Published in final edited form as:
Pers Soc Psychol Rev. 2011 February ; 15(1): 28–50. doi:10.1177/1088868310366253.

Internal Consistency, Retest Reliability, and their Implications For Personality Scale Validity

Robert R. McCrae,
Laboratory of Personality and Cognition, National Institute on Aging, NIH, DHHS, Baltimore, MD

John E. Kurtz,
Department of Psychology, Villanova University, Villanova, PA

Shinji Yamagata, and
Department of Sociology, Keio University, Tokyo, Japan

Antonio Terracciano
Laboratory of Personality and Cognition, National Institute on Aging, NIH, DHHS, Baltimore, MD

Abstract

We examined data (N = 34,108) on the differential reliability and validity of facet scales from the NEO Inventories. We evaluated the extent to which (a) psychometric properties of facet scales are generalizable across ages, cultures, and methods of measurement; and (b) validity criteria are associated with different forms of reliability. Composite estimates of facet scale stability, heritability, and cross-observer validity were broadly generalizable. Two estimates of retest reliability were independent predictors of the three validity criteria; none of three estimates of internal consistency was. Available evidence suggests the same pattern of results for other personality inventories. Internal consistency of scales can be useful as a check on data quality, but appears to be of limited utility for evaluating the potential validity of developed scales, and it should not be used as a substitute for retest reliability. Further research on the nature and determinants of retest reliability is needed.

Keywords

Reliability; validity; cross-national; Five-Factor Model; personality traits

Scale reliability is commonly said to limit validity (John & Soto, 2007); in principle, more reliable scales should yield more valid assessments (although of course reliability is not sufficient to guarantee validity). For a given set of scales, such as the 30 facets of the NEO Inventories (McCrae & Costa, in press), there is differential reliability: Some facets are more reliable than others. That fact makes it possible to test the maxim that reliability limits validity, provided that criteria of validity are chosen that are comparable across all facet scales: More reliable facets ought to be more valid. We will argue that three relevant criteria are longitudinal stability, heritability, and cross-observer agreement. Each of the 30 NEO facets is known to be more or less stable (Costa, Herbst, McCrae, & Siegler, 2000) and heritable (Jang, McCrae, Angleitner, Riemann, & Livesley, 1998), and to show evidence of some degree of cross-observer agreement (McCrae et al., 2004); however, other things being equal, more reliable facets should be more stable and heritable, and show stronger evidence

Address correspondence to Robert R. McCrae, 809 Evesham Avenue, Baltimore, MD 21212. RRMcCrae@gmail.com.
Robert R. McCrae is now retired.

of consensual validity. There are, however, different forms of reliability, of which internal consistency and retest reliability are the most prominent. In the present article we (a) assemble evidence on the stability, heritability, and cross-observer validity of NEO facets from the published literature; and (b) predict these values from estimates of internal consistency and retest reliability. These analyses allow us to assess the relative importance of these two forms of reliability.

In this article we construe *validity* broadly to refer to the quality of a scale as a measure of its intended construct. However, our discussion is limited to a consideration of convergent validity; readers should recall that discriminant validity is also an essential attribute of a good scale.

Coefficient Alpha and Retest Reliability in Contemporary Research

It would be a mistake to say that reliability is one of the fundamentals of personality assessment, because reliability is not one thing. Internal consistency, which reflects the coherence (or redundancy) of the components of a scale, is conceptually independent of retest reliability, which reflects the extent to which similar scores are obtained when the scale is administered on different occasions separated by a relatively brief interval. A scale composed of one's date of birth, height in inches, and the last two digits of one's social security number would have dismal internal consistency, but near-perfect retest reliability. Conversely, a measure of mood might have excellent internal consistency but poor retest reliability when administered on different occasions. Empirically, Chmielewski and Watson (2009) recently reported that these two coefficients were only weakly related (mean $r = .25$) in a set of personality measures. The two forms of reliability have rather different statuses in contemporary research.

It has become routine in psychological research to report coefficient alpha (Cronbach, 1951), the most commonly used measure of internal consistency. John and Soto (2007) suggested that this is true in part because it is so convenient: Whenever a multi-item scale is administered, alpha can be easily calculated. Alphas are also widely used because influential texts have suggested that they are necessary and perhaps sufficient to assess reliability. Nunnally and Bernstein (1994), for example, stated that "Coefficient α usually provides a good estimate of reliability because sampling of content is usually the major source of measurement error for static constructs" (p. 252), and they argued that it "should be applied to all new measurement methods" (p. 251-252). In contrast, they recommended "that the retest method generally not be used to estimate reliability" (p. 255). Nunnally and Bernstein are one of the sources of the rule-of-thumb that states that alphas below .70 indicate poor reliability and imply poor predictive validity.

For decades, however, psychometricians have pointed out limitations of coefficient alpha as a sufficient measure of reliability. Loevinger (1954) discussed the attenuation paradox: Higher reliability may be attained by narrowness of content that can limit predictive utility. Contemporary statisticians (e.g., McDonald, 1999), have shown that this is not really a paradox, but a consequence of certain (probably unrealistic) assumptions in classical test theory. However, the fact remains that high alphas may indicate undue narrowness or item redundancy. For this reason, Streiner (2003) cautioned against alphas greater than .90. Schmitt (1996) disputed the view that alpha should necessarily be above .70. He also noted that coefficient alpha is not a measure of unidimensionality and may underestimate reliability if a scale is multidimensional. Cronbach himself came to have misgivings about the sufficiency of coefficient alpha and moved on to a broader view of reliability embodied in generalizability theory (Cronbach, Nageswari, & Gleser, 1963; Cronbach & Shavelson,

2004). Sijtsma (2009) offered statistical reasons for his claim that alpha has very limited usefulness as a measure of reliability.

Retest reliability plays a considerably smaller role in contemporary research. Watson (2004) lamented both the sparseness and the superficiality of retest reliability studies: When researchers conducted them at all, they “almost invariably concluded that their stability correlations were ‘adequate’ or ‘satisfactory’ regardless of the size of the coefficient or the length of the retest interval” (p. 326). Roberts and DelVecchio (2000) were unable to correct for retest unreliability in their meta-analysis of longitudinal stability because so few studies reported the necessary data. However, Schmidt, Le, and Ilies (2003) renewed the call for attention to retest reliability and proposed the use of a general reliability measure that combined internal consistency and retest reliability.

Aside from the inconvenience of testing the same sample twice, retest reliability has been neglected because many researchers apparently assume that different measures of reliability are interchangeable. Conley (1984), for example, disattenuated longitudinal stability coefficients using a formula in which reliability was interpreted as “the internal consistency or period-free reliability of the measuring instrument” (pp. 12-13). Guadiano (2006), in calculating Reliable Change Index values, argued that “other indices of reliability can be used instead if test-retest reliability estimates are unavailable” (p. 15). Citing Jacobson, Wilson, and Tupper (1988) in support, Guadiano used interrater reliability as a substitute. Snell, Mallinckrodt, Hill, and Lambert (2001) calculated reliable change using the mean of the published retest value and the internal consistency in their own sample.

In all these instances, researchers attempted to understand change across administrations by correcting for unreliability, but it should be clear that only test-retest reliability is necessarily relevant to studies of longitudinal stability or change. Retest reliability sets an upper limit to longitudinal stability, because stability is reduced both by retest unreliability and by true score change. In contrast, there is no necessary association of internal consistency with longitudinal stability: the internally inconsistent sum of date of birth, height, and social security number might be constant for many years. Of course, in the usual case, internal consistency might also affect longitudinal stability because it is an index of measurement error, and error might certainly impair observed stability. This appears to be the basis of Schmidt and colleagues' (2003) view that internal consistency and retest reliability should be combined to evaluate reliability per se. Our point is that these two are conceptually different properties of scales, and “a reliability coefficient or standard error of measurement based on one approach should not be interpreted as interchangeable with another derived by a different technique” (AERA, 1999, p. 32).

A Conceptual Analysis of Reliabilities in Personality Assessment

To understand differences between the two, it is necessary to consider the factors that affect reliability (and validity) coefficients. Table 1 lists six conditions that can affect unreliability. For each, the table suggests whether it affects the absolute level of internal consistency or retest reliability (i.e., the magnitude of the coefficient compared to a fixed standard, such as . 70), and whether it affects the differential internal consistency or retest reliability (i.e., the magnitude compared to that of other scales administered to the same sample under the same conditions). The last two columns of the table suggest the effect of these sources of unreliability on absolute and relative (i.e., differential) validity coefficients.

The first determinant of reliability is the appropriateness or relevance of the item content. An Extraversion (E) scale that includes items assessing Neuroticism (N) and Conscientiousness (C) as well as E is unlikely to be internally consistent, because these three factors are largely unrelated. It is because of potential item irrelevance that internal

consistency analyses are routinely used to eliminate “bad” (i.e., invalid) items in scale development; this is one of the applications of coefficient alpha that has proven most useful. Note, however, that item irrelevance need not affect retest reliability: Including an N item in an E scale should not reduce retest reliability, because N, like E, is a stable trait and N items are likely to yield the same response on two occasions. When a set of scales, like the facets of the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992), are examined, the relative internal consistency may vary because the degree of item relevance varies across scales; again, however, differential retest reliability should not be affected by such variation across scales. Scales with low alphas due to item irrelevance will presumably show lowered validity coefficients with external criteria, because they contain content unrelated to the construct of interest.

A second determinant of reliability is item content heterogeneity: whether the items in a scale cover many different aspects of a trait or focus on only a few. An E scale with three items assessing friendliness, energy level, and optimism will show lower internal consistency than one with three friendliness items. As with item invalidity, item heterogeneity will affect alpha both absolutely and differentially, but it should not affect retest reliability. Whether it leads to lower scale validity has long been debated (e.g. Boyle, 1991), and remains an empirical question (Eichler & Kurtz, 2008).

Retest reliability is affected by a third property of scales, their state variation over time, called *transient error* by Schmidt and colleagues (2003). What intrinsic properties of a scale would lead to different scores on different occasions? The timeframe of items is one obvious candidate. Items such as “As a child, I enjoyed make-believe” ought to be more stable than “Nowadays I enjoy make-believe,” because the answer to the former ought not to change, whereas the answer to the latter might. The use of single adjective items versus statements may also affect retest reliability, although it is not clear why (Watson, 2004). Other, subtler, and perhaps more psychologically interesting properties may have to do with the traits themselves. Fleeson (2001) argued that the state expression (or perception) of traits varies around a central tendency that represents the trait level. It is possible that such fluctuations affect responses to questionnaires, and it is also possible that some traits show intrinsically higher state variation than others. More temporally reliable traits may be those that are more central to the individual's identity, or more observable, or more evaluatively neutral (John & Robins, 1993). Unless one wishes to argue that the respondent's state may change between completion of the first and last items in a scale, there is no reason to think that state variation will affect internal consistency. State variation does, however, affect validity: Although stable scores may or may not be valid, scores that vary across time cannot always be true scores.

Unreliability is often characterized as the result, and thus an indication, of error of measurement. Systematic biases that in effect add a constant to each item (e.g., perhaps, self-enhancement) do not generally affect either internal consistency or retest reliability; this is one reason that reliability is traditionally distinguished from validity. But random error certainly affects reliability, and it has several sources. Characteristics of the respondents' test-taking behavior can affect response accuracy, and represent a fourth source of unreliability, respondents' error. Alpha for any test will likely be reduced if respondents have limited literacy or intelligence (Allik, Laidra, Realo, & Pullmann, 2004), if they are responding in a second language in which they are not fully fluent, or if they are uncooperative and respond randomly. Such random responding will also affect retest reliability. But note that it should not affect the differential reliability of a set of scales; all facets should suffer equally from limitations in the accuracy of the respondents.[1]

A fifth source of unreliability—item ambiguity—is due to problems with the items, or the items in interaction with the respondents. Items that are difficult to understand because of obscure vocabulary, ambiguous or double-barreled phrasing, or the use of negations or complex sentence structure may confuse respondents and reduce both the internal consistency and retest reliability of scores. There may also be interactions of item and respondents' characteristics (Schmidt et al.'s, 2003, *specific factor error*): A sophisticated vocabulary may not affect the reliability of tests administered to college students, but may do so when given to adolescents or adults with limited literacy. Item phrasing is scale-specific; for example, the scales of the NEO-PI-R are known to differ in readability (Schinka & Borum, 1994). Thus, problematic items may contribute to the differential reliability of scales. They may also affect the validity and differential validity of scales; an item is unlikely to be valid if it is misunderstood.

Reliability and validity coefficients are also affected by characteristics of the sample, most notably trait variance. In a sample of clinically depressed patients there might be a restriction of range on N scores (all might score high), but not on other personality factors. Observed coefficients for N would then be attenuated, whereas those of other factors would not; both absolute and differential values would thus be affected.

The relative importance of these six sources of unreliability is unknown and will presumably vary across different samples and instruments. But it should be clear from Table 1 that internal consistency and retest reliability are not interchangeable measures of the quality of a scale. The two indicators will be similar only if both are affected by the same sources of unreliability. With regard to absolute reliability, only respondents' error, item ambiguity, and sample variance affect both coefficients. With regard to differential reliability, only item ambiguity and sample variance will tend to make the two similar. Other sources of unreliability are unique to each form of reliability and should contribute to differential effects.

Reliability in the NEO Inventories

In this article we examine the relation of reliability to validity using data from the NEO Inventories (McCrae & Costa, in press). The NEO-PI-R is a 240-item measure of the five major factors of personality, Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C), and of 30 specific traits, or facets, that define the factors (facet labels are given in Table 2). There are two forms of the instrument: Form S, for self-reports, is phrased in the first person; Form R, for observer ratings, is phrased in the third person. Recently, the NEO-PI-3 was developed; it utilizes replacements for 37 of the original items to lower the Flesch-Kincaid reading level (McCrae, Costa, & Martin, 2005).

These NEO Inventories are convenient instruments to use for the present analyses for two reasons. First, the facet scales are internally comparable in many respects: All have eight items, all are roughly balanced in keying to minimize the effects of acquiescent responding, and items from all facets are evenly distributed across the 240 items of the test. The fact that scales are of equal length is particularly important, because coefficient alpha is a function of both the number of items and the mean interitem correlation. Our design controls for—but does not investigate—the influence of scale length on differential validity. Second, there is a wealth of validity data on three relevant criteria across cultures, ages, and methods of

[1]Scale statistics are determined by properties of the sample as a whole, not the individual respondents, so alpha offers no clue as to which individuals have responded inconsistently. Attempts to identify such respondents (and thus to eliminate invalid protocols) have had limited success (Kurtz & Parrish, 2001).

measurement. (Useable data from other personality inventories are much scarcer; see Appendix A.)

Estimates of internal consistency have been available since the publication of the NEO-PI (Costa & McCrae, 1985). The Manual for the NEO-PI-R includes coefficient alphas for both Form S and Form R, though these values have sometimes been a point of criticism (e.g., Cheung et al., 2008). A common rule-of-thumb (e.g., Nunnally & Bernstein, 1994) states that coefficient alpha should be at least .70, but 13 of the 30 NEO-PI-R facets failed to reach that standard (alphas = .56 to .68). Because all 30 scales have amassed a large body of validity evidence (e.g., Costa, McCrae, & Dye, 1991; Jang et al., 1998; McCrae & Costa, 1992), the wisdom of the rule-of-thumb may be called into question, and a major concern of this article is the impact of differences in internal consistency on the magnitude of validity coefficients for the three criteria we examine here. The wide range of internal consistencies among NEO scales (.56 to .81) is advantageous for the present study; if all alphas were uniformly high, they could not show differential effects on validity.

Less information is available about the retest reliability of NEO facets. The NEO-PI-R and its precursors were first used in a longitudinal study, and six-year stability coefficients were published twenty years ago (Costa & McCrae, 1988). Because the observed values were quite high (.63 to .83 in the total group), and because long-term stability sets a lower limit to retest reliability, the test authors asserted that retest reliability must be satisfactory, and they did not conduct short-term retest studies. They did report six-month retest reliability in a sample of 31 men and women who completed only the N, E, and O scales, and, as expected, observed values were generally good (.66 to .92).

Subsequent research has examined retest reliability or stability in clinical samples (Carter et al., 2001; Costa, Bagby, Herbst, & McCrae, 2005; Yang et al., 1999), but it is not clear that those data would be relevant to non-clinical samples; in particular, retest unreliability in psychiatric samples would likely be confounded with real change due to treatment or the natural course of the disorder (Costa et al., 2005; Piedmont, 2001). As detailed below, retest reliability for the NEO-PI-R has usually been estimated from multiple administrations of the instrument in a longitudinal sample (Terracciano, Costa, & McCrae, 2006). The only study that conducted a short-term test-retest study in an American sample appears to be Kurtz and Parrish (2001), and that article did not report reliabilities for the individual facet scales. In this article, we present facet-level retest reliability estimates from Terracciano and colleagues and from Kurtz and Parrish, and we provide some evidence of their generalizability by comparing them to composite retest data from cross-national studies.

Three Validity Criteria

The design employed in the present study is based on the hypothesis that, in general, more reliable scales will be more valid. To test that hypothesis, reliability and validity must be quantified. It is relatively simple to compute various reliability coefficients, but difficult to quantify validity. Statisticians rightly assert that there is no single measure of validity, because a scale may be valid in some populations, but not others, or more valid for some purposes than others. These issues are secondary here; more reliable scales are presumably more valid across populations and applications. The major problem here is the comparability of validity estimates across a range of scales with different reliabilities.

Validity must be judged against some external criterion; the most familiar criteria are alternative measures of the same construct (in studies of construct validity) and theoretically relevant outcomes (in studies of predictive validity). In much personality research, new scales are correlated with established scales purporting to measure the same or similar traits to provide evidence of convergent validity. Table 6 of the NEO-PI-R Manual reports such

correlations for each of the 30 facets. For example, Assertiveness correlates .64 with the Dominance scale of the Personality Research Form (PRF; Jackson, 1984), whereas Modesty correlates .32 with the Abasement scale of the PRF. It would be premature, however, to conclude from these findings that the Assertiveness scale is more valid than the Modesty scale, because it is possible that the constructs of Assertiveness and Dominance are more closely related theoretically than the constructs of Modesty and Abasement, or that the PRF Dominance scale—the criterion—is itself more reliable and valid than the PRF Abasement scale. In short, these two PRF scales may not provide comparable criteria for assessing the validity of NEO facets. The same problem would arise if we considered behavioral measures or life outcomes as validity criteria, because these would presumably vary across traits in their relevance and reliability. Validity is normally assessed in terms of a nomological network of associations that is often difficult to compare quantitatively to the nomological networks of other scales.

We will return in the Discussion to a consideration of one possible solution to this problem: Meta-analyses that combine data across a wide range of criteria. By averaging across a variety of different operationalizations of validity, differences in the appropriateness of the criteria for different facets is likely to be minimized. If the NEO Assertiveness scale was (on average) strongly correlated with five different dominance scales, whereas the NEO Modesty scales was only modestly correlated with seven measures of humility, we could be more confident that the NEO Assertiveness scale is indeed the more valid.

In this article we adopt a different solution to the problem of identifying validity criteria that are comparable across facets: We examine longitudinal stability, heritability, and cross-observer agreement. Provided the same retest intervals, the same twin pairs, and the same sets of observers are used for all facets, these statistics can provide a basis for comparing the validities of scales.

Agreement across observers (e.g., correlations between self-reports and spouse ratings; McCrae, 1982) is widely construed as evidence of validity, especially valuable because it avoids some of the problems introduced by using the same method of measurement for the scale of interest and its criterion. But stability and heritability are not usually thought to provide evidence of validity. In fact, however, they are useful and appropriate criteria in the present context. Personality traits are, by definition, enduring dispositions; measures that fail to show long-term stability cannot be valid trait measures. There are both empirical (Loehlin, 1992) and theoretical (McCrae & Costa, 2008) reasons to argue that personality traits are heritable; from this perspective, valid personality measures must show evidence of heritability.[2]

It is essential to note that, for the purposes of this article, the observed validity coefficients are of primary interest, rather than the true-score values. There is no reason to think that the traits more reliably measured by the NEO Inventories will be inherently more stable, heritable, or observable than other traits, but there is reason to think that poor measurement will consistently attenuate the observed stability, heritability, or cross-observer correlations of traits. Across a broad range of traits, those more reliably assessed ought to show higher validity coefficients, and conversely, those that show higher validity coefficients must have been more reliably assessed. This is the rationale for using these validities as criteria for evaluating alternative ways of assessing reliability.

[2]Some theorists would not consider heritability a necessary attribute of a personality trait, and thus would dispute the claim that heritability is a form of validity. Most, however, would acknowledge that many traits are in fact heritable. If so, then heritability is a characteristic that can be used (as we do here) to investigate the reliability of trait measures: Other things being equal, reliable scales should lead to higher observed estimates of heritability than do unreliable scales.

As a criterion, longitudinal stability is different from heritability and cross-observer agreement in one important respect: It is conceptually and operationally linked to retest reliability. To assess either, a scale must be administered twice; the correlation between the two administrations is interpreted as retest reliability if the retest interval is short enough that no real change in trait level is to be expected (Watson, 2004), and as stability if the retest interval is longer. In this article, we regard retests from one week to one year as assessments of reliability. Because of this operational link, we have a strong expectation that differential retest reliability will predict differential stability. However, longitudinal stability might also be influenced by other sources of error that are indexed by internal consistency; this is the assumption behind the use of coefficient alpha to evaluate reliable change (e.g., Snell et al., 2001). In the usual studies of heritability, relatives (usually twins) each provide self-reports on a single occasion, and in studies of cross-observer agreement, each observer completes the assessment once. These two criteria are thus operationally independent of retest reliability.

Behavior geneticists routinely acknowledge that their estimates of heritability are limited by measurement unreliability (Loehlin, 1992), but there is no accepted procedure for estimating true heritability net of reliability. Rushton, Fulker, Neale, Nias, and Eysenck (1986) used coefficient alpha to estimate corrected heritabilities, as did Beatty, Heisel, Hall, Levine, and La France (2002) in their meta-analysis. Jang and colleagues (1998) reported corrections for retest reliability; Riemann, Angleitner, and Strelau (1997) used a structural equation modeling (SEM) approach that corrected for interrater reliability. In this article we compare the associations of internal consistency and retest reliability with heritability, to determine which, if either, should be used to correct estimates of true score heritability.

Internal consistency should predict differential heritability if the principal sources of unreliability are item irrelevance, item ambiguity, or sample variance (see Table 1); the effect of item heterogeneity on correlations with heritability is an empirical question. Retest reliability should predict differential heritability if state variations in test responses, item ambiguity, and sample variance chiefly determine reliability.

Cross-observer validity coefficients form the third set of criteria. Although some studies compare ratings of the same trait made with different instruments (e.g., a self-report questionnaire with global ratings by peers; Jackson, 1984), the requirement of the present study is that both observers use the same instrument (e.g., NEO Form S and NEO Form R), so that the criteria will be comparable across facets. Interrater agreement is an important indicator of *reliability* for some assessments (e.g., behavior codings), but it is not usually considered as a measure of reliability in personality assessment. Cross-observer agreement (e.g., between two peers who both know the target well, or between self-reports and spouse ratings) is taken more often as evidence of consensual validity (e.g., McCrae et al., 2004). Different judges normally know the target in quite different circumstances or from different perspectives; their judgments may be based on very different information. When two such judges are asked to describe the same person, what we are interested in is not their ability to make similar personality inferences from the same cues (interrater reliability), but the convergence of their inferences based on different cues, a convergence that presumably says something about the true characteristics of the target. In this article we treat cross-observer agreement not as a form of reliability, but as a validity criterion that might be attenuated by unreliability.

As with heritability and longitudinal stability, it is reasonable to hypothesize that cross-observer agreement will be limited by retest unreliability. If markedly different scores are obtained on different occasions for characteristics (like traits) that are stable over relatively short intervals, it must mean that the observed score on any given occasion is likely to miss

the true score. Certainly, high retest reliability does not necessarily imply that the assessments are valid, but low retest reliability implies a degree of error that should limit cross-observer agreement. Low internal consistency should attenuate differential consensus among observers, if alpha is mostly a function of item invalidity or item ambiguity. If alpha is chiefly an index of item heterogeneity, it is not clear a priori whether it will affect differential validity.

The Reliability of the Criteria

We have argued that a particular strength of our design is that our criteria are comparable across the 30 NEO facets. In contrast to the miscellaneous validity coefficients that might be gathered in a meta-analysis, our data are in each comparison based on the same sample and method of measurement, using criterion scales that are of identical length and format for each of the 30 facets. Perhaps most crucially, each criterion is of equal theoretical relevance: PRF Dominance may be conceptually closer to NEO Assertiveness than PRF Abasement is to NEO Modesty, but surely NEO Assertiveness and NEO Modesty as criteria (in the longitudinal retest, the report of the second twin, or the observer rating) are equally related NEO Assertiveness and NEO Modesty, respectively, as predictors.

The astute reader, however, may have noticed that the criteria used here are not comparable across facets in one respect: The NEO scales used as criteria themselves differ in reliability. However, this is not a problem for our design; it is in fact a source of power in the present analysis, essentially squaring the effect of unreliability on observed validity. For example, Openness to Actions has a coefficient alpha of .58 in self-reports and .60 in observer ratings (Costa & McCrae, 1992), whereas Depression has corresponding coefficients of .81 and .81. If internal consistency limits validity, one would expect that cross-observer agreement should be much lower for Openness to Actions, where an unreliable (by the criterion of alpha < .70) scale is used to predict an unreliable criterion, than for Depression, where a reliable scale predicts a reliable criterion.

We tested this intuitive notion in a small simulation study that examines the effect of the reliability of the criterion on the association between predictor reliability and observed validity. In each of 50 replications, we created a random variable r_p with a mean of .75 and standard deviation of .06 to represent the reliabilities of the 30 predictor facets; a second random variable r_c (M = .75, SD = .06) to represent the reliabilities of the 30 criteria; and a third random variable, V_T (M = .55, SD = .06) to represent the 30 true-score validity coefficients. (The numerical values for the simulation were suggested by the data in Table 2.) To estimate results of a hypothetical meta-analysis in which a variety of criteria with different reliabilities are used, we first calculated an observed validity coefficient, V_O. For this we used the standard formula for the attenuation of validity due to unreliability in both the predictor and the criterion:

$$V_O = V_T * \left(r_p * r_c\right)^{1/2}.$$

We then correlated r_p with V_O across the 30 simulated facets to assess the degree to which reliability in the predictor affects observed validity coefficients. The median value across the 50 replications was a modest .32. Next, to estimate results of the present design, in which the reliability of each criterion (e.g, informant rated NEO Depression) is equal to the reliability of its predictor (e.g., self-reported NEO Depression)—that is, $r_c = r_p$—we calculated $V_{O(c \equiv p)}$ as

$$V_{O(c\equiv p)} = V_T * \left(r_p * r_p\right)^{1/2} = V_T * r_p.$$

The median correlation of r_p with $V_{O(c \equiv p)}$ across 50 simulation replications was .60. Clearly, the present design is likely to be very sensitive to the effects of reliability on observed validity. This should allow us to determine which form or forms of reliability in fact attenuate validity.

Universal Psychometrics?

The preceding discussion has been predicated on the assumption that it makes sense to consider reliability and validity coefficients as characteristics of trait measures themselves, whereas psychometricians are careful to point out that these are properties of measures used in particular samples or populations (Streiner, 2003). Table 1 acknowledges that the variance of traits in a sample can affect these coefficients, but it is conceivable that they will also vary substantially with the age or gender of the respondent or the method of measurement (e.g., self-report vs. observer rating). Certainly, it cannot be assumed that psychometric properties will be retained when instruments are translated into foreign languages. It might be argued that the relation of reliability to validity can only properly be evaluated when both are assessed within the same sample (or population). Although that would be in some respects an ideal design, it would not be directly relevant to the general practice of psychological assessment. Researchers may, and probably should, examine reliability (at least internal consistency) in their own samples, but they must depend on available assessments of reliability when choosing instruments to administer. Individual clinicians may never have a sufficient database to examine reliability in their own client populations. For practical purposes, it is necessary to characterize the reliability and validity of scales independent of sample characteristics. At the same time, it is essential to assess the generalizability of these estimates: Is a given retest reliability estimate applicable to children as well as adults, in Chinese- as well as English-language versions of the test?

Although “local norms” for reliability may be ideal, there are empirical reasons to think that estimates of reliability and validity for NEO scales may be widely generalizable. Many properties of personality traits, such as factor structure and gender differences, appear to be universal (McCrae et al., 2005a). The same may prove true for differential reliabilities and validities. For example, in the Russian version of the NEO-PI-R (Martin, Costa, Oryol, Rukavishnikov, & Senin, 2002), the facet with the lowest alpha (.43) was Tender-Mindedness, which also had the lowest alpha (.56) in the American sample (Costa & McCrae, 1992). McCrae and colleagues (2004) explicitly addressed the generalizability of differential reliability and cross-observer validity by examining data from American, German, Russian, Chinese, and Czech samples. They reported that the European samples (but not the Chinese) showed similar profiles of self-other agreement, and that all samples showed similar patterns of internal consistency. This article extends those analyses to larger samples and other criteria.[3]

We focus on three indicators of internal consistency (from American self-reports and international observer ratings) and two estimates of retest reliability (from three measurements in a longitudinal study and from a one-week test-retest design). We assess the generalizability of these estimates by comparisons with other published data, to see if differences in facet scale internal consistency and retest reliability are similar across

[3]Some of the data analyzed in McCrae et al. (2004) are included in the Cross-Observer values in Table 2 and in Table 4.

different cultures, languages, ages, and methods of measurement. We then compile data on stability, heritability, and cross-observer validity from available sources; assess consistency of facet scale differences in these criteria across samples; and create composite validity criteria. These composite scores offer evidence on the universality of the criteria and provide the very large *N*s that may be needed to yield precise estimates of validity coefficients (cf. Watson, 2004, on the need for large samples in retest reliability studies). Reducing error in these criteria should allow us to detect relatively subtle differential effects of interest. Finally, we relate our estimates of reliability to the three criteria by correlating reliability indicators with validity criteria across the 30 facets of the NEO-PI-R;[4] this allows us to assess how and to what extent internal consistency and retest reliability affect validity. It must be stressed that the present article is concerned with differential reliabilities of NEO-PI-R facets; we will return in the Discussion to a consideration of the interpretation of absolute levels of reliability.

Development of Facet Reliability Indicators

Internal Consistency

Data on internal consistency were obtained from American adults (Costa et al., 1991), as reported in the NEO-PI-R Manual; college students (Kurtz & Parrish, 2001); and a multi-national project (McCrae et al., 2005a, -b). Self-report data from the college study provide new normative information for American college-age respondents. Observer rating data from the multi-national study provide a test of the generalizability of differential reliability across methods, cultures, and languages.

The NEO-PI-R Manual gives coefficient alphas for Form S domains and facets from a sample of 1,539 adults aged 21 to 64 who had at least a high school education (Costa et al., 1991). These data are reproduced in the first column (Manual) of Table 2.

Form S internal consistency data from a study by Kurtz and Parrish (2001) are shown in the second data column (College) of Table 2. Kurtz and Parrish administered the self-report version of the NEO-PI-R to 132 college students (84 males) twice. Coefficient alphas were calculated on both occasions with similar results ($r = .88$, $N = 30$ facets, $p < .001$); the table reports the mean of the two estimates.

For the Personality Profiles of Cultures Project (PPOC; McCrae et al., 2005a, -b), Form R of the NEO-PI-R was administered to 12,156 respondents who described a member of their culture whom they knew well. Raters were randomly assigned to select a college-age male or female or an adult (age 40+) male or female target. Data were obtained from 51 cultures, using the instrument in English or one of 28 translations. *N*s for the samples ranged from 106 to 919. Analyses in this and other samples (e.g., McCrae et al., 2004) suggest that the NEO Inventories maintain their factor structure, gender differences, age differences, and cross-observer validity in translation.

Cross-cultural personality research poses several potential problems. Translations may be imperfect, items may have less relevance in different cultural contexts, and respondents, particularly in non-Western cultures, may be unfamiliar with questionnaires (Marsella, Dubanoski, Hamada, & Morse, 2000). Although such problems cannot be avoided, they can be tracked by assessing data quality in each culture. In PPOC, data quality was assessed in

[4]Because the six facets that define each NEO domain are intercorrelated, the 30 observations are not independent, and the resulting statistical tests are not exact. This is unlikely to affect the interpretation of results, because our emphasis is on the relative strength of the associations of the two forms of reliability with validity indicators, and these are equally affected by non-independence. Correlations across the 30 NEO facet scales have been widely used, and appear to yield robust and meaningful results (e.g., McCrae et al., 1999; McCrae et al., 2005b).

each sample by a Quality Index based on repetitive responding, missing data, acquiescence or naysaying, whether the test was administered in the respondents' native language, whether the translation was published, and a judgment by the administrator concerning problems with the task (McCrae et al., 2005a, -b). These indicators formed an internally consistent index (alpha = .76), suggesting that all provide operationalizations of data quality. Coefficient alphas for the NEO-PI-R facets might also be expected to reflect in part data quality (to the extent to which they are determined by respondents' error and item ambiguity); relating alphas to the Quality Index across cultures allows us to assess whether and to what extent facet internal consistencies are affected by data quality.

Coefficient alpha was computed for each culture for all 30 facets. Eight of the 1,530 coefficients were negative (Self-Consciousness in Nigeria, Impulsiveness in Morocco and Nigeria, Openness to Fantasy in Nigeria, Openness to Actions in Uganda and Morocco, and Openness to Values in Nigeria and Botswana); these were recoded as zero. A factor analysis using facet alphas as variables and cultures as cases showed a single, general factor accounting for 80.4% of the variance, on which all facets had loadings greater than .74. This factor reflects the fact that data were consistently more reliable in some cultures than in others. As expected, across the 51 cultures, scores on this factor correlated $r = .76$, $p < .001$, with the data Quality Index. Internal consistency does not appear to be related to the language of the test per se, but to the conditions of testing and the appropriateness of the language to the population tested: The mean coefficient alpha using the French version was .78 in Switzerland but .63 in Burkina Faso; using the English version it was .78 in England but only .25 in Nigeria.

Although absolute internal consistency varies across languages and cultures, differential internal consistency of facets may be generalizable. To evaluate this, we examined interculture agreement across the profile of traits. For example, we correlated the 30 alphas for the U.S. with the 30 corresponding alphas for Canada, and found $r = .92$, $p < .001$. There are 1,275 such correlations between pairs of cultures, but these can be summarized by a scale reliability analysis, in which each culture is considered an item, and each facet a case. The 51-item analysis showed an intraclass correlation (under the SPSS consistency model) for pairs of cultures of .57, $p < .001$, alpha = .99. Thus, there is substantial agreement across cultures on which facets are more and less internally consistent. All cultures, including those with low absolute levels of reliability (such as Nigeria), showed similar patterns, with all corrected item/total correlations greater than .55. For this reason, it is reasonable to collapse the samples to provide a single estimate of internal consistencies for Form R NEO-PI-R domains and facets for the multi-national PPOC sample. These are reported in the third data column of Table 2.

In all three samples, alpha is lowest (< .70) for Self-Consciousness, Activity, Openness to Actions and Values, and Tender-Mindedness; alpha is highest (≥ .75) for Angry Hostility, Vulnerability, Openness to Aesthetics and Ideas, Trust, and Self-Discipline. Internal consistency does not vary systematically by domain, and all five domains are highly consistent.

Table 3 shows intercorrelations across the 30 facets for estimates of internal consistency in the three indicator samples (Manual, College, and PPOC) and five additional samples: PPOC male and female targets separately, adult observer ratings (Costa & McCrae, 1992,Table 5), and observer ratings and self-reports from adolescents (14-20 years) using the more readable modification of the NEO-PI-R, the NEO-PI-3 (McCrae, Costa, & Martin, 2005). Moderately strong correlations are seen across samples, genders, methods of measurement, age groups, and languages, suggesting that the differential internal consistency of NEO Inventory facets is highly generalizable.

Retest Reliability

To date, there are no published short-term retest reliability values for NEO-PI-R facet scales in non-clinical American samples. Previous research (e.g., McCrae, Yik, Trapnell, Bond, & Paulhus, 1998) has relied on two-year retest data, or on estimates of retest reliability based on Heise's (1969) path analytic formula (Terracciano, Costa, & McCrae, 2006). Heise argued that observed stability coefficients are a function of the intrinsic retest reliability of the instrument and the decay of true stability; if three time points are observed, reliability can be estimated as $(r_{12} * r_{23}) / r_{13}$. Terracciano and colleagues (2006) analyzed data from 520 adult American participants in an ongoing longitudinal study of aging (initial *M* age = 64.0, *SD* = 10.8 years; 57% males) who had three computer administrations of the NEO-PI-R with an interval of at least six years (*M* = 10.1 years) between the first and last administration. Heise estimates of these reliabilities are given in the fourth data column of Table 2.

More conventional assessments of retest reliability are available from the data collected for the Kurtz and Parrish (2001) study, which also allow a comparison of internal consistency and retest reliability estimates derived from the same sample. Self-reports were obtained twice, with a retest interval of 7 to 14 days (*M* = 8.5 days). These (roughly) One-Week values are given in the fifth data column of Table 2. The correlation of these values with the Heise estimates for the 30 facets was $r = .40$, $p < .05$, suggesting modest generalizability over samples, ages, and methods of estimating retest reliability.

Data on short-term retest reliability from non-clinical samples outside the U.S. are available for several translations of the NEO-PI-R. Ostendorf and Angleitner (2004) reported one- and two-month retest reliabilities for the German self-report version (*N*s = 70, 119). Martin and colleagues (2002) reported one-year retest reliabilities for both self-report and observer rating forms of the Russian NEO-PI-R (*N*s = 60). In Hong Kong, 41 undergraduates took a Chinese translation of the NEO-PI-R with a two-week retest interval, and 40 completed the English version twice (McCrae, Yik, et al., 1998). Mainland Chinese students (*N* = 66) aged 16-20 took the Mandarin version of the NEO-PI-R twice over an interval of three months (Dai & Wu, 2005). Bleidorn, Kandler, Riemann, and Angleitner (2009) reported longitudinal retests for successive five-year intervals in a sample of 172 German twin pairs; Heise estimates of reliability can be calculated from these data. Treating these samples as items in a scale analysis showed intraclass agreement of .31 between pairs of samples, $p < .001$, alpha = .78. The mean of these eight retest samples correlated $r = .39$ with the Heise estimates of reliability, $p < .05$, and $r = .57$ with the One-Week estimates, $p < .001$. These data provide evidence of some generalizability of the differential retest reliability estimates across highly diverse languages and cultures.[5, 6]

[5]The only other relevant sample we could identify consisted of data collected by Piedmont, Bain, McCrae, and Costa (2002), who administered a Shona translation of the NEO-PI-R to 42 Zimbabweans twice over a one-week interval. A number of problems were noted in the back-translation of this instrument, but it was not feasible to postpone data collection due to the deteriorating political situation. As a result, there were several indications that the quality of the Shona translation was sub-optimal. Scale reliability analyses showed that including the Shona retest data decreased alpha of the composite to .74. If the Shona data are included in the composite measure of reliability, its correlations are .35, *n.s.*, with Heise reliability and .58, $p < .01$, with One-Week reliability. Shona reliabilities themselves are unrelated to either Heise or One-Week reliabilities (rs = −.05, .17, *n.s.*).

[6]Kurtz, Lee, and Sherker (1999) provide data on six-month retest reliability of Form R ratings of an adult target in a sample of 109 American undergraduates. Although comparable in magnitude to other retest estimates (*Mdn* = .80 for 30 facets), differential reliability was unrelated to either Heise or One-Week Form S reliabilities. These American Form R data are in contrast to the Russian Form R data (Martin et al., 2002), which were significantly related to One-Week reliabilities ($r = .45$, $p < .01$) and marginally related to Heise reliabilities ($r = .36$, $p < .06$). Additional short-term retest reliability studies of informant ratings are needed.

Development of Facet Validity Criteria

Longitudinal Stability (Rank-Order Consistency)

Several American samples provided longitudinal data. Six- to nine-year retest stability data were taken from a study of 1,779 men and 495 women; five-year data were from a supplementary sample of 367 spouses of these participants (Costa, Herbst, McCrae, & Siegler, 2000). Ten-year stability data from adults initially aged 30-50 (N = 151), 51-65 (N = 259), and over 65 (N = 266) were taken from Terracciano and colleagues (2006). Data from a sample intended to be representative of East Baltimore were available from 505 respondents initially aged 30 to 88 and retested after about 8 years (Löckenhoff et al., 2008). In addition, five-year longitudinal data were available from a study of German twins (N = 754; 148 males) aged 21 to 74 (Ostendorf & Angleitner, 2004). The means of stability coefficients for Twin A and Twin B were used. Treated as items of a scale, these eight samples showed considerable agreement on the 30-facet profile, with an intraclass correlation between pairs of samples of .60, $p < .001$, alpha = .92. The mean five-to-ten-year stability coefficients are given in the sixth data column of Table 2. As assessed by the NEO-PI-R, Dutifulness, Tender-Mindedness, and Competence appear to be the least stable traits, whereas Assertiveness and Openness to Ideas and Aesthetics are the most stable.

Heritability

Additive heritability (the proportion of variability in a population attributable to the summed effects of individual genes, ignoring genetic dominance effects) was estimated from twin studies in three countries: Canada (N = 450 pairs), Germany (N = 806 pairs), and Japan (N = 646 pairs; Yamagata et al., 2006). For each, heritability (h^2) was estimated from an additive genetic and non-shared environment model. Estimates of heritability from an additive model were also available from a study of 5,657 family members in Sardinia (Pilia et al., 2006) who completed the Italian version of the NEO-PI-R. Over 30,000 relationship pairs were identified (parent/child, uncle/nephew, etc.), and estimates of h^2 were obtained by comparing the scores of these relatives. Treated as items of a scale, these four samples showed considerable agreement on the 30-facet profile, with an intraclass correlation between pairs of samples of .47, $p < .001$, alpha = .78; all samples showed significant corrected item/total correlations. There is thus evidence of differential heritability of facets across nations, languages, and family designs. The mean h^2 values across the four samples are given in the seventh data column of Table 2. As assessed by the NEO-PI-R, the least heritable traits ($h^2 < .33$) are Impulsiveness, Trust, Altruism, Tender-Mindedness, Competence, and Deliberation; the most heritable traits ($h^2 > .39$) are Warmth, Gregariousness, Assertiveness, Openness to Aesthetics, Ideas, and Values, and Self-Discipline.

Cross-Observer Agreement

The NEO-PI-R offers both self-report (Form S) and observer rating (Form R) versions; the latter consists of the Form S items rephrased in the third person. Studies summarized here correlate Form S with Form R data, and the resulting cross-observer correlations are presumably limited by the unreliability of both versions. Table 3 shows that differential internal consistency is quite similar for the two versions; it is not yet clear whether short-term retest reliability is also similar (see Footnote 6).

The NEO-PI-R Manual (Costa & McCrae, 1992) provides cross-observer correlations for peer/peer (N = 193), peer/self (N = 250), and spouse/self (N = 68) ratings. German studies (Ostendorf & Angleitner, 2004) provided peer/peer (N = 750) and self/mean peer (N = 750) ratings. Self/observer correlations were available for Russian (N = 800) and Czech (N = 714) samples (McCrae et al., 2004). Data from the NEO-PI-3 came from two samples: adult (self/

spouse, $N = 532$; McCrae, Martin, & Costa, 2005) and adolescent (self/sibling, $N = 180$; McCrae, Costa, & Martin, 2005) samples. Piedmont (1994) provided data on peer/peer ($N = 97$) and mean peer/self ($N = 101$) agreement. McCrae, Stone, Fagan, and Costa (1998) reported data from 94 adults with self/spouse agreement. Scale reliability analyses of these samples showed that the Piedmont peer/peer data did not fit the pattern of the other samples, and they were discarded.[7] The remaining 11 samples showed moderate agreement, with an intraclass correlation of .35 between pairs of samples, $p < .001$, alpha = .86. The mean values are presented in the eighth data column of Table 2. Observers agree least ($r < .38$) on Self-Consciousness, Vulnerability, Straightforwardness, Tender-Mindedness, Competence, Dutifulness, and Deliberation; they agree most ($r > .50$) on Warmth, Assertiveness, Excitement Seeking, Openness to Aesthetics, and Compliance.

Predicting and Correcting Validity Criteria

Table 4 reports correlations among the first eight columns in Table 2. The most striking findings are the large correlations between the two indicators of retest reliability and the three validity criteria. These values are close to the median value (.60) predicted by the simulation study. It is not surprising that retest reliability is related to longitudinal stability, but it is remarkable that it also predicts the differential heritability and cross-observer validity of traits. For example, the temporal consistency of responses across two occasions among American undergraduates is a predictor of differential trait heritability in Japan and Sardinia, five-year stability in Germany, and self/observer agreement in Russia and the Czech Republic (rs = .42 to .56, $N = 30$, ps < .05).

A more complete matrix of intercorrrelations is presented in Table B1, Appendix B. That table shows that the Heise estimates of retest reliability from Bleidorn and colleagues' (2009) German data predict differential stability and heritability, and that the mean values from the seven non-U.S. test-retest studies predict stability, heritability, and cross-observer validity. These consistent correlations—like the intercorrelations in Table 3—suggest that the composites we have created across samples are meaningful.

Internal consistency is largely unrelated to the three criteria. Coefficient alphas from American self-reports and from multi-national observer ratings show no significant correlations with validity. In addition, Table B1 shows that none of the correlations of the validity criteria with the estimates of internal consistency from the five Additional Samples listed in Table 3 was significant. Internal consistency in the College sample does predict differential stability and heritability; however, the strongest correlate of College internal consistency is One-Week retest reliability, suggesting that the associations with the criteria may be due to variance shared with retest reliability. To assess this possibility, we conducted three stepwise multiple regressions, predicting each of the criteria from the five reliability estimates across the 30 facets (i.e., we predicted columns 6, 7, and 8 in Table 2 from columns 1 through 5). In all three regressions, only the first two steps were significant. Table 5 shows that both retest reliability indicators were significant and independent predictors of cross-observer validity, whereas none of the internal consistency indicators was a significant predictor of any criterion. These data suggest that internal consistency adds nothing to the prediction of the differential validity of NEO Inventory facets.[8]

[7] If the Piedmont peer/peer data are retained in the cross-observer composite, the correlations for both Heise and One-Week reliabilities in Table 4 remain significant.

[8] A reviewer asked for results if we analyzed composites consisting of the averages of the three internal consistency estimates, the two retest reliability estimates, and the three validity indicators. In this analysis, composite retest reliability was related to composite validity ($r = .79$, $p < .001$), whereas composite internal consistency was not ($r = .30$, *n.s.*). In a multiple regression, only composite retest reliability was a significant predictor.

The concept of an attenuation paradox (e.g., Loevinger, 1954) suggests that internal consistency may show a curvilinear relation to validity, with lower validity associated with both low and very high internal consistency. To test that hypothesis, we repeated the regression analyses summarized in Table 5, adding as potential predictors the squares of the three internal consistency indicators. None contributed significantly to the prediction of the three criteria, providing no support for that hypothesis in these data.

Heise and One-Week estimates were independent predictors of the three criteria, suggesting that they tap somewhat different aspects of retest reliability—an issue we return to below in a discussion of macro- and micro-state variation. If so, the best estimate of facet reliability may combine information from these two. The ninth data column in Table 2 gives retest reliability, r_{tt}, defined as the mean of the two estimates. These values predict differential heritability in all four samples (*r*s = .41 to .60, *p*s < .05), stability in all eight samples (*r*s = .51 to .78, *p*s < .01), and cross-observer validity in nine of eleven studies (*r*s = .44 to .60, *p* < .05).

Table 4 shows that all three criteria are strongly intercorrelated, which is due in some degree to the fact that all three are related to retest reliability. The true associations among these criteria could better be estimated by correlating disattenuated values, obtained by dividing each validity coefficient by its respective r_{tt}. After this correction, stability was still significantly related to cross-observer agreement (see Table 4), perhaps because raters come to understand their targets better if their targets are more consistent over time. However, disattenuated heritability was unrelated to either stability or cross-observer agreement, and, as one would expect, all three adjusted criteria were unrelated to r_{tt}. Harris, Vernon, and Jang (1995) had noted an association between cross-observer agreement and heritability in the PRF, and suggested that accurate person perception had a genetic basis. The present analyses suggest that the association is more likely due to the influence of retest reliability on both heritability and cross-observer agreement.

The rationale for combining data from multiple samples into the three criteria was based on agreement among samples on the 30-facet profiles, with alphas of .78 to .92. Yet agreement between samples must have been due in part to the operation of differential retest reliability. Is there coherence across samples when corrected for r_{tt}? We repeated the scale reliability analyses on disattenuated values for the three sets of samples, and found that alphas were substantially reduced, but remained significant for stability (.76), heritability (.59), and cross-observer validity (.77; all $p < .001$). The last three data columns of Table 2 give disattenuated values for the three criteria, and thus an estimate of the five-to-ten-year stability, heritability, and cross-observer validity free from the effects of measurement occasion.

Table 6 reports the highest and lowest scoring facets after correction. The high heritability of O facets and Excitement Seeking may be due to assortative mating for these traits (McCrae et al., 2008), which, in the long run, increases genetic variance in the population.[9] The low stability of N facets may reflect the operation of acute clinical depression in a subset of respondents, because depression is known to affect trait level and trait stability (Costa et al., 2005). The high cross-observer validity of E facets is well-known and widely attributed to the greater observability of these traits (Borkenau & Liebler, 1995;John & Robins, 1993). In an earlier analysis, McCrae and colleagues (2004) tested the hypothesis that the differential cross-observer validity of E facets might be due to differential reliability, but found no

[9]In the usual MZ/DZ twin design, assortative mating tends to cause an underestimate of heritability, because with assortative mating, DZ twins share, on average, more than one-half their genes. This effect, however, may be overshadowed by the long-term effects of assortative mating.

association between agreement and internal consistency. The present analyses suggest that retest reliability is a more plausible moderator of agreement, but that even when corrected for retest unreliability, facets of E are among the most accurately rated.

Relations between Internal Consistency and Retest Reliability

The data in Tables 4 and 5 make it clear that—at least for the NEO Inventories—retest reliability is strongly related to differential validity, whereas internal consistency is essentially unrelated. These two forms of reliability are conceptually distinct, and when retest reliability is estimated by the Heise procedure, it is also statistically independent of internal consistency, as the first data column of Table 4 shows. This is a robust finding: Appendix Table B1 shows that the Heise reliabilities are also unrelated to internal consistencies seen in the five Additional Samples shown in Table 3, *r*s = −.10 to .26, *n.s.* Further, a Heise-type estimate of retest reliability was calculated for Bleidorn et al.'s (2009) German twin sample; it too was unrelated (*r*s = −.06 to .23, *n.s.*) to the eight estimates of internal consistency from Table 3. These data suggest that internal consistency and retest reliability are unrelated variables.

That simple picture is complicated by the fact that internal consistencies are systematically related to retest reliabilities when the latter are assessed by the short-term test-retest method. That is seen in the second data column of Table 4, and in several other results: (a) One-Week reliability is significantly associated with coefficient alphas for the five Additional Samples (Table B1, *r*s = .40 to .58); (b) the composite of seven non-U.S. test-retest coefficients is significantly associated with coefficient alphas for all eight samples (Table B1, *r*s = .44 to .72); and (c) test-retest reliabilities are significantly related to internal consistencies for other personality inventories (Appendix Table A1, third data column, *r*s = 47 to .76).

The largest of these correlations, .78, is between the College and One-Week indicators (see Table 4), where exactly the same 132 cases were used to compute both estimates. These data are not independent: An individual who misread an item at Time 1 will detract from both the internal consistency and the retest reliability of the facet score. Such dependencies in the data doubtless inflate the correlation, but they do not explain why One-Week retest reliability is also related to the other, independent estimates of internal consistency.

According to the analysis summarized in Table 1, differential internal consistency and retest reliability should be related in independent samples only if there is a substantial influence from item ambiguity that differed across scales, because only item ambiguity affects both kinds of differential reliability (sample variability is not directly relevant when independent samples are considered). To explore that possibility, we examined item content in the English version of the NEO-PI-R. For each facet, we calculated four indices that might indicate problematic features: the Flesch-Kincaid Grade reading level; the number of items phrased as negations (e.g., “I am not a worrier,” “I seldom give in to my impulses”); the number of items that included conditional phrases (“If I don't like people, I let them know it,” “When I make a commitment I can always be counted on to follow through”); and the number of items with words unfamiliar to some adolescents (e.g., *lackadaisical*, *panhandler*; see McCrae, Costa, et al., 2005). Two judges rated the items for negations and conditional phrases and showed significant agreement across the 30 facets (*r*s = .86, .59, $p < .001$); differences were resolved by consensus. We correlated each of these indicators of problematic items with the five reliability indicators in Table 2, and with r_{tt}. The correlation between the number of items with unfamiliar words and PPOC internal consistency approached significance ($r = -.36$, $N = 30$, $p < .10$), but none of the other correlations did.

Thus, if there are problematic items in the NEO-PI-R that differentially affect both the internal consistency and the retest reliability of the 30 facets, they do not seem to result from the usual causes, such as readability or negations. Instead, there may be something substantive about some traits that makes them elicit less consistent responses. Openness to Values and Tender-Mindedness are two of the facets that show lowest reliability of both forms; they are attitudinal scales, and it may be that attitudes are less coherent and more changeable than traits such as Openness to Ideas and Self-Discipline, which are among the most reliable of the 30 facets by both indicators.

There is another possible explanation for the correlation between alpha and test-retest reliability: The psychological state of the respondent may in fact change during the course of the test administration, perhaps because a consideration of the items brings different aspects of the self-concept into focus. In this scenario, Table 1 would need to be revised to indicate an effect of micro-state (i.e., within testing session) variability on both absolute and differential internal consistency. When that micro-state component is removed, as in the multiple regressions reported in Table 5, coefficient alpha is apparently unrelated to the validity criteria.

This explanation, however, does not explain the independence of Heise-method estimates of retest reliability from internal consistencies. One possible explanation for this phenomenon is that the Heise method, which uses data from longitudinal intervals, is sensitive mainly to macro-state changes that endure for a period of weeks or months (what Terracciano et al., 2006, p. 10, called “medium-term stability”). Such changes would represent aberrations from the enduring true score, and thus affect scale validities, but they would not affect short-term retest reliability or patterns of responding within one testing session. Episodes of clinical or subclinical depression, which may last several months, provide an example of this kind of medium-term stability; it is suggestive that the largest difference between Heise (.73) and One-Week (.90) estimates of retest reliability in Table 2 is for N3: Depression. The differential sensitivity of test-retest and Heise reliability estimates to micro- and macro-state variability would explain why they are independent predictors of validity criteria. Multi-wave longitudinal studies that incorporate short-term retest studies would be useful for testing these hypotheses.

Discussion

The present analyses, drawing on data available from many previous studies in many nations, lead to relatively clear conclusions. First, the differential psychometric properties of NEO facet scales are robust, being generalizable across genders, ages, and methods of measurement. They also appear to be similar across many different cultural contexts, contributing evidence to the view that personality traits are part of a universal human nature (McCrae et al., 2005a). Second, retest reliability is strongly related to validity, including not only differential stability, but also heritability and cross-observer validity. Third, internal consistency, the most widely used measure of reliability, is essentially unrelated to differential validity (at least when scales are of equal length). Finally, even when corrected for unreliability, differential stability, heritability, and cross-observer validity each shows a consistent pattern across diverse samples. Why some traits should be intrinsically more heritable, or more stable, or more consensually valid than others should be the subject of future personality research.

Issues in Generalization

If the present results were applicable only to differential validity, or only to NEO Inventory scales, they would be of fairly narrow interest. But there is reason to think that they speak to a much wider range of circumstances.

Normally researchers are interested in the validity of scales for the prediction of specific, often behavioral, criteria, such as job performance, health behavior, educational achievement, or marital adjustment. They do not ask whether Scale X is better for predicting educational achievement than Scale Y is for predicting marital adjustment, although it is just such comparative validity that was the focus of research in the present article. We examined differential validity because it allowed us to use criteria that were comparable across a range of scales. With that design, we were able to isolate the contribution to validity of different forms of reliability in the predictor scales; we found that only retest reliability was relevant to differential validity.

Does this conclusion generalize to the absolute validity that is more commonly of interest? The conceptual analysis in Table 1 suggests that it does. For five of the six sources of unreliability, the effects on differential and absolute validity are the same. It is only in the case of Respondents' Error that internal consistency might have an impact on absolute validity that our analysis did not detect. In samples in which there is considerable error introduced by uncooperative or careless respondents, for example, both alpha and validity would be reduced; across a range of samples in which respondent cooperation varied widely, alpha should be a significant predictor of validity. However, this association is a function of the samples, not the measures, and is not normally of concern to applied researchers. If two measures of achievement motivation are both used to predict grade point average *in the same sample*, it is the measure with higher retest reliability (not alpha) that ought—other things being equal—to yield stronger predictive validity. Of course, the two measures of achievement motivation may not be equal in terms of construct validity or conceptual relevance to academic performance, and the researcher or educational psychologist who must choose between them must also consider these issues. But with respect to reliability, our results point to the importance of retest data; at a minimum, this is an attractive hypothesis to test in future research.

We examined differential reliability and validity because it allowed us to use criteria that were comparable for all predictor scales. As noted above, one alternative design that might be used to test our conclusions would be a meta-analysis of traditional validity studies. Across a wide range of studies, one might argue that the criteria chosen would, on average, be equally relevant to the predictor variables and equally well assessed. If so, the internal consistency and retest reliability of predictor scales could be meaningfully related to the criteria and the relative importance of these two reliabilities could be determined. At least one such meta-analysis, of marketing research scales, has been reported (Churchill & Peter, 1984; Peter & Churchill, 1986). It showed that reliability predicted convergent validity, but unfortunately, it did not distinguish among various forms of reliability. That distinction would be crucial in any future meta-analysis, which would also need to track the number of items in the predictor scales, to separate the effects of scale length from internal consistency per se. Results of our simulation study suggest that the associations of reliability with validity depend to a considerable extent on the reliability of the criteria; future meta-analyses should record both the internal consistency and retest reliability of the criteria, and compare the effects of correcting for each on the relations of the predictor's reliability to validity.

The present article analyzed data from the NEO Inventories because large quantities of relevant data were available. But there is no reason to think that the conclusions apply only to NEO scales. The Appendix provides additional evidence from studies of several other personality inventories, and it is consistent with the NEO findings. We know from studies of personality structure that measures of normal and abnormal personality are closely related, so it is likely that similar findings would hold for measures of clinically-relevant attributes such as narcissism, identity disturbance, dependency, and eccentric perceptions (Markon,

Krueger, & Watson, 2005). It is less clear whether retest reliability would be especially relevant for tests of ability, aptitude, or achievement, but that is now an intriguing hypothesis.

The Use and Utility of Internal Consistency Measures

John and Soto (2007) describe coefficient alpha as a "misunderstood giant" (p. 467)—a measure of reliability that is easily assessed and ubiquitously used, despite the fact that for decades, psychometricians have expressed misgivings about it (e.g., Loevinger, 1954). To date, most arguments over its merits have been conceptual or statistical (e.g., Sijtsma, 2009). The present study provides empirical data, and the results are striking: Internal consistency seems to have little to do with the validity of NEO Inventory facets. We argued above that, if alpha restricts validity, cross-observer agreement should be much lower for Openness to Actions than for Depression, yet Table 2 shows that the observed correlations (.43 and .44) are essentially identical. This pattern is seen throughout. The first data column of Table 2 shows that 17 of the NEO facets have coefficient alphas of .70 or greater, whereas 13 have alphas below .70. Conventional wisdom would suggest that validity coefficients for the former should be markedly higher than for the latter. Yet the mean of the validity coefficients for stability, heritability, and cross-observer agreement is .50 for the 17 "reliable" facets and .49 for the 13 "unreliable" facets—a trivial difference. The facet scales of the NEO-PI-R are not, of course, random collections of items. They have a clear conceptual basis, profited from several cycles of item selection, and have documented validity over a wide range of criteria (Costa & McCrae, 1992). In these respects, they resemble most published personality scales (Grucza & Goldberg, 2007). Internal consistency does not seem to be relevant to the evaluation of such scales.

These findings suggest that some current uses of coefficient alpha may need to be reconsidered. As noted in the Introduction, researchers sometimes treat internal consistency as the equivalent of retest reliability for the measurement of reliable change (Snell et al., 2001) or the estimation of the stability of true scores (Conley, 1984). Neither of these applications appears to be supported by the present analyses. Although coefficient alpha is positively related to test-retest reliability estimates, it is not related to long-term stability; and if low internal consistency does not attenuate stability, then it cannot properly be used to disattenuate it.

Internal consistency is also routinely used in SEM measurement models to estimate the correlations of latent variables that are intended to represent true scores. DeShon (1998) cautioned that such a procedure may underestimate true correlations, because they may also be attenuated by retest unreliability and interrater unreliability. He proposed that all sources of unreliability be modeled. But if internal consistency is not relevant to validity, it should not be used to estimate latent variables; such a procedure would tend to overestimate the true correlations among them. Our results suggest not only a different analytic procedure, but also a different strategy for data collection: If researchers are interested in the relations among latent trait variables, they ought to measure them on two or more occasions and use these multi-occasion measures as indicators of latent variables.

We do not mean to suggest that internal consistency has no utility. It is useful first descriptively, as an indicator of the degree to which constituent parts of a whole cohere. For example, it was meaningful to average heritability estimates from Canada, Germany, Japan, and Sardinia because they showed a significant and substantial coefficient alpha (.78) across the 30 NEO facets. Conversely, we excluded Shona data from our non-U.S. retest reliability composite because it detracted from internal consistency (Footnote 5). It is in addressing this basic question of whether items do or do not go together that analyses of internal

consistency are useful in eliminating irrelevant and invalid items in the early stages of scale development.

The utility of internal consistency analyses at this stage was demonstrated by Ashton and Goldberg (1973; see also Jackson, 1975). They asked 15 psychology graduate students and 15 non-psychologists to construct a 20-item scale to measure sociability, achievement, or dominance, and correlated scores on these intuitive scales with peer ratings of the traits in a sample of 168 college females. Internal consistencies for the intuitive scales ranged from .15 to .81, with a median value of .59. Given these very low values (especially for 20-item scales), it can be assumed that many of the item-writers either did not understand the assigned construct or were unable to write relevant items. Under these circumstances, it is not surprising that internal consistency correlated .71 with validity.[10] Our review, however, suggests that that association is not found when more carefully developed scales are examined; most published scales have already benefited from internal consistency analyses (or the related technique of item factor analysis), and there are diminishing returns from further consideration of internal consistency.

Coefficient alpha is also sensitive to random error attributable to respondents' carelessness or misunderstanding of items. A comparison of the observed alpha in a given sample with normative values of alpha for the scale can be a useful indicator of data quality in that sample. Assuming that the sample variability is similar to that in the normative sample, comparatively low alphas suggest a high proportion of random or otherwise errorful responding. We demonstrated that in PPOC data by correlating a general internal consistency factor with independently assessed data quality, $r = .76$. The same factor score is also a predictor of factorial validity, $r = .81$, $N = 51$, $p < .001$, as assessed by total congruence coefficients between the sample structure and the normative American structure (McCrae et al., 2005a, b). Researchers should be encouraged to calculate alphas in their own samples and compare them to normative values to help assess the quality of the data they have collected.

Implications for the Assessment of Retest Reliability

Our data suggest that retest reliability is a powerful determinant of long-term stability. Students of longitudinal development who wish to estimate true score stability and clinical researchers interested in assessing reliable change should employ measures of retest reliability rather than internal consistency in their calculations. Our data do not address the question of whether interrater reliability is an adequate substitute for retest reliability, as Guadiano (2006) argued, but our findings do point to the need to assess that possibility rather than merely to assume that all reliability estimates are interchangeable.

Retest reliability predicted not only stability, to which it is operationally tied, but also heritability and cross-observer validity. Behavior geneticists who wish to take unreliability into account are now well advised to use retest reliability rather than internal consistency to do so. However, the work of Riemann and colleagues (1997) shows that interrater disagreement may also powerfully affect estimates of heritability. Disattenuating for retest unreliability does not yield the results that would be given by true scores in an absolute sense; it yields results that are unaffected by the circumstances of particular test administrations. Such scores may still be biased by the views of individual informants.

[10]This correlation is inflated by dependencies in the data: Both internal consistency and validity are computed from the same responses; by chance, some scales will accumulate more errors, which will decrease both their reliability and validity relative to other scales, and thus increase the correlation between differential reliability and validity in this sample. This effect will be more pronounced in small samples. More accurate and generalizable estimates of the relations between reliability and validity would be based on large, independent samples, such as those in Table 4.

Correcting for interrater disagreement allows one to estimate the heritability of that portion of trait variance that is shared by ratings from two or more observers. As DeShon (1998) argued, it may be optimal to correct simultaneously for both retest and interrater reliability.

Most major personality inventories (e.g., Conn & Rieke, 1994; Millon, 1994) report retest reliabilities in one or more samples, but, judging from the difficulty with which retest data can be identified in meta-analyses (e.g., Beatty et al., 2002; Roberts & DelVecchio, 2000), many ad hoc scales do not. New scale development should routinely include studies of retest reliability, with appropriate care given to the size and representativeness of the retest samples, and, ideally, with a range of retest intervals (Watson, 2004). Watson also wisely recommended that studies of retest reliability be incorporated into the early phases of scale development, so that their results can inform item selection. Researchers undertaking a large-scale study in a distinctive sample (e.g., twins, longitudinal participants) should probably obtain sample-specific retest reliability estimates even when standard instruments with published reliabilities are used.

Retest reliability is clearly an important psychometric property that should be considered in all assessments of scale quality, but it is also an indicator of state variations in trait perception that are themselves worthy of more serious research attention. Why do respondents choose different answers on different occasions (cf. Vul & Pashler, 2008), and why do they do so more with some traits than others? Are state variations more common or more marked in adolescents than in older adults? How do micro- and macro-state variations affect different traits? These are fundamental questions for personality assessment that remain to be addressed.

Acknowledgments

Robert R. McCrae receives royalties from the Revised NEO Personality Inventory. This research was supported in part by the Intramural Research Program, NIH, National Institute on Aging. We thank Juko Ando, Yutaka Ono, Fritz Ostendorf, Alois Angleitner, Rainer Riemann, Frank M. Spinath, W. John Livesley, and Kerry L. Jang for providing heritability data; Jian Yang for assistance with the Chinese literature; David M. Greenberg for coding NEO-PI-R items; Dorret I. Boomsma for advice on assortative mating; and the members of the Personality Profiles of Cultures Project for collecting PPOC data.

Appendix A

Evidence from Other Inventories

To provide some idea of the generalizablity of our findings to other inventories, we conducted a literature review. This was not intended to be a full-scale meta-analysis, but rather an attempt to use the differential validity design on available data from other measures. More extensive analyses, with more liberal rules for the inclusion of studies, would be desirable.

The differential design employed here requires that each set of reliabilities or validities be available for the same scales, and that each kind of coefficient (e.g., all coefficient alphas) be obtained in a single sample. In practice, this limits analyses to multi-scale personality inventories. Table 4 shows that the median retest reliability/validity correlation is .59 for the NEO-PI-R; to replicate these findings, power analysis (alpha = .05, beta = .20, one-tailed) suggested a required sample size of 16 scales, ruling out shorter instruments such as the Comrey (1970) Personality Scales or the Guilford-Zimmerman Temperament Survey (Guilford, Zimmerman, & Guilford, 1976). We identified five longer instruments, the California Psychological Inventory (CPI; Gough, 1987), Adjective Check List (ACL; Gough & Heilbrun, 1983), Temperament and Character Inventory (TCI; Cloninger, Przybeck, Svrakic, & Wetzel, 1994), Personality Research Form (PRF; Jackson, 1984), and Sixteen

Personality Factor Questionnaire, Form A (16PF; Cattell, Eber, & Tatsuoka, 1970)—although one of the 16PF scales assesses intelligence, not personality.

Estimates of internal consistency from non-clinical samples were obtained from the manual for each inventory, except the 16PF; internal consistencies for Form A of the 16PF were taken from Matthews (1989). If more than one estimate was provided, they were averaged. Separate internal consistency estimates were used for PRF Forms AA and E, because these did not agree across PRF scales. Averaged estimates of retest reliability were taken from the manual for the CPI, ACL, PRF, and 16PF. The TCI manual provides six-month retest reliability estimates only for psychiatric inpatients and outpatients. We identified one TCI retest study with a normal sample, 130 college students who completed a Korean translation of the TCI twice over a three-month interval (Sung, Kim, Yang, Abrams, & Lyoo, 2002), and used those data to estimate TCI retest reliability.

We conducted PsycINFO searches, combining each inventory (by name or acronym) successively with “heritability or genetic,” “longitudinal or stability,” and “cross-observer or spouse ratings or peer ratings or informant ratings.” We also consulted Roberts and DelVecchio (2000) for studies on stability. We identified 422 potentially relevant articles from these sources: 114 for the CPI, 74 for the ACL, 107 for the 16PF, 64 for the PRF, and 63 for the TCI. In many cases, studies were excluded because data were not presented for the full set of scales from the inventory. For example, Helson and Moane (1987) provided longitudinal data on only 10 CPI scales. In other cases, the same data were reanalyzed: Carey, Goldsmith, Tellegen, and Gottesman (1978) combined data from earlier studies by Gottesman (1966), Nichols (1966), and Horn, Plomin, and Rosenman (1976); in this case, only the study by Carey and colleagues was included in our analysis. Many of the TCI studies were excluded because they concerned molecular genetics rather than behavior genetics. After screening, we identified only 13 studies with sufficient data to include in the analysis. Table A1 summarizes the studies, with zero-order correlations for internal consistency and retest reliability and stepwise multiple regressions including both as predictors.

Table A1

Results from Studies Using Other Inventories.

					Correlation with Reliability		Beta, Stepwise Multiple Regression	
Inventory	**Study**	***N***	***k***	r_{att}	**Alpha**	**Retest**	**Alpha**	**Retest**
				Heritability				
CPI	Carey et al. (1978)	2,386	18	.51*	.12	.05	—	—
CPI	Bouchard et al. (1998)	322	23	.58**	−.02	.54**	−.51*	.83***
ACL	Scarr (1966)	104	18	.14	−.16	.02	—	—
TCI	Ando et al. (2004)	1,234	25	.69***	.46*	.57**	(.14)	.57**
PRF–E	Johnson et al. (2004)	494	20	.47*	−.28	.02	—	—
PRF–E	Vernon et al. (1997)	418	20	.47*	.07	.27	—	—
				Stability				
16PF	Costa & McCrae (1978)[a]	139	16[b]	.76***	.65**	.75***	(.20)	.75***
16PF	Nichols (1967), Males	432	16[b]	.76***	.55*	.72***	(.02)	.72***
16PF	Nichols (1967), Females	204	16[b]	.76***	.68**	.85***	(.07)	.85***

					Correlation with Reliability		Beta, Stepwise Multiple Regression	
Inventory	**Study**	***N***	***k***	$r_{\alpha tt}$	**Alpha**	**Retest**	**Alpha**	**Retest**
			Cross-Observer Agreement					
PRF–AA	Jackson (1984)	51	20	.75***	−.30	−.28	—	—
PRF–AA	Jackson (1984)	202	20	.75***	.03	−.09	—	—
PRF–E	Fekken et al. (1987)	394	20	.47*	.49*	.64***	(.24)	.64***
PRF–E	Paunonen, in Jackson (1984)	90	20	.47*	−.01	.36	—	—

Note. Betas in parentheses are for excluded variables. k = number of scales. $r_{\alpha tt}$ = correlation between coefficient alphas and retest reliabilities. CPI = California Psychological Inventory. ACL = Adjective Checklist. TCI = Temperament and Character Inventory. 16PF = Sixteen Personality Factor Questionnaire. PRF-E = Personality Research Form E. PRF-AA = Personality Research Form AA.

[a] Ten–year stability of scales composed of items shared by 1962 and 1967 editions for 15 scales; for full 1962 scales, *N* = 403, for scales I, M, and Q1.

[b] Multiple regression analyses omitting scale B, Intelligence, showed the same pattern of results.

* $p < .05$;

** $p < .01$;

*** $p < .001$, one–tailed.

Our review suggests that there are scant data available for differential validity analysis from other inventories. However, the general pattern of results is similar to that found for the NEO Inventories: Multiple regression shows that retest reliability is a significant predictor of the validity criterion in 6 of the 13 studies, whereas internal consistency is significant in only one, and in that case it is negatively related.

Appendix B

Intercorrelations Among Variables

Table B-1 provides a matrix of intercorrelations among the variables assessing internal consistency, retest reliability, and the three validity criteria.

Table B1

Intercorrelations Among Variables.

	Internal Consistency									Retest Reliability			Validity		
	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**	**12.**	**13.**	**14.**	**15.**
Internal Consistency															
Indicator Samples															
1. Manual															
2. College	.81*														
3. PPOC	.62*	.73*													
Additional Samples															
4. PPOC Males	.56*	.70*	.98*												
5. PPOC Females	.64*	.74*	.99*	.96*											

	Internal Consistency								Retest Reliability				Validity		
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
6. Adult Form R	.69*	.62*	.71*	.67*	.72*										
7. NEO-PI-3 Form R	.49*	.66*	.82*	.80*	.80*	.53*									
8. NEO-PI-3 Form S	.65*	.82*	.69*	.63*	.68*	.50*	.86*								
Retest Reliability															
9. Heise	.15	.21	.20	.24	.20	.26	−.10	−.03							
10. One-Week	.56*	.78*	.46*	.43*	.46*	.46*	.40*	.58*	.40*						
11. r_{tt}	.42*	.59*	.40*	.40*	.40*	.43*	.18	.33	.84*	.83*					
12. German Heise	.04	.23	.13	.19	.12	−.06	.22	.12	.16	.39*	.32				
13. Non-U.S. Test-Retest	.66*	.72*	.69*	.68*	.72*	.44*	.50*	.52*	.39*	.55*	.56*	.34			
Validity Criteria															
14. Stability	.28	.43*	.28	.28	.28	.20	.20	.30	.71*	.65*	.82*	.52*	.59*		
15. Heritability	.31	.38*	.11	.12	.12	.23	.07	.24	.57*	.60*	.70*	.37*	.41*	.71*	
16. Cross-Observer	.12	.20	.10	.08	.14	.19	−.04	.06	.57*	.52*	.65*	.27	.40*	.77*	.60*

Note. $N = 30$ facets. See text for sources and sample sizes.

* $p < .05$.

References

AERA. APA. NCME. Standards for educational and psychological testing. American Educational Research Association; Washington, DC: 1999.

Allik J, Laidra K, Realo A, Pullmann H. Personality development from 12 to 18 years of age: Changes in mean levels and structures of traits. European Journal of Personality 2004;18:445–462.

Ando J, Suzuki A, Yamagata S, Kijima N, Maekawa H, Ono Y, et al. Genetic and environmental structure of Cloninger's temperament and character dimensions. Journal of Personality Disorders 2004;18:379–393. [PubMed: 15342324]

Ashton SG, Goldberg LR. In response to Jackson's challenge: The comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by experts, novices, and laymen. Journal of Research in Personality 1973;7:1–20.

Beatty MJ, Heisel AD, Hall AE, Levine TR, La France BH. What can we learn from the study of twins about genetic and environmental influences on interpersonal affiliation, aggressiveness, and social anxiety?: A meta-analytic study. Communication Monographs 2002;69:1–18.

Bleidorn W, Kandler C, Riemann R, Angleitner A. Patterns and sources of adult personality development: Growth curve analyses of the NEO-PI-R scales in a longitudinal twin study. Journal of Personality and Social Psychology 2009;97:142–155. [PubMed: 19586245]

Borkenau P, Liebler A. Observable attributes as manifestations and cues of personality and intelligence. Journal of Personality 1995;63:1–25.

Bouchard TJ Jr. McGue M, Hur Y-M, Horn JM. A genetic and environmental analysis of the California Psychological Inventory using adult twins reared apart and together. European Journal of Personality 1998;12:307–320.

Boyle GJ. Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? Personality and Individual Differences 1991;12:291–294.

Carey G, Goldsmith HH, Tellegen A, Gottesman II. Genetics and personality inventories: The limits of replication with twin data. Behavior Genetics 1978;8:299–313.

Carter JA, Herbst JH, Stoller KB, King VL, Kidorf MS, Costa PT Jr. et al. Short-term stability of NEO-PI-R personality trait scores in opioid-dependent outpatients. Psychology of Addictive Behaviors 2001;15:255–260. [PubMed: 11563805]

Cattell, RB.; Eber, HW.; Tatsuoka, MM. The handbook for the Sixteen Personality Factor Questionnaire. Institute for Personality and Ability Testing; Champaign, IL: 1970.

Cheung F, Cheung SF, Zhang J, Leung K, Leong F, Yeh K-H. Relevance of Openness as a personality dimension in Chinese culture: Aspects of its cultural relevance. Journal of Cross-Cultural Psychology 2008;39:81–108.

Chmielewski M, Watson D. What is being assessed and why it matters: The impact of transient error on trait research. Journal of Personality and Social Psychology 2009;97:186–202. [PubMed: 19586248]

Churchill GA Jr. Peter JP. Research design effects on the reliability of rating scales: A meta-analysis. Journal of Marketing Research 1984;21:360–375.

Cloninger, CR.; Przybeck, TR.; Svrakic, DM.; Wetzel, RD. The Temperament and Character Inventory (TCI): A guide to its development and use. Author; St. Louis, MO: 1994.

Comrey, AL. Manual for the Comrey Personality Scales. EdITS; San Diego, CA: 1970.

Conley JJ. The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality, and self-opinion. Personality and Individual Differences 1984;5:11–26.

Conn, SR.; Rieke, ML., editors. 16PF Fifth Edition technical manual. Institute for Personality and Ability Testing; Champaign, IL: 1994.

Costa PT Jr. Bagby RM, Herbst JH, McCrae RR. Personality self-reports are concurrently reliable and valid during acute depressive episodes. Journal of Affective Disorders 2005;89:45–55. [PubMed: 16203041]

Costa PT Jr. Herbst JH, McCrae RR, Siegler IC. Personality at midlife: Stability, intrinsic maturation, and response to life events. Assessment 2000;7:365–378. [PubMed: 11151962]

Costa, PT., Jr.; McCrae, RR. Objective personality assessment. In: Storandt, M.; Siegler, IC.; Elias, MF., editors. The clinical psychology of aging. Plenum Press; New York: 1978. p. 119-143.

Costa, PT., Jr.; McCrae, RR. The NEO Personality Inventory manual. Psychological Assessment Resources; Odessa, FL: 1985.

Costa PT Jr. McCrae RR. Personality in adulthood: A six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. Journal of Personality and Social Psychology 1988;54:853–863. [PubMed: 3379583]

Costa, PT., Jr.; McCrae, RR. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Psychological Assessment Resources; Odessa, FL: 1992.

Costa PT Jr. McCrae RR, Dye DA. Facet scales for Agreeableness and Conscientiousness: A revision of the NEO Personality Inventory. Personality and Individual Differences 1991;12:887–898.

Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.

Cronbach LJ, Nageswari R, Gleser GC. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology 1963;16:137–163.

Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement 2004;64:391–418.

Dai X-Y, Wu Y-Q. A study on NEO-PI-R used in 16-20 years old people [in Chinese]. Chinese Journal of Clinical Psychology 2005;13:14–18.

DeShon RP. A cautionary note on measurement error corrections in structural equation models. Psychological Methods 1998;3:412–423.

Eichler, WC.; Kurtz, JE. Coefficient alpha and the convergent validity of multi-item personality scales; Poster presented at the annual meeting of the Society of Personality Assessment; New Orleans, LA. Mar. 2008

Fekken GC, Holden RR, Jackson DN, Guthrie GM. An evaluation of the validity of the Personality Research Form with Filipino university students. International Journal of Psychology 1987;22:399–407.

Fleeson W. Toward a structure- and process-integrated view of personality: Traits as density distributions of states. Journal of Personality and Social Psychology 2001;80:1011–1027. [PubMed: 11414368]

Gottesman II. Genetic variance in adaptive personality traits. Journal of Child Psychology and Psychiatry 1966;7:199–208.

Gough, HG. California Psychological Inventory administrator's guide. Consulting Psychologists Press; Palo Alto, CA: 1987.

Gough, HG.; Heilbrun, AB, Jr.. Adjective Check List manual. Consulting Psychologists Press; Palo Alto, CA: 1983.

Grucza RA, Goldberg LR. The comparative validity of 11 modern personality inventories: Behavioral acts, informant reports, and clinical indicators. Journal of Personality Assessment 2007;89:167–187. [PubMed: 17764394]

Guadiano BA. Is symptomatic improvement in clinical trials of cognitive-behavior therapy for psychosis clinically significant? Journal of Psychiatric Practice 2006;12:11–23. [PubMed: 16432441]

Guilford, JS.; Zimmerman, WS.; Guilford, JP. The Guilford-Zimmerman Temperament Survey Handbook: Twenty-five years of research and application. EdITS Publishers; San Diego, CA: 1976.

Harris JA, Vernon PA, Jang KL. A possible genetic basis of accuracy in personality perception. Personality and Individual Differences 1995;18:791–792.

Heise DR. Separating reliability and stability in test-retest correlation. American Sociological Review 1969;34:93–101.

Helson R, Moane G. Personality change in women from college to midlife. Journal of Personality and Social Psychology 1987;53:176–186. [PubMed: 3612488]

Horn JM, Plomin R, Rosenman R. Heritability of personality traits in adult male twins. Behavior Genetics 1976;6:17–30. [PubMed: 943159]

Jackson DN. The relative validity of scales prepared by naive item writers and those based on empirical methods of personality scale construction. Educational and Psychological Measurement 1975;35:361–370.

Jackson, DN. Personality Research Form manual. 3rd. ed.. Research Psychologists Press; Port Huron, MI: 1984.

Jacobson NS, Wilson L, Tupper C. The clinical significance of treatment gains resulting from exposure-based intervention for agoraphobia: A reanalysis of outcome data. Behavior Therapy 1988;19:539–554.

Jang KL, McCrae RR, Angleitner A, Riemann R, Livesley WJ. Heritability of facet-level traits in a cross-cultural twin sample: Support for a hierarchical model of personality. Journal of Personality and Social Psychology 1998;74:1556–1565. [PubMed: 9654759]

John OP, Robins RW. Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. Journal of Personality 1993;61:521–551. [PubMed: 8151500]

John, OP.; Soto, CJ. The importance of being valid: Reliability and the process of construct validation. In: Robins, RW.; Fraley, RC.; Krueger, RF., editors. Handbook of research methods in personality psychology. Guilford; New York: 2007. p. 461-494.

Johnson AM, Vernon PA, Harris JA, Jang KL. A behavior genetic investigation of the relationship between leadership and personality. Twin Research 2004;7:27–32. [PubMed: 15053851]

Kurtz JE, Lee PA, Sherker JL. Internal and temporal reliability estimates for informant ratings of personality using the NEO-PI-R and IAS. Assessment 1999;6:103–113. [PubMed: 10335016]

Kurtz JE, Parrish CL. Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. Journal of Personality Assessment 2001;76:315–332. [PubMed: 11393463]

Löckenhoff CE, Terracciano A, Bienvenu OJ, Particiu NS, Nestadt G, McCrae RR, et al. Ethnicity, education, and the temporal stability of personality traits in the East Baltimore Epidemiologic Catchment Area study. Journal of Research in Personality 2008;42:577–598. [PubMed: 19122849]

Loehlin, JC. Genes and environment in personality development. Sage; Newbury Park, CA: 1992.

Loevinger J. The attenuation paradox in test theory. Psychological Bulletin 1954;51:493–504. [PubMed: 13204488]

Markon KE, Krueger RF, Watson D. Delineating the structure of normal and abnormal personality: An integrative hierarchical approach. Journal of Personality and Social Psychology 2005;88:139–157. [PubMed: 15631580]

Marsella AJ, Dubanoski J, Hamada WC, Morse H. The measurement of personality across cultures: Historical, conceptual, and methodological issues and considerations. American Behavioral Scientist 2000;44:41–62.

Martin, TA.; Costa, PT., Jr.; Oryol, VE.; Rukavishnikov, AA.; Senin, IG. Applications of the Russian NEO-PI-R. In: McCrae, RR.; Allik, J., editors. The Five-Factor Model of personality across cultures. Kluwer Academic/Plenum Publishers; New York: 2002. p. 253-269.

Matthews G. The factor structure of the 16PF: Twelve primary and three secondary factors. Personality and Individual Differences 1989;10:931–940.

McCrae RR. Consensual validation of personality traits: Evidence from self-reports and ratings. Journal of Personality and Social Psychology 1982;43:293–303.

McCrae RR, Costa PT Jr. Discriminant validity of NEO-PI-R facets. Educational and Psychological Measurement 1992;52:229–237.

McCrae, RR.; Costa, PT, Jr.. The Five-Factor Theory of personality. In: John, OP.; Robins, RW.; Pervin, LA., editors. Handbook of personality: Theory and research. 3rd ed.. Guilford; New York: 2008. p. 157-180.

McCrae, RR.; Costa, PT, Jr.. Professional manual for the NEO Inventories: NEO-PI-3, NEO-PI-R, and NEO-FFI-3. Psychological Assessment Resources; Odessa, FL: in press

McCrae RR, Costa PT Jr. Lima MP, Simões A, Ostendorf F, Angleitner A, et al. Age differences in personality across the adult life span: Parallels in five cultures. Developmental Psychology 1999;35:466–477. [PubMed: 10082017]

McCrae RR, Costa PT Jr. Martin TA. The NEO-PI-3: A more readable Revised NEO Personality Inventory. Journal of Personality Assessment 2005;84:261–270. [PubMed: 15907162]

McCrae RR, Costa PT Jr. Martin TA, Oryol VE, Rukavishnikov AA, Senin IG, et al. Consensual validation of personality traits across cultures. Journal of Research in Personality 2004;38:179–201.

McCrae RR, Martin TA, Costa PT Jr. Age trends and age norms for the NEO Personality Inventory-3 in adolescents and adults. Assessment 2005;12:363–373. [PubMed: 16244117]

McCrae RR, Martin TA, Hřebíčková M, Urbánek T, Boomsma DI, Willemsen G, Costa PT Jr. Personality trait similarity between spouses in four cultures. Journal of Personality 2008;76:1137–1163. [PubMed: 18665894]

McCrae RR, Stone SV, Fagan PJ, Costa PT Jr. Identifying causes of disagreement between self-reports and spouse ratings of personality. Journal of Personality 1998;66:285–313. [PubMed: 9615420]

McCrae RR, Terracciano A, 78 Members of the Personality Profiles of Cultures Project. Universal features of personality traits from the observer's perspective: Data from 50 cultures. Journal of Personality and Social Psychology 2005a;88:547–561. [PubMed: 15740445]

McCrae RR, Terracciano A, 79 Members of the Personality Profiles of Cultures Project. Personality profiles of cultures: Aggregate personality traits. Journal of Personality and Social Psychology 2005b;89:407–425. [PubMed: 16248722]

McCrae RR, Yik MSM, Trapnell PD, Bond MH, Paulhus DL. Interpreting personality profiles across cultures: Bilingual, acculturation, and peer rating studies of Chinese undergraduates. Journal of Personality and Social Psychology 1998;74:1041–1055. [PubMed: 9569658]

McDonald, RP. Test theory: A unified treatment. Erlbaum; Mahwah, NJ: 1999.

Millon, T. Millon Index of Personality Styles manual. Psychological Corporation; San Antonio: 1994.

Nichols RC. The resemblance of twins in personality and interests. National Merit Scholarship Corporation Research Reports 1966;2:1–23.

Nichols RC. Personality change and college. American Educational Research Journal 1967;4:173–190.

Nunnally, JC.; Bernstein, I. Psychometric theory. 3rd ed.. McGraw-Hill; New York: 1994.

Ostendorf, F.; Angleitner, A. NEO-Persönlichkeitsinventar, revidierte Form, NEO-PI-R nach Costa und McCrae [Revised NEO Personality Inventory, NEO-PI-R of Costa and McCrae]. Hogrefe; Göttingen, Germany: 2004.

Peter JP, Churchill GA. Relationships among research design choices and psychometric properties of rating scales: A meta-analysis. Journal of Marketing Research 1986;23:1–10.

Piedmont RL. Validation of the NEO-PI-R observer form for college students: Toward a paradigm for studying personality development. Assessment 1994;1:259–268.

Piedmont RL. Cracking the plaster cast: Big Five personality change during intensive outpatient counseling. Journal of Research in Personality 2001;35:500–520.

Piedmont, RL.; Bain, E.; McCrae, RR.; Costa, PT, Jr.. The applicability of the Five-Factor Model in a Sub-Saharan culture: The NEO-PI-R in Shona. In: McCrae, RR.; Allik, J., editors. The Five-Factor Model of personality across cultures. Kluwer Academic/Plenum Publishers; New York: 2002. p. 155-173.

Pilia G, Chen W-M, Scuteri A, Orrú M, Albai G, Deo M, et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genetics 2006;2:1207–1223.

Riemann R, Angleitner A, Strelau J. Genetic and environmental influences on personality: A study of twins reared together using the self- and peer report NEO-FFI scales. Journal of Personality 1997;65:449–475.

Roberts BW, DelVecchio WF. The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. Psychological Bulletin 2000;126:3–25. [PubMed: 10668348]

Rushton JP, Fulker DW, Neale MC, Nias DKB, Eysenck HJ. Altruism and aggression: The heritability of individual differences. Journal of Personality and Social Psychology 1986;50:1192–1198. [PubMed: 3723334]

Scarr S. The origins of individual differences in adjective check list scores. Journal of Consulting Psychology 1966;30:354–357.

Schinka JA, Borum R. Readability of normal personality inventories. Journal of Personality Assessment 1994;62:95–101.

Schmidt FL, Le H, Ilies R. 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 2003;8:206–224. [PubMed: 12924815]

Schmitt N. Uses and abuses of coefficient alpha. Psychological Assessment 1996;8:350–353.

Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 2009;74:107–120. [PubMed: 20037639]

Snell MN, Mallinckrodt B, Hill RD, Lambert MJ. Predicting counseling center clients' response to counseling: A 1-year follow-up. Journal of Counseling Psychology 2001;48:463–473.

Streiner DL. Starting at the beginning: An introduction to coefficient alpha and internal consistency. Journal of Personality Assessment 2003;80:99–103. [PubMed: 12584072]

Sung MS, Kim JH, Yang E, Abrams KY, Lyoo IK. Reliability and validity of the Korean version of the Temperament and Character Inventory. Comprehensive Psychiatry 2002;43:236–243.

Terracciano A, Costa PT Jr. McCrae RR. Personality plasticity after age 30. Personality and Social Psychology Bulletin 2006;32:999–1009. [PubMed: 16861305]

Vernon PA, Jang KL, Harris JA, McCarthy JM. Environmental predictors of personality differences: A twin and sibling study. Journal of Personality and Social Psychology 1997;72:177–183. [PubMed: 9008379]

Vul E, Pashler H. Measuring the crowd within: Probabilistic representations within individuals. Psychological Science 2008;19:645–647. [PubMed: 18727777]

Watson D. Stability versus change, dependability versus error: Issues in the assessment of personality over time. Journal of Research in Personality 2004;38:319–350.

Yamagata, S.; Ando, J.; Ostendorf, F.; Angleitner, A.; Riemann, R.; Spinath, F., et al. Cross-cultural differences in heritability of personality traits: Using behavioral genetics to study culture; Paper presented at the 4th CEFOM/21 International Symposium; Tokyo. Sep. 2006

Yang J, McCrae RR, Costa PT Jr. Dai X, Yao S, Cai T, et al. Cross-cultural personality assessment in psychiatric populations: The NEO-PI-R in the People's Republic of China. Psychological Assessment 1999;11:359–368.

Table 1

Sources of Unreliability and Their Effects.

	Reliability Indicator Affected?					
	Absolute		Differential		Validity Affected?	
Source	**Alpha**	**Retest**	**Alpha**	**Retest**	**Absolute**	**Differential**
Item Irrelevance	Yes	No	Yes	No	Yes	Yes
Item Heterogeneity	Yes	No	Yes	No	?	?
State Variation	No	Yes	No	Yes	Yes	Yes
Respondents' Error	Yes	Yes	No	No	Yes	No
Item Ambiguity	Yes	Yes	Yes	Yes	Yes	Yes
Sample Variance	Yes	Yes	Yes	Yes	Yes	Yes

Note. This table summarizes conceptual arguments offered in the text, not empirical findings. Absolute = magnitude for a given test compared to a fixed standard (e.g., .70); Differential = magnitude compared to that of other tests administered to the same sample. Item irrelevance = inappropriateness of item content. Item heterogeneity = statistical independence of trait indicators. State variation = effects of time-of-administration. Respondents' error = item misunderstanding or carelessness. Item ambiguity = lack of clarity or intelligibility of items; also includes effects due to the interaction of respondent and item characteristics. Sample variability = range of variation of traits in the sample. ? = Effect unknown.

Pers Soc Psychol Rev. Author manuscript; available in PMC 2011 May 1.

Table 2

Reliabilities, Validity Criteria, and Corrected Criteria for NEO-PI-R Domains and Facets.

	Internal Consistency			Retest Reliability								
NEO-PI-R Scale	**Manual (1,539)**	**College (132)**	**PPOC (12,156)**	**Heise (520)**	**One-Week (132)**	**S (4,576)**	**h^2 (9,461)**	**CA (4,338)**	**r_{tt}**	**$\hat{S}$**	**$\hat{h}^2$**	**$\hat{C}A$**
N: Neuroticism	.92	.92	.89	.83	.91	.76	.44	.48	.87	.87	.51	.55
E: Extraversion	.89	.92	.89	.92	.92	.81	.48	.59	.92	.88	.52	.64
O: Openness	.87	.89	.87	.90	.93	.83	.52	.54	.92	.90	.57	.59
A: Agreeableness	.86	.89	.91	.87	.92	.78	.41	.52	.90	.87	.46	.58
C: Conscientiousness	.90	.93	.93	.88	.92	.79	.43	.45	.90	.88	.48	.50
N1: Anxiety	.78	.79	.72	.77	.85	.70	.36	.48	.81	.86	.45	.59
N2: Angry Hostility	.75	.77	.76	.80	.83	.67	.36	.47	.82	.82	.44	.58
N3: Depression	.81	.83	.74	.73	.90	.65	.39	.44	.82	.80	.48	.54
N4: Self-Consciousness	.68	.67	.58	.70	.79	.66	.38	.35	.75	.89	.51	.47
N5: Impulsiveness	.70	.72	.62	.67	.77	.65	.30	.39	.72	.90	.42	.54
N6: Vulnerability	.77	.79	.76	.81	.85	.66	.35	.36	.83	.79	.42	.44
E1: Warmth	.73	.73	.77	.84	.86	.72	.41	.47	.85	.85	.48	.55
E2: Gregariousness	.72	.84	.75	.83	.89	.74	.42	.52	.86	.86	.49	.61
E3: Assertiveness	.77	.82	.71	.82	.91	.78	.41	.52	.87	.90	.48	.60
E4: Activity	.63	.64	.62	.86	.78	.70	.37	.49	.82	.86	.45	.60
E5: Excitement Seeking	.65	.64	.70	.83	.78	.72	.40	.52	.81	.89	.49	.65
E6: Positive Emotions	.73	.81	.76	.83	.86	.71	.36	.47	.85	.84	.43	.56
O1: Fantasy	.76	.82	.73	.80	.82	.71	.37	.41	.81	.87	.45	.51
O2: Aesthetics	.76	.87	.77	.86	.91	.78	.47	.54	.89	.88	.53	.61
O3: Feelings	.66	.74	.65	.75	.82	.67	.35	.39	.78	.85	.45	.49
O4: Actions	.58	.52	.53	.85	.78	.70	.38	.43	.82	.86	.46	.53
O5: Ideas	.80	.83	.81	.82	.87	.78	.43	.46	.85	.92	.51	.54
O6: Values	.67	.62	.50	.80	.80	.66	.40	.45	.80	.83	.50	.57
A1: Trust	.79	.85	.78	.78	.83	.67	.36	.40	.81	.83	.45	.50
A2: Straightforwardness	.71	.74	.73	.77	.86	.65	.32	.37	.82	.80	.40	.46
A3: Altruism	.75	.68	.77	.72	.75	.65	.29	.40	.74	.88	.39	.54

	Internal Consistency			Retest Reliability								
NEO-PI-R Scale	Manual (1,539)	College (132)	PPOC (12,156)	Heise (520)	One-Week (132)	S (4,576)	h^2 (9,461)	CA (4,338)	r_{tt}	$\hat{S}$	$\hat{h}^2$	$\hat{C}A$
A4: Compliance	.59	.63	.71	.77	.83	.70	.33	.51	.80	.88	.41	.64
A5: Modesty	.67	.82	.76	.81	.86	.69	.38	.39	.84	.82	.46	.47
A6: Tender-Mindedness	.56	.50	.57	.74	.70	.63	.31	.37	.72	.87	.43	.51
C1: Competence	.67	.71	.70	.72	.80	.63	.32	.34	.76	.83	.42	.44
C2: Order	.66	.77	.73	.80	.90	.74	.35	.49	.85	.87	.41	.58
C3: Dutifulness	.62	.66	.76	.69	.75	.62	.36	.37	.72	.86	.49	.52
C4: Achievement Striving	.67	.82	.72	.75	.88	.70	.38	.44	.82	.86	.46	.54
C5: Self-Discipline	.75	.81	.82	.86	.83	.71	.40	.40	.85	.84	.47	.47
C6: Deliberation	.71	.76	.79	.80	.77	.67	.28	.36	.79	.85	.35	.46
Facet:												
Mdn	.71	.77	.73	.80	.83	.70	.37	.44	.82	.86	.45	.54
Range	.56-.81	.50-.87	.50-.82	.67-.86	.70-.91	.62-.78	.28-.47	.34-.54	.72-.89	.79-.92	.35-.53	.44-.65

Note. *Ns* in parentheses. NEO-PI-R = Revised NEO Personality Inventory. Manual = data from NEO-PI-R Manual. College = data from College sample. PPOC = data from Personality Profiles of Cultures project. Heise = retest reliability estimated from longitudinal data. One-Week = test-retest coefficients in the College sample. S = stability. h^2 = heritability. CA = cross-observer agreement. r_{tt} = retest reliability ([Heise + One-Week]/2). $\hat{S}$ = disattenuated stability. $\hat{h}^2$ = disattenuated heritability. $\hat{C}A$ = disattenuated cross-observer agreement.

Table 3

Intercorrelations among Internal Consistency Estimates across 30 NEO Facets.

Sample	2.	3.	4.	5.	6.	7.	8.
Indicator Samples							
1. Manual (Form S)	.81***	.62***	.56***	.65***	.69***	.49**	.65***
2. College (Form S)		.74***	.70***	.74***	.62***	.66***	.82***
3. PPOC (Form R)			.98***	.99***	.71***	.82***	.69***
Additional Samples							
4. PPOC Males (Form R)				.96***	.67***	.80***	.63***
5. PPOC Females (Form R)					.72***	.80***	.68***
6. NEO-PI-R Adults (Form R)						.53**	.50**
7. NEO-PI-3 Adolescents (Form R)							.86***
8. NEO-PI-3 Adolescents (Form S)							

Note. Correlations across $N = 30$ facets. Data from sample sizes 1,539; 312; 12,156; 5,852; 6,304; 277; 500; and 500 for Samples 1 through 8, respectively.

** $p < .01$.

*** $p < .001$.

Table 4

Intercorrelations among Reliability Estimates and Validity Criteria.

Sample	4.	5.	6.	7.	8.
Internal Consistency					
1. Manual	.15	.56***	.28	.31	.12
2. College	.21	.78***	.43*	.38*	.20
3. PPOC	.20	.46*	.28	.11	.10
Retest Reliability					
4. Heise		.40*	.71***	.57***	.57***
5. One-Week			.65***	.61***	.52**
Validity Criteria					
6. Stability				.71*** (.24)	.77*** (.45*)
7. Heritability					.60*** (.32)
8. Cross-Observer					

Note. Correlations across $N = 30$ facets. Data from sample sizes 1,539; 132; 12,156; 520; 132; 4,576; 9,461; and 4,338 for Samples 1 through 8, respectively. Values in parentheses are correlations among disattenuated criteria.

* $p < .05$.

*** $p < .001$.

Table 5

Stepwise Multiple Regressions Predicting Criteria from Internal Consistency and Retest Reliability Indicators.

	Stability		Heritability		Cross–Observer	
Predictor	**Adjusted R^2**	**β**	**Adjusted R^2**	**β**	**Adjusted R^2**	**β**
			Selected Variables			
One-Week Retest	.64	.44***	.46	.45**	.39	.35*
Heise Retest		.54***		.40*		.43*
			Excluded Variables			
Manual Alpha		−.07		.00		−.20
College Alpha		−.05		−.12		−.42
PPOC Alpha		−.05		−.22		−.19

Note. For excluded variables, β = standardized regression coefficient if the variable were included with selected variables.

Table 6

Highest and Lowest Scoring NEO-PI-R Facets for Retest Reliability and the Disattenuated Criteria.

Retest Reliability (r_{tt})	Heritability	5-10 year Stability	Cross-Observer Agreement
		Highest	
O2: Aesthetics	O2: Aesthetics	O5: Ideas	E5: Excitement Seeking
E3: Assertiveness	O5: Ideas	E3: Assertiveness	A4: Compliance
E2: Gregariousness	N4: Self-Consciousness	N5: Impulsiveness	O2: Aesthetics
C2: Order	O6: Values	E5: Excitement Seeking	E2: Gregariousness
E1: Warmth	C3: Dutifulness	N4: Self-Consciousness	E4: Activity
C5: Self-Discipline	E5: Excitement Seeking	C2: Order	E3: Assertiveness
O5: Ideas	E2: Gregariousness	A4: Compliance	
E6: Positive Emotions		A3: Altruism	
		O2: Aesthetics	
		Lowest	
C1: Competence	C1: Competence	A1: Trust	C5: Self-Discipline
N4: Self-Conscientiousness	N6: Vulnerability	O6: Values	A5: Modesty
A3: Altruism	N5: Impulsiveness	A5: Modesty	N4: Self-Consciousness
C3: Dutifulness	C2: Order	N2: Angry Hostility	C6: Deliberation
A6: Tender-Mindedness	A4: Compliance	A2: Straightforwardness	A2: Straightforwardness
N5: Impulsiveness	A2: Straightforwardness	N3: Depression	C1: Competence
	A3: Altruism	N6: Vulnerability	N6: Vulnerability

Note. For retest reliability, highest values = .84 to .88, lowest values = .72 to .76; for heritability, highest values = .49 to .53, lowest values = .35 to .42; for stability, highest values = .88 to .92, lowest values = .79 to .83; for cross-observer agreement, highest values =.60 to .65, lowest values = .44 to .47.