.x

Gomà-i-Freixanet, M., Valero, S., Puntí1, J., & Zuckerman, M. (2004). Psychometric properties of the Zuckerman-Kuhlman Personality Questionnaire in a Spanish sample. *European Journal of Psychological Assessment, 20*, 134-146. doi:10.1027/1015-5759.20.2.134

Gorostiaga, A., Belluerka, N., Alonso-Arbiol, I., & Haranburu, M. (2011). Validation of the Basque Revised NEO Personality Inventory (NEO PI-R). *European Journal of Psychological Assessment, 27*, 193-205. doi:10.1027/1015-5759/a000067

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504–528. doi:10.1016/S0092-6566(03)00046-1

Harris, T. C., Tracey, A. J., & Fisher, G. G. (2011). *Predictive Index Technical Overview*. Wellesley Hills, MA: PI Worldwide.

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. doi:10.1037/0021-9010.91.1.9

Hendriks, A. A. J, Hofstee, W. K. B., & De Raad, B. (1999). The Five-Factor Personality Inventory (FFPI). *Personality and Individual Differences, 27*, 307-325. doi:10.1016/S0191-8869(98)00245-1

Herzberg, P. Y., & Brähler, E. (2008). Assessing the Big-Five personality domains via short forms: A cautionary note and a proposal. *European Journal of Psychological Assessment, 22*, 139-148. doi:10.1027/1015-5759.22.3.139

Hjelle, L. A., & Bernard, M. (1994). Private self-consciousness and the retest-reliability of self-reports. *Journal of Research in Personality, 28*, 52-67. doi:10.1006/jrpe.1994.1006

Hogan, R., & Hogan, J. (1995). *Manual for the Hogan Personality Inventory*. Tulsa, OK: Hogan Assessment Systems.

Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5-120 on Amazon's mechanical Turk. *Computers in Human Behavior, 29,* 1749-1754. doi:10.1016/j.chb.2013.02.020

Hossiep, R., Paschen, M., & Mühlhaus, O. (2003). *Manual zum Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung* [Business-Focused Inventory of Personality]. Göttingen, Germany: Hogrefe.

Hyphantis, T., Antoniou, K., Floros, D. K., Valma, V., Pappas, A. I., Douzenis, A., …, Kuhlman, M. (2013). Assessing personality traits by questionnaire: Psychometric properties of the Greek version of the Zuckerman-Kuhlman personality questionnaire and correlations with psychopathology and hostility. *Hippokratia, 17*, 342-350.

Inscape. (2005). *DiSC validation* (Research report). Inscape Publishing.

Karanci, A. N., Dirik, G., & Yorulmaz, O. (2007). Reliability and validity studies of Turkish translation of Eysenck Personality Questionnaire Revised-Abbreviated. *Turkish Journal of Psychiatry, 18*, 1-7.

Karwowski, M., Lebuda, I., Wisniewska, E., & Gralewski, J. (2013). Big Five personality traits as the predictors of creative self-efficacy and creative personal identity: Does gender matter? *Journal of Creative Behavior, 47*, 215–232. doi:10.1002/jocb.32

Kroner, D. G., Reddon, J. R., & Beckett, N. (1991). Basic Personality Inventory clinical and validity scales: Stability and internal consistency. *Journal of Psychopathology and Behavioral Assessment, 13*, 147-154. doi:10.1007/BF00961428

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology, 22*, 251-259. doi:10.1007/s10869-008-9064-2

Laguna, M. (2011). *A Polish version of the TIPI.* Unpublished manuscript, Institute of Psychology, KUL Lublin, Poland.

Lang, F. R. (2005). *Erfassung des kognitiven Leistungspotenzials und der „Big Five" mit Computer-Assisted-Personal-Interviewing (CAPI)* [Assessment of cognitive competencies and the „Big Five" with computer-assisted personal interviewing]. Berlin, Germany: DIW.

Langford, P. H. (2003). A one-minute measure of the Big Five? Evaluating and abridging Shafer's (1999a) Big Five markers. *Personality and Individual Differences, 35*, 1127-1140. doi:10.1016/S0191-8869(02)00323-9

Lubin, B., & Van Whitlock, R. (2002). Development of a measure that integrates positive and negative affect and personality: The Comprehensive Personality and Affect Scales. *Journal of Clinical Psychology, 58*, 1135-1156. doi:10.1002/jclp.10042

Mascaro, N., & Rosen, D. H. (2005). Existential meaning's role in the enhancement of hope and prevention of depressive symptoms. *Journal of Personality, 73*, 985-1013. doi:10.1111/j.1467-6494.2005.00336.x

Matthews, G., Stanton, N., Graham, N. C., & Brimelow, C. (1990). A factor analysis of the scales of the Occupational Personality Questionnaire. *Personality and Individual Differences, 11*, 591-596. doi:10.1016/0191-8869(90)90042-P

Matz, P. A., Altepeter, T. S., & Perlman, B. (1992). MMPI-2: Reliability with college students. *Journal of Clinical Psychology, 48*, 330-334. doi:10.1002/1097-4679(199205)48:3<330::AID-JCLP2270480310>3.0.CO;2-3

McCrae, R. R., Yik, M. S. M., Trapnell, P. D., Bond, M. H., & Paulhus, D. L. (1998). Interpreting personality profiles across cultures: Bilingual, acculturation, and peer rating studies of Chinese undergraduates. *Journal of Personality and Social Psychology, 74*, 1041-1055. doi:10.1037/0022-3514.74.4.1041

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*, 28-50. doi:10.1177/1088868310366253

Navrátil, M., & Lewis, C. A. (2006). Temporal stability of the Czech translation of the Eysenck Personality Questionnaire Revised-Abbreviated: Test-retest data over one-week. *Individual Differences Research, 4*, 208-212.

Oshio, A., Abe, S., & Cutrone, P. (2012). Development, reliability, and validity of the Japanese version of Ten Item Personality Inventory (TIPI-J). *Japanese Journal of Personality, 21*, 40-52. doi:10.2132/personality.21.40

Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R – NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung*. Göttingen, Germany: Hogrefe.

Peterson, J. B. (2010). *Caliper assessment process*. Caliper Assessment.

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. L., Houston, J. S., Toquam, J. L., & Wing, H. (1990). Project A: Specification of the predictor domain and development of

new selection / classification tests. *Personnel Psychology, 43*, 247-276. doi:10.1111/j.1744-6570.1990.tb01558.x

Piedmont, R. L., Bain, E., McCrae, R. R., & Costa, Paul T. Jr. (2002). The applicability of the five-factor model in a sub-Saharan culture. The NEO-PI-R in Shona. In R. R. McCrae & J. Allik (Eds.), *The Five-Factor model of personality across cultures* (pp. 155–174). New York, NY: Kluwer Academic Publisher.

Putnam, S. H., Kurz, J. E., & Houts, D. C. (1996). Four-month test-retest reliability of the MMPI-2 with normal male clergy. *Journal of Personality Assessment, 67*, 341-353. doi:10.1207/s15327752jpa6702_9

Profiles International (2010). *ProfileXT: Technical manual*. Waco, TX: Profiles International.

Rammstedt, B., & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K) [A short version of the Big Five Inventory]. *Diagnostica*, *51*, 195-206. doi:10.1026/0012-1924.51.4.195

Rammstedt, B., & John, O. P. (2007). 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*, *41*, 203–212. doi:10.1016/j.jrp.2006.02.001

Renau, V., Oberst, U., Gosling, S. D., Rusinol, J., & Chamarro, A. (2013). Translation and validation of the Ten Item-Personality Inventory into Spanish and Catalan, *Aloma, Revista de Psicologia, Ciències de l'Educació i de l'Esport, 31*, 85-97.

Robins, R. W., Fraley, R. C., Roberts, B. W., & Trzesniewski, K. H. (2001). A longitudinal study of personality change in young adulthood. *Journal of Personality*, *69*, 617-640. doi:10.1111/1467-6494.694157

Romero, E., Villar, P., Gómez-Fraguela, K. A., & López-Romero, L. (2012). Measuring personality traits with ultra-short scales: A study of the Ten Item Personality Inventory

(TIPI) in a Spanish sample. *Personality and Individual Differences, 53*, 289-293. doi:10.1016/j.paid.2012.03.035

Ruch, W. (1999). Die revidierte Fassung des Eysenck Personality Questionnaire und die Konstruktion des deutschen EPQ-R bzw. EPQ-RK [The Eysenck Personality Questionnaire-Revised and the construction of German standard and short versions]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 20*, 1-24. doi:10.1024//0170-1789.20.1.1

Sato, T. (2005). The Eysenck Personality Questionnaire Brief Version: Factor structure and reliability. *Journal of Psychology, 135*, 545-552. doi:10.3200/JRLP.139.6.545-552

Saville, P., Holdsworth, R., Nyfiled, G., Cramp, L., & Mabey, W. (1996). *Occupational Personality Questionnaire: Manual and user's guide*. Boston, MA: SHL.

Sukigara, M. (1996). Equivalence between computer and booklet administrations of the new Japanese version of the MMPI. *Educational and Psychological Measurement, 56*, 570-584. doi:10.1177/0013164496056004002

Sun, L., Kosinski, M., Stillwell, D., & Rust, J. (2011, July). *On the test-retest reliability of the 100-item IPIP scales: Differential temporal stability of the Big Five personality traits*. Paper presented at the International Meeting of the Psychometric Society, Hong Kong.

Thompson, E. R. (2008). Development and validation of an international English Big-Five Mini-Markers. *Personality and Individual Differences, 45*, 542-548. doi:10.1016/j.paid.2008.06.013

Trickey, G., & Hyde, G. (2009). *Profile:Match Manual*. Turnbridge Wells, England: Psychological Consultancy.

Tsaousis, I. (1999). The traits personality questionnaire (TPQue): A Greek measure for the five factor model. *Personality and Individual Differences, 26*, 271-283. doi:10.1016/S0191-8869(98)00131-7

Tsaousis, I., & Georgiades, S. (2009). Development and psychometric properties of the Greek Personality Adjective Checklist (GPAC). *European Journal of Psychological Assessment, 25*, 164-174. doi:10.1027/1015-5759.25.3.164

Tsaousis, I., & Kerpelis, P. (2004). The traits personality questionnaire 5 (TPQue5). *European Journal of Psychological Assessment, 20*, 180-191. doi:10.1027/1015-5759.20.3.180

Van der Heijden, P. T., Egger, J. I. M., & Derksen, J. J. L. (2008). Psychometric evaluation of the MMPI-2 restructured clinical scales in two Dutch samples. *Journal of Personality Assessment, 90*, 456-464. doi:10.1080/00223890802248745

Watson, D. (2003). Investigating the construct validity of the dissociative taxon: Stability analyses of normal and pathological dissociation. *Journal of Abnormal Psychology, 112*, 298-305. doi:10.1037/0021-843X.112.2.298

Yang, J.-F. (2010). Cross-cultural personality assessment: The Revised NEO Personality Inventory in China. *Social Behavior and Personality, 38*, 1097-1104. doi:10.2224/sbp.2010.38.8.1097

Zuckerman, M. (2002). Zuckerman-Kuhlman Personality Questionnaire (ZKPQ): An alternative five-factorial model. In B. de Raad, & M. Perugini (Eds), *Big Five Assessment* (pp. 377-396). Göttingen, Germany: Hogrefe.