Personality traits below facets

Citation for published version:
Mõttus, R, Kandler, C, Bleidorn, W, Riemann, R & McCrae, RR 2017, 'Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances' Journal of Personality and Social Psychology, vol. 112, no. 3, pp. 474-490. DOI: 10.1037/pspp0000100

Digital Object Identifier (DOI):
10.1037/pspp0000100

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Personality and Social Psychology

Personality Traits Below Facets: The Consensual Validity, Longitudinal Stability, Heritability, and Utility of Personality Nuances

Date of submission: 2nd March 2015

Abstract

It has been argued that facets do not represent the bottom of the personality hierarchy—even more specific personality characteristics, nuances, could be useful for describing and understanding individuals and their differences. Combining two samples of German twins, we assessed the consensual validity (correlations across different observers), rank-order stability, and heritability of nuances. Personality nuances were operationalized as the 240 items of the Revised NEO Personality Inventory (NEO-PI-R). Their attributes were examined by analyzing item residuals, controlling for the variance of the facet the item had been assigned to and all other facets. Most nuances demonstrated significant ($p < .0002$) cross-method agreement and rank-order stability. A substantial proportion of them (48% in self-reports, 20% in informant ratings, and 50% in combined ratings) demonstrated a significant ($p < .0002$) component of additive genetic variance, whereas evidence for environmental influences shared by twins was modest. Applying a procedure to estimate stability and heritability of true scores of item residuals yielded estimates comparable to those of higher-order personality traits, with median estimates of rank-order stability and heritability being .77 and .52, respectively. Few nuances demonstrated robust associations with age and gender, but many showed incremental, conceptually meaningful, and replicable (across methods and/or samples) predictive validity for a range of interest domains and body mass index. We argue that these narrow personality characteristics constitute a valid level of the personality hierarchy. They may be especially useful for providing a deep and contextualized description of the individual, but also for the prediction of specific outcomes.

Keywords: personality hierarchy; nuances; heritability; stability; prediction

Personality Traits Below Facets: The Consensual Validity, Longitudinal Stability, Heritability, and Utility of Personality Nuances

The characteristics often used to describe human personality variation are traits: relatively consistent patterns of emotion, cognition, and behavior in which individuals differ from one another. Traits are thought to be arranged hierarchically. At the highest level, personality trait variation can be described along only a few overarching dimensions. For example, one- (Rushton, Bons, & Hur, 2008) and two- (DeYoung, 2006) factor levels of the hierarchy have been proposed (but also challenged; see Ashton, Lee, Goldberg, & de Vries, 2009; McCrae et al., 2008). The most widely endorsed model of broad traits, however, is the Five-Factor Model of personality (FFM; McCrae & John, 1992) or the Big Five (Goldberg, 1990). At lower levels, the broad traits are thought to split into increasingly narrow constructs. For example, each FFM domain can be split into two aspects (DeYoung, Quilty, & Peterson, 2007) or more numerous facets (Costa & McCrae, 1992).

These different levels of generality/specificity reflect a trade-off. Like all sciences, personality research strives for simplicity and parsimony, which is why researchers sometimes prefer to describe and, perhaps, explain personality variation using as few and comprehensive constructs as possible. And indeed, this has proven a useful strategy. For example, broad traits such as those of the FFM can be used to map personality variability across demographic and cultural groups (e.g., Bleidorn et al., 2013; Costa, Terracciano, & McCrae, 2001; Rentfrow et al., 2013; Schmitt, Realo, Voracek, & Allik, 2008; Terracciano, McCrae, Brant, & Costa, 2005) and predict important life outcomes such as occupational success, divorce, and longevity (Ozer & Benet-Martínez, 2006; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). They are systematically linked to several personality-relevant characteristics, such as self-esteem, control beliefs, motives, values, interests, attitudes, and subjective well-being (Kandler, Zimmermann, & McAdams, 2014). Likewise, they provide a framework for diagnostic distinctions in psychopathology (Krueger & Markon, 2014).

On the other hand, science also strives for comprehensiveness and precision, and therefore using a few broad personality dimensions may not always be optimal. Narrower, lower-order traits such as aspects and facets can often provide useful incremental information for the purpose at hand. Facets have been shown to outperform broad traits in the prediction of specific behaviors (Paunonen & Ashton, 2001), general job performance (Judge, Rodell, Klinger, Simon, & Crawford, 2013), and personality disorder diagnoses (Reynolds & Clark, 2001). The specific variance in facet scores—assessed by statistically controlling for variance from the five broad FFM domains—has been shown to be consensually valid (agreed upon by different observers; Kandler, Riemann, Spinath, & Angleitner, 2010; Costa & McCrae, 2008; Mõttus et al., 2014) and genetically influenced (Briley & Tucker-Drob, 2012; Jang, McCrae, Angleitner, Riemann, & Livesley, 1998). This evidence supports the hypothesis that the specific variance reflects something substantive about personality.

Facet-level traits are therefore not simply interchangeable operationalizations of the broad factor they were designed to reflect or blends of two or more factors (Hofstee, DeRaad, & Goldberg, 1992). Instead, they are traits in their own right and make unique contributions to the depiction of personality. In this article, we examine whether the same conclusions can be drawn for characteristics at an even lower level of the personality hierarchy, using cross-sectional and longitudinal multi-method data from twins.

Nuances as Traits

McCrae (2015) proposed that there is a meaningful level of the trait hierarchy below facets, called *nuances*, which correspond roughly to single items (or groups of very similar items) in a facet scale. For example, bitterness and touchiness may be different nuances of angry hostility, a facet of Neuroticism. Nuances may specify either the eliciting situation (e.g., fear of heights as a source of anxiety or inability to accept criticism as source of anger) or the characteristic response to a range of situations (e.g., a nervous tic as an expression of anxiety across different circumstances or feeling offended as a result of criticism of any kind). Initially, nuances were evoked to account for the fact that retest reliability is generally a better predictor of facet scales' differential stability, heritability, and consensual validity than is internal consistency (McCrae, Kurtz, Yamagata, & Terracciano, 2011). McCrae (2015) argued that this occurs because retest reliability reflects both the variance common to items in a facet scale and the item-specific variance that distinguishes different nuances of the same facet. Here we seek to directly investigate the properties of nuances and establish at least some of them as substantive personality characteristics.

A number of psychometricians have acknowledged that facet-level traits are composed of more specific elements, but it has not been clear whether these elements are themselves lower-level traits. Eysenck (1967, 1991) argued that traits are composed of habitual responses, and habits are usually distinguished from traits. Jackson (1971) regarded scale items as manifestations of the trait that "covered the gamut of situations and modes of response appropriate for sampling broadly a personality construct" (p. 234), a formulation that suggests that items are merely samples of the broader trait, not traits in their own right. Goldberg (1993) recognized traits at many levels of the hierarchy, but he conceived of traits as phenotypic descriptors that might or might not be temporally stable and may not reflect underlying causal influences (Saucier & Goldberg, 1996)—views far removed from the classic formulation of traits as enduring underlying dispositions. Also, some researchers have conceived of traits at higher levels of personality hierarchy as emergent from direct associations between specific characteristics that could potentially be represented by single items (Cramer et al., 2012) or from the direct links of these characteristics with beliefs and motivational constructs (Wood, Hensler Gardner, & Harms, 2015).

Five-Factor Theory (FFT; McCrae & Costa, 1999) states that traits are basic tendencies, "organized hierarchically from narrow and specific to broad and general dispositions" (p. 145), but it does not specify how many levels the trait hierarchy has, or what lower limit (if any) it has. Instead, it specifies definitional criteria for traits. Specifically, FFT holds that traits are not groundless attributions, but enduring dispositions, and adds that they are also genetically based[1]. This conceptualization provides criteria for determining whether nuances are traits: If so, they must be detectable by different observers, stable over substantial periods of time, and have some demonstrable

[1] FFT requires some biological basis for traits, but not necessarily a genetic one: diet, disease, or drugs might also affect traits. However, with the present data we are only able to address genetic influences.

genetic foundation. Evidence that nuances meet these criteria would expand what have typically been considered traits in FFT-based research from domains and facets to narrower characteristics.

Individual items are likely to show some degree of consensual validity, stability, and heritability, if only because they share variance with, or reflect, facets known to have those properties. However, the claim that nuances constitute a distinct level of the trait hierarchy requires evidence that they contribute something unique to the characterization of the individual—evidence of the consensual validity, stability, and heritability of that portion of the items' variance that is not shared with facets. Also, the unique variance in single items could demonstrate associations with external variables (i.e., variables not explicitly included in personality frameworks such as FFM domains and their facets).

Extending the personality hierarchy below facets to nuances could have important conceptual and practical implications. Conceptually, this would refine our understanding of the phenotypic architecture of traits (McCrae, 2015). For example, if there are nuances that do not appear to be merely interchangeable reflections of ostensible higher-level domains or facets, then this would suggest that domains and even facets aggregate numerous specific etiological mechanisms. Empirical and theoretical work on this hypothesis could lead to refinement of trait theories such as the FFT. Practically, nuances might have unique links with etiological factors such as genetic variants or brain parameters, offer incremental validity in the prediction of important outcomes, and have implications for test construction and data analytic practices. For example, tests that incorporate different sets of nuances of facets could not be expected to measure exactly the same facets, which has implications for how their links with other variables can be interpreted (Mõttus, 2015).

Cross-Method Agreement on Nuances

The first direct evidence for the proposal that nuances have trait-like characteristics was provided in a study by Mõttus, McCrae, Allik, and Realo (2014). They examined data from the Estonian version of the NEO Personality Inventory-3 (NEO-PI-3; McCrae & Costa, 2010) and found that item residual scores—from which the common variance of the facet had been removed—showed significant cross-rater (i.e., cross-method) agreement. Individual items are likely to include substantial error variance; as a result, the magnitude of agreement on item-specific variance was smaller than that typically found in studies of domains and facets (Mõttus et al., 2014). The mean correlation between raters on individual items was .31; when item residuals were analyzed, this mean agreement was reduced to .19. However, statistically significant agreement was found for most of the 240 residualized items (range = .06 to .47). This suggests that the different ways in which the same facet can be expressed (or, alternatively, their more or less autonomous constituents) are real in the sense of being detectable by different raters. In the present study, we attempted to replicate this finding with German multi-method data.

Stability of Nuances

Cross-method agreement would appear to be a necessary condition for considering nuances to be traits, but it is not sufficient. Small variations in the ways facet-level traits are expressed might be transient, perhaps reflecting current situational pressures that are visible to all observers. For example, a person who is normally high in the activity facet of Extraversion may have suffered a recent physical injury which temporarily alters his or her response to the item “I act forcefully and energetically;” the response of an informant to this item may also be temporarily altered. If nuances are traits, they—and their specific variance—should show substantial rank-order stability over an interval of years. To the best of our knowledge, this has not yet been investigated. In this article, therefore, we examined whether nuances also show rank-order stability, a fundamental requirement for considering them traits.

Genetic and Environmental Variance in Nuances

Nuances as enduring dispositions might be considered traits, but it is possible that they have a distinct type of etiology from facets and domains. Higher order traits are known to be substantially heritable (Bleidorn, Kandler, & Caspi, 2014), and in some theories (e.g., Five-Factor Theory; McCrae & Costa, 1999) traits are *defined* as dispositions with a heritable basis. Similarly, Turkheimer, Pettersson, and Horn (2014) asserted that “all traits are heritable” (p. 532). From this perspective, the question is: Are nuances really traits? It is possible that the distinguishing features of nuances are entirely learned, as experience shapes the expression of a given facet. That is, the mechanisms that make nuances co-vary and thereby give rise to higher-order traits may be heritable, whereas influences that make them different may reflect purely environmental effects (cf. Borkenau, Riemann, Angleitner, & Spinath, 2001). If this were true, nuances would be qualitatively different from facets and domains; by the standards of the FFT, they would be characteristic adaptations (McCrae & Costa, 1999). To examine this issue, we conducted behavioral genetic analyses of items and their specific variance. Evidence of substantial heritability would suggest that nuances are qualitatively similar to higher-level traits. Furthermore, evidence for either genetic or shared environmental effects (i.e., shared by family members such as twins reared together) on the specific variance of items would provide further evidence that nuances reflect 'signal' (some kind of substantive variance) rather than mere 'noise' (artifact or random error).

Demographic Variations in Nuances

Nuances could provide incremental value for our understanding of personality variability across demographic groups. For example, Mõttus and colleagues (2015) showed that a substantial proportion of age-group differences in personality characteristics could be ascribed to unique variance of single personality test items. None of the 30 NEO-PI-3 facet scales used in the study met the criteria for strong (scalar) measurement invariance across age groups, suggesting that items of the same scales varied in age-trajectories. Furthermore, 46% of items had significant ($p < .0002$, i.e., after Bonferroni correction) correlations with age when residualized for the variance of their respective facet scores. To the best of our knowledge, similar analyses have not yet been carried out for gender. In this paper, we will examine the extent to which residual variance in single items is linked with age and gender.

Predictive Utility of Nuances

Personality characteristics predict important outcomes, such as subjective wellbeing, occupational success, divorce, and even mortality (Roberts et al., 2007). Hence, the case for nuances as personality traits would be strengthened by showing that they also predict incremental variance in outcomes (although we believe that nuances need not predict outcomes in order to broaden our conceptual understanding of how personality is organized). To date, there is very little evidence that nuances predict outcomes—but there is an obvious reason for such paucity of evidence: With some notable exceptions (e.g., Buss, Block, & Block, 1980; Kolar, Funder, & Colvin, 1996) researchers almost never examine individual items as predictors of outcomes.

A full demonstration of the utility of examining nuances as predictors of outcomes would require a lengthy program of research, examining a wide range of outcomes, both broad and narrow. Picking one or a few outcomes and finding these to have limited associations with nuances would not be a conclusive test, because not every nuance is likely to have relevance for every outcome. Broad traits—at least their operationalizations as personality test scores—summarize wide ranges of behaviors and broad outcomes also summarize the cumulative effects of multiple behaviors, which makes some associations between them almost inevitable (Mõttus, 2015). Nuances do not summarize wide ranges of behaviors and are, for that reason alone, likely to have fewer links with outcomes, either broad or specific. They might, however, offer some small incremental validity in the prediction of broad criteria, and if these criteria were of great importance—such as happiness or mortality—even a small contribution would be welcome. Perhaps more importantly, nuances might have substantial utility in the prediction of narrow criteria to which they are directly relevant.

Because personality inventories typically include many items, an unstructured search for correlates would likely yield many false positive associations. For broad criteria, systematic exploration of the full set of item predictors might be reasonable if large samples are used, stringent significance levels are chosen, and, most importantly, replication is required before reaching conclusions. For narrow criteria, it may be more reasonable to test specific a priori hypotheses for a selected subset of items (although replication is important in this case, too). This is similar to the two strategies for identifying the molecular genetic correlates of phenotypic traits: genome-wide and candidate gene approaches (Van Gestell & Van Broeckhoven, 2003).

In this article, we illustrate these kinds of research by exploring the value of individual items in predicting the broad criteria of conservative attitudes (typically correlated with openness and conscientiousness; Jost, Federico, & Napier, 2009) and satisfaction with life (typically associated with variation in neuroticism, extraversion, agreeableness, and conscientiousness; Steel, Schmidt, & Shultz, 2008). We also employed the second strategy of rationally matching specific nuances with specific outcomes for the narrower criteria of body mass index (BMI) and interests in various domains of life.

The Analysis of Single Items

Analyses of single items—and item residuals—pose special challenges. Item scores may be less reliable than aggregate scale scores, so effect sizes are likely to be small, and large samples may therefore be needed for appropriate statistical power. Individual items are also subject to the same artifacts of method (e.g., evaluative bias, extreme responding) as longer scales, so multi-method analyses may be advisable. Likewise, item responses are often based on an ordinal scale with a limited number of possible values and the responses are often strongly skewed, both of which artificially constrain the available variance. Of particular importance is that all single items are inherently unbalanced scales, potentially distorted by acquiescent responding. We address these concerns below.

Rank-Order Stability of Nuances

There is a large literature on the stability of domains (Roberts & DelVecchio, 2000; Terracciano, Costa, & McCrae, 2006). In general, stability increases with age, and by mid-adulthood, traits typically show stability coefficients ranging from .60 to .90 over intervals as long as 10 or more years (Roberts & DelVecchio, 2000). Stability coefficients of this magnitude can be found in both self-report and informant rating data (Costa & McCrae, 1992). They are, however, probably underestimates of true stability, because the observed stability coefficients are attenuated by retest unreliability. When corrected for unreliability and method-specific variance, stability coefficients are often greater than .90 (Kandler, Bleidorn, Riemann, Spinath, Thiel, & Angleitner, 2010).

In this study, we examined self-reports and aggregated informant ratings of twin pairs assessed on two occasions five years apart using the German version of the NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992; Ostendorf & Angleitner, 2004). Long-term retest correlations for single items represent the stability of nuances, but in principle that stability might be attributable solely to the common variance: Perhaps each item is stable simply because it assesses a stable facet. Therefore, we also present more informative analyses based on item residuals, from which the common facet variance has been removed.

McCrae (2015) offered an analysis of the typical variance components in single NEO Inventory items; not surprisingly, most (51%) of the variance was attributable to random error. Consequently, one should expect that the average observed stability of single items would not exceed about .50. Item residuals consist of specific variance plus random error, and McCrae's (2015) analysis suggests that about two-thirds of this total is error. We would thus expect the observed stability of item residuals to be no higher than .33.

Ideally, a comparison of the stability of factors, facets, and nuances would be based on true scores, or on stability coefficients disattenuated for retest unreliability. Data on the retest reliability of individual NEO items—and item residuals—are not yet available, but it is possible to estimate true score stability by using multi-method longitudinal data. For example, if one calculates the correlation between self-reports of a characteristic at the first measurement occasion and informant ratings of the same characteristic at the second measurement occasion, the resulting stability coefficient should be relatively free from both self- and informant-report method biases, because these biases are not likely to be shared. However, this correlation would be attenuated by imperfect agreement between self- and informant-reports even when collected at the same time. The concurrent cross-method correlation thus represents an upper limit to observed cross-method stability. It then follows that true score stability can be estimated as the *ratio* of the cross-lagged, cross-method correlation to this upper limit (McCrae, 1994). This is similar to the widely-used procedure of correcting a correlation for unreliability, if we think of cross-rater agreement as a reliability estimate. We used this method to estimate the true-score stability of items and item residuals.

Genetic and Environmental Variance in Nuances

We applied standard behavioral genetic analyses to item scores and their residuals to estimates the genetic and environmental sources of variance in nuances. It is possible to partition individual differences in item scores and item residuals into components that estimate additive genetic effects (a^2), genetic dominance effects within gene loci (d^2), epistatic gene × gene interaction effects between gene loci (b^2), twins' shared environmental effects (c^2), and environmental effects not shared by twins reared together, including random error variance (e^2). Whereas variance due to additive genetic effects reflects narrow-sense heritability, variance due to both additive and nonadditive genetic influences (such as dominance or epistatic effects) reflects broad-sense heritability. However, with a simple twin design, environmental effects shared by twins and nonadditive genetic influences cannot be estimated in the presence of each other. Therefore, only three of the five effects can be estimated in a single model. The usual procedure in such cases is to examine different models ($\sigma^2 = a^2 + c^2 + e^2$, $\sigma^2 = a^2 + d^2 + e^2$, and $\sigma^2 = a^2 + b^2 + e^2$) and select the best-fitting model. Our data allowed us to conduct such analyses with twins' self-reports and with averaged informant ratings. A comparison of results would show how robust the findings are across different methods of measurement, which is an interesting question in its own right. Additionally, we also carried out the analyses in combined self- and informant-reports.

It is also possible to estimate the true score heritability of item residuals using a procedure similar to that employed to estimate the longitudinal stability of true scores. In this analysis, the disattenuated, random and nonrandom error-reduced estimate of the similarity of twins was calculated as the ratio of the cross-twin, cross-informant correlation to the within-twin, cross-informant correlation. For example, one can correlate Twin A's self-report with informant ratings of Twin B, and divide this by the correlation between Twin A's self-report and informant ratings of Twin A. This estimate can be calculated for both monozygotic (MZ) and dizygotic (DZ) twins, and standard behavioral genetic formulas can then be used to estimate a^2, d^2, b^2, c^2, and e^2 components. These estimates may provide a rough indication of genetic and environmental influences on individual differences in true scores of nuances.

Estimates of the stability and heritability of true scores are derived from a series of calculations using item data that is intrinsically less reliable than aggregated data. Thus, the estimate obtained for any given item is likely to have a very wide margin of error. Our focus, however, was on the distribution of estimates and not on any single one of them; the median and interquartile range are probably relatively robust, providing a sense of the stability or heritability of the true score for a typical nuance. Recall, however, that these are ideal values, essentially the upper limit of what might be observed if very reliable measures of nuances, based on ratings of many items made by many observers, were analyzed. Therefore, the estimates that we provide are of more theoretical than practical value.

Variations Across Items

Although operationalizing nuances as single items is a reasonable way to begin the study of this level of the trait hierarchy, it is unlikely that each of the 240 NEO Inventory items corresponds to a unique nuance. Two or more items might tap the same nuance—a likely possibility, given that Mõttus and colleagues (2014) found residual correlations between many items after controlling for factor and facet variance. Some items might be relatively pure measures of the facet they reflect, with little specific variance. Other items, however, may have a more substantial specific component, and it would be useful to identify these items, because these are the items most likely to add incremental validity in the prediction of outcomes or otherwise contribute to our understanding of personality differences and their development.

We hypothesized that the cross-observer agreement, longitudinal stability, and genetic component of item residuals would be larger for items with high saturations of nuance-specific variance beyond random error and method variance. If so, observed self-informant agreement, rank-order stability, and heritability estimates should have been positively intercorrelated across the set of 240 items. Such regularity would support the claim that there is substantive residual variance in particular items. Items could then be ranked by combining these three indicators of the quantity of specific variance. We examined these rankings to identify and characterize the items with most nuance-specific variance.

Facets, too, will vary in the amount of item-specific variance they contain. McCrae (2015) estimated these values by subtracting the internal consistency from the retest reliability of each of the 30 NEO Inventory facets, using American and international data. In the present German data, it is possible to make a corresponding estimate by calculating the mean cross-observer agreement, longitudinal stability, and heritability of item residuals across the eight items in each facet. Correlations of these means with the estimates provided by McCrae allowed for a direct test of the hypothesis that retest reliability exceeds internal consistency because it includes item-specific variance.

Method

Participants

Cross-sectional sample. Combined data from the third wave of the Bielefeld Longitudinal Study of Adult Twins (BiLSAT; Kandler, Riemann, Spinath, Bleidorn, Thiel, & Angleitner, 2013) and from the Jena Twin Study of Social Attitudes (JeTSSA; Stößel, Kämpfe, & Riemann, 2006) were used. Only participants with at least two sets of personality ratings (either data for both twins and/or for self- and informant-reports) were used in this study. The resulting sample consisted of 1599 individuals (1231 females), who were between 17 and 82 years old (*M* = 36.02; *SD* = 13.41). Among these participants were 690 complete twin pairs: 432 monozygotic (MZ) pairs (74 male pairs and 358 female pairs) and 258 dizygotic (DZ) pairs (34 male, 134 female, and 90 opposite-sex pairs). Both self- and informant ratings were available for 1,492 people (93%). Among the twins, informant ratings were not available for 107 people; for 26 twin pairs, informant ratings were not available for either pair-member. For genetic analyses that required informant ratings to be available for both siblings of a twin pair, data from 609 twin pairs (88%) could be used. Most participants were instructed to ask acquaintances, who knew them well but preferably did not know their twin sibling, to provide the informant ratings. Thus, there were different, quasi-independent informants for twin siblings. Most informants were friends and spouses. Reports from multiple informants were averaged when available (in 88% of cases).

Longitudinal sample. For longitudinal analyses, data from the third and fourth waves of BiLSAT was used (*N* = 400; 337 females). In particular, the sample consisted of 200 twin pairs tested twice over a period of about five years. At wave three, participants' ages ranged from 22 to 74 with a mean of

41.47 (*SD* = 13.80), whereas at wave four the average age was 47.14 years (*SD* = 13.81). Among the twins were 138 MZ pairs (12 male pairs and 126 female pairs) and 62 DZ pairs (12 male, 35 female and 15 opposite-sex pairs). For all twins a pair of informant ratings was available at each measurement occasion, although the informants were not always the same on both occasions.

Estonian Genome Bank. The associations between BMI and item residual variances were replicated in the data published by Vainik et al. (2015; *N* = 2,581) with 982 added cases (total *N* = 3,563; age range 18 to 91 years with a mean of 46.82 and *SD* of 17.01; 1,434 men). In these data, both self- and informant-ratings of personality were available (for details see Vainik et al., 2015).

Measures

Personality nuances. For the measurement of personality characteristics, the German version of the NEO-PI-R was administered (Costa & McCrae, 1992; Ostendorf & Angleitner, 2004). The NEO-PI-R contains 240 items, grouped into 30 facet scales, which are hierarchically organized under the five domain scales of the FFM. Responses were made on a 5-point Likert scale ranging from *strongly disagree* to *strongly agree*. The domains of the NEO Inventory correspond closely to the five factors found in analyses of many personality inventories (see Markon, Krueger, & Watson, 2005), and its facets were chosen to represent important traits within each domain. For these reasons, the NEO Inventory item pool seems to provide a broad sample of nuances; clearly, however, it does not exhaust the population of nuances. In the Estonian Genome Bank data, personality ratings were obtained using the Estonian translation of the NEO-PI-3 (McCrae & Costa, 2010), which is a slightly modified version of the NEO-PI-R.

Conservatism. At the third wave of BiLSAT, twins' self-reported political attitudes were captured with a German 35-item version of the Wilson-Patterson C Scale (Schiebel, Riemann & Mummendey, 1984; see also Riemann, Grubich, Hempel, Mergl, & Richter, 1993). This measure was actualized with respect to catchphrases including current political topics (e.g., promotion of alternative energy resources, homosexual marriage, or German contribution to UN). Responses were made on a 7-point Likert scale ranging from *strongly disagree* to *strongly agree*. Items endorsing liberal positions were reverse coded. Internal consistency was good (α = .86).

Life satisfaction. At the fourth wave of BiLSAT, twins completed the German version of the Satisfaction With Life Scale (Diener, Emmons, Larson, & Griffin, 1985; Glaesmer, Grande, Braehler, & Roth, 2011). Responses were made on a 7-point Likert scale ranging from *strongly disagree* to *strongly agree*. Internal consistency was good (α = .86).

Interests. Interests were measured with the German General Interest Scale (GIS; Brickenkamp, 1990), a questionnaire including 48 unipolar items developed to map the intensity and breadth of individual interests across diverse fields of interest covering music, computer technology, arts, science, architecture, biology, literature, nutrition, politics, agriculture, business, fashion, education, sport, medicine, and entertainment. Each field is represented by three items that reflect different types of activities in the specific field of interest: (a) consuming and receptive activities (e.g., watching sports), (b) producing, participating, and imitating activities (e.g., getting exercise), and (c) creative and inventive activities (e.g., developing new methods of training). Both self-reports and informant-ratings were available.

Body mass index (BMI). In the German twin data, self-reported height and weight were used. The weight and height of the members of the Estonian Genome Bank were measured by trained medical staff at the recruitment. BMI was calculated as the ratio of weight (kg) to height (m^2).

All these data were collected in a manner consistent with ethical standards for the treatment of human subjects.

Results

Cross-Method Agreement on Nuances

Cross-method agreement on raw item scores and item residuals was estimated in the cross-sectional sample by pooling data from all individuals for whom self- and informant-reports were available (*N* = 1,492). Items were residualized for their respective facet scores (with the item itself excluded) and scores of all other facets using linear regression. Note that this procedure also means that items were residualized for the FFM domains, because these are linear combinations of facets. Here and henceforth, unless otherwise noted, we used a conservative Bonferroni-corrected *p*-value threshold of .0002 for testing the significance of item-level associations to minimize type 1 error rate.

Across the 240 items, self-informant correlations for raw (unresidualized) item scores ranged from .07 to .59, with a median of .28 [interquartile range (*IQR*): .22 to .34]. When items were residualized, the correlations ranged from .01 to .41, with a median of .12 (*IQR*: .08 to .16), and 144 (60%) of the correlations were significant. The two vectors of 240 cross-method correlations (for raw item scores and residuals) correlated at .69. These findings suggest that, as a tendency, the items that showed higher cross-method agreement were agreed upon not only because they reflected facets (i.e., common variance across particular items), but to a substantial extent because of their distinctive aspects not shared with other items (i.e., nuance-specific item variance). Likewise, items with lower cross-method agreement were not agreed upon regardless of whether we considered their trait-related or unique variance.[2]

Stability of Nuances

Rank-order stability of nuances over the period of about 5 years was estimated based on the 400 members of the longitudinal sample, treating twin pair members as independent individuals. As in analyses of cross-method agreement, items were residualized for their respective facet scores (with the item itself excluded form it) as well as all other facet scores. In self-ratings, the correlations of raw item scores ranged from .22 to .75 with a median of .53 (*IQR*: .48 to .58). The correlations ranged from .12 to .61 (*Mdn* = .34; *IQR*: .29 to .39) for item residuals, with 232 (97%) being significant. These values are consistent with the variance components of NEO Inventory items estimated by McCrae (2015). Thus, residualizing the item scores for facet variance reduced the median rank-order stability by 36%, but there remained a non-trivial level of stability. In informant ratings, stability was confounded with inter-rater agreement, because the informants did not always overlap at the two time points;

[2] Because of the possible effects of dependency in the pooled data, all analyses here, as well as similar analyses elsewhere, were repeated with only one twin from each pair. Results were similar to those reported. Analyses on the predictive validity and demographic associations of item residuals were carried out only on the pooled sample, because maximum power was required.

correlations were therefore predictably lower than in self-reports. Raw item correlations ranged from .14 to .60 with a median of .37 (*IQR*: .31 to .44), and residual item correlations ranged from -.01 to .48 (*Mdn* = .15; *IQR*: .11 to .22), with 85 (35%) being significant. We also calculated these estimates for combined self- and informant-ratings: the median agreement estimates were .51 and .27, respectively for raw and residualized item scores (193, or 80%, were statistically significant for the latter).

The items that had highest rank-order stability when unresidualized tended to be the same items showing highest rank-order stability when residualized for facet variance: Across 240 items, the vectors of rank-order stabilities for raw and residualized items correlated at .66 and .64, respectively, for self-reports and informant ratings (the correlation was .63 for the combined ratings). There was also a substantial level of consistency in rank-order stabilities of items across self-reports and informant ratings, with the correlations across 240 items being .70 for raw item score-based and .59 for item residual-based rank-order stabilities.

To the extent that method variance was stable across time, these estimates were inflated by method variance, but they were also attenuated by measurement error. In order to address this possibility, we estimated the stability of true scores by dividing cross-lagged, cross-method correlations by concurrent, cross-method correlations. This resulted in rank-order stabilities of raw item scores ranging from .54 to 1.00+ with a median of .89 (*IQR*: .81 to .95). For item residuals, the estimated median rank-order stability was .77 (*IQR*: .58 to 1.00+). These estimates suggest that nuances demonstrate trait-like stability.

Genetic and Environmental Variance in Nuances

Based on the data from 690 twin pairs, we fitted three models (Figure 1) to raw scores and residuals of each of the 240 items, separately for self- and informant ratings. The ACE model specified latent factors representing additive genetic influences (A), shared environmental influences (C) and unique environmental influences (E). The ADE model specified latent factors representing additive genetic influences (A), nonadditive genetic influences due to dominance effects within gene loci (D) and unique environmental influences (E). Finally, the ABE model specified nonadditive genetic (epistatic) influences due to dominance effects between gene loci (B) instead of genetic dominance effects within gene loci. For each item analysis, the three models were compared in terms of comparative model fit based on Akaike Information Criterion (AIC) with the smallest value indicating the best fitting model.

We were first interested in the degree to which the preference for any of the three models was replicable across methods. Assuming that the variance captured in self- and informant-reports reflects to some degree the same substantive variance of personality characteristics, we expected to find a substantial degree of convergence. This would have supported the robustness of preferring one model over the others. However, the convergence was modest. In raw item scores, the same type of model was preferred for 97 (40%) items, whereas for item residuals the preference overlapped for 92 (38%) items. Furthermore, the modest convergence could not be ascribed to difficulties between differentiating between ABE and ADE models. If the latter were collapsed into one category of non-ACE models and opposed to the category of ACE models, the overlap in preference across rating types was still poor, with the same category being preferred for 125 (52%) and 112 (47%) items, respectively for raw scores and residuals. Clearly, preferring one model over another was very inconsistent across methods, as the observed level of convergence could be expected by chance. This prompted us to select only one type of model for further analyses. For the selection of the best model type, we compared intra-class correlations (*ICC*) in MZ (ICC_{mz}) and DZ twins (ICC_{dz}). Because the ICC_{dz} -s were generally about half the size of ICC_{mz}-s, we opted for, and will only present results of, the ACE models.

In raw item scores, the heritability estimates (a^2) varied from .00 to .50 (*Mdn* = .26; *IQR*: .21 to .33) for self-ratings and from .00 to .48 (*Mdn* = .19; *IQR*: .11 to .25) for informant ratings. In item residuals, a^2 estimates varied from .00 to .37 (*Mdn* = .14; *IQR*: .06 to .19) for self-ratings and from .00 to .27 (*Mdn* = .04; *IQR*: .00 to .09) for informant ratings. In combined self- and informant-ratings, the median a^2 estimates varied from .00 to .56 (*Mdn* = .32; *IQR*: .24 to .37) and .00 to .41 (*Mdn* = .13; *IQR*: .05 to .19), respectively for raw and residualized item scores. Therefore, residualizing the items for facet variance reduced the median a^2 estimate by 50%, 79% and 59%, for self-reports, informant ratings and combined ratings, respectively. The items that showed higher levels of heritability unresidualized for trait variance tended to be the same that showed higher levels of heritability after residualization. Across the 240 items, raw item-based and residuals-based a^2 estimates correlated at .61, .54 and .62, respectively for self-reports, informant ratings and combined ratings.

The heritability estimates tended to be higher for self-rated than for informant-rated item scores both before and after residualizing for facet variance. Nevertheless, in both types of ratings a non-negligible proportion of item variance could be ascribed to genetic influences. For example, a^2 estimates for raw item scores and residuals were significant at $p < .0002$ for 150 (63%) and 116 (48%)

items, respectively, based on self-ratings; the corresponding numbers were 121 (50%) and 48 (20%) for informant ratings, and 169 (70%) and 121 (50%) for combined self- and informant-ratings. Heritability estimates for the 240 self-rated items correlated with heritability estimates for the 240 informant-rated items at .36 and .32, respectively for raw item scores and residuals. The items showing significant a^2 estimates overlapped across methods for 80 (33%) and 24 (10%) items, respectively for raw score- and residual-based estimates. Overall, thus, the residual variances of a tenth of the items showed highly significant evidence for genetic influences simultaneously in self- and informant-reports, suggesting these items consistently reflected some unique signal beyond their role in the measurements of any NEO-PI-R facet (and thereby any FFM domain).

In raw item scores, the c^2 estimates (i.e., variance component due to environmental effects shared by twins) varied from .00 to .37 (*Mdn* = .00; *IQR*: .00 to .08) for self-ratings and from .00 to .31 (*Mdn* = .00; *IQR*: .00 to .08) for informant ratings. In item residuals, c^2 estimates varied from .00 to .25 (*Mdn* = .00; *IQR*: .00 to .07) for self-ratings, and from .00 to .19 (*Mdn* = .00; *IQR*: .00 to .06) for informant ratings; 27 estimates were statistically significant in both cases. In combined self- and informant-ratings, the median c^2 estimates varied from .00 to .34 (*Mdn* = .00; *IQR*: .00 to .09) and .00 to .22 (*Mdn* = .00; *IQR*: .00 to .08; 30 estimates were significant), respectively for raw and residualized item scores. In summary, shared environmental influences were generally small and residualizing items for facet variance had only a small effect on the estimates; if anything, the number of items with significant shared environmental influences increased. Across the 240 items, raw item-based and residuals-based c^2 estimates correlated at .64, .38 and .66, respectively for self-reports, informant ratings and combined ratings. However, c^2 estimates for the 240 self-rated items correlated only modestly with c^2 estimates for the 240 informant-rated items (r = .26 and .08, respectively for raw item scores and residuals). Furthermore, the items showing significant c^2 estimates overlapped across the two rating types for only one and four items, respectively for raw score- and residuals-based estimates. These data provided little consistent evidence of shared environmental effects on nuances.[3]

Genetic and Environmental Variance in Nuances after Controlling for Method Effects

Estimates of nuance heritability were higher for self-reports than for informant ratings. However, raters of different members of a twin pair were independent, but the twins themselves were not. If rater biases, such as acquiescence, were heritable or influenced by shared (family) experience, twin resemblance may have been inflated in self-reports by shared method artifacts (see Riemann & Kandler, 2010, or Nelling, Kandler, & Riemann, 2015, for a detailed discussion). We therefore created an index of acquiescence from NEO-PI-R items (see Supplement 1) and examined the extent to which genetic and shared environmental effects contribute to acquiescence in self-reports. We found the effects to be non-negligible (a^2 = .39, c^2 = .14; see Supplement 1 for details), so we conducted additional analyses of heritability in self-reports controlling for acquiescence (see Supplement 1). This, however, resulted in only minor effects on the median heritability estimates for item residuals (across

[3] In order to rule out the possibility that the observed and arguably small a^2 and c^2 effects were flukes, we re-ran the analyses on self-reported item residuals 240 times (once for each item) such that the residual scores were reshuffled across individuals and thereby any resulting a^2 and c^2 estimates reflected random noise. Only three a^2 estimates and one c^2 estimates were significant at $p < .0002$. The corresponding numbers of significant findings had been substantially higher (respectively, 116 and 27) with real data, suggesting that they are meaningful.

240 items, *Mdn* a^2 = .13; *IQR*: .07 to .19). Of course, other shared method artifacts, such as social desirability, might also account for the higher values in self-reports.

In order to circumvent the problem of artifacts, we estimated the genetic and environmental effects on item true scores by employing a procedure similar to that used for estimating the stability of true scores. By calculating ICC_{dz}-s and ICC_{mz}-s across methods, we first assessed the similarity of MZ and DZ twins in a way that eliminated shared method variance (e.g., twin A's self-ratings was correlated with twin B's informant ratings and vice versa; we averaged the *ICC*s across twin-combinations). We then disattenuated these cross-method, cross-twin correlations for imperfect agreement across methods and for random error by dividing them by cross-method, within-twin *ICC*s (e.g., twin A's self-ratings correlated with twin A's informant ratings and the same for twin B; we used the average correlations across twin-combinations). These estimated twins' true score similarity coefficients are designated $ICC_{TRUE\text{-}MZ}$ and $ICC_{TRUE\text{-}DZ}$.

As the next step, we decomposed the true score variance in raw item scores and residuals into estimates of heritable (a^2) and shared environmental (c^2) variance, using the classic formulas according to which $a^2 = 2^*(ICC_{TRUE\text{-}MZ} – ICC_{TRUE\text{-}DZ})$ and $c^2 = 2^*ICC_{TRUE\text{-}DZ} – ICC_{TRUE\text{-}MZ.}$ Because the *ICC*s had been derived via a series of calculations, the resulting a^2 and c^2 values were often outside their natural boundaries (i.e., lower than zero or above one); in these cases, the values were set to zero or one. For item raw scores, a^2 estimates varied from .00 to 1.00 with a median of .68 (*IQR*: .36 to .97). The corresponding c^2 estimates varied from .00 to 1.00 with a median of .00 (*IQR*: .00 to .26). For item residuals, a^2 estimates ranged from .00 to 1.00 with a median of .52 (*IQR*: .00 to 1.00). The c^2 estimates varied from .00 to 1.00 with a median of .00 (*IQR*: .00 to .75). Based on these calculations, thus, residualizing items for facet variance reduced median a^2 estimates by only 24% but did not really change c^2 estimates.

Results of these behavioral genetic analyses are summarized in Table 1. Although the true score estimates are based on a series of calculations and must therefore be interpreted with considerable caution, they do suggest that nuances, if accurately measured, may be substantially heritable, largely because the item-specific variance is itself substantially heritable. As with domains and facets, it appears that variability in nuances is chiefly influenced by genetic factors and the non-shared environment; however, there is some evidence to suggest that environmental influences shared by twins may also play a non-trivial role for some nuances.

Magnitude of Specific Effects

Supplement 2 lists the results for each individual item from analyses of raw scores and residuals (26 columns in total). Five of the columns for item residuals are of particular interest: the observed cross-observer agreement, stability in self-reports and informant ratings, and heritability in self-reports and informant ratings (we do not focus on combined ratings here because they are not independent of self-reports and informant ratings). These provide an indication of the size and distribution of the effects due to specific variance, and show that it is appreciable in magnitude and widespread in distribution. The five columns contain 1200 coefficients, ranging from .00 to .61. Of these, 584 (49%) are between .10 and .30, and would be regarded as small by Cohen's (1988) rule-of-thumb; 187 (16%) are between .30 and .50 (or medium in size); and 6 (0.5%) exceed the threshold for a large effect (> .50). However, given typical effect sizes in personality psychology (e.g., heritability estimates and cross-rater agreement of *facets* being around .30 to .50), these numbers are certainly not negligible.

In these five columns, there are 40 coefficients for each of the 30 NEO Inventory facets; averaging across them shows that specific variance is lowest in N3: Depression and N6: Vulnerability (*Ms* = .12) and highest in O2: Aesthetics (*M* = .26) and E5: Excitement Seeking (*M* = .27); at the domain level, specific variance is lowest in Neuroticism (*M* = .14) and highest in Extraversion and Openness (*M*s = .18). Analyses of other instruments would be needed to determine whether these rather small differences are due to the nature of different facets and domains, or are simply peculiarities of the NEO-PI-R item pool.

Convergence of the Findings

Table 2 shows the correlations (below the diagonal) among cross-method agreement, rank-order stability, and heritability estimates for the residual variances of the 240 NEO-PI-R items. The correlations were moderate to relatively high (range = .31 to .72; *Mdn* = .51), suggesting consistency in different ways of assessing the presence of valid variance in item residuals. Put differently, these correlations suggest that some items relatively consistently showed more specific variance than others. In order to identify the items most likely to contain substantive residual variance, we averaged the five indicators of substance (shown in Table 2); the items with highest average were those that had most consistently displayed signal. Inspection of the 30 items with highest mean scores (see Supplement 3 for the paraphrased items) showed that they included items expressing a characteristic "mode of response" (Jackson, 1971, p. 234), such as using sarcasm (from the A4: Compliance facet, reverse scored), being the most talkative in a group (E3: Assertiveness), and being entranced by music or looking for patterns in art and nature (O2: Aesthetics). Some items tapped beliefs (e.g., in the need to follow religious authorities; O6: Values). However, most of the items represented responses to particular situations and contexts, such as enjoying roller coaster rides or attending sporting events (from E5: Excitement Seeking), trying different kinds of food (O4: Actions), or making detailed preparations for a trip (C6: Deliberation).

Mõttus and colleagues (2014) assessed cross-method agreement on the residual variances of 240 NEO-PI-3 items in a large Estonian sample. Although the present study used NEO-PI-R, which has 37 differently worded items, the estimates obtained in the Estonian sample correlated .64 with those obtained in this study. Likewise, the Estonian estimates of cross-method agreement on item residuals correlated with the stability (.45 and .52) and heritability estimates (.36 and .33) of item residuals obtained in this study (for self- and informant-reports, respectively). Therefore, evidence of items showing potentially valid unique variance is not specific to a particular sample, but seems at least moderately replicable even across cultures.

Finally, we calculated the mean cross-method agreement, rank-order stability, and heritability estimates of the eight items in each facet. Their intercorrelations, ranging from .48 to .87 (*Mdn* = .68), are given above the diagonal in Table 2. The last column shows correlations between these indicators of specific variance in the 30 NEO facets with estimates based on internal consistency and retest reliability (McCrae, 2015). They tend to support the hypothesis that the superiority of retest reliability to internal consistency as a predictor of facet-level consensual validity, stability, and heritability is due to the presence of substantive item-specific variance.

Associations with Age and Gender

We examined the relationships of item scores with age and gender in the cross-sectional sample. Of the 240 self-reported raw item scores, 80 (33%) correlated significantly with age, with correlations ranging from -.32 to .24. When residualized for facet variance, 14 (6%) items showed significant correlations with age, ranging from -.20 to .16. In informant ratings, 71 (30%) raw item scores correlated significantly with age (range: -.34 to .30); when items were residualized, nine (4%) correlated significantly with age (range: -.17 to .17). The same items tended to correlate with age when unresidualized and residualized (across 240 items, the item-age correlations were correlated at .74 in both self- and informant-reports). Likewise, the item-age correlations were quite similar across methods (r = .85 and .76, respectively for raw and residualized item scores), although none of the significant age-correlations overlapped. In combined self- and informant-reports, scores of 100 (42%) raw items significantly correlated with age (range: -.36 to .31), whereas the number decreased to 20 (8%) in item residuals (range: -.20 to .19).

Raw scores of 44 self-report items (18%) showed significant associations with gender (according to Welch t-test), with Cohen d's ranging from -0.62 to 0.52. Of item residuals, 6 (3%) were significantly associated with gender, with d's varying from -0.42 to 0.35. Raw scores of 68 (28%) informant-rated items were significantly associated with gender (d's ranged from -1.03 to 0.68), whereas the number was 9 (4%) for residualized scores (range: -0.58 to 0.49). Again, item-gender associations tended to be similar for raw and residualized item scores (across 240 items, r = .70 and .60, respectively for self- and informant-reports) and across methods (r = .82 and .53, respectively for raw and residualized item scores), although none of the significant sex-difference overlapped across methods. In combined self- and informant-reports, 77 (32%) raw items scores were significantly associated with sex (Cohen *d*'s range: -0.96 to 0.67), whereas the number decreased to 11 (5%) in item residuals (range: -0.50 to 0.63).

Predictive Validity of Nuances: Broad Criteria

We correlated the 240 items with two broad criteria, self-reported conservatism and satisfaction with life. Raw scores of 45 self-report items were significantly (p < .0002) correlated with the aggregate scores of the conservatism scale, with correlations ranging from -.31 to .18. Most of these items were from the Openness, Agreeableness and Conscientiousness domains. Aggregate scores of the life satisfaction scale were significantly correlated with raw scores of 56 items (r = -.38 to .37), mostly from the Neuroticism and Extraversion domains. In contrast, the residual variance of only three items correlated significantly (-.15 to -.17) with the conservatism scale, whereas no item residuals correlated significantly with life satisfaction scale. In informant-ratings, raw scores of 49 and 7 items significantly (p < .0002) correlated with the scores of the self-reported conservatism and life satisfaction scales, with correlations ranging from -.33 to .19 and from -.21 to .20, respectively. For the residual variance of informant-report items, three items correlated significantly (-.14 to -.16) with the conservatism scale, whereas no items correlated significantly with the satisfaction with life scale. Two items residuals that correlated with conservatism overlapped across the methods and referred to permitting students to hear about controversial ideas and not following religious authorities. The two non-overlapping items also came from the same facet (A6: Tendermindedness). In combined self- and informant-reports, 74 and 51 raw items scores were significantly correlated with conservatism and life satisfaction scales, respectively (respective ranges: -.39 to .20 and -.37. to .36). Of item residuals, four correlated significantly with the conservatism scale (-.19 to -.15; the items that had significantly correlated with conservatism in either self-reports or informant ratings or in both), whereas none was significantly associated with the life satisfaction scale.

Predictive Validity of Nuances: Narrow Criteria

Among the 30 items with the strongest evidence for substantive unique variance (Supplement 3), several referred to people's interests (e.g., in sports or food). In order to test the predictive validity of nuances, we therefore sought to link the residual variances of these items to interests in various life domains measured with independent scales and using both self- and informant-reports. Also, as three items referred to eating and food preferences, we correlated the residuals of these with BMI—a possible outcome of dietary choices that has wide health consequences and has been previously linked to higher-order traits (e.g., Terracciano et al., 2009; Sutin et al., 2013; Vainik et al., 2015). In doing so, we rationally matched specific NEO-PI-R items with BMI, particular interest scales and items of these scales to form directional hypotheses. For example, the residual variance of the item that referred to liking (sport) games was hypothesized to be positively associated with scores of the sports interest scale and specifically with the item of this scale that referred to receptive interest in sports ("Watching sports, e.g., on TV or live from the stands"). Likewise, the residual variance of the item that referred to trying different foods was linked with high BMI as well as with food interest scale and its items reflecting receptive ("Going out for a meal, e.g. gourmet restaurants"), productive ("Cooking and baking according to proven recipes"), and creative ("Trying new cooking and baking recipes") interest in food because the personality item may refer to both consuming or cooking food. These and other specific hypotheses, arguably rather straightforward at face value, are given in Table 3.

A hypothesis that is perhaps less straightforward than others pertained to a positive link between BMI and “plans before travel.” This hypothesis was derived from the counter-intuitive finding of Sutin et al. (2013) that increases in the deliberation facet scores of NEO-PI-R were positively correlated with increases in BMI: We hypothesized that the link might have been driven by the item that refers to planning ahead of travel, because high BMI may complicate traveling for some people. We also hypothesized that the residual variance of the item that refers to giving up on self-improvement programs might be related to BMI, because some people with high BMI may be interested in self-improvements (e.g., adhering to some forms of diet) but their high BMI may at least in some occasions attest the lack of progress in such attempts.

We did not specify hypotheses for 16 of the 30 item residuals (see Supplement 3), because they did not seem to be relevant to the available outcomes. For example, the item residuals referring to particular ways of spending leisure time (in crowds or on roller-coasters) or being easily frightened did not seem relevant to any of the interest scales or BMI.

The hypotheses were tested in the pooled data of twins' self-reports (N = 843 for interest scales and 871 for BMI) and then cross-validated in their informant-reports (N = 840 for interest scales and 861 for BMI); we also present the associations based on combined self- and informant-reports. In addition, the associations pertaining to BMI were replicated in an independent sample, which also included both self-reports and informant-ratings (Estonian Genome Bank; N = 3,563). The reported p-values have been adjusted for False Discovery Rate (Benjamin & Hochberg, 1995), separately in findings based on self-reports, informant ratings and combined ratings. The specified hypotheses along with results are given in Table 3; the results from the Estonian Genome Bank are only given in Supplement 3.

With respect to BMI, three of the five predictions were confirmed in both self- and informant-reports of German twins, as well as in the combined reports. High BMI was linked with the unique variance of items referring to giving up on self-improvement programs, planning ahead for a trip, and eating too much of one's favorite food, with effect sizes varying from .08 to .21 in self-reports and .09 to .32 in informant-reports. The fourth hypothesis that linked trying different foods with BMI yielded a significant association in self-reports and combined ratings but not in informant-reports. The fifth hypothesis pertaining to the residual of "eating excessively" was not confirmed in German data, but it was confirmed in the Estonian data in both self- and informant-reports (r = .18 and .15, respectively, p < .001; see Supplement 3). The hypotheses pertaining to giving up on self-improvement programs and over-eating favorite foods were also replicated in the Estonian data with the respective effect sizes being .06 and .14 in self-reports and .07 and .19 in informant-reports. However, the residual variance of the trying new foods item was not significantly correlated with BMI according to neither of the methods nor their combination, and the correlation with the trip-planning item residuals was only significant in self-ratings (r = .04, p = .011) and combined ratings (r = .04, p = .021). For comparison, the strongest facet-level effect sizes were .18 and .21 in the German data and .23 and .13 in the Estonian data (for self- and informant-ratings, respectively). Therefore, in terms of effect size the hypothesized associations between item residuals and BMI were generally in par with the *strongest* facet-level associations (note that the facet scores also contained FFM domain variance—had the facets been residualized for the FFM scores, their predictive validities would probably have been lower).

With respect to interests, of the 32 hypotheses tested, 22 yielded significant associations ($p < .05$) in both self-reports and informant ratings (for 19 associations, $p < .01$); all of these associations were also significant in the combined ratings and they were always in the hypothesized direction. Five more hypothesized associations were significant ($p < .05$) in informant ratings. Most of the correlations were small, but some were substantial—for example, residual variance of the item that refers to liking to attend games predicted interest in sports with effect sizes (.48 and .58, respectively for self- and informant-reports) that are relatively high in the context of personality-outcome research. Consistently with predictions, interests in food, music, politics, gardening/farming, arts, nature and architecture were significantly predicted by residual variances in one or more of the hypothesized items. However, not all predictions were confirmed. For example, the residual variance of the item that referred to liking expressive dance did not correlate with the interests in art scale as had been expected, although it did correlate significantly with the receptive interest item of the scale[4]. Also, the residual of the item that referred to finding philosophy boring did generally not correlate significantly with the interest in science scale or its items, and liking showy styles did not correlate with productive interest in fashion in neither of the methods. In evaluating the correlations and their magnitude, it must be recalled that

[4] In hindsight, it may have been reasonable to hypothesize that most people think specifically of consuming rather than also producing or creating expressive art when responding to such a personality questionnaire item, in which case the observed association pattern makes sense.

these are associations over and above whatever was attributable to their own facet and the 29 other facets.

Discussion

The main finding of this study is that the personality trait hierarchy can be extended further downwards below FFM domains and facets. Characteristics that are narrower than facets—nuances (McCrae, 2015)—can make unique contributions to describing and perhaps understanding individual differences in personality. Nuances capture variance beyond that of any facet or FFM domain and yet they have the hallmark properties of traits as these are specified by FFT: temporal stability, observability across raters, and a likely genetic basis. Further, the unique variance of nuances is at least occasionally related to age and sex and can often predict outcomes in theoretically meaningful ways.

Most of the items demonstrating the highest levels of cross-rater agreement, rank-order stability, and heritability in their item-specific variances referred to rather specific and contextualized behavioral tendencies and preferences, such as enjoying roller coaster rides, attending sporting events, trying different kinds of food, or making detailed preparations for a trip. Their level of heritability may thus appear surprising: Our true score-based heritability estimates, in particular, suggested that many residual item variances contained levels of heritable variance similar to those of facets and FFM domains. It would not have been unreasonable to expect that the specific manifestations of personality dispositions reflected habits acquired through exposure to specific (and thereby perhaps less genetically preselected) environments, but this did not appear to be the case. Even if two people had equal scores on all facets (and thereby also on all FFM domains), the tendencies for one person to seek fun from sporting events and another to prefer roller-coasters seemed to reflect, to a substantial extent, genetic differences between them. The apparent possibility that there is a similar degree of genetic variance at all levels of the personality hierarchy is consistent with the observation of Turkheimer and colleagues (2014) who noted that "when reliability is accounted for, the proportion of heritable variance does not seem to vary substantially by level of analysis" (p. 521). This is a non-trivial finding, which we cannot explain at this point. However, such pervasiveness of genetic influences needs to be heeded in theories on the genetic architecture of personality variance.

Measurement of Nuances

We operationalized nuances as single NEO-PI-R items—particular indicators of a broader facet, aspect, or domain. In order to determine whether nuances themselves are traits, it was necessary to examine what, if anything, they assess beyond the facet and domain of which they are indicators. Therefore, we residualized item scores, controlling for the facet to which they were assigned as well as all other facets. The analyses of these residuals provided crucial evidence that nuances are consensually valid, longitudinally stable, and heritable in their own right. However, although operationalizing nuances as single test items may be a good start to understanding variance at lower and more specific levels of the personality hierarchy, this approach has some important limitations, which probably resulted in underestimating the amount of unique yet valid variance in specific personality characteristics.

First, NEO-PI-R items were not designed to measure nuances, but the FFM domains and their facets. That is, the items were *created to assess common variance*, and only those that did this sufficiently well were selected into the questionnaire. The eight items in a NEO-PI-R facet scale surely do not exhaust the nuances that could define the facet, just as the 30 facets do not exhaust the facet-level definers of the five domains of Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. In fact, the range of specific personality characteristics that could constitute nuances with unique and valid variance is inevitably limited when one operationalizes nuances *ad hoc*, using any personality questionnaire that is carefully designed for measuring broader traits.

Second, any attempts to find valid variance in single items are constrained by the overall variance available for analysis. Scores of single NEO-PI-R items display an artificially limited range of variance, constrained by the five levels of the Likert response scale. Likewise, responses to items are often heavily skewed (Mõttus et al., 2015), which further reduces their variance. Consistently with this, Mõttus and colleagues (2014) found that cross-rater agreement was higher for items with higher variance.

Third, scores of single items contain a higher degree of random measurement error than aggregate scores of facets and domains. This is not a problem intrinsic to nuances—this is a measurement problem. When we estimated the stability and heritability of true scores for nuances, the values were much higher than uncorrected estimates. One way to overcome the problem of random error would be to construct multi-item measures of nuances. Mõttus and colleagues (2015) showed that unidimensional confirmatory factor analysis models for most NEO-PI-3 facets required correlated residuals, suggesting that unique variances of single items were often not independent. Thus, even in questionnaires which are not designed to measure nuances, single items can, in some cases at least, be combined to yield aggregate nuance scores. For example, the two eating-related item residuals of the N5: Impulsiveness scales (Table 3) could probably be combined into a single nuance.

Fourth, it is likely that not all items reflect nuances. Many if not most of them are likely to do exactly what they were designed for—reflect facets and FFM domains as purely as possible. As a result, the typical estimates of cross-method agreement, stability, and heritability as well as relations with external variables in the residual scores could be attenuated because the items reflect predominantly higher-order traits.

The results of this study suggest that there are valid specific personality characteristics below facets, despite the study design limitations that were likely to suppress the emergence of these findings. Future studies can improve on both the specification of the full range of nuances and their assessment. For example, data from large item-pools could be used, and item responses residualized for common traits could be subjected to clustering procedures such as ICLUST (Revelle, 1979).

The Utility of Nuances

Is there a sufficient incentive to pursue a program of further research on nuances? Their nature and existence is of considerable theoretical interest, because they help explain the way different forms of reliability operate and lead to new conceptions of trait measures and traits themselves (McCrae, 2015). For example, evidence for nuances may suggest that facets do not represent reflective-type latent traits (Bollen & Lennox, 1991) that exist independently of their indicators (McCrae, 2015), which has been the fundamental assumption of much of current personality trait research. If so, this may have wide-spread theoretical and practical implications for how personality characteristics can be linked with either etiological factors or purported consequences (Mõttus, 2015; Wood et al., 2015). However, although it could be argued that the present study has adequately established the existence of nuances as a lower level of traits, is there any practical use for nuances?

Our attempt to predict broad criteria using the specific variance in items—to add even a small quantity of incremental validity—largely failed. At least for conservatism and life satisfaction, it was chiefly the variance at the domain and/or facet level that seemed to matter, although two nuances did make a significant and robust (across methods) contribution to the prediction of conservatism. This may surprise some psychometricians, who often suppose that more predictors will yield stronger predictions: "for huge samples it would be silly even to amalgamate the items into scales because one would inevitably lose some specific variance at the item level that could serve to increase predictive accuracy" (Goldberg, 1990, pp. 181-182). Our results suggest that for very broad criteria, the specific variance in nuances may often be irrelevant, and analyses may safely rely on amalgamated scales.

However, specific variance in nuances proved to be useful in the prediction of narrow criteria to which they were directly relevant. Tests of hypotheses regarding BMI and interests were generally confirmed, and were replicated across methods and, in the case of BMI, across samples. Some of the associations were modest in size, whereas others were relatively strong, considering typical effect sizes in research that links personality traits with outcomes. Moreover, the hypothesized item residual-BMI associations were often as strong as or even stronger than the strongest facet-BMI association. These are certainly not trivial findings and suggest that nuances can meaningfully contribute to the prediction—and perhaps even explanation—of personal characteristics and life outcomes. For example, BMI is an important variable with wide health consequences that has previously been linked to FFM domains and facets in numerous studies (e.g., Terracciano et al., 2009; Vainik et al., 2015).

Occasionally, some nuances may be conceptually so close to the criteria as to make the observed associations nearly tautologous. The fact that people who try various foods have an interest in food is unsurprising, and the finding that over-eating is correlated with being over-weight may not be not truly informative. However, not all of the associations between nuances and outcomes are trivial. For example, the residual variance in an item concerned with giving up easily on self-improvement attempts was a significant predictor of BMI above and beyond general levels of self-discipline, conscientiousness and indeed any other facet or FFM domain. This is a finding that might have implications for health interventions (e.g., frequent follow-ups to sustain a dieter's commitment). The generally replicable association between planning ahead for a trip and BMI is also not immediately obvious and may potentially point to one of the specific consequences of increased body weight for personality functioning.

Considering nuances might have additional implications for personality-outcome research. For example, observed trait-outcome associations may occasionally be driven solely by specific nuances, as was likely for the link between impulsiveness and obesity (Terracciano et al., 2009; Vainik et al., 2015) and has also been reported for impulsiveness and diabetes (Čukić et al., 2016). If we attribute to a facet an effect that is due only to some of its nuances, we may be misled about the underlying mechanism.

Perhaps testing whether facet- or domain-outcome associations are driven by specific items might become a standard practice (Mõttus, 2015).

Finally, considering nuances may help to elucidate our description and understanding of personality variance across genders and age groups. For example, a substantial proportion of age-trends are specific to nuances rather than higher-order traits such as facets or domains (Mõttus et al., 2015); this may indicate that the mechanisms responsible for personality development operate at least partly on specific manifestations or constituents of personality (Soto & John, 2012).

Nuances Make the Individual

Nuances may prove to be useful predictors of relevant variables, but they may also be important in understanding the individual. A number of personality psychologists have argued that traits offer only the "psychology of the stranger" (McAdams, 1994, p. 145): Common traits are so abstract that they do not let us understand the individual except in the most general terms. Allport (1931, 1966) and Cattell (1946) distinguished between *common traits*, which are a basis for comparison among people, and *personal dispositions* or *unique traits*, which are the distinctive form in which traits appear in any given individual. The former are statistical abstractions, but the latter are real psychological structures in living human beings, and thus should be the primary concern of personality psychologists (Allport, 1966). We argue that nuances may offer a bridge between these seemingly unbridgeable constructs.

From the perspective of Five-Factor Theory (McCrae & Costa, 1999), traits are hypothetical constructs. They are basic tendencies that must be inferred from observed patterns of thoughts, feelings, and actions. These patterns—habits, roles, interests, coping mechanisms, and so on—are concrete characteristic adaptations acquired in particular life circumstances, and they are the stuff of which personality items are made. Normally personality scales are interpreted by summing across a series of items, a procedure that essentially eliminates the context in which traits are expressed. A raw score of 16 on a facet of the NEO-PI-R can be achieved by many different patterns of item responses, so facet scores tell us only about common traits. However, if personality psychologists—or clinicians—read the individual item responses, they may gain insight into the personal dispositions that show how the common trait is manifested in this particular person. The respondent is no longer a stranger, but an individual with his or her own distinctive manifestations of personality traits based on his or her unique constellation of personality nuances.

Many items are phrased conditionally: I make detailed plans *when* I go on a trip; *if* I am insulted, I forgive and forget. Mischel has called characteristic if-then response patterns *behavioral signatures* (Mischel & Shoda, 1995), and these seem to parallel nuances. From the perspective of FFT, nuances are basic tendencies, whereas behavioral signatures are characteristic adaptations. But behavioral signatures may have been acquired precisely because they expressed observable, stable, and even heritable personality nuances. The gap between social-cognitive and trait-dispositional approaches to personality may therefore be less daunting at the level of nuances than at the level of broad domains or even facets.

At the level of facets or domains, it is easy to see that scales assess abstract tendencies, but the concrete content of individual items may suggest that they are characteristic adaptations, like Eysenck's

(1967) *habitual responses*. However, we found that even the unique variance in items also displays trait-like properties, including heritable variance. This might suggest that the valid variance in a single item can perhaps be decomposed into that due to basic tendencies (at broader and sometimes also quite specific levels of the trait hierarchy) and the external influences (i.e., life experiences) that contribute to the creation of a characteristic adaptation. However, we think that such attempts may not be the most useful way to conceptualize the nature of items. Within the framework of FFT, all personality items are most directly interpreted as characteristic adaptations, because basic tendencies can never be observed, only inferred from patterns of multiple co-existing thoughts, feelings, and behaviors—patterns that are crystallized as the beliefs, attitudes, habits, styles, and so on that personality items typically address. Even items that appear to assess traits directly, such as "I am an introvert," are in fact assessing the self-concept, which has evolved over time as the person reacts to the environment; the self-concept is a major component of characteristic adaptations.

At the same time, to the extent that they are valid indicators of traits item responses reveal basic tendencies. Indeed, Costa and McCrae (in press) have argued that, somewhat as in quantum physics (where particles are seen as having both wave and particle-like properties), there is a duality principle in personality assessment. Viewed from one perspective, items can be seen as acquired characteristic adaptations, whereas from another perspective, they reflect endogenous basic tendencies. (Quantum duality is a genuine paradox, because particles and waves are quite different and seemingly incompatible; FFT duality is intelligible, merely calling attention to the fact that phenomena can be construed differently by adopting different conceptual perspectives.) Psychologists are familiar with this notion with respect to cognitive testing. No one inherits a knowledge of particular words, yet scores on a vocabulary test are good indicators of heritable intelligence. Similarly, no one is born with a love of riding roller coasters, but one may be born with a predisposition to learn to enjoy it (given the opportunity). It is this predisposition that constitutes the nuance-as-trait.

Of course, the nuance which is assessed by the roller coaster item might actually represent a broader construct such as "Excitement-seeking-through-going-fast," and also be assessed by items concerning drag racing, downhill skiing, and motorcycle riding. A scale combining such items would measure the nuance with greater reliability and more abstraction, and clearly it would assess a personality trait—albeit a very narrow one.

Limitations and Future Directions

Above, we discussed the limitations of our operationalization of nuances. However, some additional limitations must be heeded, too. First, a substantial proportion of variance in nuances—that is, item residuals—is likely to reflect measurement error, which makes effect sizes small. This, in turn, requires that very large samples be used for studying their properties. Although the current study used a relatively large (main) sample, especially given the fact that it contained twins, even bigger ones would be desirable. For example, in a sample of nearly 2,200 people, Mõttus and colleagues (2015) found that item residuals of nearly half of NEO-PI-3 items were significantly correlated with age, whereas in the present study only 6% of self-reported item residuals demonstrated significant age-correlations. In other words, the degree of substance in item residuals may be underestimated in the current study.

Second, our findings call for replication. Although we used conservative thresholds for statistical significance, it is possible that some findings reflect type I errors. Replications in other cultures and using different types of samples would be especially valuable. However, we should note that the items that were more likely to reflect valid nuances in the current German data tended to be the same showing highest cross-method agreement in an independent Estonian sample, suggesting that successful cross-cultural and cross-language replications are likely (likewise, the item residual-BMI associations generally replicated in the Estonian data). Finally, our twin sample was largely female and gender is closely related to personality traits. Future studies with balanced distributions of sexes may find more and larger sex differences in personality nuances.

Conclusions

Personality characteristics can be represented hierarchically, and facets have generally been thought to constitute the lowest level of the hierarchy. Here, we have shown that there might be a valid level of yet more specific personality characteristics below facets. The unique variance in these specific characteristics, or nuances, is often consensually valid, temporally stable, and heritable, is sometimes linked to major demographic variables, and can predict important outcomes in conceptually meaningful ways. We argue that nuances can provide a richer description of individuals and may serve to better understand the nature and developmental mechanisms of personality dispositions. Moreover, nuances may correspond to the long-held concepts of unique traits or personal dispositions and may provide a bridge between the trait and social-cognitive approaches to personality. As a result, considering nuances is likely to make personality psychology richer.

References

Allport, G. W. (1931). What is a trait of personality? *The Journal of Abnormal and Social Psychology*, *25*, 368–372. http://doi.org/10.1037/h0075406

Allport, G. W. (1966). Traits revisited. *American Psychologist*, *21*, 1–10. http://doi.org/10.1037/h0023295

Ashton, M. C., Lee, K., Goldberg, L. R., & de Vries, R. E. (2009). Higher order factors of personality: Do they exist? *Personality and Social Psychology Review*, *13*, 79-91. http://doi.org/10.1177/1088868309338467

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, *57*, 289–300. http://doi.org/10.2307/2346101

Bleidorn, W., Klimstra, T. A., Denissen, J. J. A., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2013). Personality maturation around the world: A cross-cultural examination of Social Investment Theory. *Psychological Science, 24, 2*530-2540. http://doi.org/10.1177/0956797613498396

Bleidorn, W., Kandler, C., & Caspi, A. (2014). The behavioral genetics of personality development in adulthood: Classic, contemporary, and future trends. *European Journal of Personality*, *28*, 244–255. http://doi.org/10.1002/per.1957

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305–314. http://doi.org/10.1037/0033-2909.110.2.305

Borkenau, P., Riemann, R., Angleitner, A., & Spinath, F. (2001). Genetic and environmental influences on observed personality: Evidence from the German Observational Study of Adult Twins. *Journal of Personality and Social Psychology*, *80*, 655-668. http://dx.doi.org/10.1037/0022-3514.80.4.655

Brickenkamp, R. (1990). *Die Generelle Interessen-Skala (GIS): Handanweisung [The General Interest Scale (GIS): Manual]*. Göttingen, Germany: Hogrefe.

Briley, D. A., & Tucker-Drob, E. M. (2012). Broad bandwidth or high fidelity? Evidence from the structure of genetic and environmental effects on the facets of the five factor model. *Behavior Genetics*, *42*, 743–763. http://doi.org/10.1007/s10519-012-9548-8

Buss, D. M., Block, J. H., & Block, J. (1980). Preschool activity level: Personality correlates and developmental implications. *Child Development, 51*, 401-408.

Cattell, R. B. (1946). Personality structure and measurement. *British Journal of Psychology. General Section*, *36*, 88–103. http://doi.org/10.1111/j.2044-8295.1946.tb01110.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Costa, P. T., Jr., & McCrae, R. R. (in press). The NEO Inventories as instruments of psychological theory. In T. A. Widiger (Ed.), *Oxford handbook of the Five-Factor Model*. New York: Oxford University Press.

Costa, P. T., Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*, 322–331. http://doi.org/10.1037/0022-3514.81.2.322

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. http://doi.org/10.1002/per.1866

Čukić, I., Mõttus, R., Realo, A., & Allik, J. (2016). Elucidating the links between personality traits and diabetes mellitus: Examining the role of facets, assessment methods, and selected mediators. *Personality and Individual Differences*, *94*, 377–382. http://doi.org/10.1016/j.paid.2016.01.052

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology*, *91*, 1138–1151. http://doi.org/10.1037/0022-3514.91.6.1138

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. http://doi.org/10.1037/0022-3514.93.5.880

Diener, E., Emmons, R. A., Larson, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Psychological Assessment*, *49*, 71-75.

Eysenck, H. J. (1967). *The biological basis of personality.* Springfield, IL: Charles C Thomas.

Eysenck, H. J. (1991). Dimensions of personality: 16, 5 or 3?—Criteria for a taxonomic paradigm. *Personality and Individual Differences*, *12*, 773–790. http://doi.org/10.1016/0191-8869(91)90144-Z

Van Gestel, S., & Van Broeckhoven, C. (2003). Genetics of personality: are we making progress? *Molecular Psychiatry*, *8*(10), 840–852. http://doi.org/10.1038/sj.mp.4001367

Glaesmer, H., Grande, G., Braehler, E., & Roth, M. (2011). The German version of the Satisfaction with Life Scale – Psychometric properties and population based norms. *European Journal of Psychological Assessment*, *27*, 127-132. DOI: http://dx.doi.org/10.1027/1015-5759/a000058

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216–1229. http://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1993). The structure of personality traits: Vertical and horizontal aspects. In D. C. Funder, R. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), *Studying lives through time: Approaches to personality and development* (pp. 169-188). Washington, DC: American Psychological Association.

Hofstee, W.K.B., De Raad, B., & Goldberg, L.R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*, 146–163. http://dx.doi.org/10.1037/0022-3514.63.1.146

Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review, 78,* 229-248. http://dx.doi.org/10.1037/h0030852

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. http://doi.org/10.1037/0022-3514.74.6.1556

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual Review of Psychology*, *60*, 307–337. http://doi.org/10.1146/annurev.psych.60.110707.163600

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. *The Journal of Applied Psychology*, *98*, 875–925. http://doi.org/10.1037/a0033901

Kandler, C., Bleidorn, W., Riemann, R., Spinath, F. M., Thiel, W., & Angleitner, A. (2010). Sources of cumulative continuity in personality: A longitudinal multiple-rater twin study. *Journal of Personality and Social Psychology*, *98*, 995-1008. http://dx.doi.org/10.1037/a0019558

Kandler, C., Riemann, R., Spinath, F. M., Bleidorn, W., Thiel, W., & Angleitner, A. (2013). The Bielefeld Longitudinal Study of Adult Twins. *Twin Research and Human Genetics*, *16*, 167-172. doi: 10.1017/thg.2012.67

Kandler, C., Riemann, R., Spinath, F., & Angleitner, A. (2010). Sources of variance in personality facets: A twin study of self-self, peer-peer, and self-peer (dis)agreement. *Journal of Personality*, *78*, 1565-1594. http://doi.org/10.1111/j.1467-6494.2010.00661.x.

Kandler, C., Zimmermann, J., & McAdams, D. P. (2014). Core and surface characteristics for the description and theory of personality differences and development. *European Journal of Personality*, *28*, 231-243. http://dx.doi.org/10.1002/per.1952

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality*, *64*, 311-338. http://doi.org/10.1111/j.1467-6494.1996.tb00513.x

Krueger, R. F., & Markon, K. E. (2014). The role of the DSM-5 personality trait model in moving toward a quantitative and empirically based approach to classifying personality and psychopathology. *Annual Review of Clinical Psychology, 10,* 477-501. http://doi.org/10.1146/annurev-clinpsy-032813-153732

Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the structure of normal and abnormal personality: An integrative hierarchical approach. *Journal of Personality and Social Psychology, 88*, 139-157. http://doi.org/10.1037/0022-3514.88.1.139

McAdams, D. P. (1994). A psychology of the stranger. *Psychological Inquiry*, *5*, 145-148. http://doi.org/10.1207/s15327965pli0502_12

McCrae, R. R. (1994). The counterpoint of personality assessment: Self-reports and observer ratings. *Assessment*, *1*, 159–172. http://doi.org/10.1177/107319119400100206

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, *19*, 97–112. http://doi.org/10.1177/1088868314541857

McCrae, R. R., & Costa, P. T., Jr. (1999). A Five-Factor Theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd. ed., pp. 139-153). New York: Guilford.

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

McCrae, R. R., & John, O. P. (1992). An introduction to the Five-Factor Model and its applications. *Journal of Personality*, *60*, 175–215. http://doi.org/10.1111/j.1467-6494.1992.tb00970.x

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. http://doi.org/10.1177/1088868310366253

McCrae, R. R., Stone, S. V., Fagan, P. J., & Costa, P. T., Jr. (1998). Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality*, *66*, 285-313. http://doi.org/10.1111/1467-6494.00013

McCrae, R. R., Yamagata, S., Jang, K. L., Riemann, R., Ando, J., Ono, Y., Angleitner, A., & Spinath, F. (2008). Substance and artifact in the higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 95,* 442-455. http://doi.org/10.1037/0022-3514.95.2.442

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*, 246–268. http://doi.org/10.1037/0033-295X.102.2.246

Mõttus, R. (2015). Towards more rigorous personality trait-outcome research. *European Journal of Personality.* http://dx.doi.org/10.1002/per.2041

Mõttus, R., McCrae, R. R., Allik, J., & Realo, A. (2014). Cross-rater agreement on common and specific variance of personality scales and items. *Journal of Research in Personality*, *52*, 47–54. http://doi.org/10.1016/j.jrp.2014.07.005

Mõttus, R., Realo, A., Allik, J., Esko, T., Metspalu, A., & Johnson, W. (2015). Within-trait heterogeneity in age group differences in personality domains and facets: Implications for the development and coherence of personality traits. *PLoS ONE*, *10*(3), e0119667. http://doi.org/10.1371/journal.pone.0119667

Nelling, A., Kandler, C., & Riemann, R. (2015). Substance and artifact in interest self-reports: A multiple-rater twin study. *European Journal of Psychological Assessment*, *31*, 166-173. DOI: 10.1027/1015-5759/a000222

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

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*, 401–421. http://doi.org/10.1146/annurev.psych.57.102904.190127

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. http://doi.org/10.1037/0022-3514.81.3.524

Rentfrow, P. J., Gosling, S. D., Jokela, M., Stillwell, D. J., Kosinski, M., & Potter, J. (2013). Divided we stand: Three psychological regions of the United States and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology*, *105*, 996–1012. http://doi.org/10.1037/a0034434

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research,14,* 57-74.

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. http://doi.org/10.1111/1467-6494.00142

Riemann, R., Grubich, C., Hempel, S., Mergl, S., & Richter, M. (1993). Personality and attitudes towards current political topics. *Personality and Individual Differences, 15,* 313-321. doi:10.1016/0191-8869(93)90222-O

Riemann, R., & Kandler, C. (2010). Construct validation using multitrait-multimethod-twin data: The case of a general factor of personality. *European Journal of Personality*, *24*, 258-277. http://doi.org/10.1002/per.760

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

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*, 313–345. http://doi.org/10.1111/j.1745-6916.2007.00047.x

Rushton, J. P., Bons, T. A., & Hur, Y.-M. (2008). The genetics and evolution of the general factor of personality. *Journal of Research in Personality*, *42*, 1173–1185. http://doi.org/10.1016/j.jrp.2009.01.005,

Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the Five-Factor Model. In J. S. Wiggins (Ed.), *The Five-Factor Model of personality: Theoretical perspectives* (pp. 21-50). New York: Guilford Press.

Schiebel, B. Riemann, R. & Mummendey. H. D. (1984). Eine aktualisierte deutschsprachige Form der Konservatismusskala von Wilson & Patterson [An actualized German version of Wilson and Patterson's conservatism scale]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *5*, 311-321.

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. http://doi.org/10.1037/0022-3514.94.1.168

Soto, C. J., & John, O. P. (2012). Development of Big-Five domains and facets in adulthood: Mean-level age trends and broadly versus narrowly acting mechanisms. *Journal of Personality*, *80*, 881–914. http://doi.org/10.1111/j.1467-6494.2011.00752.x

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

Stößel, K., Kämpfe, N., & Riemann, R. (2006). The Jena Twin Registry and the Jena Twin Study of Social Attitudes (JeTSSA). *Twin Research and Human Genetics, 9,* 783-786. doi:10.1375/twin.9.6.783

Sutin, A. R., Costa, P. T. J., Chan, W., Milaneschi, Y., Eaton, W. W., Zonderman, A. B., … Terracciano, A. (2013). I know not to, but I can't help it: Weight gain and changes in impulsivity-related personality traits. *Psychological Science*, *24*, 1323–1328. http://doi.org/10.1177/0956797612469212

Terracciano, A., Costa, P. T., Jr., & McCrae, R. R. (2006). Personality plasticity after age 30. *Personality and Social Psychology Bulletin*, *32*, 999–1009. http://doi.org/10.1177/0146167206288599

Terracciano, A., McCrae, R. R., Brant, L. J., & Costa, P. T., Jr. (2005). Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging*, *20*, 493–506. http://doi.org/10.1037/0882-7974.20.3.493

Terracciano, A., Sutin, A. R., McCrae, R. R., Deiana, B., Ferrucci, L., Schlessinger, D., … & Costa, P. T., Jr. (2009). Facets of personality linked to underweight and overweight. *Psychosomatic Medicine*, *71*, 682–689. http://doi.org/10.1097/PSY.0b013e3181a2925b

Turkheimer, E., Pettersson, E., & Horn, E. E. (2014). A phenotypic null hypothesis for the genetics of personality. *Annual Review of Psychology, 65,* 515–540. http://doi.org/10.1146/annurev-psych-113011-143752

Vainik, U., Mõttus, R., Allik, J., Esko, T., & Realo, A. (2015). Are Trait–Outcome Associations Caused by Scales or Particular Items? Example Analysis of Personality Facets and BMI. *European Journal of Personality*. http://doi.org/10.1002/per.2009

Wood, D., Gardner, M. H., & Harms, P. D. (2015). How functionalist and process approaches to behavior can explain trait covariation. *Psychological Review*, *122*(1), 84–111. http://doi.org/10.1037/a0038423

Table 1. *Median Values for Variance Components of NEO-PI-R Items*

	Raw Scores				Residuals			
	Self-Reports	Informant Ratings	Combined Ratings	True Scores	Self-Reports	Informant Ratings	Combined Ratings	True Scores
a^2	.26	.19	.32	.68	.14	.04	.13	.52
c^2	.00	.00	.00	.00	.00	.00	.00	.00
e^2	.74	.81	.68	.32	.86	.96	.87	.48

Note. a^2 = SEM-based variance components due to additive genetic effects and *ICC*-based heritability estimates in true scores; c^2 = SEM-based variance components due to shared environmental effects and *ICC*-based shared environmental components in true scores; e^2 = SEM-based variance components due to unique environmental effects and *ICC*-based estimates for unique environmental components in true scores.

Table 2. *Correlations among cross-method agreement, rank-order stability and heritability estimates for the residual variances at the item level (below diagonal) and the facet level (above diagonal).*

	1	2	3	4	5	Specific Variance
1. Cross-method agreement	-	.82	.87	.55	.70	.42
2. Stability (self-reports)	.67	-	.82	.48	.61	.51
3. Stability (informant ratings)	.72	.59	-	.56	.72	.57
4. Heritability (self-reports)	.54	.39	.46	-	.66	.33
5. Heritability (informant ratings)	.53	.31	.49	.32	-	.25

Note. Below diagonal, all correlations are significant at *p* < .001*, N* = 240; above diagonal, correlations .37 or higher are significant at *p* < .05, *N = 30*, whereas correlations .49 or higher are significant at *p* < .01. Specific Variance = estimate of specific variance in a facet based on retest reliability minus internal consistency.

Table 3. *Item residuals predicting BMI and interests.*

Item	Item Description	Criterion	Self-reports		Informant Ratings		Combined Ratings	
			r	*p*	*r*	*p*	*r*	*p*
BMI								
O4.4	Tries different foods	BMI	.12	.001	.06	.105	.11	.002
C6.7	Plans before travel	BMI	.12	.001	.09	.013	.13	.000
C4.3	Gives up on self-improvements	BMI	.08	.022	.10	.007	.10	.005
N5.4	Overeats favorite foods	BMI	.21	.000	.32	.000	.30	.000
N5.6	Eats excessively	BMI	.00	.943	.04	.210	.02	.655
Interest								
E5.8	Likes attending games	Sports	.48	.000	.58	.000	.55	.000
E5.8	Likes attending games	Sports (R)	.41	.000	.58	.000	.48	.000
O4.4	Tries different foods	Food	.37	.000	.40	.000	.41	.000
O4.4	Tries different foods	Food (R)	.24	.000	.26	.000	.23	.000
O4.4	Tries different foods	Food (P)	.25	.000	.30	.000	.30	.000
O4.4	Tries different foods	Food (C)	.34	.000	.35	.000	.38	.000
O2.2	Is entranced by music	Music	.27	.000	.41	.000	.34	.000
O2.2	Is entranced by music	Music (R)	.35	.000	.47	.000	.40	.000

O2.3	Doesn’t like expressive dance	Art (–)	–.04	.227	–.06	.119	-.02	.664
O2.3	Doesn’t like expressive dance	Art (R, –)	–.11	.003	–.10	.007	-.08	.025
C1.2	Ignores civic obligations	Politics (–)	–.25	.000	–.30	.000	-.27	.000
C1.2	Ignores civic obligations	Politics (R, –)	–.29	.000	–.32	.000	-.28	.000
C1.2	Ignores civic obligations	Politics (P, –)	–.18	.000	–.25	.000	-.22	.000
C1.2	Ignores civic obligations	Politics (C, –)	–.16	.000	–.20	.000	-.19	.000
O2.6	Finds music fascinating	Music	.23	.000	.35	.000	.29	.000
O2.6	Finds music fascinating	Music (R)	.30	.000	.39	.000	.33	.000
O2.6	Finds music fascinating	Music (P)	.14	.000	.22	.000	.19	.000
O2.6	Finds music fascinating	Music (C)	.08	.027	.25	.000	.13	.000
O4.6	Likes redecorating	Gardening	.08	.041	.07	.047	.08	.030
O4.6	Likes redecorating	Gardening (P)	.05	.141	.07	.047	.06	.109
O4.6	Likes redecorating	Gardening (C)	.09	.017	.11	.002	.10	.005
E5.7	Likes showy styles	Fashion	.04	.332	.10	.007	.05	.220
E5.7	Likes showy styles	Fashion (R)	.02	.594	.09	.013	.04	.344
E5.7	Likes showy styles	Fashion (P)	.01	.790	.06	.108	.04	.344
E5.7	Likes showy styles	fashion (C)	.05	.147	.09	.016	.04	.272
O5.2	Finds philosophy boring	Science (–)	.05	.157	–.05	.183	.02	.655
O5.2	Finds philosophy boring	Science (R, –)	.06	.095	.02	.636	.04	.272

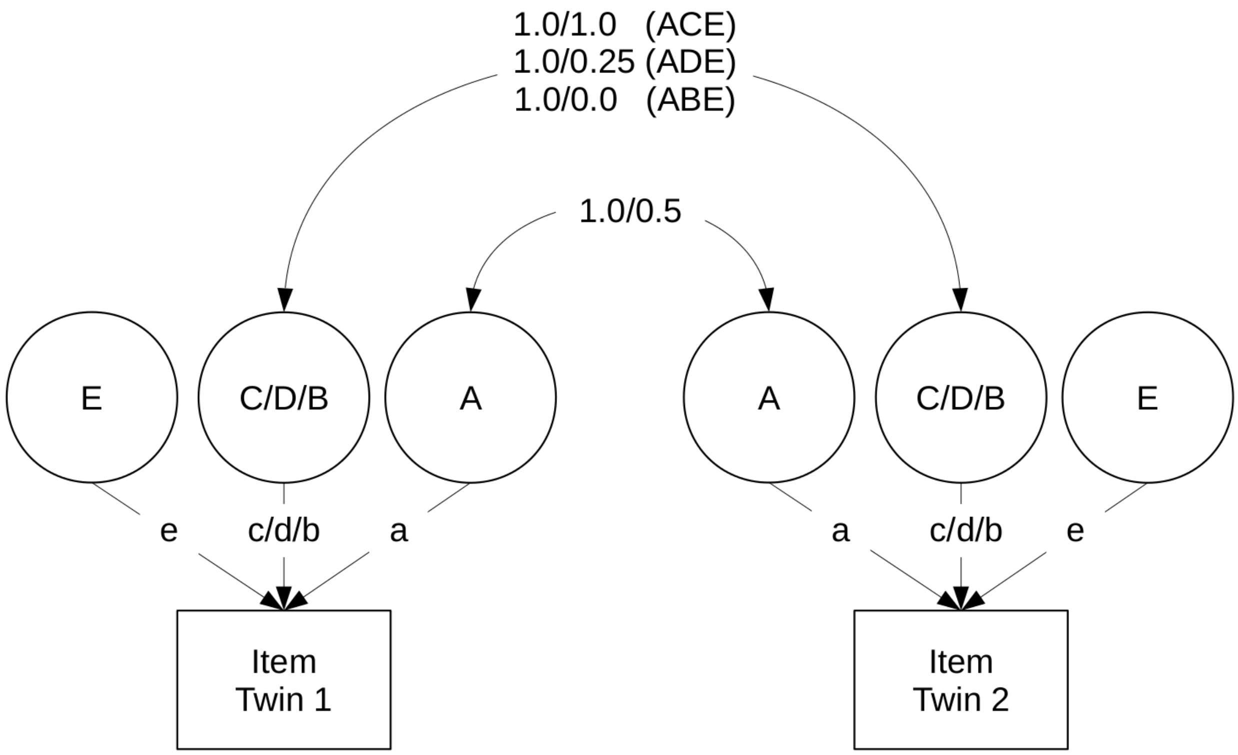
O5.2	Finds philosophy boring	Science (P, –)	.00	.983	–.09	.011	-.04	.303
O5.2	Finds philosophy boring	Science (C, –)	.07	.052	–.05	.135	.04	.311
O2.4	Is interested in patterns	Art (R)	.20	.000	.23	.000	.25	.000
O2.4	Is interested in patterns	Architecture (R)	.16	.000	.19	.000	.18	.000
O2.4	Is interested in patterns	Nature (R)	.28	.000	.30	.000	.28	.000

Note. *N*s = 871/861 for BMI, 843/840 for interests (self-reports/infomant-ratings); *p*-values have been adjusted using False Discovery Rate. For NEO-PI-R, item is indicated by item order within the designated facet (see Costa & McCrae, 1992, Appendix A. *N.B*.: Item order differs from that for NEO-PI-3.) For interests, R = Receptive interest item; P = Productive interest item; C = Creative interest item; minus sign indicates an inverse correlation was hypothesized.

Figure Captions

Figure 1. Behavior genetics models tested using structural equation modeling. NOTE: A = additive genetic variance; C = shared environmental variance; E = unique environmental variance; D = non-additive genetic variance due to interactions within genetic loci; B = non-additive genetic variance due to interactions between different genetic loci. Fixed path coefficients before slash are for monozygotic twins, coefficients after slashes are for dyzogotic twins.

Figure 1


1.0/1.0 (ACE)
1.0/0.25 (ADE)
1.0/0.0 (ABE)
1.0/0.5
E
C/D/B
A
A
C/D/B
E
e
c/d/b
a
a
c/d/b
e
Item Twin 1
Item Twin 2