Facets of Measurement Error for Scores of the Big Five: Three Reliability Generalizations

Timo Gnambs

Osnabrück University

Word count: 4,919

Author Note

Timo Gnambs, Institute of Psychology, Osnabrück University, Germany.

Correspondence concerning this article should be addressed to Timo Gnambs, Institute of Psychology, Osnabrück University, Seminarstr. 20, 49069 Osnabrück, Germany. Tel.: +49 (0)541 / 969–4417, Fax: +49 (0)541 969-14200. E-mail: timo.gnambs@uni-osnabrueck.de

Accepted for publication in *Personality and Individual Differences*

Highlights

- Different reliability coefficients reflect different facets of measurement error.
- Three reliability generalizations for measures of the Big Five are presented.
- Estimates of five reliability coefficients are derived.
- Four facets of measurement error accounted for up to half of the score variance.

Abstract

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

Keywords: Big Five, reliability, measurement error, meta-analysis, generalizability theory

Facets of Measurement Error for Scores of the Big Five:

Three Reliability Generalizations

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

Measurement Error in Self-Reports

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

Sources of Measurement Error

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

Indices of Measurement Error

Although measurement error can be analyzed using various latent variable techniques (cf. Gnambs & Batinic, 2011; Gnambs, Appel, Schreiner, Richter, & Isberner, 2014; Steyer, Mayer, Geiser, & Cole, 2014), it is more commonly quantified by forms of reliability. Reliability is defined as the ratio of true score variability to total score variability in classical test theory (Lord & Novick, 1968). While several methods have been proposed to calculate test score reliabilities, they differ in the way they define and measure the true score variance. As a result, different measures of reliability quantify different sources of measurement error (cf. Schmidt et al., 2003): Coefficients of equivalence (CE) focus on the shared variance between different items at a single measurement occasion. They quantify measurement error in terms of random and item-specific factor error because these cancel each other out across different items. On the other hand, correlations of test scores across two measurement occasions obtained from the same scale are typically used as measures of test-retest reliabilities (coefficient of stability, CS). These assess random measurement error and transient error, but do not reflect item-specific error. All three forms of measurement error are incorporated in the coefficient of equivalence and stability (CES), which results from

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

correlating two parallel forms of a measure that have been administered on separate occasions. Moreover, Le and colleagues (2009) proposed extensions of CE and CES that also acknowledge scale-specific factor errors. The generalized coefficient of equivalence (GCE) and the generalized coefficient of equivalence and stability (GCES) represent the correlations of test scores from different scales measuring the same construct, each either administered on the same (GCE) or on separate occasions (GCES). Of these coefficients, the GCES represents the most general indicator of reliability that accounts for all four sources of measurement error (see Table 1).

The Present Study

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

Method

Meta-Analytic Procedure

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

Meta-analytic model. For each trait of the Big Five the individual effect sizes were synthesized with a random effects meta-analysis using the *metaSEM* software (Cheung, 2014a). Dependencies between effects that resulted from studies reporting multiple reliability indices were accounted for by specifying a multilevel model (cf. Cheung, 2014b). This approach models three hierarchical levels that refer to the individual effect sizes (Level 1), differences between effect sizes within a sample (Level 2), and difference between samples (Level 3). To correct for sampling error each effect was weighted by the inverse of its variance.

Development of the Meta-Analytic Database

Inclusion criteria. Studies had to meet the following criteria to be included in the meta-analyses. First, the study must have administered a validated measure of personality according to the Big Five taxonomy. To avoid biased estimates due to imperfect construct validities (cf. Salgado, 2003) the analyses were limited to the four most frequently used Big Five instruments (cf. Gnambs, 2013, 2014; Sibley et al., 2012): Costa and McCrae's (1992) NEO scales, the Big Five Inventory (John, Naumann, & Soto, 2008), Goldberg's (1999) statements from the International Personality Item Pool, and various trait-descriptive adjective lists (e.g., Goldberg's, 1992, Big Five Markers). Second, the study must have reported a relevant effect size (see above). Third, following prevalent recommendations (cf. Gnambs,

2014) test-retest reliability studies must have adopted retest intervals that did not exceed two months. Fourth, to guard against potential cross-temporal changes (cf. Twenge, 2001: Twenge, Konrath, Foster, Campbell, & Bushman, 2008) that might have affected the reliability estimates, only studies published in 2000 or later were considered. Finally, studies must have reported on samples of healthy adults. Studies on children or participants with psychopathological symptoms were excluded.

Literature search. Several research strategies were employed to identify relevant studies for the series of meta-analyses. First, relevant articles were identified from database searches in PsycINFO and Psyndex using search strings including the terms *measurement error*, *composite reliability*, *coefficient omega*, *retest reliability*, *transient error,* or *coefficient of stability* in combination with the names of the considered Big Five instruments. Second, similar searches were conducted in Google Scholar. Finally, additional studies were identified from existing meta-analyses on measurement error in scores of the Big Five (Connelly & Ones, 2010; Gnambs, 2014; Pace & Brannick, 2010; Salgado, 2002; Viswesvaran & Ones, 2000). This search process identified 63 primary articles that reported on 71 independent samples.

Coded variables. From the identified primary studies the following information was extracted: (a) the respective effect sizes, that is, CE, CS and GCE (see above), (b) the sample sizes, (c) the length of the administered instrument (i.e. the number of included items), (d) for CS the length of the retest interval (in weeks), (e) and several socio-demographic information (e.g., mean age, sex ratio).

Results

The meta-analyses included a total of 38,944 individuals. The sample sizes ranged from 17 to 14,348 (*Mdn* =216). Approximately 58% of the participants were female; their reported mean age was 27.65 years (*SD* = 11.09). Most samples came from Europe (32%) and North America (44%).

Coefficient of Equivalence (CE)

The meta-analysis included between 13 and 17 ω_h coefficients for the Big Five (see Table S1 of the online supplement). For each of the five traits, the mean unweighted and inverse-variance weighted CEs that reflect item-specific and random error are reported in Table 2. The median true CE fell at .82 which clearly exceeded the threshold of .70 that many authors use as a rule of thumb to evaluate reliabilities (cf. McCrae et al., 2011). However, for all traits the random level 3 variance $\tau^2_{(3)}$ that indicates between-sample heterogeneity was significant at $p < .05$. Because the number of items per scale influences the degree of random error (Schmidt et al., 2003), longer instruments tend to exhibit larger reliabilities than shorter instruments with fewer items. In the present study, the median number of items per trait scale fell between 8 and 9 items (*Min* = 2, *Max* = 20). To examine the effects of scale length on CE the ω_h coefficients were regressed on the number of items included in the administered instrument. This reduced $\tau^2_{(3)}$ for openness and extraversion by 10% and 14%, respectively, whereas it had no effect on the other trait scales. Moreover, after accounting for the number of items the respective estimates of CE hardly changed, mean ΔCE = .01. Thus, the scale length had a rather modest impact on CE for the included instruments.

Coefficient of Stability (CS)

The meta-analysis included 53 to 54 test-retest correlations for the five traits (see Table S2 of the online supplement). The aggregated CSs (*Mdn* = .84) that acknowledges transient and random error were slightly larger than the respective CE (see Table 2). In line with previous studies (Gnambs, 2014; Viswesvaran & Ones, 2000) extraversion scales resulted in a somewhat larger CS of .88 than agreeableness scales, CE = .80. Again, the significant $\tau^2_{(3)}$ indicated unaccounted between-study heterogeneity. Because CSs are sensitive to the adopted interval between test and retest, the CSs were regressed on the length of the retest interval in weeks. In the present study, the median interval between test and retest was four weeks (*Min* = 1, *Max* = 8). Although controlling for differences in the retest interval reduced $\tau^2_{(3)}$ for openness and neuroticism by about 5% and 9%, it had negligible impact on the other trait variances, $R^2 < .02$.

Coefficient of Equivalence and Stability (CES)

A direct meta-analysis of CES that incorporates item-specific factor error in addition to random and transient error was infeasible because respective reliability indices are rarely reported in primary studies. However, an estimate of CES can be derived indirectly from the two previous meta-analyses. CES can be calculated as the difference of CE and the proportion of transient error variance (TEV; see Schmidt et al., 2003, eq. B10). The former is readily available from the previous meta-analysis on CE, whereas the latter can be derived from the meta-analysis on CS by including CE as a moderator. Gnambs (2014) showed that the intercept in this regression model (more precisely, 1 – intercept) represents an estimate of TEV after accounting for random error. In the respective analyses coefficient alpha was used as an indicator of CE because no study could be identified that reported both test-retest correlations and ω_h. As summarized in Table 2, between 8% to 10% of the observed test score variances can be attributed to TEV alone. As a consequence, CES fell between .64 (agreeableness) and .77 (extraversion) for the five traits (see Figure 1).

Generalized Coefficient of Equivalence (GCE)

The meta-analysis of correlations between different measures of the same Big Five traits included 28 to 30 effect sizes (see Table S3 of the online supplement). The aggregated GCE for the five traits (see Table 2) were .64 for openness, .74 for conscientiousness and extraversion, .62 for agreeableness, and .76 for neuroticism. Thus, about 24% to 38% of the observed score variance in measures of the Big Five reflect measurement error when acknowledging factor-specific error in addition to item-specific and random error (cf. Table 1). Most of the observed differences in GCE were accounted for by sampling error; as a consequence, all but one random variance components τ^2 were insignificant.

Generalized Coefficient of Equivalence and Stability (GCES)

It was not possible to conduct a direct meta-analysis of GCES because no studies could be identified that reported respective reliability coefficients. However, estimates of GCES can be readily derived from the previous meta-analyses. Le and colleagues (2009, eq. 6) showed that GCES that incorporates all four sources of measurement error (i.e. random, transient, item-, and factor-specific error) can be calculated as the difference of GCE and TEV. Both components were already presented in the previous sections (see Table 2). For the five traits GCES fell between .49 and .67 (see Figure 1). Thus, about a half to two thirds of observed score variance in measures of the Big Five reflect true score differences, whereas the remaining variance is due to measurement error.

Publication Bias

To determine whether systematically missing studies might have distorted the accuracy of the synthesized effects, the fail-safe number of missing studies with unreliable test scores that would be needed to alter the conclusions from the meta-analyses was estimated. Following Howell and Shields (2008), the number of file drawer studies required to lower the population reliabilities for conscientiousness, extraversion, and neuroticism below .70 was estimated to be at least as large or even larger than the number of available studies (see Table 2). Thus, for these traits measures of the Big Five seem to produce reliable test scores of at least .70. For openness and agreeableness, the respective Fail-Safe *N*s was considerably smaller, indicating somewhat less confidence in the identified effects.

Discussion

Measurement error typically attenuates scale scores in psychological research and, thus, results in observed correlations that underestimate the true relationship between constructs (Schmidt, 2010). Therefore, corrections using the instrument’s reliability are necessary to derive unbiased relationships between constructs. Unfortunately, proper reliability estimates are frequently not available in many applied situations. Particularly general reliability indices such as CES or GCES that have been advocated for use in artifact corrections (e.g., Le et al., 2009; Schmidt et al., 2003) are rarely readily at hand. However, observed statistics need to be corrected by all sources of error to obtain the true relationship between constructs. In these cases, respective estimates from reliability generalizations might be substituted. The aggregated reliabilities presented above account for all four sources of measurement error. Hence, for measures of the Big Five estimates of five types of reliability, CE, CS, CES, GCE, and GCES, are now readily available (see Figure 2). About 25% to 30% of the variance in observed scores can be attributed to random, transient and item-specific error (CES). If factor-specific errors are acknowledged as well (GCES), nearly half of the observed score variance is the result of measurement error.

Implications

The significant proportion of error in measures of the Big Five has non-trivial effects for the examination of construct-level relationships. For example, Judge, Higgins, Thoresen, and Barrick (1999, Table 4) derived a longitudinal correlation of $r = .40$ between conscientiousness scores assessed in childhood and measures of job satisfaction that were obtained over 30 years later. A commonly used approach to correct observed score relationships (i.e. by estimating true score correlations) are bivariate corrections for attenuation due to measurement error, that is, a division of the observed correlation by the square root of the product of the two reliabilities (Ree & Carretta, 2006). Use of the CE or the CS presented above to derive the construct-level effect would result in an artifact-adjusted[2] true score correlation of $\rho = .44$. Thus, assuming the administered scale represents a valid operationalization of conscientiousness the proportion of explained variance in job satisfaction is about $\Delta R^2 = .03$ higher than the uncorrected correlation would suggest. However, in fact, this represents an underestimation of the true relationship because the corrections adjusted for only two sources of errors (i.e. random and either item-specific or transient error). Using the estimate GCES that acknowledges all four facets of measurement

[2] For simplicity of presentation it was assumed that job satisfaction was measured without error.

error (see Table 1) results in an artifact-adjusted true score effect of ρ = .50. Thus, the proportion of job satisfaction variance explained by childhood conscientiousness is about ΔR^2 = .09 larger after correcting for measurement error in the observed scores.

Conclusion

Research hypotheses typically refer to relationships between constructs and not measures. Because most measures cannot operationalize the construct of interest without error, it is necessary to adjust observed correlations for the biasing influence of measurement error (Ree & Carretta, 2006). The assessment of all sources of measurement error, that is, random, transient and factor errors, would require at least two different measures for each construct to be administered at two separate occasions (cf. Le et al., 2009). Such research designs seem infeasible for most practical research scenarios. In such cases, researchers require profound *a priori* knowledge on the proportion of error in their measures. For the Big Five of personality, one of the most influential models of personality to date (cf. John et al., 2008), the present study extended previous generalizations and derived estimates for five types of reliability. Thus, researchers using the Big Five of personality now have enough information at hand to flexibly correct observed correlations for different sources of measurement error: random response error, transient error, item-specific factor error and scale-specific factor error.

References

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review, 18*, 139-152. doi:10.1177/1088868314523838

Cheung, M. W.-L. (2014a). 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

Cheung, M. W.-L. (2014b). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*, 211-229. doi:10.1037/a0032968

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*, 1092-1122. doi:10.1037/a0021212

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

Cronbach, L. J. (1947). Test reliability: Its meaning and determination. *Psychometrika, 12*, 1-16. doi:10.1007/BF02289289

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*, 399-412. doi:10.1111/bjop.12046

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment, 30*, 130-139. doi:10.1027/1015-5759/a000181

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. (2014). A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five. *Journal of Research in Personality, 52*, 20-28. doi:10.1016/j.jrp.2014.06.003

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

Gnambs, T., Appel, M., Schreiner, C., Richter, T., & Isberner, M.-B. (2014). Experiencing narrative worlds: A latent state-trait analysis. *Personality and Individual Differences, 69*, 187-192. doi:10.1016/j.paid.2014.05.034

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.

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

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.

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The Big Five personality traits, general mental ability, and career success across the life span. *Personnel Psychology, 52*, 621-652. doi:10.1111/j.1744-6570.1999.tb00174.x

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin, 136*, 768-821. doi:10.1037/a0020327

Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods, 12*, 165-200. doi:10.1177/1094428107302900

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of test scores*. Reading, England: Addison-Wesley.

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

Pace, V. L., & Brannick, M. T. (2010). How similar are personality scales of the "same" construct? A meta-analytic investigation. *Personality and Individual Differences*, *49*, 669-676. doi:10.1016/j.paid.2010.06.014

Ree, M. J., & Carretta, T. R. (2006). The role of measurement error in familiar statistics. *Organizational Research Methods, 9*, 99-112. doi:10.1177/1094428105283192

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*, 353-387. doi:10.1037/a0026838

Salgado, S. F. (2002). The Big Five personality dimensions and counterproductive behaviors. *International Journal of Selection and Assessment, 10*, 117-125. doi:10.1111/1468-2389.00198

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

Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science, 5*, 233-242. doi:10.1177/1745691610369339

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

Sibley, C. G., Osborne,D., & Duckitt, J. (2012). Personality and political orientation: Meta-analysis and test of a threat-constraint model. *Journal of Research in Personality*, *46*, 664-677. doi:10.1016/j.jrp.2012.08.002

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120. doi:10.1007/s11336-008-9101-0

Steel, P., Schmidt, J., & Schultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134*, 138-161. doi:10.1037/0033-2909.134.1.138

Steyer, R., Mayer, A., Geiser, C., & Cole (2014). A theory of states and traits: Revised. *Annual Review of Clinical Psychology.* Advance online publication. doi:10.1146/annurev-clinpsy-032813-153719

Twenge, J. M. (2001). Changes in women's assertiveness in response to status and roles: A cross-temporal meta-analysis, 1931-1993. *Journal of Personality and Social Psychology, 81*, 133-145. doi:10.1037/0022-3514.81.1.133

Twenge, J. M., Konrath, S., Foster, J. D., Campbell, W. K., & Bushman, B. J. (2008). Egos inflating over time: A cross-temporal meta-analysis of the Narcissistic Personality

Inventory. *Journal of Personality, 76*, 875-902. doi:10.1111/j.1467-6494.2008.00507.x

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

Table 1.

Sources of Measurement Error and Reliability Indices

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

Table 2.

Meta-Analyses of Reliability Coefficients for Big Five Scales

				Unweighted		Weighted					Fail Safe *N*	
	k_1	k_2	*N*	*r*	SD_r	ρ	90% CRI	$\tau^2_{(2)}$	$\tau^2_{(3)}$	TEV	$\rho_{.70}$	$N_{.70}$
Coefficient of equivalence (CE)												
Openness	16	16	5,802	.79	.15	.79*	[.56, 1.00]	.00	.02*		.59	13
Conscientiousness	13	13	4,818	.83	.08	.83*	[.71, .95]	.00	.01*		.64	29
Extraversion	17	17	6,941	.84	.11	.85*	[.67, 1.00]	.00	.01*		.62	29
Agreeableness	13	13	5,356	.77	.10	.77*	[.62, .92]	.00	.01*		.63	12
Neuroticism	13	13	4,821	.82	.10	.82*	[.67, .98]	.00	.01*		.63	21
Coefficient of stability (CS)												
Openness	53	31	9,938	.81	.07	.84*	[.73, .94]	.00	.00*	.12	.65	143
Conscientiousness	53	31	9,938	.83	.06	.84*	[.74, .94]	.00	.00*	.10	.65	147
Extraversion	53	31	9,938	.86	.06	.88*	[.79, .96]	.00	.00*	.08	.66	235
Agreeableness	53	31	9,938	.78	.09	.80*	[.68, .93]	.00	.01*	.13	.64	89
Neuroticism	54	32	9,971	.82	.08	.84*	[.73, .95]	.00*	.00*	.09	.65	136
Generalized coefficient of equivalence (GCE)												
Openness	29	24	22,118	.64	.11	.64*	[.47, .81]	.00	.01*		.62	
Conscientiousness	28	23	21,983	.74	.05	.74*	[.66, .82]	.00	.00		.66	28
Extraversion	29	23	22,432	.74	.07	.74*	[.64, .84]	.00	.00		.65	24
Agreeableness	30	24	22,722	.62	.10	.62*	[.47, .76]	.01	.00		.63	
Neuroticism	29	23	22,432	.76	.07	.76*	[.66, .86]	.00	.00		.65	35

Note. k_1 = Number of effect sizes; k_2 = Number of independent samples; *N* = Total sample size; *r* = Unweighted reliability coeffiecient; ρ = Weighted reliability coefficient; τ^2 = Random level 2 and level 3 variance of ρ_{tt}; CRI = 90% credibility interval; TEV = Transient error variance; $\rho_{.70}$ = 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

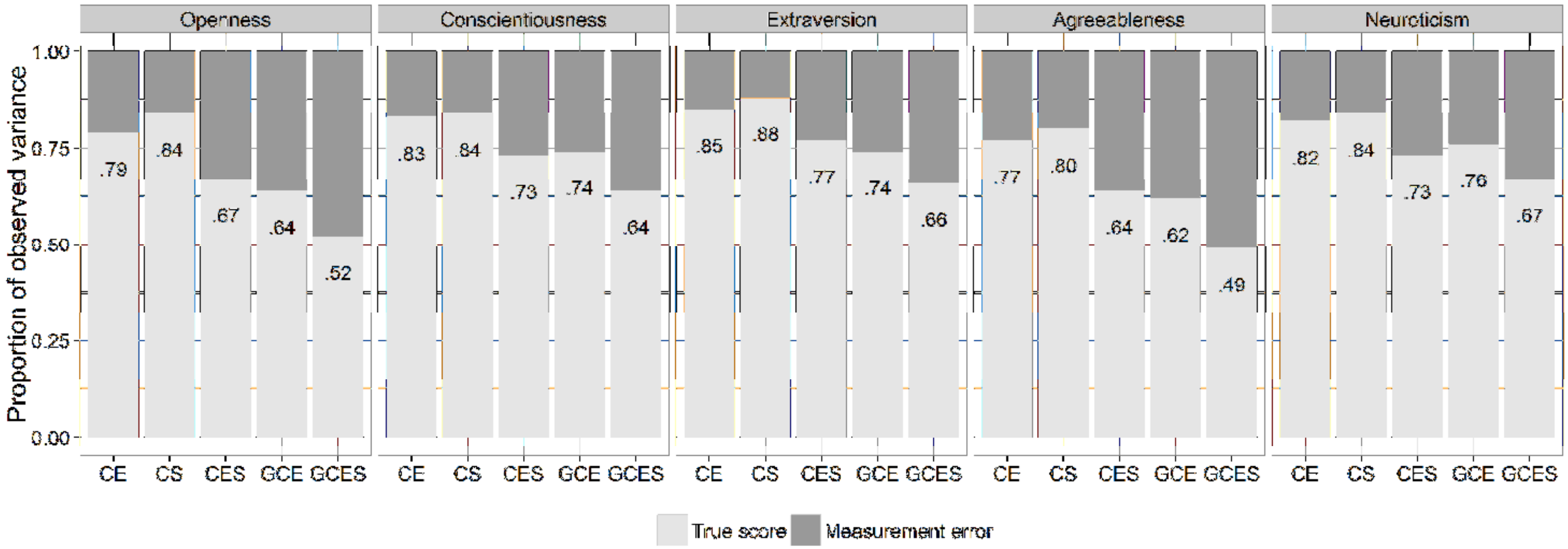

Figure 1. Indices of measurement error for measures of the Big Five; CE = Coefficient of equivalence, CS = Coefficient of stability, CES = Coefficient of equivalence and stability, GCE = Generalized coefficient of equivalence, GCES = Generalized coefficient of equivalent and stability

Online Supplement for

"Facets of Measurement Error for Scores of the Big Five: Three Reliability Generalizations"

Supplemental Tables

Table S1.

Summary of Effects for the Reliability Generalization on CE

Study	*N*	Instrument	Country	OP	CO	EX	AG	NE
Balaji & Chakrabarti (2010)	227	BFI	India			0.90		
Becker (2006)	206	NEO-FFI	Canada	0.74	0.85	0.77	0.78	0.86
	225	NEO-FFI	Canada	0.76	0.80	0.8	0.76	0.85
Davis & Yi (2012)	230	IPIP	US	0.91	0.92	0.95	0.85	0.94
Fuller et al. (2008)	550	BFI	Austria	0.88		0.92		
Huang et al. (2012)	468	TDA	England	0.84	0.83	0.82	0.79	0.75
Hull et al. (2010)	1,021	NEO-FFI	Jamaica	0.29	0.72	0.47	0.53	0.60
Kang & Johnson (2013)	319	BFI	US	0.74	0.70	0.77	0.70	0.69
Kautish (2010)	264	BFI	India	0.87		0.94		
Korzaan & Boswell (2008)	230	IPIP	US	0.82	0.89	0.91	0.89	0.91
Matzler & Mueller (2011)	124	NEO-FFI	Germany	0.81	0.84			
Matzler et al. (2011)	662	BFI	Austria			0.91	0.80	
Mehmetoglu (2012)	1,000	BFI	Norway	0.85	0.79	0.82	0.78	0.82
Salimian & Hosainian (2012)	170	BFI	Iran	0.97				
Terzis et al. (2012)	117	BFI	Greece	0.91	0.90	0.94	0.91	0.91
Tsai et al. (2012)	544	IPIP	China			0.85		
Wang et al. (2012)	228	NEO-FFI	China	0.75	0.86	0.89	0.79	0.85
Yap & Lee (2013)	512	BFI	New Zealand	0.74	0.96	0.87	0.67	0.90
Ying & Norman (2014)	138	BFI	US	0.80	0.72	0.83	0.74	0.73
Zhao (2011)	127	IPIP	Canada					0.85

Note. OP = Openness, CO = Conscientiousness, EX = Extraversion, AG = Agreeableness, NE = Neuroticism. BFI = Big Five Inventory, NEO-FFI = NEO Five Factor Inventory, IPIP = International Personality Item Pool, TDA = Trait-descriptive adjectives.

Table S2.

Summary of Effects for the Reliability Generalization on CS

Study	*N*	Instrument	Country	OP	CO	EX	AG	NE
Adebayo & Arogundade (2011)	45	BFI	Nigeria	0.96	0.84	0.96	0.92	0.77
Al-Jurany (2013)	33	BFI	Iraq					0.82
	17	BFI	Iraq	0.81	0.80	0.82	0.78	0.84
Anusic et al. (2012)	199	IPIP	US	0.74	0.82	0.85	0.65	0.76
	199	IPIP	US	0.73	0.82	0.85	0.67	0.76
	199	IPIP	US	0.70	0.81	0.78	0.66	0.75
	199	IPIP	US	0.73	0.82	0.78	0.65	0.69
	199	IPIP	US	0.68	0.77	0.76	0.59	0.72
	199	IPIP	US	0.68	0.79	0.76	0.62	0.71
	199	IPIP	US	0.69	0.80	0.79	0.62	0.73
Biesanz & West (2004)	339	TDA	US	0.80	0.77	0.80	0.72	0.68
	339	TDA	US	0.82	0.76	0.80	0.73	0.70
	339	TDA	US	0.78	0.69	0.75	0.64	0.66
Buhrmeister et al. (2011)	70	BFI	US	0.90	0.86	0.94	0.87	0.92
Caldwell-Andrews et al. (2000)	42	NEO-PI-R	US	0.89	0.91	0.93	0.74	0.81
Chmielewski & Watson (2009)	447	TDA	US	0.81	0.78	0.89	0.69	0.83
	447	BFI	US	0.84	0.81	0.83	0.78	0.83
Donnellan et al. (2006)	216	IPIP	US	0.83	0.79	0.89	0.72	0.87
	216	IPIP	US	0.77	0.75	0.87	0.62	0.80
Fossatti et al. (2011)	70	BFI	Italy	0.86	0.83	0.84	0.87	0.82
	141	BFI	Italy	0.81	0.90	0.85	0.76	0.75
Gorostiaga et al. (2011)	178	NEO-PI-R	Spain	0.90	0.86	0.91	0.83	0.90
Gosling et al. (2003)	114	BFI	US	0.80	0.76	0.82	0.76	0.83
Heggestad et al. (2006)	139	IPIP	US	0.80	0.88	0.91	0.87	0.84
	139	IPIP	US	0.78	0.82	0.78	0.77	0.80
	139	NEO-FFI	US	0.82	0.82	0.86	0.84	0.84
Holden et al. (2013)	46	IPIP	US	0.91	0.82	0.90	0.86	0.79
Karwowski et al. (2013)	94	BFI	Poland	0.63	0.74	0.61	0.67	0.68
Kulas et al. (2008)	118	IPIP	US	0.90	0.88	0.95	0.89	0.91
Lang (2005)	115	BFI	Germany	0.82	0.76	0.84	0.76	0.80
Langford (2003)	237	TDA	Australia	0.88	0.92	0.91	0.79	0.88
	237	TDA	Australia	0.89	0.86	0.89	0.71	0.85
Mascara & Rosen (2005)	191	IPIP	US	0.74	0.65	0.80	0.74	0.72

McCrae et al. (2011)	132	NEO-PI-R	US	0.93	0.92	0.92	0.92	0.91
Ostendorf & Angleitner (2004)	70	NEO-PI-R	Germany	0.89	0.91	0.91	0.88	0.91
	119	NEO-PI-R	Germany	0.82	0.90	0.88	0.88	0.90
Peterson (2010)	117	TDA	US	0.88	0.81	0.88	0.81	0.84
Piedmont et al. (2002)	42	NEO-PI-R	Simbabwe	0.77	0.81	0.92	0.80	0.97
	44	NEO-PI-R	Simbabwe	0.87	0.93	0.95	0.93	0.92
Rammstedt & John (2005)	57	BFI	Germany	0.83	0.88	0.93	0.78	0.80
	57	BFI	Germany	0.85	0.85	0.93	0.76	0.77
Rammstedt & John (2007)	178	BFI	US	0.65	0.70	0.79	0.69	0.76
	57	BFI	Germany	0.78	0.83	0.87	0.66	0.71
Robins et al. (2001)	107	NEO-FFI	US	0.88	0.90	0.86	0.86	0.89
Sun et al. (2011)	5,759	IPIP	Various	0.84	0.87	0.90	0.88	0.89
	2,827	IPIP	Various	0.85	0.88	0.89	0.88	0.89
	2,159	IPIP	Various	0.85	0.86	0.90	0.87	0.87
	2,102	IPIP	Various	0.84	0.86	0.89	0.87	0.87
	2,839	IPIP	Various	0.83	0.87	0.89	0.86	0.87
	1,703	IPIP	Various	0.81	0.85	0.87	0.84	0.85
	1,482	IPIP	Various	0.83	0.86	0.89	0.84	0.84
	1,410	IPIP	Various	0.78	0.85	0.86	0.83	0.85
Watson (2003)	465	BFI	US	0.81	0.79	0.89	0.79	0.83
Yang (2010)	30	NEO-PI-R	China	0.80	0.94	0.89	0.84	0.92

Note. OP = Openness, CO = Conscientiousness, EX = Extraversion, AG = Agreeableness, NE = Neuroticism. BFI = Big Five Inventory, NEO-FFI = NEO Five Factor Inventory, NEO-PI-R = NEO Personality Inventory – Revised, IPIP = International Personality Item Pool, TDA = Trait-descriptive adjectives.

Table S3.

Summary of Effects for the Reliability Generalization on GCE

Study	*N*	Instruments		Country	OP	CO	EX	AG	NE
Adebayo & Arogundade (2011)	40	BFI	NEO-FFI	Nigeria	0.78	0.80	0.82	0.74	0.92
Aluja et al. (2002)	429	NEO-PI-R	TDA	Spain	0.47	0.75	0.75	0.52	0.72
Ashton & Lee (2005)	449	IPIP	TDA	US			0.77	0.69	0.69
DeYoung et al. (2007)	480	IPIP	BFI	Canada	0.67	0.77	0.78	0.68	0.80
	481	IPIP	BFI	US	0.77	0.71	0.76	0.59	0.75
Dilchert (2007)	380	NEO-PI-R	IPIP	US	0.86	0.83	0.89	0.83	0.87
Donnellan et al. (2006)	300	IPIP	BFI	US	0.74	0.73	0.84	0.64	0.86
	300	IPIP	BFI	US	0.68	0.66	0.81	0.49	0.80
Fossati et al. (2011)	500	BFI	IPIP	Italy	0.67	0.71	0.71	0.51	0.70
	318	BFI	IPIP	Italy	0.73	0.82	0.71	0.71	0.73
	223	BFI	IPIP	Italy	0.69	0.81	0.56	0.74	0.77
Gow et al. (2005)	207	NEO-FFI	IPIP	England	0.59	0.76	0.69	0.49	0.83
Hahn et al. (2012)	598	BFI	NEO-PI-R	Germany	0.58	0.60	0.76	0.44	0.66
Heggestad et al. (2006)	303	NEO-FFI	IPIP	US	0.76	0.81	0.67	0.70	0.68
Jensen-Campbell et al. (2002)	113	IPIP	BFI	US	0.77	0.73	0.81	0.70	0.82
Lang (2005)	119	BFI	NEO-FFI	Germany	0.43	0.81	0.71	0.67	0.71
	116	BFI	NEO-FFI	Germany	0.48	0.72	0.69	0.67	0.74
Lim & Ployhart (2006)	353	NEO-FFI	IPIP	US	0.71	0.72	0.69	0.50	0.76
Miller et al. (2011)	290	BFI	NEO-PI-R	US				0.76	
Miller et al. (2013)	368	BFI	NEO-FFI	US	0.46	0.73	0.68	0.64	0.79
Mlačić & Goldberg (2007)	513	TDA	IPIP	Croatia	0.60	0.72	0.77	0.63	0.70
	513	TDA	IPIP	Croatia	0.58	0.69	0.74	0.56	0.67
Mõttus et al. (2013)	804	NEO-FFI	IPIP	England	0.59	0.75	0.62	0.56	0.79
Rammstedt & John (2005)	184	BFI	NEO-PI-R	Germany	0.71	0.74	0.73	0.63	0.82
	184	BFI	NEO-PI-R	Germany	0.72	0.80	0.78	0.63	0.86
Rammstedt & John (2007)	726	BFI	NEO-PI-R	US	0.63	0.70	0.69	0.51	0.73
	457	BFI	NEO-PI-R	Germany	0.61	0.70	0.79	0.65	0.71
Silvia & Sanders (2010)	135	IPIP	BFI	US	0.57				
Vianello et al. (2013)	14,348	IPIP	TDA	Various	0.42	0.71	0.73	0.59	0.70
Zehng et al. (2008)	300	BFI	IPIP	China	0.59	0.67	0.72	0.47	0.70
	300	BFI	IPIP	China	0.61	0.71	0.75	0.58	0.72

Note. OP = Openness, CO = Conscientiousness, EX = Extraversion, AG = Agreeableness, NE = Neuroticism. BFI = Big Five Inventory, NEO-FFI = NEO Five Factor Inventory, IPIP = International Personality Item Pool, TDA = Trait-descriptive adjectives.

Articles included in the Reliability Generalizations

[+#]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

[#]Aluja, A., García, Ó., & García, L. F. (2002). A comparative study of Zuckerman's three structural models for personality through the NEO-PI-R, ZKPQ-III-R, EPQ-RS and Goldberg's 50-bipolar adjectives. *Personality and Individual Differences, 33*, 713-725. doi:10.1016/S0191-8869(01)00186-6

[+]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

[#]Ashton, M. C., & Lee, K. (2005). Honesty-Humility, the Big Five, and the Five-Factor Model. *Journal of Personality, 73*, 1321-1354. doi:10.1111/j.1467-6494.2005.00351.x

[*]Balaji, M. S., & Chakrabarti, D. (2010). Student interactions in online discussion forum: Empirical research from 'media richness theory' perspective. *Journal of Interactive Online Learning, 9*, 1-22.

[*]Becker, G. (2006). NEO-FFI scores in college men and women: A view from McDonald's unified treatment of test theory. *Journal of Research in Personality, 40*, 911-941. doi:10.1016/j.jrp.2005.09.009

[+]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

+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

+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

+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

*Davis, J. M., & Yi, M. Y. (2012). User disposition and extent of Web utilization: A trait hierarchy approach. *International Journal of Human-Computer Studies, 70*, 346-363. doi:10.1016/j.ijhcs.2011.12.003

#DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93*, 880. doi:10.1037/0022-3514.93.5.880

#Dilchert, S. (2007). Peaks and valleys: Predicting interests in leadership and managerial positions from personality profiles. *International Journal of Selection and Assessment, 15*, 317-334. doi:10.1111/j.1468-2389.2007.00391.x

+#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

*Füller, J., Matzler, K., & Hoppe, M. (2008). Brand community members as a source of innovation. *Journal of Product Innovation Management, 25*, 608-619. doi:10.1111/j.1540-5885.2008.00325.x

+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

#Gow, A. J., Whiteman, M. C., Pattie, A., & Deary, I. J. (2005). Goldberg's 'IPIP'Big-Five factor markers: Internal consistency and concurrent validation in Scotland. *Personality and Individual Differences, 39*, 317-329. doi:10.1016/j.paid.2005.01.011

#Hahn, E., Gottschling, J., & Spinath, F. M. (2012). Short measurements of personality: Validity and reliability of the GSOEP Big Five Inventory (BFI-S). *Journal of Research in Personality, 46*, 355-359. doi:10.1016/j.jrp.2012.03.008

+#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

+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

*Huang, H. H., Mitchell, V. W., & Rosenaum-Elliott, R. (2012). Are consumer and brand personalities the same? *Psychology & Marketing, 29*, 334-349. doi:10.1002/mar.20525

*Hull, D. M., Beaujean, A. A., Worrell, F. C., & Verdisco, A. E. (2010). An item-level examination of the factorial validity of neo five-factor inventory scores. *Educational and Psychological Measurement, 70*, 1021-1041. doi:10.1177/0013164410378091

#Jensen-Campbell, L. A., Rosselli, M., Workman, K. A., Santisi, M., Rios, J. D., & Bojan, D. (2002). Agreeableness, conscientiousness, and effortful control processes. *Journal of Research in Personality, 36*, 476-489. doi:10.1016/S0092-6566(02)00004-1

*Kang, J.-Y., & Johnson, K. K. P. (2013). M-consumer segmentation: M-communication, M-distribution, and M-accessibility. *International Journal of Marketing Studies; 5*, 86-95. doi:10.5539/ijms.v5n1p86

+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

*Kautish, P. (2010). Empirical study on influence of extraversion on consumer passion and brand evangelism with word-of-mouth communication. *Review of Economic and Business Studies, 6*, 187-198.

*Korzaan, M. L., & Boswell, K. T. (2008). The influence of personality traits and information privacy concerns on behavioral intentions. *Journal of Computer Information Systems, 48*, 15-24.

+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

+#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

#Lim, B. C., & Ployhart, R. E. (2006). Assessing the convergent and discriminant validity of Goldberg's International Personality Item Pool: A multitrait-multimethod examination. *Organizational Research Methods, 9*, 29-54. doi:10.1177/1094428105283193

+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

*Matzler, K., & Mueller, J. (2011). Antecedents of knowledge sharing–Examining the influence of learning and performance orientation. *Journal of Economic Psychology, 32*, 317-329. doi:10.1016/j.joep.2010.12.006

*Matzler, K., Pichler, E., Füller, J., & Mooradian, T. A. (2011). Personality, person-brand fit, and brand community: An investigation of individuals, brands, and brand communities. *Journal of Marketing Management, 27*, 874-890. doi:10.1080/0267257X.2010.543634

+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

*Mehmetoglu, M. (2012). Personality effects on experiential consumption. *Personality and Individual Differences, 52*, 94-99. doi:10.1016/j.paid.2011.09.008

#Miller, J. D., Gaughan, E. T., Maples, J., & Price, J. (2011). A comparison of agreeableness scores from the Big Five Inventory and the NEO PI-R: Consequences for the study of narcissism and psychopathy. *Assessment, 18*, 335-339. doi:10.1177/1073191111411671

#Miller, J. D., MacKillop, J., Fortune, E. E., Maples, J., Lance, C. E., Keith Campbell, W., & Goodie, A. S. (2013). Personality correlates of pathological gambling derived from Big Three and Big Five personality models. *Psychiatry Research, 206*, 50-55. doi:10.1016/j.psychres.2012.09.042

#Mlačić, B., & Goldberg, L. R. (2007). An analysis of a cross-cultural personality inventory: The IPIP Big-Five factor markers in Croatia. *Journal of Personality Assessment, 88*, 168-177. doi:10.1080/00223890701267993

#Mõttus, R., Luciano, M., Starr, J. M., Pollard, M. C., & Deary, I. J. (2013). Personality traits and inflammation in men and women in their early 70s: the Lothian birth cohort 1936 study of healthy aging. *Psychosomatic Medicine, 75*, 11-19. doi:10.1097/PSY.0b013e31827576cc

+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.

+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.

[+#]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

[+]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

[*]Salimian, M. A., & Hosainian, R. (2012). The effects of optimism and openness to experience on employees' happiness. *Journal of Basic and Applied Scientific Research, 2*, 10876-10882.

[#]Silvia, P. J., & Sanders, C. E. (2010). Why are smart people curious? Fluid intelligence, openness to experience, and interest. *Learning and Individual Differences, 20*, 242-245. doi:10.1016/j.lindif.2010.01.006

[+]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.

[*]Terzis, V., Moridis, C. N., & Economides, A. A. (2012). How student's personality traits affect computer based assessment acceptance: Integrating BFI with CBAAM. *Computers in Human Behavior, 28*, 1985-1996. doi:10.1016/j.chb.2012.05.019

[*]Tsai, H. T., Huang, H. C., & Chiu, Y. L. (2012). Brand community participation in Taiwan: Examining the roles of individual-, group-, and relationship-level antecedents. *Journal of Business Research, 65*, 676-684. doi:10.1016/j.jbusres.2011.03.011

[#]Vianello, M., Schnabel, K., Sriram, N., & Nosek, B. (2013). Gender differences in implicit and explicit personality traits. *Personality and Individual Differences, 55*, 994-999. doi:10.1016/j.paid.2013.08.008

[*]Wang, W., Ngai, E. W., & Wei, H. (2012). Explaining instant messaging continuance intention: the role of personality. *International Journal of Human-Computer Interaction, 28*, 500-510. doi:10.1080/10447318.2011.622971

[+]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

[*]Yap, S. F., & Lee, C. K. C. (2013). Does personality matter in exercise participation? *Journal of Consumer Behaviour, 12*, 401-411. doi:10.1002/cb.1442

[*]Ying, T., & Norman, W. C. (2014). Personality effects on the social network structure of boundary-spanning personnel in the tourism industry. *Journal of Hospitality & Tourism Research*. Advance online publication. doi:10.1177/1096348014538047

[*]Zhao, B. (2011). Learning from errors: The role of context, emotion, and personality. *Journal of Organizational Behavior, 32*, 435-463. doi:10.1002/job.696

[#]Zheng, L., Goldberg, L. R., Zheng, Y., Zhao, Y., Tang, Y., & Liu, L. (2008). Reliability and concurrent validation of the IPIP Big-Five factor markers in China: Consistencies in factor structure between Internet-obtained heterosexual and homosexual samples. *Personality and Individual Differences, 45*, 649-654. doi:10.1016/j.paid.2008.07.009

[*] Included in the reliability generalization on CE

[+] Included in the reliability generalization on CS

[#] Included in the reliability generalization on GCE