Article

A More Nuanced View of Reliability: Specificity in the Trait Hierarchy

Personality and Social Psychology Review
2015, Vol. 19(2) 97–112

DOI: 10.1177/1088868314541857
pspr.sagepub.com
SAGE

Robert R. McCrae[1]

Abstract

Retest reliability is a better predictor of validity than is internal consistency. One explanation for this is item-specific variance, which distinguishes different nuances of a facet and contributes to retest reliability but not internal consistency. Specific variance at the facet level is temporally stable, consensually validated, and heritable; a consideration of the role of specific variance in personality measures leads to a distinction between traits as the intersection ($\cap$) versus the union ($\cup$) of their constituents. I discuss specific variance at the item level and its implications for scale development and argue that retest reliability outpredicts internal consistency because item-specific variance has been shown to be observable, and is probably heritable and stable. I consider some implications of these ideas for the use of single-item scales, the causal interpretation of traits, and the notion of scalar equivalence. Finally, I note that the sources of random error in scales are still poorly understood.

Keywords

Five-Factor Model, consensual validation, scale development, scalar equivalence, reliability

In 2011, McCrae, Kurtz, Yamagata, and Terracciano published an article on the effects of reliability on the differential validity of personality measures. They thought the findings—that retest reliability predicted validity whereas internal consistency did not—were striking, and they expected the article to be provocative. So far, however, it has provoked few responses from statisticians (but see Holden & Bernstein, 2013). Many readers (like most of the original reviewers) seem to be uncomfortable with the conclusions, but no one so far has suggested obvious errors in the article or its conclusions. In this article, I argue that one possible explanation for the findings lies in the nature of the trait hierarchy, in which assessed traits may be better envisioned as the union, rather than as the intersection, of their subtraits.

Conventional wisdom has it that reliability (loosely, how well a test measures something) sets limits to validity (how well a test measures what it is supposed to). Other things being equal, most psychologists would choose a scale with a coefficient alpha of .80 over another with an alpha of .60. Internal consistency is the most widely used measure of reliability because it is the easiest to obtain, but retest reliability is often considered a substitute, and occasionally other measures (interrater reliability, parallel form reliability) are used.

For decades, writers such as Cattell (1973), Watson (2012), and Cronbach himself (Cronbach & Shavelson, 2004) have cautioned that this is an oversimplification, but that has had little impact on psychometric practice. McCrae, Kurtz, and colleagues (2011) hoped to revive this controversy by comparing validity coefficients for the 30 facet scales of the NEO Inventories, which differ in reliability. In large data sets, coefficient alpha did not consistently predict differential validity (operationalized as long-term stability, heritability, and cross-observer agreement)—despite a wide range of variation (αs ≈ .5-.9). Surprisingly, retest reliability did, despite a restricted range (r_{tt}s ≈ .7-.9). It was not clear at the time why that should be so.

Revising the Classical Perspective on Reliability

Coefficient alpha is beloved of statisticians because it makes simple and elegant predictions—under a particular set of assumptions. Suppose (in the very simplest case) that a test consists of k items, equally valid measures of the trait, and that responses to these items are determined completely by the trait (T) and random (i.e., uncorrelated) error (ε). Then the proportion of trait variance in the total scale (the sum of the standardized items) is just alpha. If there is a parallel form with identical item properties, its correlation with the original scale will be alpha, and, if we assume that the trait does not change over a short time period, so will retest reliability. Finally, if we assume that observer ratings are as

[1]Baltimore, Maryland, USA

Corresponding Author:
Robert R. McCrae, 809 Evesham Avenue, Baltimore, MD 21212, USA.
Email: RRMcCrae@gmail.com

valid as self-reports, then, under the classical assumptions, the cross-observer correlation will also be alpha.[1]

Anyone with any acquaintance with the empirical literature on self/other agreement knows that this is nonsense. Cross-observer correlations are typically substantial (≈ .4 to .6) but almost invariably lower than reliabilities. We explain this by saying that there is also method variance (*M*), a consistent bias that differs across observers. For example, two respondents describing the same target may vary in their level of socially desirable responding or leniency bias. Acquiescent (vs. naysaying) responding—the tendency to endorse items regardless of content—is a consistent individual difference variable (McCrae, Herbst, & Costa, 2001) that systematically biases scale scores, particularly if most items are keyed in the same direction. Finally, respondents may differ simply in their perceptions of the trait: Mary may think that she is *very assertive*, whereas John thinks she is only *moderately assertive*.

Because method variance is likely to influence all items in a scale in the same way, it inflates alpha: Alpha becomes a measure of the variance due to $T + M$. If this modified model were correct, and if the amount of method variance relative to trait variance were constant, then alpha would be proportional to validity, and the scale with alpha = .80 would indeed be better than the scale with alpha = .60. However, there is no reason to think that scales are equally affected by method variance—for example, scales differ in evaluativeness and might be differentially sensitive to desirability bias—so the relation between alpha and validity is diminished to an unknown extent. That could provide an explanation for the finding that alpha does not predict differential validity for NEO Inventory facet scales (McCrae, Kurtz, et al., 2011).

Retest reliability—that is, consistency of scale scores across two occasions when the time interval and circumstances make it unlikely that any true change has occurred in the trait[2]—is also surely affected by method variance. The simplest assumption is that method variance is stable across occasions (Mary always thinks she is *very assertive*). In this case, retest reliability will consist of the variance common to the two administrations, namely, $T + M$—and that is simply alpha. In general, alpha increases as the number of items in the scale increases, but this should also lead to similar increases in retest reliability. For example, the retest reliability of the 48-item NEO Inventory domains is uniformly higher than the retest reliability of the 8-item facets (McCrae, Kurtz, et al., 2011).

However, it is possible that method variance changes across administrations. Clearly, this can happen when retests are given under substantially different conditions, as when participants in an experiment are asked first to respond honestly, and then to fake good or bad. However, changes in method variance are also possible in ostensibly similar circumstances: Today, Mary may feel very assertive, but next week she may feel a bit more inhibited. This is a very plausible prediction, because Fleeson (2001) has shown that most traits have corresponding states that fluctuate from day to day. It is reasonable that these states might affect responses to personality scales framed as trait measures, and thus attenuate retest reliability. Schmidt, Le, and Ilies (2003) discuss this as *transient error*. In this scenario, trait variance is preserved but method variance (or some of it) is not.[3] Retest reliability ought then to be lower than alpha. Thus, according to the $T + M + \varepsilon$ model, $r_{tt} < \alpha$.

However, that is not what happens—at least not for NEO Inventory facets. McCrae, Kurtz, and colleagues (2011, Table 2) reported 3 independent estimates of facet internal consistency and 2 estimates of retest reliability, so it is possible to make 6 comparisons for each of 30 facets, 180 in total. In 160 of these comparisons (89%), retest reliability was higher than alpha. The median values were .73 for alpha and .81 for retest.[4] Clearly, something needs to be added to the $T + M + \varepsilon$ model.

McCrae, Kurtz, and colleagues (2011) pointed out that α and r_{tt} are conceptually distinct, and illustrated that point by noting that "the internally inconsistent sum of date of birth, height, and social security number might be constant for many years" (p. 29), and yield a very high r_{tt}. I do not, of course, mean to suggest that NEO Inventory facet scales include irrelevant items that spuriously inflate retest reliability. All items were selected from a pool of conceptually relevant trial items through a series of item factor analyses designed to maximize convergent and discriminant validity (Costa, McCrae, & Dye, 1991; McCrae & Costa, 1983), and the overall success of these procedures was demonstrated in subsequent analyses of larger, independent samples (McCrae & Costa, 2008). However, it is possible that there are meaningful sources of variance in individual items that contribute to retest reliability but not internal consistency.

McCrae, Kurtz, and colleagues (2011) discussed item heterogeneity: "whether the items in a scale cover many different aspects of a trait or focus on only a few" (p. 30). They noted that item heterogeneity should diminish internal consistency but not retest reliability, but they made no predictions about its effect on validity, perhaps because they did not conceptualize heterogeneity in terms of item-specific variance (*s*). What is it that distinguishes different aspects or *nuances*[5] of a trait? It must be something peculiar to each nuance and contributing uniquely to the variance of each item. By definition, this specific variance in an item is not shared by other items in the scale, so it detracts from alpha. However, in retest designs, the same items, with the same specific variance, are readministered, and they may elicit the same response. Item-specific variance could thus account for the fact that retest reliability is greater than alpha, especially if we also assume that method variance is stable over short intervals.[6]

A plausible (although still highly simplified) model is thus represented by the formula

$$\text{Scale} = T + M + s_{1-k} + \varepsilon_{1-k}, \quad (1)$$

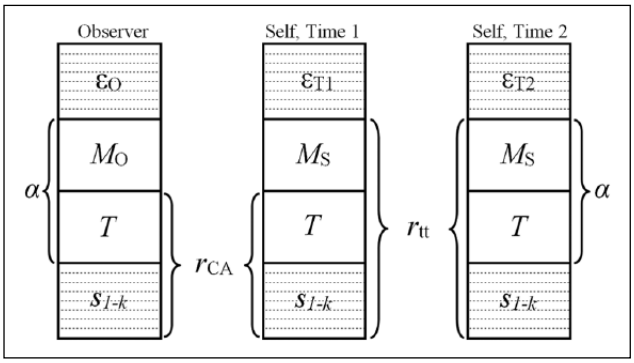

Figure 1. Schematic representation of the proportions of variance attributable to error (ε), method (*M*), common trait (*T*), and item specifics ($s_{1\text{-}k}$) in scale scores from an observer and from a self-report test and short-term retest. Dashed lines indicate distinct contributions from different items. Internal consistency (α) is attributable to method and trait variance shared by items. Cross-observer agreement (r_{CA}) and retest reliability (r_{tt}) reflect shared components of variance across observers and occasions, respectively.

where all *k* items share trait and method variance, and each item has its own specific and error variance. Let us assume that this model applies equally to any self-report or rating from a knowledgeable informant. This assumption is based on evidence that internal consistency is roughly equivalent for self-reports and observer ratings (McCrae, Kurtz, et al., 2011) and that self/informant ratings agree at about the same level as informant/informant ratings (McCrae & Costa, 1989; although this can vary by trait: see Vazire & Carlson, 2011). As Figure 1 illustrates, internal consistency reflects the proportion of variance due to trait and method ($T + M$); retest reliability the proportion due to trait, method, and item specifics ($T + M + s_{1\text{-}k}$); and cross-observer agreement (r_{CA}) the proportion due to trait and item specifics ($T + s_{1\text{-}k}$). If we have data on these three coefficients, we can calculate the total variance attributed to each component:

$$M = r_{tt} - r_{CA}. \tag{2}$$

$$T = [\alpha - M] = [\alpha - (r_{tt} - r_{CA})]. \tag{3}$$

$$s_{1-k} = r_{tt} - \alpha. \tag{4}$$

$$\varepsilon_{1-k} = 1 - r_{tt}. \tag{5}$$

Table 1 illustrates these values with data on NEO Inventory facets taken from McCrae, Kurtz, and colleagues (2011, Table 2). Internal consistency is the mean of three estimates and retest reliability is the mean of two estimates. The components of variance refer to scores for an individual self-report or single observer rating; presumably *T* and *s* variance would increase relative to *M* and ε in scores based on the mean of multiple ratings. The last line of Table 1 suggests that, at the facet level, about one third of the variance is common trait variance and one tenth is item-specific variance. The largest component is method variance, which in turn is a major contributor to internal consistency. Thus, it is not surprising that coefficient alpha is a poor predictor of differential validity. Retest reliability, in contrast, is the inverse of random error variance, which would explain why it predicts such criteria as stability and heritability: Scales with less random error are more valid.

The Issue of Specific Variance

Item-specific variance differentiates retest from internal consistency reliability, and thus may account for the superiority of the former in predicting scale validity. However, very little attention has been paid to this component, and its adequate conceptualization is challenging. At first glance, it appears to be a form of error, because, in the classical model, it is independent of the common trait variance, and trait variance is usually presumed to be the source and substance of validity. This initial impression requires careful consideration.

It is useful to begin with some comments on the nature of the personality trait hierarchy. Almost all trait theorists recognize at least two levels, corresponding to first- and second-order factors (Cattell, Eber, & Tatsuoka, 1970) or the facets and domains of the Five-Factor Model (FFM; Costa & McCrae, 1995). Costa and McCrae (1995) made it clear that they believed this simple two-level architecture is merely a conceptual convenience. For example, Figure 2 shows elements of Neuroticism, grouped into possible facets (of course, a different grouping is actually used for the NEO Inventory Neuroticism facets). This figure illustrates the point that facets are not the absolute bottom of the trait hierarchy; they are themselves divisible into nuances. Being characteristically tense and characteristically worried are both nuances of Anxiety, but they are differentiable. McCrae, Harwood, and Kelly (2011) asked how specific measures should be: "Do we need separate scales for anxiety, test anxiety, math test anxiety, or advanced calculus test anxiety?" (p. 255). For the present purposes, it suffices to combine all trait-like characteristics below the level of facets into the level of nuances.

Personality psychologists understand something about specific variance at the level of facets. In the NEO Inventories, facets are relatively narrow traits at a lower order; six facets are summed to create each of the five broad domain scales. The higher order traits are also, somewhat more precisely, computed as five factor scores. It is possible to examine the specific variance in the 30 facet scales by simply calculating residual scores for each from which the variance in the five factors has been removed.[8] It is somewhat difficult to conceptualize the specific components: What is left of N1: Anxiety after all traces of Neuroticism (and the other factors) have been removed? It is easier to conduct empirical research, however; one simply correlates the residual scores with

Table 1. Estimated Components of Variance in NEO Inventory Scales.

	Observed values			Estimated proportion of variance			
Facet	α	r_{tt}	CA	Trait	Method	Specific	Error
N1: Anxiety	.76	.81	.48	.43	.33	.05	.19
N2: Angry hostility	.76	.82	.47	.42	.35	.05	.19
N3: Depression	.79	.82	.44	.42	.38	.02	.19
N4: Self-consciousness	.64	.75	.35	.25	.40	.10	.26
N5: Impulsiveness	.68	.72	.39	.35	.33	.04	.28
N6: Vulnerability	.77	.83	.36	.30	.47	.06	.17
E1: Warmth	.74	.85	.47	.36	.38	.11	.15
E2: Gregariousness	.77	.86	.52	.43	.34	.09	.14
E3: Assertiveness	.77	.87	.52	.42	.35	.10	.14
E4: Activity	.63	.82	.49	.30	.33	.19	.18
E5: Excitement seeking	.66	.81	.52	.38	.29	.14	.20
E6: Positive emotions	.77	.85	.47	.39	.38	.08	.16
O1: Fantasy	.77	.81	.41	.37	.40	.04	.19
O2: Aesthetics	.80	.89	.54	.46	.35	.09	.12
O3: Feelings	.68	.79	.39	.29	.40	.10	.22
O4: Actions	.54	.82	.43	.16	.39	.27	.19
O5: Ideas	.81	.85	.46	.43	.39	.03	.16
O6: Values	.60	.80	.45	.25	.35	.20	.20
A1: Trust	.81	.81	.40	.40	.41	.00	.20
A2: Straightforwardness	.73	.82	.37	.28	.45	.09	.19
A3: Altruism	.73	.74	.40	.40	.34	.00	.27
A4: Compliance	.64	.80	.51	.35	.29	.16	.20
A5: Modesty	.75	.84	.39	.31	.45	.09	.17
A6: Tender-mindedness	.54	.72	.37	.19	.35	.18	.28
C1: Competence	.69	.76	.34	.27	.42	.07	.24
C2: Order	.72	.85	.49	.36	.36	.13	.15
C3: Dutifulness	.68	.72	.37	.33	.35	.04	.28
C4: Achievement striving	.74	.82	.44	.36	.38	.08	.19
C5: Self-discipline	.79	.85	.40	.35	.45	.05	.16
C6: Deliberation	.75	.79	.36	.33	.43	.03	.22
M	0.72	0.81	0.43	0.34	0.37	0.09	0.19

Note. Data summarized from McCrae, Kurtz, Yamagata, and Terracciano (2011), Table 2. CA = cross-observer agreement for single raters.

external criteria. McCrae and Costa (1992) showed that residual facet scores in self-reports were significantly related to residual facet scores in peer or spouse ratings for 28 of the traits; in a later and larger study (Costa & McCrae, 2008), cross-observer agreement was found for all 30 facet residuals (*Mdn r* = .33, *N* = 532). Jang, McCrae, Angleitner, Riemann, and Livesley (1998) reported that residual scores showed modest retest reliability (*Mdn*s = .65 for 1-week retest, *n* = 58, and .61 for 2-year retest, *n* = 338) and 26 of them showed evidence of heritability. McCrae and colleagues (1999) reported that residual facet scores showed parallel developmental trends in German, Italian, Portuguese, Croatian, and South Korean samples. McCrae and Costa (1992) argued that the discriminant validity of facet scales was due chiefly to the presence of specific variance, and it is the surplus information that such scales provide that makes them superior to domains as predictors of outcomes of interest, such as behaviors (Paunonen & Ashton, 2001) and personality disorders (Reynolds & Clark, 2001).

Although it is statistically easy to remove the factor variance from facet scales, it is essentially impossible to remove facet-specific variance from the higher order factors (except as unobservable latent variables). NEO Inventory domain scales are simply the sum of six facets, so clearly, they include not only the common dimension that accounts for the covariation of the facets but also specific variance contributed by each facet. However, because common variance aggregates across facets, whereas specific variance does not, the more distinct facets one combines, the greater the proportion of common variance relative to the total variance of the scale. Factor scores (or component scores) give a somewhat purer representation of the common variance, because factor scoring weights are chosen to maximize it and to suppress variance attributable to other factors. However, ultimately a

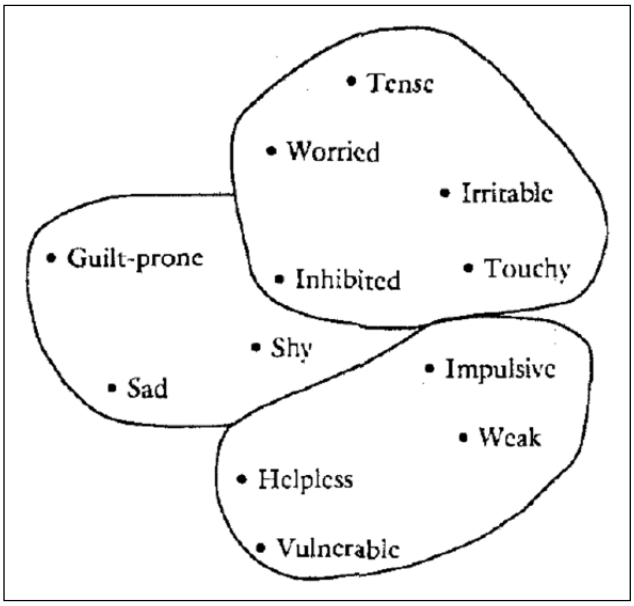

Figure 2. Nuances of Neuroticism, showing one possible organization into facets (but not the one used in the NEO Inventories).
Source. Adapted from Costa and McCrae (1995).

factor score is a linear combination of facet scores and inevitably incorporates some facet-specific variance.

Trait purists might see this as a problem: Assessed factor scores are at best rough approximations to the latent scores. However, it is equally possible to see the inclusion of specific variance as a strength, because it increases the range of criteria the factor or scale will predict. For example, a Neuroticism scale that omitted specific variance from N2: Angry Hostility would probably be a weaker predictor of Borderline Personality Disorder than one that included it. Global domain-level scales rarely provide optimal prediction (Judge, Rodell, Klinger, Simon, & Crawford, 2013); instead, they are perhaps best viewed as screening devices: If a domain scale is significantly correlated with a criterion, it suggests that one or more of its facets will probably be more strongly related to the criterion. The screening correlations are likely to be larger (and more apt to reach statistical significance) if the domain scale includes specific variance from the relevant facets. Soto and John (2009), for example, suggest that Big Five Inventory (BFI; John, Donahue, & Kentle, 1991) Extraversion emphasizes just two facets, Assertiveness and Activity. Other things being equal, BFI Extraversion should therefore be somewhat less likely to predict a criterion such as subjective well-being than is the NEO Inventory Extraversion scale, which also includes a Positive Emotions facet, a stronger predictor of well-being (Costa & McCrae, 1984).

Higher-order personality traits can thus be conceptualized in two different ways. The first, corresponding to domain scales, is as the sum of the component facets. In this view, higher order traits carry all the variance (or, in principle, all the non-error variance) of their constituents. A second view is that higher-order traits are best viewed as that which is *common* to the constituent traits, and thus by definition excludes their specific variance. Borrowing terminology loosely from set theory, we might characterize these as *union* ($\cup$) versus *intersection* ($\cap$) conceptions of traits. The same conceptions can be applied to the relation between scales and their items, as illustrated in Figure 3.

There is an important difference between the mathematical sense of the terms *union* and *intersection* and the psychometric sense I am proposing here. In mathematics, sets and their unions are characterized only by the presence or absence of elements. In scales, the *quantity* of each element is crucial; because of aggregation, longer scales include a higher proportion of variance from the common element. For example, the union of {a, b} and {b, c} is simply {a, b, c}. However, if one personality item has elements A and B, and a second has elements B and C, a scale composed of the sum of the two items has elements A, 2B, and C, and the variance attributable to B is four times the variance attributable to A or C.

From the perspective of sources of variance, union and intersection views of traits are very different. Within classical test theory, there is a single source of variance for any $\cap$Trait (represented by the latent variable that accounts for the covariation of the subtraits), whereas there are necessarily many different sources of variance in a $\cup$Trait, because in addition to the core latent variable, it includes specific variance passed on by each subtrait. The variance component in the classical model presented above and labeled T is, of course, $\cap T$, as are the factors sought in common factor analysis. As Spearman noted long ago (see Widaman, 2007), this intersection model implies that particular indicators are interchangeable: The latent Agreeableness factor common to A1: Trust, A2: Straightforwardness, and A3: Altruism ought to be identical to the Agreeableness factor common to A4: Compliance, A5: Modesty, and A6: Tender-Mindedness. This is clearly not the case from the $\cup$Trait perspective. We would expect the sum of Trust, Straightforwardness, and Altruism to be substantially correlated with the sum of Compliance, Modesty, and Tender-Mindedness, but they would certainly not be identical. Although theorists commonly deal with $\cap$Traits, most (if not all) trait measures used to assess individuals (including factor scores) actually assess $\cup$Traits.[9]

Bollen (Bollen, 2002; Bollen & Lennox, 1991) distinguished between latent variables whose indicators were reflective (i.e., caused by the latent variable) versus formative (i.e., constituting and thus causing the latent variable). The distinction between $\cap$Traits and $\cup$Traits is somewhat different. $\cap$Traits correspond to latent variables with reflective indicators, but the indicators of $\cup$Traits are both

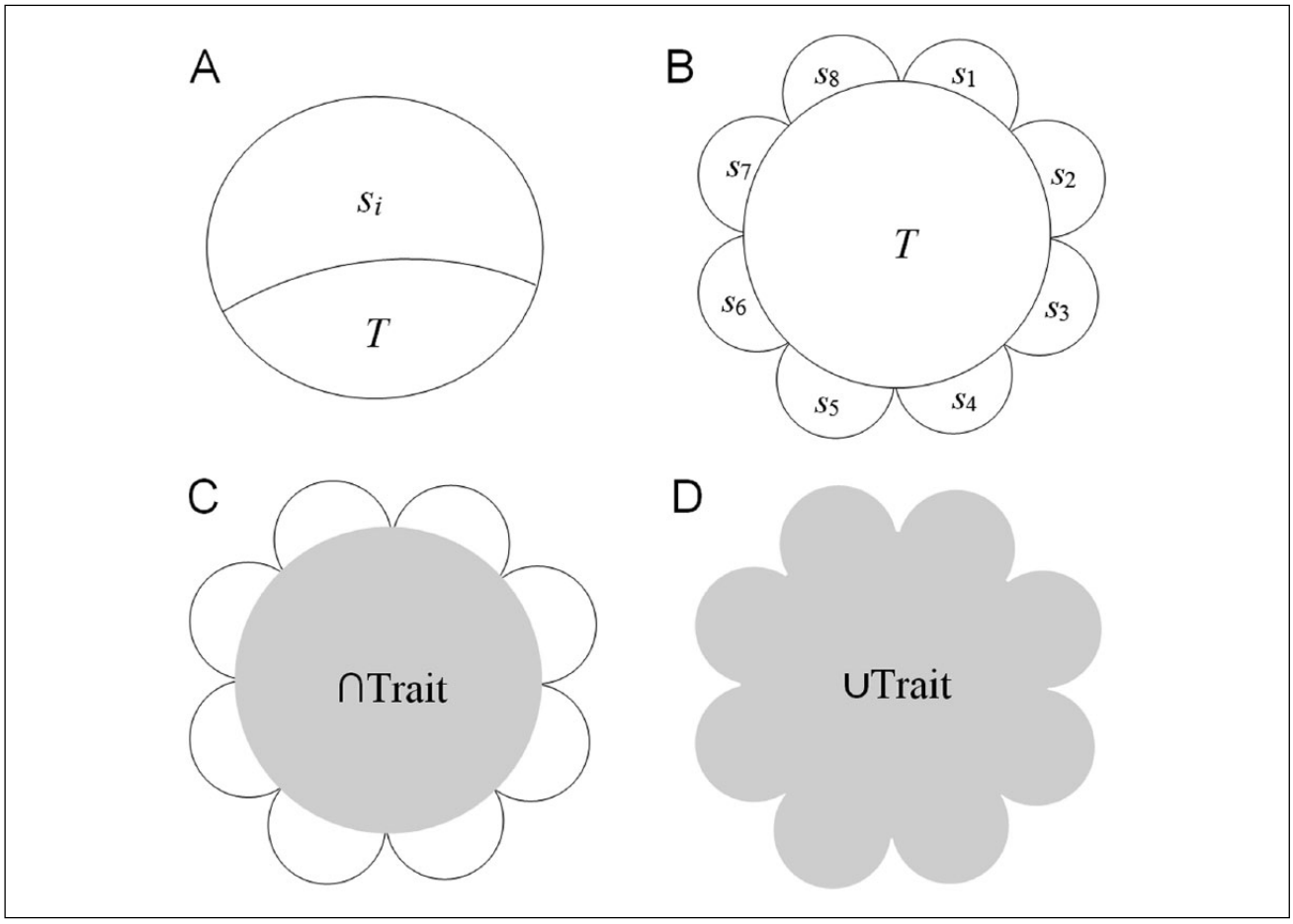

Figure 3. Schematic representation of proportions of true-score variance in (A) an item and (B) an eight-item scale. Because effects aggregate and variance is a squared term, the common trait variance, *T*, forms a much larger portion of the variance in the scale than in the item, and this would be increasingly true as the number of items with distinct specific variances, s_i, increases. Shaded area represents the variance in the scale construed as an intersection trait (C) and as a union trait (D).

reflective (insofar as they are assumed to be caused by the core ∩Trait) and formative (insofar as they contribute specific variance to the assessed trait).

Nuances of Facets and Scale Development

Specific variance at the item level is less commonly discussed than specific variance at the facet level, and is rarely researched. Even psychologists who adopt a ∪Trait view of domains may not appreciate its importance for facets. For example, Costa and McCrae (1994) wrote, "For narrow constructs, the higher the internal consistency, the better. For broad constructs, however, higher internal consistency is not necessarily better, because it may be purchased with a loss of generality" (p. 130). However, of course, the same loss of generality may occur at the facet level. Instead of seeking to maximize the internal consistency of facets, it may be better to adopt a ∪Facet perspective and attempt to include as many nuances of each facet as possible.

Recall that the term *nuance* refers to different forms of a facet, corresponding operationally to individual items (or groups of conceptually redundant items).[10] For example, one of the NEO Inventory N2: Angry Hostility items concerns feelings of bitterness and resentment; another concerns being hot-blooded and quick-tempered. These are clearly related affective dispositions (they are characteristic negative emotions prompted by and focused on the perceived hostile actions of others), but they are also discriminable: Bitterness is a far more passive reaction than temper. An Angry Hostility scale ought arguably to include both these nuances, but in consequence, the scale will have relatively more specific and less common variance.

In the classical model, trait variance is identified with common variance, and that suggests that scales should be developed with internal consistency as the overriding goal. Ignoring for the moment the fact that much or most of the common variance is in fact trait-irrelevant method variance (see Table 1), this strategy would promote the suppression of

extraneous specific variance. There are two feasible ways to do this. The first is to increase the number of items, so that the common variance (which is aggregated across items) would progressively outweigh the specific variances of individual items. Here there is a trade-off between efficiency and purity of assessment that individual researchers and assessors must weigh.

The second method is to refine the item pool by selecting items that most fully correspond to the common variance—for example, have the highest loadings on the first item factor. If the common variance were in fact the trait variance of interest, this would be an impeccable strategy. However, a frequent result of this process is the selection of items that essentially ask the same question again and again. The common variance of the resulting scale is not composed purely of trait variance but of trait variance plus the specific variance associated with the particular nuance of the trait that the repeated item assesses. An Angry Hostility scale consisting solely of bitterness items would in fact be a Bitterness scale and a relatively poor predictor of, say, spouse ratings of temper tantrums. By maximizing alpha, specific variance has not been eliminated; instead, it has infiltrated the operationalization of the trait. Of course, this is not a new insight; Cattell (1973) famously called scales with excessively high internal consistencies "bloated specifics."

If we abandon the classical model and adopt a ∪Facet approach to scale development, the ideal scale is one in which all important nuances of the facet are represented equally. As Watson (2012) pointed out, this is traditionally viewed as the issue of content validity. The obvious difficulty this strategy poses is that of identifying all important nuances of each facet—indeed, researchers do not even agree on the facet-level components of the broader and better-understood domains (McCrae & Costa, 2008; Roberts, Bogg, Walton, Chernyshenko, & Stark, 2004). However, in that case, some progress has been made. The facets selected by Costa and McCrae for the NEO Inventories show some correspondence to those independently identified by other researchers—see, for example, the alignment (in McCrae, 2009) of NEO Inventory facets with the scales of the 16 Personality Factors Questionnaire (16PF; Cattell et al., 1970) and the Eysenck Personality Profiler (Eysenck, Barrett, Wilson, & Jackson, 1992), and the identification of NEO-like facets in the BFI (Soto & John, 2009). Thus, it is reasonable to suggest that deliberate efforts at maximizing diversity in item content might also be effective at the level of items. Perhaps a panel of psychologists could be asked to write items that exemplify a given facet; a content analysis could sort these rationally into distinct nuances.

Empirical item selection would be different here than in the classical model. It would not, in general, be wise to submit all the items suggested by the panel to an item analysis, because such a strategy would favor the selection of the particular nuance or nuances that most item writers happened to identify. Instead, a stratified approach would be preferred: A single item (or perhaps pair of items) should be chosen for each conceptually distinct nuance; item analyses would then test whether the candidate items in fact shared the common variance of the facet.

Structural equation modeling could also be useful. These models can identify items with what are known as *correlated errors*—that is, associations that cannot be accounted for by the latent dimension common to the pool of items. One cause of correlated errors might be redundancy in specific variance. For example, a rational analysis might suggest that *bitterness* and *resentment* are distinct nuances of Angry Hostility, but very high correlations between items assessing these two would suggest that they are better combined into a single nuance.[11] The specific variance, s_i, that distinguishes different nuances of the same facet is not, from the ∪Trait perspective, really error, so the term *correlated error* is a misnomer here; instead, we might designate this as *redundant specifics*. However, if the goal is to have a scale in which each nuance is represented by exactly one item, conventional indicators of correlated error might be a useful tool.

It is easier to check for redundancy within an item pool than for comprehensiveness. One way to evaluate the representativeness of a given set of nuances would be to compare items with those of other scales purporting to assess the same facet-level trait: Do different scale developers converge on the same nuances of the facet? If not, the new item pool might need to be expanded.

Parallel Forms: Uses and Limitations

Psychologists who develop cognitive or achievement tests routinely create parallel forms so that respondents who retake the test do not benefit from prior exposure to the items. Parallel forms consist of different sets of items, but ideally have identical psychometric properties, so that the correlation between one form and another ought (under the classic model) to be equal to coefficient alpha. The model developed here (Equation 1) requires a different conceptualization of parallel forms, in which specific variance must also be considered: Truly parallel forms ought also to contain items representing the same set of nuances.

How does one determine whether alternate forms of X and Y are strictly parallel in this sense? Conceptually, one can conduct a content analysis of the two sets of items to see whether they contain the same distribution of nuances. Statistically, there are also some ways to assess this. Most obvious is that the retest reliabilities (which in both cases should equal $T + M + s_{I\text{-}k}$) should equal the cross-form correlation:

$$r_{tt} = r_{XX} = r_{YY} = r_{XY}. \quad (6)$$

Because it is unlikely that two forms will be perfectly parallel, it would be useful to get a sense of how nearly they are

matched on content. Let us assume that two forms are parallel in the sense that they contain equal proportions of *T, M*, and *s* variance, but that it is not known to what extent the specific variances are the same in the two forms—for example, each form might assess three nuances of the facet, but none, one, two, or three of the nuances might be the same across the two forms. Then it might appear that observed parallel form reliability will be intermediate between internal consistency, where no nuances are shared (= $T + M$) and retest reliability, where all nuances are shared (= $T + M + s_{1\text{-}k}$):

$$\alpha < r_{XY} < r_{tt}, \tag{7}$$

and that the proportion of overlapping specific variance, *P*, would be given by the following:

$$P = \left[r_{XY} - \alpha\right] / \left[r_{tt} - \alpha\right]. \tag{8}$$

The problem with this formula is that *T* (and thus alpha) may be spuriously inflated by specific variance if one or both of the forms is a "bloated specific." For example, if Form X of an Angry Hostility scale consists entirely of bitterness items, whereas Form Y consists solely of temper items, alpha is likely to exceed parallel form reliability. In fact, violations of the Equation 8 would suggest just such a scenario.

Although they are occasionally encountered in personality measures—perhaps most notably in alternate forms A, B, C, and D of the 16PF—most trait researchers do not bother to develop parallel forms. The practice effects that bedevil cognitive testers do not seem to matter much in personality assessment. In consequence, longitudinal stability is routinely assessed by administering the same instrument twice. Similarly, in the usual practice, heritability is assessed by asking twins to complete identical questionnaires, and cross-observer validity is frequently based on the responses of two different raters of a target to the same set of items (Vazire, 2006), or to parallel forms that differ only in the use of first-person versus third-person phrasing (McCrae et al., 2004).

The advantage of this strategy is obvious: It obviates the need to develop truly parallel forms. However, there is also a potential drawback, because when identical forms are administered to the same person, estimates of reliability could be inflated by what Schmidt and colleagues (2003) called *specific factor error*, an artifactual form of specific variance. For example, in the usual retest design, the respondent is faced with an identical set of items on two occasions. If the respondent misinterprets one of the words in an item, he or she may respond incorrectly, and may do so on both occasions. This systematic error—although it would not be psychologically meaningful—would inflate retest reliability. Disattenuating with such a reliability estimate would undercorrect for error (Schmidt et al., 2003). Use of a parallel form would minimize this possibility.

It is unlikely, however, that both a self-report respondent and an informant who rated that target would misunderstand a term in the same way; artifactual specific variance would not in general be shared. In the model presented above (Equation 1), cross-observer agreement is attributed to shared trait and specific variance ($T + s$); if different observers did not share specific variance, r_{CA} would be simply *T*. If specific variance were nothing but error, r_{CA} would be better predicted by alpha ($T + M$) than by retest reliability ($T + M + s$), because the latter includes more error. This is clearly not the case (McCrae, Kurtz, et al., 2011). In other words, the specific variance of interest seems to be a substantive characteristic of the item, apparent to independent observers and contributing to cross-observer agreement on the facet. This would be the counterpart on the item level of the consensual validation of facet-level specific variance already demonstrated for the NEO Inventories (McCrae & Costa, 1992).

There is a long chain of inference between the observation that $r_{tt} > \alpha$ for most NEO Inventory facets, and the prediction that the specific variance in items is a substantive property of nuances of personality that can be consensually validated, so it is of particular interest that a recent study has tested that hypothesis. Mõttus, McCrae, Realo and Allik (2013) examined agreement between self-reports and informant ratings using the Estonian translation of the NEO Personality Inventory–3 in a large ($N > 2{,}500$) sample. They calculated correlations across raters for each of the 240 single items, and also for item residual scores, controlling for the facet scale to which the item belonged. Under the assumptions of the present model, this residual consisted solely of specific variance and error. Correlations for the raw items ranged from .13 to .56 ($M = .31$); correlations for the residual scores ranged from .06 to .47 ($M = .19$, all ps < .001). Every item had valid variance net of the facet to which it contributed, and the magnitude of the residual correlations was not much less than that of raw scores. Item-specific variance is indeed an observable characteristic of personality.

The usual design used to assess stability, heritability, and cross-observer validity—administration of the same items in two conditions—ensures that both common and item-specific variances are assessed. When analyzed through correlations between observed variables (as opposed to latent variable modeling), these designs determine the stability, heritability, and cross-observer validity of *what the scale measures*. This is a ∪Facet conception, where the union includes all and only those nuances of the facet that are included in the scale. The observed values are probably an upper limit to what would be seen with ostensibly parallel forms that differed in item content, because the specific variance is likely to vary somewhat across different forms. For the same reason, differential retest reliability is probably best as a predictor of differential validity (McCrae, Kurtz, et al., 2011) when identical items are used to assess stability, heritability, or cross-observer agreement.

Schmidt and colleagues (2003) recommended that reliability be assessed as the coefficient of equivalence and

stability (CES), derived from the administration of parallel forms on different occasions. They argued that the CES controls for random, transient, and specific factor errors, and is thus an appropriately conservative estimate of reliability. From the present perspective, this view is problematic, both because of the difficulty of constructing genuinely parallel forms, and because specific variance may in fact be valid trait variance at the nuance level. Schmidt and colleagues relied entirely on self-report data and thus had no way to separate specific error from valid specific variance. The present model (Equation 1) suggests that the true reliability of a scale (what is consistently measured across occasions, observers, and parallel forms) is given by $T + s$, which is simply the cross-observer correlation, r_{CA}. When heteromethod correlations are examined, it is appropriate to use r_{CA} to disattenuate. For example, McCrae (1994) argued that true-score longitudinal stability for a trait could be calculated as the ratio of cross-lagged to concurrent cross-observer correlations—in essence, using the concurrent r_{CA} to disattenuate the cross-observer correlation observed over time. It would, however, be inappropriate to use r_{CA} to disattenuate correlations between two self-report measures—indeed, that procedure would often lead to estimated values over 1.0. That is because correlations between self-reports (or between informant ratings from the same rater and of the same target) are inflated by method variance. For monomethod correlations, a better choice for disattenuation might be $T + M + s$, which is retest reliability.[12]

Alternatives to Classical Test Theory

In this article, I adopt a revised version of classical test theory, but it is worth mentioning briefly two alternatives. One envisions different components of variance; the other adopts a different analytic approach.

In the classical model, the trait variance, T, is identical in all its indicators; stripped of method, specific, and error variance, any one of them would be a perfect measure of the trait. An alternative conception (suggested by a reviewer) is that traits are intrinsically compound, and that different indicators assess different parts of the trait. Neuroticism, for example, might have a publically observable part that self-reports and observer ratings would share, and a private, intrapsychic part accessible only via self-report. A Neuroticism measure based solely on observer ratings, no matter how many were aggregated, would never provide a perfect measure of the trait, and might be outpredicted by a single self-report when the criterion (say, suicide) depended on both observable and private components.

This more differentiated view of trait variance has some attractions, particularly when causal accounts are proposed. In a later section, I discuss such a model with respect to hypothetical genes underlying the heritable part of Neuroticism. However, it is possible in many respects to view this model as a version of ∪Traits. Observer-rated Neuroticism and Self-Reported Neuroticism might be considered facets of ∪Neuroticism, where Self-Reported Neuroticism has a specific private component not shared by Observer-rated Neuroticism.

A well-established alternative to classical test theory is Generalizability (G) theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In classical theory, scales are described in terms of the proportions of variance attributable to various causes, as if these were intrinsic and unconditional properties of the scale. In fact, reliability and validity depend on the population in which the scale is used and the circumstances under which it is administered. For example, if trait variance is restricted in a given population, T and coefficient alpha will normally be reduced. G theory begins with this perspective; it seeks to analyze proportions of variance in a set of observed scores due to such features as persons, items, occasions, observers, and their interactions. There are of course conceptual and mathematical correspondences between classical and G theories. For example, variance attributable to persons in G theory corresponds to T in classical theory: Persons have a universe score in G theory as they have a true score in classical theory.

Of particular interest here is the concept of item-specific variance. In the analysis-of-variance framework used in G theory, this corresponds to a Person × Item interaction term, which "reflects the fact that not all people find the same items easy or difficult" (Shavelson & Webb, 1991, p. 21). An Angry Hostility item about bitterness would be "easier" for a respondent high on that particular nuance than for one who was low, even if the two respondents have identical levels of Angry Hostility. This is a legitimate way to analyze specific variance, but it does not seem conducive to substantive insights into the nature of the construct, nor to the prediction that similar interactions would be found if observer ratings were substituted for self-reports.

Some Implications

Practical Consequences

The bottom row of Table 1 gives the mean values across the 30 NEO facet scales for estimated components of variance at the facet scale level. The proportion of variance in the typical item can be estimated from these data (see Appendix). Because of aggregation, we would expect that the proportion of trait and method variance would be substantially larger at the scale level than at the item level, and thus that items would have more specific variance and error than scales have. For the typical NEO Inventory item,

$$T = (\alpha - r_{tt} + r_{CA}) / (8 - 7\alpha) \approx .12, \quad (9)$$

$$M = (r_{tt} - r_{CA}) / (8 - 7\alpha) \approx .13, \quad (10)$$

$$s_i = (8(r_{tt} - \alpha)) / (8 - 7\alpha) \approx .24, \quad (11)$$

and the remainder (.51) is error.[13] The observed values are sobering: In the typical item, nearly two thirds of the variance is either random or systematic error ($\varepsilon_i + M$), which is why single items are notoriously unreliable; of the remaining true-score variance ($T + s_i$), only a third is due to the common trait. It might be possible to write items with a higher proportion of common variance, but it is probably fair to assume that most single items have a substantial portion of specific variance. The implication is that single-item scales, and even two-item scales (e.g., Gosling, Rentfrow, & Swann, 2003), are more problematic than usually supposed. In addition to considerable error, they contain what is, at least from a ∩Trait perspective, substantive bias. "I worry a lot" might prove to be a decent one-item Neuroticism measure in the sense that it correlates substantially with longer Neuroticism scales, but it is also more particularly an anxiety measure, and one that emphasizes the *apprehension* nuance of anxiety more than the *tension* nuance (Spielberger et al., 1979). Correlations of outcomes with this item might reflect the influence of general Neuroticism, or of anxiety, or of apprehension, and there is no easy way to decide which.

Pullmann, Allik, and Realo (2009) examined this issue using the individual items of the Rosenberg Self-Esteem Scale in large Internet and nationally representative samples in Estonia. They were concerned with age differences across the life span, which are relatively subtle, and found that individual items

> have quite different lifespan trajectories: some of the items are relatively stable . . . the means of other items increase across age groups, and finally, the third group of items reach the maximum value in early adulthood and start to decline afterwards. (p. 31)

They attributed these differences to "idiosyncracies about the single item" (p. 22), which we might interpret as specific variance. McCrae and colleagues (1999) reported that the specific variance in NEO Inventory facet scales showed distinct developmental trajectories; the same appears to be true for single items. This research suggests that the analysis of single items can be seriously misleading if they are viewed simply as short measures of a broader trait.

There are research contexts (e.g., telephone surveys; Mondak, 2010) in which the use of highly abbreviated scales is unavoidable, but researchers need to be aware that results may reflect the specific content of the items they use. Ideally, research on large and random samples using brief scales would be supplemented with validating research on smaller samples whose respondents were administered full-length versions of the measures. If the same pattern of results emerges in both samples, it is probably trustworthy; if not, further research is needed before conclusions can be drawn.

Once the specific variance in individual items is recognized, it is tempting to capitalize on it by using nuances as predictors in their own right. A single item might be a better predictor of some criterion than is the scale to which the item belongs, just as single facets may be better predictors than the domain scale to which they contribute. There is, however, a crucial distinction between these two levels: At the present time, there is a wealth of information on the construct validity of many facet scales, but almost none on the validity of individual items as measures of nuances. Combined with their high proportion of error ($M + \varepsilon$) and the increased probability of Type I errors if exploratory analyses of many items are conducted, this fact discourages use of individual items as predictors.

Causal Models

Two contrasting models have been proposed for traits (Bollen & Lennox, 1991; Borsboom, Mellenbergh, & van Heerden, 2003). In the first, reflective, model, the causal basis of the trait is a latent variable that gives rise to various manifestations that are then taken as *effect indicators* of the trait. In the second, formative, model, the causal order is reversed, and the trait emerges as the sum of its components, which are *causal indicators* (Watson, 2012). Socio-economic status, for example, is a summary variable that does not cause income and education, but expresses them. Unless some process of suppression is at work, effect indicators must be correlated, whereas this is not necessarily true of causal indicators. Wealth, for example, might be the sum of earned income and lottery winnings, although we would not expect these to be correlated.

Of these two models, trait theorists who take a realist position are more likely to choose the former: There is something about the individual (perhaps genes or infant attachments or peer modeling) that causes the appearance of the relevant trait indicators at a characteristic level. However, this rather vague formulation (which can hardly be otherwise in the present state of our knowledge) is not necessarily isomorphic with a latent trait model in which the covariation of indicators is due to a single causal entity. As Bartholomew, Deary, and Lawn (2009) have shown, many different causal structures are compatible with any given correlation matrix and associated latent structure. It is worthwhile considering this idea with respect to specificity in the trait hierarchy.

Cramer and colleagues (2012) argued that the usual conception of personality factors as latent traits—that is, ∩Traits—suggests a particular genetic model, in which a gene, or a set of genes, determines the level of the factor, which in turn affects all its facets. Certainly, that is a reasonable and parsimonious model, but it is not the only one consistent with the data. It is possible, for example, that the heritable portion of Neuroticism is caused by a large set of genes, and that facets of Neuroticism are caused (in part) by overlapping subsets of the full set. As long as all the intersections of these subsets are non-empty (that is, all pairs of facets share some common cause, and are thus correlated), such a model could give rise to a correlation matrix that would suggest a single latent trait (Bartholomew et al., 2009). This

can be illustrated by hypothetical sets of Genes a, b, c, and so on, associated with traits:

Neuroticism: {a, b, c}
Anxiety: {a, b}
Angry Hostility: {b, c}
Depression: {a, c}

In this example, there is no single gene that is shared by, and thus gives rise to, all three facets; in this strict sense, there is no common core, and interpreting the latent variable model (even if it perfectly fits the observed data) as the representation of a single cause is wrong. There is nothing, however, to prevent us from saying that a Neuroticism-related *set* of genes is a cause of each of the facets.

In reality, given the very small effects of individual genes (McCrae, Scally, Terracciano, Abecasis, & Costa, 2011), it seems likely that factors correspond to extremely large sets of genes, some of which may affect all facets, some many facets, and some only a few facets. A gene that affected only one facet, however, would not be part of the set; it would instead be one of the sources of specific variance in the facet (see Jang et al., 1998). A better representation of Neuroticism and its facets would thus be as follows:

Neuroticism: {a, b, c}
Anxiety: {a, b, x}
Angry Hostility: {b, c, y}
Depression: {a, c, z}

where x, y, and z are genes contributing to the discriminant validity of the three facets.

If we adopt a ∪Trait view of Neuroticism, its genetic basis would then be {a, b, c, x, y, z}; these are the genes that contribute to the *observed* scores on Neuroticism measures. The already large pool of genes underlying ∩Trait Neuroticism is thus even larger in practice; this is one reason to focus molecular genetic analyses on facet-level traits (Terracciano & McCrae, 2012).

Jang and colleagues (1998) showed that the specific variance in facet scales is heritable. The fact that retest reliability—which includes item-specific variance—is a good predictor of differential heritability suggests that nuances of facets also have heritable specific variance. The argument here parallels that for cross-observer agreement: If specific variance were not heritable, it would be error from a behavior genetics perspective, and alpha (which treats specific variance as error) would be a better predictor of differential heritability than retest reliability.[14] The heritability of specific variance in nuances of facets was envisioned by Harkness and McNulty (2002), who wrote "there may be important genetically influenced determining tendencies at all levels of trait breadth, from broad common traits to narrow facets to unique traits" (p. 397). Using anecdotal evidence from studies of monozygotic twins raised apart, they speculated that even highly specific traits unique to individuals (like preference for a particular brand of toothpaste) may have some genetic basis, and they provided a thoughtful discussion of the clinical implications of that scenario.

We would then need to hypothesize new genes at lower levels:

Neuroticism: {a, b, c}
Anxiety: {a, b, x}
Angry Hostility: {b, c, y}
Bitterness: {b, c, y, β}
Temper: {b, c, y, γ}
Depression: {a, c, z}

These additional genes (β, γ, etc.) that distinguish nuances of facets also in principle contribute to ∪Neuroticism, although their effects are presumably quite small.

Scalar Equivalence in Groups and Individuals

Scalar equivalence refers fundamentally to the idea that the same raw score represents the same underlying trait level when a scale is administered to different groups. Like construct validity, this property can only be established by a network of evidence, but statisticians have attempted to address it, at least in part, by internal analyses of factor loadings and intercepts. The basic logic is that the individual items in a scale ought to follow the same pattern across groups as the scale as a whole; if not, the item must be functioning differently in the groups that are compared. For example, women consistently score higher than men in five of the facets of Openness, but lower in Openness to Ideas (McCrae, Terracciano, & 78 Members of the Personality Profiles of Cultures Project, 2005). As Marsh and colleagues (2010) noted, this means that the O5: Ideas facet shows differential item functioning (DIF) when considered as an item in the Openness domain scale. If women are truly higher in Openness—as most facets suggest—we would expect them to score higher on O5: Ideas, too; because they score lower, something must be "wrong" with the O5: Ideas facet as an indicator of Openness. Some statisticians would recommend omitting O5: Ideas from the Openness domain or factor score, at least when researchers are interested in estimating the difference between men and women on Openness (see Church et al., 2011).

This approach is reasonable if—and only if—one adopts a ∩Trait perspective. In that view, scales are intended to represent only the core construct shared by all the items (along with an unavoidable quantity of random error). DIF suggests that some systematic bias prevents an item from faithfully representing the core construct. This bias might, of course, be an artifact that truly distorts scores—for example, Openness to Ideas might be more socially desirable for men than for women, leading men to exaggerate their true level. However, it might also be the result of group differences in

specific variance associated with the item. Men may in truth be higher in Openness to Ideas, and women may in truth be higher in all the other facets of Openness.

What does that imply for the domain or factor score of Openness—do women truly score higher than men in total Openness, or is that statement *not meaningful*, as the title of Church and colleagues' (2011) article seems to suggest? From a strict ∩Trait perspective, the meaning is surely problematic. Just as main effects may be qualified by interactions in an analysis of variance, so domain-level comparisons may need to be qualified by facet-level comparisons: "Women are higher in Openness to Experience, except for Openness to Ideas," is surely a meaningful and informative statement.

If, however, one adopts a ∪Trait perspective, then there is nothing wrong with stating that women are higher in total Openness, because Openness is the sum of its various expressions.[15] Openness manifests itself in somewhat different form in women than in men, and a statement that specifies those differences would be more informative than a blanket assertion about which group is higher. However, a sum score is most certainly not meaningless; it should allow inferences about other, extra-test manifestations and correlates. If ∪Traits are (in part) formative variables, then it is reasonable to say that the same consequences can arise from different facet-level causes; facets may be legitimately combined because they are functionally equivalent (cf. Cole, Maxwell, Arvey, & Salas, 1993, on the analysis of *emergent* constructs). People high in Openness may prove to be equally creative, whether they are especially open to Ideas (like Thomas Jefferson) or to Fantasy (like Jean-Jacques Rousseau; see McCrae & Greenberg, 2014). Similarly, people who are very high in Neuroticism are likely to be unhappy, regardless of whether they are predominantly depressed or anxious or painfully shy.

The choice of a ∪Trait or a ∩Trait perspective is particularly important in interpreting scores of individuals. Although scalar equivalence is usually discussed in the context of groups, in principle it also applies to individuals. If Mary has a total Openness domain score of 108, is she just as open as John, who also scores 108? If both score 18 on each of the six facets, it seems clear that they are equally open. However, suppose Mary scores 13 on Openness to Fantasy, Aesthetics, and Feeling but 23 on Openness to Actions, Ideas, and Values, whereas John has the opposite pattern. Using combined-sex self-report norms (McCrae & Costa, 2010), these scores imply that Mary is down-to-earth, insensitive to art, and unempathic, but enjoys novelty and intellectual challenges and has liberal views; John is imaginative, artistic, and passionate, but set in his ways, uninquisitive, and conventional. The statement that they are equally open—or indeed that they are each *average* in Openness—seems odd.

In fact, most people show similar levels for all the facets in a given domain (Allik et al., 2012), but exceptions exist, and these are not merely outliers on the distribution of errors that should tend to cancel out in summary scores. Studies of self–other agreement show that some atypical profiles are consensually validated (Allik et al., 2012), suggesting that real differences in facet-specific variance account for the unusual patterns. It is for this reason that the NEO Interpretive Report states that "to the extent that there is wide scatter among facet scores within a domain, interpretation of that domain and factor becomes more complex. In these cases, particular attention should be focused on the facet scales" (Costa & McCrae, 1992, p. 22).

This phenomenon suggests a possible test of the utility of ∩Trait versus ∪Trait conceptions. When comparing groups such as men and women, the ∩Trait requirement of scalar equivalence suggests that facets that show DIF should be omitted from the total score. By a similar logic, one could argue that individuals who show DIF—that is, whose facet scales show large within-domain scatter—should be screened out from a research sample. Would that in fact increase validity coefficients? For each domain, a large sample might be stratified by total score, and then divided into those who showed more and less variance among the facet *T*-scores. Domain scores could then be correlated with a range of theoretically meaningful criteria within each subsample, and the corresponding validity coefficients compared. The ∩Trait perspective would argue for more meaningful scores and thus higher validities in the non-DIF group, whereas the ∪Trait perspective would predict equal validities.[16]

An Unsolved Mystery

The hypothesis of specific variance associated with particular nuances of facets (and thus with individual items) can explain why retest reliability is higher than internal consistency, and why it is a better predictor of differential validity. When retest reliability is high, random error is low, and it is hardly surprising that validity is higher. However, as yet we have no clear idea why some measures of some traits are more heavily infested with error than others. Because error is the inverse of retest reliability (Equation 5), it is tempting to see it as *transient* error and attempt to explain why some items are particularly susceptible to change over time. However, the finding that $r_{tt} > \alpha$ (at least for brief personality scales) implies that time of measurement is essentially irrelevant. Some scales have lower retest reliability simply because they show more error whenever they are measured.

Watson and colleagues (Chmielewski & Watson, 2009; Watson, 2004) have called attention to the need for research on this topic. McCrae and colleagues (2011) tested a few hypotheses related to item ambiguity: NEO Inventory facets with higher reading levels, more unfamiliar words, negations, or conditional phrasings might have been confusing to some respondents, leading to more random error. However, in fact, none of these was significantly related to retest reliability. Wood and Wortman (2012) have shown that the extremeness and desirability of items may play a role. Error can be reduced by using more items in a scale; efficiency would be improved by using better items—if we knew how to write them.

Appendix

Derivation of the Components of Variance in Single Items From an Eight-Item Scale

A normalized individual item can be represented as a weighted combination of common trait, method, specific, and error variance:

$$\text{Item} = aT + bM + cs_i + d\varepsilon_i, \tag{1A}$$

where a, b, c, and d are standardized such that $a^2 + b^2 + c^2 + d^2 = 1$. The sum of eight such items is

$$\begin{aligned}\text{Scale} = {} & 8aT + 8bM + cs_1 + cs_2 + \ldots + \\ & cs_8 + d\varepsilon_1 + d\varepsilon_2 + \ldots + d\varepsilon_8.\end{aligned} \tag{2A}$$

This variable can be normalized by dividing each coefficient by the root sum of squares of the coefficients,

$$\text{SQRT}\left(64a^2 + 64b^2 + 8c^2 + 8d^2\right). \tag{3A}$$

Because the sum of the squares of the item coefficients is 1, $d^2 = 1 - a^2 - b^2 - c^2$, and this denominator can be written as

$$\begin{aligned}&\text{SQRT}\left(64a^2 + 64b^2 + 8c^2 + 8 - 8a^2 - 8b^2 - 8c^2\right) \\ &= \text{SQRT}\left(8\left(1+7\left(a^2 + b^2\right)\right)\right).\end{aligned} \tag{4A}$$

Now, coefficient alpha is the total variance in the scale due to $T + M$, because these two account for the intercorrelation of items (see main text). This total variance is the sum of the squares of the standardized scale coefficients for T and M:

$$\begin{aligned}\alpha = {} & \left[64a^2/\left(8\left(1+7\left(a^2+b^2\right)\right)\right)\right] + \\ & \left[64b^2/\left(8\left(1+7\left(a^2+b^2\right)\right)\right)\right] \\ = {} & 8\left(a^2+b^2\right)/\left(1 + 7\left(a^2+b^2\right)\right).\end{aligned} \tag{5A}$$

From this, we can calculate that

$$\left(a^2+b^2\right) = \alpha/(8-7\alpha). \tag{6A}$$

Retest reliability for the scale is due to *T, M*, and *s* (see main text) and is thus given by the sum of the squares of the standardized scale coefficients for these three:

$$\begin{aligned}r_{tt} = {} & \left(64a^2+64b^2+8c^2\right)/\left(8\left(1+7\left(a^2+b^2\right)\right)\right) \\ = {} & \left(8a^2 + 8b^2 + c^2\right)*(8-7\alpha)/8.\end{aligned} \tag{7A}$$

Cross-observer agreement for the scale is due to *T* and *s* (see main text) and is thus given by the sum of the squares of the standardized coefficients for these two:

$$\begin{aligned}r_{CA} = {} & \left(64a^2+8c^2\right)/\left(8\left(1+7\left(a^2+b^2\right)\right)\right) \\ = {} & \left(8a^2+c^2\right)*(8-7\alpha)/8\end{aligned} \tag{8A}$$

From Eqs. 7A and 8A, it follows that

$$r_{tt} - r_{CA} = 8b^2 * (8-7\alpha)/8, \tag{9A}$$

and

$$b^2 = \left(r_{tt} - r_{CA}\right)/(8-7\alpha). \quad [\text{Method Variance}] \tag{10A}$$

Because $(a^2 + b^2) = \alpha/(8 - 7\alpha)$, Equation 10A implies that

$$\begin{aligned}a^2 = {} & \left[\alpha/(8-7\alpha)\right] - \left[\left(r_{tt} - r_{CA}\right)/(8-7\alpha)\right] \\ = {} & \left(\alpha - r_{tt} + r_{CA}\right)/(8-7\alpha). \quad [\text{Trait Variance}]\end{aligned} \tag{11A}$$

Substituting this value for a^2 into Equation 8 gives

$$r_{CA} = \left(\alpha - r_{tt} + r_{CA}\right) + \left[c^2(8-7\alpha)\right]/8, \tag{12A}$$

so

$$\left[c^2(8-7\alpha)\right]/8 = r_{CA} - \left(\alpha - r_{tt} + r_{CA}\right) = r_{tt} - \alpha, \tag{13A}$$

and

$$c^2 = 8\left(r_{tt} - \alpha\right)/(8 - 7\alpha). \quad [\text{Specific Variance}] \tag{14A}$$

The rest is error.

Acknowledgments

I thank Donald Filan, John Kurtz, René Mõttus, Antonio Terracciano, and David Watson for helpful comments.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Robert R. McCrae receives royalties from the NEO Inventories.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Cronbach's (1951) formulation considered the more general case where items were not equally valid, and these assertions are only approximately true.
2. Retest reliability is normally assessed by a Pearson correlation of scores collected a few days or weeks apart. It might also be assessed by an intraclass correlation; however, if both true-score and method variance are unchanged (as I will assume here), mean levels will not change, and the two coefficients will be essentially the same. Retest reliability can also be estimated from three or more longitudinal administrations, using Heise's (1969) method. McCrae, Kurtz, and colleagues (2011) showed that Heise estimates are similar (although not identical) to simple retest correlations. The correlation between

two administrations of a scale before and after a significant intervention (e.g., Piedmont, 2001) cannot be used to assess retest reliability, because unreliability of measurement is confounded with differential change in true scores.

3. This discussion refers exclusively to trait measures; measures of moods or states would of course be expected to show changes in the true score, *T*, across occasions, further reducing r_{tt}.
4. The same pattern is seen for the Personality Research Form (Jackson, 1984; *Mdn* α = .70, *Mdn* 2-week r_{tt} = .91). It is less marked for the longer scales of the Multidimensional Personality Questionnaire (Tellegen & Waller, 2008; *Mdn* α = .85, *Mdn* 1-month r_{tt} = .89) and disappears for the 48-item domain scales of the Revised NEO Personality Inventory (NEO-PI-R; Kurtz & Parrish, 2001; *Mdn* α = .92, *Mdn* 1-week r_{tt} = .92). This is presumably because the influence of item-specific variance on a scale (described in a later section) decreases roughly as the square of the number of items.
5. DeYoung, Quilty, and Peterson (2007) have adopted the term *aspects* for a trait level between facets and domains; I will use *nuances* to refer to a level below facets.
6. Cronbach (1951) noted that alpha "treats the specific content of an item as error, but the coefficient of precision [an instantaneous retest reliability] treats it as part of the thing being measured" (p. 307). He apparently regarded specific variance as unimportant.
7. In real data, different raters may share variance that is not due to the true score of the trait—for example, in the phenomenon of false consensus. In that case, some portion of what appears as *T* is actually *M*. I disregard such complications here.
8. These residuals are not pure measures of the specific variance in a facet, because factor scores are not pure measures of the common variance. In addition to their own specific variance, facet residuals also include small contributions from the specific components of the other facets (as well as error). It is possible to analyze residual facet scores through latent variable modeling (e.g., Mõttus, McCrae, Realo, & Allik, 2013).
9. Some readers may find a musical analogy helpful: ∩Traits are like the fundamental tone of a note, the 440 vibrations-per-second A that is identical for all the instruments in an orchestra. ∪Traits are like each sounded note, including the fundamental tone but also all the overtones that distinguish the A of a violin from the A of an oboe. Factor analysts are concerned with pure tones; personologists may prefer the riches of a full orchestra.
10. Nuances might also differ by emphasizing frequency ("I often get angry") versus intensity ("I sometimes get really mad"). Research is needed to determine whether these are simply alternative phrasings of the same nuance, or truly distinct. If they are different, there should be some people who are consistently and consensually high in frequency, but not intensity, and vice versa. At present, it seems prudent to focus on differences in content.
11. It is possible that two highly correlated items still possess demonstrable discriminant validity that would justify their inclusion as separate nuances, but short of a program of validational research on each item, the assumption that high correlations reflect redundancy seems reasonable.
12. Of course, these disattenuated values are not the "true" correlations, because they are inflated by method variance. They do, however, make it possible to compare within-method correlations, at least if we assume that the contribution of method variance is equal for all traits. For example, McCrae, Kurtz, and colleagues (2011) corrected 5- to 10-year stability coefficients using retest reliability and concluded that O5: Ideas and N5: Impulsiveness were intrinsically more stable than O6: Values and N3: Depression.
13. It is of interest to compare these estimates to the cross-observer correlations reported by Mõttus and colleagues (2013). Correlations between raw item scores should be $T + s_i \approx .36$; correlations between residuals should be $s_i / (s_i + \varepsilon_i) \approx .32$. The corresponding observed values of .31 and .19 are reasonably close, although the smaller observed values of the residual correlations suggest that a more accurate model than that developed here would also add separate method terms (m_i) at the nuance level.
14. The same reasoning suggests that item-specific variance must show long-term stability, because retest reliability is a better predictor of stability than is internal consistency (McCrae, Kurtz, et al., 2011).
15. Recall, however, that even in principle, ∪Trait measures are not necessarily interchangeable. The Big Five Inventory (BFI) measure of Openness contains a much higher proportion of items related to ideas than does the NEO-PI-R (Soto & John, 2009), so it is not surprising that women do not consistently score higher than men on this scale (Schmitt, Realo, Voracek, & Allik, 2008).
16. This design presumes that scatter is due to real differences in specific variance at the facet level, and not simple error of measurement, which might be more common in the differential item functioning (DIF) subsample. Equivalent validity at the facet level could be shown by comparable cross-observer correlations for individual facets in the two subsamples.

References

Allik, J., Realo, A., Mõttus, R., Borkenau, P., Kuppens, P., & Hřebíčková, M. (2012). Person-fit to the Five-Factor Model of personality. *Swiss Journal of Psychology*, *71*, 35-45.

Bartholomew, D. J., Deary, I., & Lawn, M. (2009). A new lease of life for Thomson's bonds model of intelligence. *Psychological Review*, *116*, 567-579.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605-634.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305-314.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203-219.

Cattell, R. B. (1973). *Personality and mood by questionnaire*. San Francisco, CA: Jossey-Bass.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *The handbook for the Sixteen Personality Factor Questionnaire*. Champaign, IL: Institute for Personality and Ability Testing.

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

Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katighbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality

Inventory. *Journal of Personality and Social Psychology*, *101*, 1068-1089.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin*, *114*, 174-184.

Costa, P. T., Jr., & McCrae, R. R. (1984). Personality as a lifelong determinant of well-being. In C. Malatesta & C. Izard (Eds.), *Affective processes in adult development and aging* (pp. 141-157). Beverly Hills, CA: SAGE.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1994). Personality assessment. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior* (Vol. 3, pp. 453-460). San Diego, CA: Academic Press.

Costa, P. T., Jr., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment*, *64*, 21-50.

Costa, P. T., Jr., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. Saklofske (Eds.), *SAGE handbook of personality theory and assessment* (*Vol. 2*, pp. 179-198). Los Angeles, CA: SAGE.

Costa, P. T., Jr., McCrae, R. R., & Dye, D. A. (1991). Facet scales for Agreeableness and Conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences*, *12*, 887-898.

Cramer, A. O. J., Van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., . . . Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*, 414-431.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

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

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391-418.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*, 880-896. doi:10.1037/0022-3514.93.5.880

Eysenck, H. J., Barrett, P., Wilson, G. D., & Jackson, C. (1992). Primary trait measurement of the 21 components of the P-E-N system. *European Journal of Psychological Assessment*, *8*, 109-117.

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*, 1011-1027.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504-528.

Harkness, A. R., & McNulty, J. L. (2002). Implications of personality individual differences science for clinical work on personality disorders. In P. T. Costa, Jr. & T. A. Widiger (Eds.), *Personality disorders and the Five-Factor Model of personality* (2nd ed., pp. 391-403). Washington, DC: American Psychological Association.

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review*, *34*, 93-101.

Holden, R. R., & Bernstein, I. H. (2013). Internal consistency: Reports of its death are premature. *Behavior Research Methods*, *45*, 946-949. doi:10.3758/s13428-013-0315-4

Jackson, D. N. (1984). *Personality Research Form manual* (3rd ed.). Port Huron, MI: Research Psychologists Press.

Jang, K. L., McCrae, R. R., Angleitner, A., Riemann, R., & Livesley, W. J. (1998). Heritability of facet-level traits in a cross-cultural twin sample: Support for a hierarchical model of personality. *Journal of Personality and Social Psychology*, *74*, 1556-1565.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The "Big Five" Inventory–Versions 4a and 54*. Berkeley: Institute of Personality and Social Research, University of California, Berkeley.

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the Five-Factor Model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology*, *98*, 875-925.

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment*, *76*, 315-332.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the Big Five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471-491.

McCrae, R. R. (1994). The counterpoint of personality assessment: Self-reports and observer ratings. *Assessment*, *1*, 159-172.

McCrae, R. R. (2009). The Five-Factor Model of personality traits: Consensus and controversy. In P. Corr & G. Matthews (Eds.), *The Cambridge handbook of personality psychology* (pp. 148-161). Cambridge, UK: Cambridge University Press.

McCrae, R. R., & Costa, P. T., Jr. (1983). Joint factors in self-reports and ratings: Neuroticism, Extraversion, and Openness to Experience. *Personality and Individual Differences*, *4*, 245-255.

McCrae, R. R., & Costa, P. T., Jr. (1989). Different points of view: Self-reports and ratings in the assessment of personality. In J. P. Forgas & M. J. Innes (Eds.), *Recent advances in social psychology: An international perspective* (pp. 429-439). Amsterdam, The Netherlands: Elsevier Science.

McCrae, R. R., & Costa, P. T., Jr. (1992). Discriminant validity of NEO-PI-R facets. *Educational and Psychological Measurement*, *52*, 229-237.

McCrae, R. R., & Costa, P. T., Jr. (2008). Empirical and theoretical status of the Five-Factor Model of personality traits. In G. Boyle, G. Matthews, & D. Saklofske (Eds.), *SAGE handbook of personality theory and assessment* (*Vol. 1*, pp. 273-294). Los Angeles, CA: SAGE.

McCrae, R. R., & Costa, P. T., Jr. (2010). *NEO Inventories professional manual*. Odessa, FL: Psychological Assessment Resources.

McCrae, R. R., Costa, P. T., Jr., Lima, M. P., Simões, A., Ostendorf, F., Angleitner, A., . . . Piedmont, R. L. (1999). Age differences

in personality across the adult life span: Parallels in five cultures. *Developmental Psychology*, *35*, 466-477.

McCrae, R. R., Costa, P. T., Jr., Martin, T. A., Oryol, V. E., Rukavishnikov, A. A., Senin, I. G., . . . Urbánek, T. (2004). Consensual validation of personality traits across cultures. *Journal of Research in Personality*, *38*, 179-201.

McCrae, R. R., & Greenberg, D. M. (2014). Openness to Experience. In D. K. Simonton (Ed.), *Handbook of genius* (pp. 222-243). West Sussex, UK: Wiley-Blackwell.

McCrae, R. R., Harwood, T. M., & Kelly, S. L. (2011). The NEO Inventories. In T. M. Harwood, L. E. Beutler, & G. Groth-Marnat (Eds.), *Integrative assessment of adult personality* (3rd ed., pp. 252-275). New York, NY: Guilford.

McCrae, R. R., Herbst, J. H., & Costa, P. T., Jr. (2001). Effects of acquiescence on personality factor structures. In R. Riemann, F. Ostendorf, & F. Spinath (Eds.), *Personality and temperament: Genetics, evolution, and structure* (pp. 217-231). Berlin, Germany: Pabst Science.

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

McCrae, R. R., Scally, M., Terracciano, A., Abecasis, G. R., & Costa, P. T., Jr. (2011). An alternative to the search for single polymorphisms: Toward molecular personality scales for the Five-Factor Model. *Journal of Personality and Social Psychology*, *99*, 1014-1024.

McCrae, R. R., & Terracciano, A., & 78 Members of the Personality Profiles of Cultures Project. (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology*, *88*, 547-561.

Mondak, J. J. (2010). *Personality and the foundations of political behavior*. New York, NY: Cambridge University Press.

Mõttus, R., McCrae, R. R., Realo, A., & Allik, J. (2013). *Cross-rater agreement on common and specific variance of personality scales and items*. Manuscript submitted for publication.

Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, *81*, 524-539.

Piedmont, R. L. (2001). Cracking the plaster cast: Big Five personality change during intensive outpatient counseling. *Journal of Research in Personality*, *35*, 500-520.

Pullmann, H., Allik, J., & Realo, A. (2009). Global self-esteem across the life span: A cross-sectional comparison between nationally representative and self-selected Internet samples. *Experimental Aging Research*, *35*, 20-44.

Reynolds, S. K., & Clark, L. A. (2001). Predicting dimensions of personality disorder from domains and facets of the Five-Factor Model. *Journal of Personality*, *69*, 199-222.

Roberts, B., Bogg, T., Walton, K. E., Chernyshenko, O. S., & Stark, S. E. (2004). A lexical investigation of the lower-order structure of Conscientiousness. *Journal of Research in Personality*, *38*, 164-178.

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

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, *94*, 168-182.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: SAGE.

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO-PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*, 84-90.

Spielberger, C. D., Jacobs, G., Crane, R., Russell, S., Westberry, L., Barker, L., . . . Marks, E. (1979). *Preliminary manual for the State-Trait Personality Inventory (STPI)*. Tampa: University of Florida Human Resources Institute.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *SAGE handbook of personality theory and assessment* (*Vol. 2*, pp. 261-292). Los Angeles, CA: SAGE.

Terracciano, A., & McCrae, R. R. (2012). Why do birds flock? Causality and the structure of characteristic adaptations. *European Journal of Personality*, *26*, 449-450.

Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality*, *40*, 472-481.

Vazire, S., & Carlson, E. N. (2011). Others sometimes know us better than we know ourselves. *Current Directions in Psychological Science*, *20*, 104-108.

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

Watson, D. (2012). Objective tests as instruments of psychological theory and research. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology. Vol. 1: Foundations, planning, measures, and psychometrics* (pp. 349-369). Washington, DC: American Psychological Association.

Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177-203). Mahwah, NJ: Lawrence Erlbaum.

Wood, D., & Wortman, J. (2012). Trait means and desirabilities as artifactual and real sources of differential stability of personality traits. *Journal of Personality*, *80*, 665-701.