Personality characteristics below facets

A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability and utility of personality nuances

Abstract

Mõttus and colleagues (2017) reported evidence that the unique variance in specific personality characteristics captured by single descriptive items often displayed trait-like properties of cross-rater agreement, rank-order stability and heritability. They suggested that the personality hierarchy should be extended below facets to incorporate these specific characteristics, called personality nuances. The present study attempted to replicate these findings, employing data from 6,287 individuals from six countries (Australia, Canada, Czech Republic, Denmark, Japan, and United States). The same personality measure—240-item Revised NEO Personality Inventory—and statistical procedures were used. The present findings closely replicated the original results. When the original and current results were meta-analyzed, the unique variance of nearly all items (i.e., items' scores residualized for all broader personality traits) showed statistically significant cross-rater agreement (median = .12) and rank-order stability over an average of 12 years (median = .24), and the unique variance of the majority of items had a significant heritable component (median = .14). These three item properties were inter-correlated, suggesting that items systematically differed in the degree of reflecting valid unique variance. Also, associations of items' unique variance with age, gender, and Body Mass Index (BMI) replicated across samples and tracked with the original findings. Moreover, associations between item residuals and BMI obtained from one group of people allowed for a significant incremental prediction of BMI in an independent sample. Overall, these findings reinforce the hypotheses that nuances constitute the building blocks of the personality trait hierarchy, their properties are robust and they can be useful.

Keywords: Personality nuances; items; Five-Factor Model; Big Five; trait hierarchy

Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability and utility of personality nuances

What are the basic units of individual differences in personality? This fundamental question for personality science has implications for not only personality assessment—what should be measured?—but also for our understanding of the very architecture of personality. In many approaches to personality, the basic units are traits (e.g., Deary, 2009; Funder, 1991; Johnson, 1997): relatively enduring behavioral, cognitive, affective, and motivational tendencies. While the key feature of traits is their temporal stability, they should also become manifest in ways that are detectable by different observational methods, for otherwise we cannot infer their objective existence (i.e., construct validity) and/or they are inconsequential and thereby of little practical and scientific interest (e.g., Funder, 1995; McCrae, 1982). Additionally, several theorists (e.g., Allport, 1931; McCrae & Costa, 2008a) have proposed at least some neurological/heritable basis as a hallmark property of traits. If there is a relatively small number of characteristics that meet these trait-requirements, this would suggest that the architecture of personality can be parsimoniously conceived of as a limited set of underlying structures (for a review, see Kandler, Zimmermann, & McAdams, 2014). For example, the neuroscience-grounded Reinforcement Sensitivity Theory outlines three underlying systems for personality (Corr, 2009). However, a large number of relatively stable and heritable characteristics, which are also consensually observable by independent raters, would imply that the etiology of individual differences is commensurately more diverse.

Personality trait hierarchy

Current evidence suggests that it may be hard to pinpoint an exact and finite set of personality traits. The currently most widely employed trait model, the Five-Factor Model (FFM; McCrae &

John, 1992) or the Big Five (Goldberg, 1990), outlines five traits: Neuroticism [or Emotional (In)Stability], Extraversion (versus Introversion), Openness (or Intellect), Agreeableness (versus Aggressiveness), and Conscientiousness (or Constraint). However, the Big Five traits (often called "domains" in the FFM) do not constitute the only possible personality trait model. On one hand, the domains are inter-correlated and can therefore be merged into fewer "metatraits" such as Stability and Plasticity (DeYoung, 2006), although such higher-order traits may be less trait-like than the FFM domains in having lower cross-rater agreement and heritability (McCrae et al., 2008; Riemann & Kandler, 2010). On the other hand, each FFM domain appears to consist of at least somewhat distinct, narrower "subtraits". For example, DeYoung, Quilty and Peterson (2007) have delineated two "aspects" for each FFM domain, whereas each aspect may be further broken down into yet more specific traits, which have been termed "facets" (e.g., McCrae & John, 1992). Although there is no universally accepted taxonomy of facets yet, a widely adopted personality measurement framework embodied in the NEO Personality Inventories (NEO) proposes six facets for each FFM domain (Costa & McCrae, 1992). Of course, this multitude of broader and narrower constructs does not necessarily represent "nature carved at its joints", with boundaries between constructs being fuzzy and potentially arbitrary. Such multi-layered and fuzzy representation of traits can be called the "personality hierarchy", with facets often taken as its most basic, granular units that make up all broader traits. Or so it has been thought for some time.

Personality nuances

It was recently suggested that personality hierarchy may in fact extend below facets (McCrae, 2015). Specifically, even very narrow behavioral, cognitive, affective, and motivational tendencies that can be represented by individual personality questionnaire items tend to display the hallmark properties of traits: stability over time, agreement among raters, and some heritability (Mõttus, Kandler, Bleidorn, Riemann, & McCrae, 2017; Mõttus, McCrae, Allik, & Realo, 2014). Of course, this may be because the individual items are indicators of facets, aspects, and domains and may therefore manifest the properties of those trait constructs. However, this cannot be the full explanation. Mõttus and colleagues (2017) found that even when residualized for the variance of all 30 facets of the NEO (and thereby for the aspects and domains), the remaining individual differences in most items showed an appreciable level of stability over about five years, could be validated across rater perspectives (self-reports and two well-informed observer ratings), and often reflected a significant level of genetic influences. This suggests that specific items—or perhaps bundles of content-wise very similar items (e.g., constantly re-trying to improve oneself or quickly giving up on self-improvements)—may not only be indicators of traits such as their corresponding facets (e.g., Achievement Striving), aspects (e.g., Industriousness), and domains (e.g., Conscientiousness) but may also reflect something unique about how individuals differ from one another. These unique characteristics have been called nuances (McCrae, 2015).

In other words, one can imagine a population of people who are identical in all FFM domains, aspects and facets. Would there no longer be any valid variance as far as personality is concerned (i.e., any residual variance is just noise)? No. To the extent that the findings pertaining to nuances are valid, the individuals could still differ in many personality-relevant ways that are visible to different observers and stable over several years, and these nuanced variations could partly reflect genetic differences among these people.

Relevancy of personality nuances

To the extent that the findings showing pervasive unique variance in nuances are robust and replicable, they suggest that nuances may often be the most basic units of personality variance rather than facets, aspects or yet broader traits[1]. If so, attempts to identify etiological factors

1 Measurement-wise, of course, items are indisputably the basic units of personality. However, here we are suggesting that the specific characteristics that they reflect may also be the most basic *conceptual* units. As a

contributing to personality variance need to at least heed the possibility that any of the putative factors (e.g., a life event, intervention or genetic variant) may operate at the level of specific nuances, in addition to or even instead of broader traits, and could be studied accordingly. In principle, this may explain the relatively modest success of attempts to identify gene variants associated with broad personality traits (Lo et al., 2017; McCrae, Scally, Terracciano, Abecasis, & Costa, 2010; Vinkhuyzen et al., 2012). That is, any single gene variant may account for a negligible proportion of variance in those composite trait constructs, because it is relevant for only few of its constituents, whereas at the level of these specific constituents the effect of the genetic variant may be substantial. Also, it may be that the mechanisms underlying normative personality development drive the apparent personality change through nuances. There is indeed evidence that age differences in personality are often specific to individual items (Mõttus et al., 2015). For example, the normative (and perhaps adaptive) increase of Neuroticism in old age (Kandler, Kornadt, Hagemeyer, & Neyer, 2015; Mõttus, Johnson, & Deary, 2012) may be driven by increases in nuances pertaining to wariness of health-risks and dangers of everyday life.

Likewise, the associations between personality and life-outcomes may sometimes pertain to specific characteristics represented by individual test items rather than their facets, aspects, or domains *per se* (Mõttus, 2016). For example, Mõttus and colleagues (2017) found that the unique variance in items (e.g., trying different foods or being entranced or fascinated by music) was meaningfully linked with variables, such as Body Mass Index (BMI) and interests in various life domains. This possibility does in no way belittle the existent findings obtained at the level of broader traits, which often show substantial links to broad criteria, such as conservatism (e.g., Openness) and subjective well-being (e.g., Neuroticism), whereas nuance-specific differences may not be relevant in these instances (Mõttus et al., 2017). These broader traits-based findings have demonstrated the robustness and ubiquity of the links between personality and life outcomes. However, considering nuances *without* aggregating them may at least sometimes reveal additional information and thereby provide novel research avenues. Nuances may also provide additional ways of mapping personality variation across demographic groups or cultures.

Such findings also attest to the pervasiveness of the kinds of enduring psychological features that personality research focuses on—often in stark contrast with other fields of psychology that deal with change and malleability (e.g., developmental and social psychology). For example, even very specific behavioral patterns such as liking showy styles or redecorating, attending sports games or parties, and being entranced by music or different foods show remarkable stability over time, even *net* of the common variance of hundreds of other specific characteristics (Mõttus et al., 2017). Thus, virtually everything that people do seems to form specific patterns of their own that endure over several years.

The findings regarding nuances also tell us something about the ubiquity of genetic influences (cf. Turkheimer, Pettersson, & Horn, 2014). For instance, even when two people are identical in 30 NEO personality facets including Excitement-Seeking, one may be more inclined to listen to loud music whereas the other prefers roller-coasters (two items measuring Excitement-Seeking), and this variance appears to at least in part mirror genetic differences between these individuals. Likewise, two people may substantially *differ* on the level of facets, aspects, or domains, but be similar on the level of a specific nuance (e.g., enthusiasm for technology), possibly because they share gene variants that are somehow relevant for these nuances.[2] It seems thus possible that not only are even the most specific behavioral tendencies often distinctively stable over time, but this may be because

consequence, scores of an item and even their residuals (i.e., item scores residualized for all broader personality traits) may not only reflect measurement-invariant elements of a personality trait measure, but also a trait by itself. Some items may in fact reflect more than one yet more specific trait.

2 The same can be true for shared (or not shared) formative experiences that act to increase (or decrease) the similarity in personality nuances between two people who are significantly different (or completely similar) on the level of facets, aspects, or domains.

they reflect unique expressions of genetic variance. In fact, this possibility was already alluded to by Eaves and Eysenck (1975) who estimated the heritability of "inconsistency" in item responses at .47.

Also, nuances may provide a bridge between the trait and social-cognitive approaches to personality (e.g., Mischel, 2004). On one hand, nuances may be traits like any other construct that trait psychologists study—because they display the hallmark properties of traits. On the other hand, however, personality test items and thereby nuances often represent contextualized tendencies: "When I go on a trip, I make detailed plans" or "If I am insulted, I forgive and forget." (Mõttus et al., 2017). Such conditional tendencies (or *if… then…* contingencies) closely mirror the concept of behavioral signatures (i.e., individual situation-behavior profiles), often focused on by social-cognitive approaches (Mischel & Shoda, 1995). The stability of item residuals may reflect the context-dependent stability of individual differences that may otherwise eliminated by aggregations across situations or different items reflecting different situation-behavior expressions (Mischel, Shoda, & Mendoza-Denton, 2002).

Finally, the findings highlighting the unique variance in nuances may also have implications for personality measurement. Specifically, they suggest that instruments that are designed to measure the same constructs but contain different sets of items may not always measure the same (underlying) constructs after all. This is because items are not merely interchangeable indicators of broader traits and therefore different scales aggregate different psychological characteristics. This mostly applies to shorter scales, because in longer scales the unique variance of nuances is more effectively filtered out/balanced by other nuances. Likewise, shorter scales are not likely to capture a wide range of nuances and may therefore lack coverage of the personality feature space. The pervasiveness of unique variance in nuances therefore speaks against the use of short personality scales (McCrae, 2015). These findings may also inform debates as to whether unidimensional personality scales are feasible or desirable at all.

The present study

To date, there have been a limited number of attempts to systematically study the trait-like properties of nuances. The key findings pertaining to the stability and heritability of items' unique variance have only been studied in one study (Mõttus et al., 2017) that made use of longitudinal and multiple-rater data from a sample of German twins, whereas cross-rater agreement on items' unique variance was also studied by Mõttus and colleagues (2014) based on an Estonian sample. Clearly, the robustness and generalizability of these findings needs further empirical scrutiny, given their broad implications for the conceptualizations and operationalizations of personality.

The present study attempted to replicate the key findings of Mõttus and colleagues (2017): the rank-order stability, cross-rater agreement, and heritability of the unique variance of personality questionnaire items. It constituted an almost direct replication in that exactly the same comprehensive NEO questionnaire—consisting of 240 items mapped into 30 facets and five FFM domains—and identical statistical procedures were used. However, the replication was more comprehensive in using more samples, consisting of 6,287 people from six different cultures (Australia, Canada, Czech Republic, Denmark, Japan, and the United States of America [US]), who completed the questionnaire in four different languages (Czech, Danish, English, and Japanese). The different datasets were collected by different researchers for different purposes at different time-points and varied in their age ranges. The replication extended the re-testing interval (to assess rank-order stability of items' unique variance) from about 5 years to about 15 years. The current samples were also pooled with the German twin sample to obtain meta-analytic estimates for rank-order stability, cross-rater agreement, and heritability of the raw scores and residuals of the individual questionnaire items. Finally, prior results regarding the relationships of items' unique variance with age, gender, and BMI were replicated, and all available data were again pooled to

obtain meta-analytic estimates for these associations. In sum, this constitutes the yet most systematic and powerful scrutiny of the validity and utility of the allegedly lowest level of personality hierarchy—the nuances.

Method

Ethics Statement

The current study collates and re-analyses personality questionnaire data from a range of previously published studies that are referred to in the *Participants* section. Data were collected in a number of countries for different research projects. Most of the data were collected more than a decade ago. In each project, participation was voluntary, participants were informed that they could stop participating at any time and that their data would be treated with confidentiality. All projects were approved by appropriate institutional review boards, except for the Czech data collection (the institution did not have a review board in 2000 when the data was collected).

Participants

Participants' demographic characteristics are summarized in Table 1.

Samples from Australia, Canada, Denmark and Japan were used to study the genetic and environmental influences on personality nuances. **The Australian sample**, drawn from the Borderline Personality Disorder Study and the Genetics of Laterality, Smell, Taste and Reading Study at QIMR Berghofer Medical Institute (Distel et al., 2008, Gillespie et al., 2013), consisted of 866 twins, who had completed the personality inventory (with no more than 40 missing responses for either twin). **The Canadian sample**, drawn from the University of British Columbia Twin Project and described in Jang, Livesley, and Vernon (2002) and Jang, Taylor, and Livesley (2006) consisted of 900 twins. **The Danish sample** is described in Schousboe and colleagues (2004) and Makransky, Mortensen and Glas (2013); here, we use data from 1,190 Danish twins. **The Japanese sample** was drawn from the Keio Twin Project described in Ando and colleagues (2004); the current sample consisted of 1,278 twins. **The Czech sample** was used to study cross-rater agreement on personality nuances and was based on the dataset described in McCrae and colleagues (2004). Here, data from 709 participants for whom both self- and informant-ratings were available were used. The informants were often the participants' partners (aged between 15 and 81 with a mean age of 36.10 years; 298 of the informants were men). **The US sample** was used to study rank-order stability of personality nuances and was taken from the Baltimore Longitudinal Study of Aging (BLSA), described in Terracciano, McCrae, Brand, and Costa (2005). Here, data from 1,336 people, who had provided complete data for two measurement points, were used. The two measurement occasions were on average 14.66 years apart (*SD* = 5.64).

Measure

For the measurement of personality characteristics, a version of the Revised NEO Personality Inventory (NEO-PI-R) was administered (Costa & McCrae, 1992); in Australia, Canada and the US the original English version was used, whereas in Czech Republic, Denmark and Japan the Czech (Hřebíčková, 2002), Danish (Skovdahl Hansen & Mortensen, 2004) and Japanese (Yoshimura, 1998) versions were used, respectively (in Czech Republic, informants completed the third-person form of the questionnaire). The NEO-PI-R contains 240 items, grouped into 30 facet scales, which are hierarchically organized under the five FFM domains. The domains of the NEO-PI-R correspond to the five factors found in analyses of multiple personality inventories (John, Naumann, & Soto, 2008; Markon, Krueger, & Watson, 2005), and its facets were chosen to represent important traits within each domain (McCrae & Costa, 2008b). Therefore, the NEO-PI-R item pool provides a broad sample of nuances, even though it clearly does not exhaust the population of

nuances. Responses were given on a 5-point Likert scale ranging from *strongly disagree* to *strongly agree*. In the Australian, Canadian, Danish, Japanese and US samples, most participants reported their height and weight, allowing us to calculate their BMI.

Table 1. *Descriptive information of the samples.*

	Australia	Canada	Denmark	Japan	Czech Republic	United States
N	866	900	1190	1286	709	1336
Female	502	587	603	846	415	657
Age (M)	22.47*	33.66	38.14	20.85**	36.03***	56.73
Age (SD)	3.73	13.87	11.52	3.9	14.02	14.8
Age (range)	17-33	15-68	18-67	14-30	15-81	19-89
Identical MZ twin pairs	193	249	232	449	-	-
Same sex DZ twin pairs	124	154	238	118	-	-
Opposite sex DZ pairs	116	47	125	76	-	-

NOTE: MZ = monozygotic; DZ = dizygotic; * For two individuals age at testing was not known; ** For one twin pair age was unknown; *** For 9 participants age was unknown.

Statistical analyses

When up to 40/240 responses were missing, median imputation was used. Generally, the prevalence of missing data was low (although 702 participants had some missing data in the combined twin datasets, only 39 had more then 10 missing responses and only 22 had more than 20 missing responses; there were no missing responses in the Czech and US data). As in Mõttus and colleagues (2017), two analyses were carried out for each of the 240 items, one based on raw scores and the other based on residuals. The residuals were obtained from linear regressions of item scores against scores of all 30 NEO-PI-R facets (the score of the item being residualized was excluded from its respective facet at the time). Therefore, item residuals were independent of FFM domains, aspects, and facets and represented the unique variance in nuances.

Cross-rater agreement and rank-order stability were estimated via product moment correlations across different rater perspectives and across measurement occasions, respectively. Heritability analyses were carried out both independently in the Australian, Canadian, Danish and Japanese twin samples to estimate the consistency of findings and then in the combined sample to maximize power. For the raw scores and item residuals of each item, a structural equation model specifying additive genetic contributions to (i.e., narrow-sense heritability of) nuances and two environmental variance components reflecting non-genetic influences shared and not shared by twins; the models were fit using Robust Maximum Likelihood estimator (see Mõttus et al., 2017, for more details; in short this was the standard so-called ACE model of behavior genetic variance decomposition). In such a model, heritability (h^2) is indicated by (about two-fold) larger monozygotic (MZ) twin correlations in a trait compared to the dizygotic (DZ) twin correlation. Shared environmental influences (c^2) are indicated by substantial twin correlations but no marked differences between MZ and DZ twin correlations. The remaining variance is taken as evidence for non-shared environmental influences, but this component also includes measurement error variance. For heritability analyses, item scores were first residualized for gender because the sample included opposite-sex twins, and in the combined sample, also for the cohort to account for country differences in item means.

Age, gender and BMI associations with raw item scores and item residuals were estimated with bivariate linear models separately in each sample, with items as outcomes (twins were treated as

independent individuals as in Mõttus et al., 2017). All variables other than gender were standardized, so that age and BMI associations were in the correlation metric, whereas gender associations were in Cohen's *d* (standard deviation) metric. The age-, gender- and BMI-associations in the original study (based on combined self- and informant-ratings) were re-analyzed using exactly the same analytic approach as used in this replication a) to obtain estimates for all associations pertaining to BMI (only a set of specific hypotheses were tested in the original study), b) to obtain standard errors for age- and gender-associations (for meta-analysis), and c) by using a larger number of participants than was used in the original report for this subset of analysis (N = 1,491; this is because for age and gender-associations, only the subset of people for whom conservatism and subjective well-being ratings—other "broad criteria"—were available was used in the original study, N = 844)[3]. Associations were then meta-analyzed across the samples. For the meta-analysis, we used the standard inverse-variance based formula (e.g., Willer, Li, & Abecasis, 2010). For statistical significance, the most conservative Bonferroni-corrected *p*-value threshold (.05/240 = .0002) was used for each key type of analysis to retain consistency with Mõttus and colleagues. However, in meta-analyses we also used the more lenient False-Discovery Rate (FDR; Benjamini & Hochber, 1995), because replicable patterns across samples and types of findings would indicate that the assumption of testing a novel null hypothesis in each single association would be too stringent.

The degree of replication was estimated by comparing the distributions and rankings of the respective item properties (e.g., cross-rater correlations for the residuals of the 240 NEO-PI-R items) in the original study and the current replication. For each property, the findings are summarized in the text and/or tables [median (*Mdn*), inter-quartile range (*IQR*) and Spearman's rank-order correlation (ρ)], whereas the density distributions and scatterplots are provided in the Online Supplemental Materials. We also compared the proportions of associations that were statistically significant, although we note that this criterion directly depends on the sample size, and we compared the degrees to which residualizing items for all facets reduced the estimates of respective item properties. For BMI, the replicability of a series of specific hypotheses put forward in the original study was also estimated.

All analyses were performed in R statistical software (R Development Core Team, 2017).

Results

Cross-rater agreement, rank-order stability, heritability and shared environmental influence estimates for each item's raw and residual scores are reported in the Online Supplemental Materials; for reference, these estimates are also reported for the FFM domains and their facets.

Cross-rater agreement

Cross-rater correlations (in Czech data) for raw item scores ranged from r = .11 to r = .54 with a median correlation of r_{Mdn} = .29; see also Table 2 for *IQR*s. For item residuals, the correlations ranged from r = -.01 to r = .43 (r_{Mdn} = .13). For comparisons, the respective estimates from Mõttus and colleagues (2017) are also reported in Table 2; they are very similar. For item residual-based correlations, 43% were statistically significant (p < .0002); the percentage had been 60% in Mõttus and colleagues but this study relied on a larger sample size for cross-rater correlations. Residualizing items for all facets reduced the median cross-rater correlation by 54%; the respective estimate was 57% in Mõttus and colleagues. Across the 240 items, the cross-rater correlations also ranked similarly across the two studies, with rank-order correlations between the respective vectors of cross-rater correlations being ρ = .58 and ρ = .59 for raw item scores and residuals, respectively (p < .001; we use rank-order correlations for comparing items in their properties to guard against

3 The respective vectors of original (N = 844) and re-analyzed (N = 1,491) associations of items/their residuals with age and gender correlated highly, with Spearman's ρ = .89 to .95.

out-lier effects). Thus, the items and item residuals for which there was greater higher cross-rater agreement in the German data also tended to have higher cross-rater agreement in the Czech data, providing evidence for the robustness of the findings.

Table 2. *Cross-rater agreement on item scores.*

	Raw item scores		Residual item scores	
	Original	Replication	Original	Replication
Median	.28	.29	.12	.13
1st quartile	.22	.25	.08	.09
3rd quartile	.34	.34	.16	.18
Proportion significant	98.33%	97.50%	60.00%	42.50%

NOTE: Original = Mõttus et al. (2017); Replication = The current results based on Czech data. Proportion significant = the percentage of estimates significant at $p < .0002$.

Rank-order stability

Test-retest correlations (in US data) for raw item scores ranged from $r = .15$ to $r = .59$ ($r_{Mdn} = .37$; see Table 3 for *IQR*s), whereas those for item residuals ranged from $r = .07$ to $r = .53$ ($r_{Mdn} = .23$). For comparisons, the respective estimates (based on self-reports and combined self- and informant-ratings) from Mõttus and colleagues (2017) are also reported in Table 3; the estimates are somewhat smaller in the current study. This can be explained by, on average, three-times longer retesting intervals, because the stability of personality differences tends to decrease with time interval length (Fraley & Roberts, 2005; Terracciano, Costa, & McCrae, 2006). For item residual-based correlations, about 99% were statistically significant ($p < .0002$). Residualizing items for all facets reduced the median cross-rater correlation by 37%; the reductions were 36% to 47% in Mõttus and colleagues. Therefore, despite the absolute values of retest correlations being smaller in the current study, the ratio of the rank-order stability of residual item scores to rank-order stability of raw item scores was comparable in both studies. Also, the retest correlations of the 240 items tended to rank similarly across the two studies, with $\rho = .51$ and $\rho = .55$ for raw item scores (respectively, for self-ratings and combined self- and informant-ratings of Mõttus et al., 2017), and respectively $\rho = .49$ and $\rho = .52$ for item residuals ($p < .001$). Thus, the items and item residuals that displayed relatively higher rank-order stability in the German data also tended to demonstrate relatively higher rank-order stability in the US data, despite a substantially longer average retesting interval in the latter.

Heritability and shared environmental effects

Across the four twin samples, the median ratio of intraclass correlations (*ICC*) for dizygotic twins to these for the monozygotic twins (ICC_{DZ}/ICC_{MZ}) varied from .35 to .52 for raw item scores and from .33 to .48 for item residuals, with the median of the eight ratios being .41. Therefore, all samples indicated evidence for heritability and low evidence for shared environmental contributions for both raw item scores and item residuals.[4]

4 Although there was some evidence for non-additive genetic influences, as indicated by ICC_{DZ}/ICC_{MZ} ratios lower than 0.50, for the sake of consistency with Mõttus and colleagues (2017) we fitted models not allowing for non-additive genetic influences. However, this is not hugely problematic, because estimates of additive genetic effects derived from twin models have been shown to be good estimations of broad-sense heritability including additive and nonadditive genetic factors (Hill, Goddard, & Visscher, 2008).

Table 3. *Rank-order stability (retest correlations) of item scores.*

	Raw item scores			Residual item scores		
	Original (self-ratings)	Original (combined ratings)	Replication	Original (self-ratings)	Original (combined ratings)	Replication
Median	.53	.51	.37	.34	.27	.23
1st quartile	.49	.44	.31	.29	.20	.18
3rd quartile	.58	.58	.41	.39	.33	.28
Proportion significant	100.00%	100.00%	100.00%	96.67%	80.42%	98.75%

NOTE: Original = Mõttus et al. (2017); Replication = The current results based on the US data. Proportion significant = the percentage of estimates significant at $p < .0002$.

In both Australian and Japanese samples, the median heritability estimates were $h^2_{Mdn} = .22$ for raw item scores and $h^2_{Mdn} = .10$ for item residuals; in the Danish sample, the respective estimates were $h^2_{Mdn} = .23$ and $h^2_{Mdn} = .09$, and in the Canadian sample they were $h^2_{Mdn} = .24$ and $h^2_{Mdn} = .11$. Thus, the typical estimates were very similar across the samples. For the Canadian, Danish and Japanese samples, the rankings of heritability estimates of the 240 items correlated at between $\rho = .22$ and $\rho = .33$ in case of raw scores and between $\rho = .21$ and $\rho = .24$ for residuals ($p < .001$ for all). The estimates from the Australian (smallest) sample correlated between $\rho = .15$ ($p = .022$) and $\rho = .25$ ($p < .001$) with the corresponding estimates from other samples for raw scores and between $\rho = .04$ ($p = .541$) and $\rho = .11$ ($p = .093$) for item residuals; however, the Australian estimates correlated at $\rho = .28$ and $\rho = .25$ with the averaged estimates of the other three samples, respectively for raw and item residual scores ($p < .001$ for both). Overall, there was thus clear evidence for replicability across the four countries and we reran the analyses in the pooled dataset.

Table 4. *Heritability estimates of item scores.*

	Raw item scores			Residual item scores		
	Original (self-ratings)	Original (combined ratings)	Replication (combined sample)	Original (self-ratings)	Original (combined ratings)	Replication (combined sample)
Median	.26	.32	.24	.14	.13	.14
1st quartile	.21	.24	.20	.06	.05	.09
3rd quartile	.33	.37	.28	.19	.19	.18
Proportion significant	62.50%	70.42%	85.83%	48.33%	50.42%	66.67%

NOTE: Original = Mõttus et al. (2017); Replication = The current results based on the combined sample of Australian, Canadian, Danish, and Japanese twins. Proportion significant = the percentage of estimates significant at $p < .0002$.

In the combined sample, heritability estimates for raw item scores ranged from $h^2 = 0$ to $h^2 = .43$ ($h^2_{Mdn} = .24$; see Table 4 for *IQRs*). For item residuals, the estimates ranged from $h^2 = 0$ to $h^2 = .41$ ($h^2_{Mdn} = .14$). For comparison, the respective estimates (based on self-reports and combined self- and informant-reports) from Mõttus and colleagues (2017) are reported in Table 4; the estimates are relatively similar, especially for item residuals. For item residual-based heritability estimates, 67% were statistically significant in the combined sample ($p < .0002$); the respective estimates had been 48% and 50% in Mõttus and colleagues (the current sample was substantially larger). Residualizing

items for all facets reduced the median heritability estimate by 42%; the respective estimates were 46% and 59% in Mõttus and colleagues. Moreover, the heritability estimates of the 240 items tended to rank similarly across the two studies, with ρ = .30 and ρ = .36 for raw item scores and ρ = .29 and ρ = .34 for item residuals (respectively, for combined-sample heritability estimates from the current study and self-ratings and combined self- and informant-ratings based estimates from Mõttus et al.; $p < .001$ for all correlations). Thus, there was a tendency for items and item residuals that displayed relatively higher heritability estimates in the German data to demonstrate relatively higher heritability estimates in the current Australian/Canadian/Danish/Japanese combined sample.

In contrast, there was less consistency across the Australian, Canadian, Danish, Japanese and German (Mõttus et al., 2017) samples in the estimates for shared environmental influences (c^2), with rankings of estimates correlating between ρ = -.12 and ρ = .19 (*Mdn* = .02). The estimates (Australian/Canadian/Danish/Japanese/Combined sample) ranged from c^2 = 0 to c^2 = .27/.33/.22/.28/.19 for the raw item scores (c^2_{Mdn} = 0/.02/0/0/0, c^2_{IQR}: 0 to .05/.12/.05/.02/.03) and from c^2 = 0 to c^2 = .21/.24/.17/.19/.15 for item residuals (c^2_{Mdn} = 0/.01/0/0/0, c^2_{IQR}: 0 to .05/.09/.05/.05/.03). In the combined sample, the estimates were significant ($p < .0002$) for 3% and 6% of items, respectively for raw and residual scores. These generally low estimates are consistent with Mõttus and colleagues (2017), suggesting modest systematic effects of shared (e.g., family) environment on even the most specific personality characteristics.

Convergence of the findings

Were items and their unique variances that displayed highest cross-rater agreement also those that demonstrated highest rank-order stability and heritability? This could suggest that these trait-like properties are systematic features of the personality characteristics that the items reflect. There had been evidence for this in Mõttus and colleagues (2017), but in this study all estimates had been based on the same sample. In the present study that relied on different samples for each type of estimates the same applied: The estimates correlated between ρ = .36 and ρ = .47 for both raw and residual item scores (Table 5; $p < .001$). However, estimates for shared environmental effects did not clearly track with either cross-rater agreement nor rank-order stability, although items with higher heritability estimates tended to be those with smaller shared environmental effect estimates. Items with relatively higher estimates for raw scores tended to be the items with relatively higher estimates for residual scores (diagonal of Table 5).

Table 5. *Correlations between the item-level estimates for cross-rater agreement, rank-order stability, heritability and shared environmental influences.*

	Cross-rater agreement	Rank-order stability	Heritability	Shared environment
Cross-rater agreement	.58	.36	.39	-.11
Rank-order stability	.43	.68	.47	-.07
Heritability	.42	.46	.73	-.45
Shared environmental influences	-.06	-.12	-.52	.59

NOTE: Correlations for estimates of item residuals are below the diagonal; correlations for estimates of raw items scores are above the diagonal. On the diagonal are the correlations between respective estimates from items' raw and residual scores.

Age and gender effects

There was cross-study consistency in how items were associated with gender. Across the six samples, rankings of items in the associations with gender correlated between ρ = .23 and ρ = .67 (*Mdn* = .50) for raw and ρ = .22 and ρ = .58 (*Mdn* = .38) for residual item scores, respectively (p < .001 for all). Likewise, there was consistency across Canadian, Czech, Danish and US samples (mean age > 30) in how items were associated with age; the rankings of items in the associations with age correlated between ρ = .45 and ρ = .65 (*Mdn* = .56) for raw and ρ = .28 and ρ = .44 (*Mdn* = .38) for residual item scores, respectively (p < .001 for all). Therefore, the findings were meta-analyzed across the four samples. The Japanese and Australian samples were substantially younger and had a relatively limited variance in age; their rankings of items' associations with age correlated at ρ = .58 (p < .001) and ρ = .15 (p = .020), respectively for raw and residual item scores and their associations with the estimates from other samples varied from ρ = -.08 to ρ = .63. Because of generally different age ranges and associations with age that tracked less consistently those in older samples, the age-associations were separately meta-analyzed for the two younger samples (i.e., Japanese and Australians), and replicability (in relation to the findings of the original study) was primarily expected for the older sample-based meta-analytic results (because the original sample more closely matched the age of the older meta-analytic sample).

The meta-analytic (across the four older samples) effect sizes were generally modest. Absolute correlations with age ranged from 0 to .30 (*Mdn* = .07; *IQR*: .04 to .10) for raw items scores and from 0 to .13 (*Mdn* = .03; *IQR*: .01 to .04) for item residuals. In the original study data (Mõttus et al., 2017), the respective median estimates were .09 (*IQR*: .05 to .16) and .04 (*IQR*: .02 to .08). Cohens' *d*s in relation to gender ranged from 0 to .69 (*Mdn* = .11; *IQR*: .07 to .22) for raw items scores and from 0 to .44 (*Mdn* = .05; *IQR*: .02 to .08) for item residuals. In the original study data, the respective median estimates were .20 (*IQR*: .10 to .32) and .08 (*IQR*: .03 to .13). For both age and gender, replication effect sizes were somewhat smaller than in the original study probably because large sample sizes generally entail more realistic (i.e., small) effect sizes. For items residuals, 15% of associations were significant (p < .0002) for both age and gender; in Mõttus and colleagues (2017) the percentages were 16% and 8%, respectively for age and gender. Notably, the rankings of items in terms of their associations with age and gender in the current meta-analysis correlated with respective rankings obtained in the German sample (combined self- and informant-ratings): ρ = .74 and ρ = .63 for associations with age (raw item scores and item residuals, respectively), and ρ = .76 and ρ = .54 for associations with gender (raw item scores and item residuals, respectively); p < .001 for all. Thus, although the associations of items' unique variance with age and gender were small in magnitude and often statistically non-significant, their general patterns clearly replicated across studies and countries. Therefore, even small effect sizes are likely to reflect valid signal.

The estimates for the associations with age were also small in the meta-analysis of Australian and Japanese samples, with median absolute correlations with age *Mdn* = .05 and *Mdn* = .02, for raw and residual item scores. Interestingly, the age-associations in the meta-analysis of younger samples tended to rank similarly to associations in the meta-analysis of older samples and original German data for raw item scores (ρ = .47 and ρ = .46, p < .001, respectively), but not for item residuals (ρ = .01, p = .924, and ρ = .03, p = .690, respectively). It may thus be that age-differences are particularly non-linear for the unique variance in items; or it may be that the generally weaker associations of item residuals with age could not emerge within the limited age range in the younger samples.

Associations with BMI

Items' associations with BMI were tested in Australian, Canadian, Danish, Japanese and US samples; the rankings of items in their associations with BMI correlated between $\rho = .04$ and $\rho = .48$ for raw item scores and between $\rho = .11$ and $\rho = .34$ for item residuals (all but two associations were significant at $p < .05$). At a closer inspection, it appeared that all correlations between rankings of items in BMI-associations were significant at $p \leq .001$ for Canadian, Danish and US samples ($\rho \geq .21$), whereas all of the somewhat lower correlations pertained to the Australian and Japanese samples (the youngest of the five samples), suggesting that BMI-personality associations may be moderated by age. However, BMI-personality associations in the Australian sample correlated with the averaged associations from the other samples at $\rho > .28$ ($p < .001$) for both raw and residual item scores and the same applied to the associations in the Japanese data. Given this overall consistency, we meta-analyzed the associations across the five samples.

The median absolute correlations with BMI were .02 (*IQR*: .01 to .04) for raw items and .02 (*IQR*: .01 to .03) for item residuals; 10% and 3% of the respective associations were statistically significant ($p < .0002$). In the original study data (Mõttus et al., 2017), the respective estimates were .05 (*IQR*: .02 to .09), .04 (*IQR*: .02 to .07), 14% and 1%. Mõttus and colleagues (2017) put forward five hypotheses for particular item residuals to be associated with BMI: two of them were significant in the current study ("Overeats favorite foods", $r = .14$, $p < .001$; "Eats excessively", $r = .07$, $p < .001$). The residuals of items "Gives up on self-improvements" ($r = -.02$, $p = .101$; reverse-keyed item), "Plans before travel" ($r = .03$, $p = .057$) and "Tries different food" ($r = .02$, $p = .092$) were not significantly associated with BMI. However, the rankings of the 240 items in their meta-analytic associations with BMI tracked respective associations in the German twins in Mõttus and colleagues, with $\rho = .65$ and $\rho = .42$ for the associations with raw item scores and item residuals, respectively ($p < .001$ for both). Again, although individual effect sizes of item-BMI associations were small and their level of significance varied across studies, collectively they did reflect a replicable association pattern.

In order to quantify the extent to which items' unique variance could contribute to the prediction of BMI we took the meta-analytic associations across the Australian, Canadian, Japanese and US samples (i.e., we excluded Danish results at this step from the meta-analysis) and used these to predict BMI in the Danish sample. Specifically, standardized scores of the 240 items in the Danish sample were multiplied by the respective items' meta-analytic correlations with BMI and these products where then summed for each individual; we call these predicted-from-items BMI values *polyitem scores*. This procedure resembles the creation of polygenic scores, widely used in quantitative genetics to predict phenotypes based on (often many thousands of very small) genomic associations found in independent samples (Dudbridge, 2013). We then carried out a similar procedure for item residuals, by multiplying standardized residuals in the Danish sample with the meta-analytic associations between item residuals and BMI; we call the results *polyresidual scores*. Both polyitem and polyresidual scores correlated significantly with BMI in the Danish sample ($r = .25$, $p < .001$, for both). In a multiple regression, where polyresidual scores were entered as predictors of BMI alongside age, gender and the scores of the 30 NEO-PI-R facets, they still significantly predicted BMI ($\beta = .15$, $p < .001$). This prediction was not entirely driven by the N5: Impulsivenss items referring to eating (see Figure 1), because removing all items of this facet from the polyresidual scores still resulted in them predicting BMI ($\beta = .11$, $p < .001$), controlling all facets, age and gender. Therefore, item residuals allowed for an incremental prediction of BMI even in an independent sample tested in another language[5].

5 The Danish sample was selected for prediction, because the Australian and Japanese samples were notably younger than the other samples (their somewhat lower consistency in the item-BMI associations indicated that these links may not have developed to the full extent in early adulthood) and the US data had been analyzed separately from other samples (by the third author; all other analyses were carried out by the first author). However, the same pattern replicated, when polyitem and polyresidual scores were calculated in the Canadian sample (e.g., correlations

Meta-analysis of cross-rater agreement, rank-order stability, heritability and shared environmental effects

Finally, we calculated the meta-analytic estimates for cross-rater agreement, rank-order stability, heritability and shared environmental effect estimates, collating findings from the present samples and Mõttus and colleagues (2017). That is, we meta-analyzed Czech and German data for cross-rater agreement and US and German data for rank-order stability, wheres as for heritability and shared environmental influences, we used the combined Australian, Canadian, Danish, and Japanese samples-based estimates from this study and estimates based on combined self- and informant-ratings of German twins from Mõttus and colleagues. For age, gender and BMI-associations, all available data from this study was meta-analysed with the German data (there was no BMI-data for the Czech participants, and age-associations were not taken from the younger Australian and Japanese samples).

Table 6. *Meta-analytic estimates for cross-rater agreement, rank-order stability, heritability and shared environmental influences.*

	Raw item scores				Residual item scores			
	Agree-ment	Stability	Herita-bility	Shared environ-ment	Agree-ment	Stability	Herita-bility	Shared environ-ment
Median	.28	.41	.28	.00	.12	.24	.14	.00
1st quartile	.23	.34	.23	.00	.09	.20	.07	.00
3rd quartile	.33	.45	.33	.00	.16	.29	.18	.00
Proportion significant	100.00%	100.00%	95.00%	4.58%	81.67%	100.00%	68.33%	7.50%
Proportion significant (FDR)	100.00%	100.00%	97.50%	14.12%	97.08%	100.00%	74.17%	12.91%

NOTE: Proportion significant = the percentage of estimates significant at $p < .0002$. Proportion significant (FDR) = the percentage of estimates that were significant when corrected for false discovery rate across all 960 estimates of the kind (based on either raw or residual item scores).

The summary of the estimates is given in Table 6. For cross-rater agreement, the median meta-analytic estimate was $r_{Mdn} = .28$ for raw item scores and $r_{Mdn} = .12$ for item residuals; removing the common variance of all facets reduced the median estimate by 59%. Estimates for all raw scores and 82% of item residuals were statistically significant ($p < .0002$). In the Online Supplemental Materials, we also added the Estonian data reported in Mõttus and colleagues (2014) to the meta-analysis, which resulted in almost all item residuals displaying significant cross-rater agreement. The rankings of the 240 items in our meta-analytic cross-rater correlations of residuals correlated with the respective Estonian estimates at $\rho = .64$ ($p < .001$). For retest correlations, the median meta-analytic estimate was $r_{Mdn} = .41$ for raw item scores and $r_{Mdn} = .24$ for item residuals; removing the common variance of all facets reduced the median estimate by 41% but retest correlations of all item residuals were still statistically significant. The sample size-weighted average retest interval for these meta-analytic estimates was about 12 years. For heritability, the median meta-analytic estimate was $h^2_{Mdn} = .28$ for raw item scores and $h^2_{Mdn} = .14$ for item residuals; removing the

of observed BMI with polyitem and polyresidual scores were $r = .21$ and .20, $p < .001$, respectively; for calculations of these scores, Canadian data was swapped with Danish data in the meta-analysis).

common variance of all facets reduced the median estimate by 50%. Estimates for 95% of raw scores and 68% of item residuals were statistically significant ($p < .0002$). The meta-analytic estimates for the effect of shared environment were small, but significant for 8% of item residuals. Figure 1 displays the Manhattan plots of the *p*-values for cross-rater agreement, heritability and shared-environmental influences, grouped by facets and domains: for the former two, highly significant estimates pertain to items throughout the set of thirty facets.

Meta-analytic estimates for cross-rater agreement and rank-order stability were highly correlated ($\rho = .64$ and $\rho = .63$, for raw item scores and item residuals). Also, both of these item properties were correlated with items' heritability estimates (ranging from $\rho = .49$ to $\rho = .53$). Specifically, given their high correlations, rank-order stability and cross-rater agreement may have indexed a common property of items, which we can call items' *nuancedness*. Plotting the nuancedness (average of rank-order stability and cross-rater agreement) against heritability (Figure 2), it appears that several items had zero heritability (5 for raw items and 51 for item residuals) despite an often appreciable level of nuancedness. Did these items reflect nuances, but a kind of nuances that were not subject to systematic influences as detectable by behavior genetic models? Not always, because several of the items with zero heritability were the ones for which models had estimated significant ($p < .0002$) levels of shared environmental effects: one of the five items in case of raw scores and 17 of the 51 items in case of item residuals (the shared environment effects for these items are shown with crosses in Figure 2; the items with stronger shared environmental influences also tended to be more nuanced in item residuals). Therefore, some of specific personality characteristics that are stable over time and observable consistently across raters may be more subject to shared environmental than genetic influences, despite the overall level of shared environmental influences being modest. This leaves raw scores of four and residual scores of 34 items with only non-shared environmental influences (or reflecting only measurement error).

Given the clear patterns in the findings outlined above, using Bonferroni correction in null hypothesis testing may have been too stringent as it assumed that a new null hypothesis was tested for each item. In other words, null was an unlikely default estimate for the item [residual] properties of cross-rater agreement, rank-order stability and heritability. We recalculated the meta-analytic *p*-values using less stringent FDR correction across the 960 *p*-values of a kind (e.g., raw item score-based estimates for cross-rater agreement, rank-order stability, heritability and shared environmental influences of 240 items). According to the FDR-corrected *p*-values, for example, the unique variance in 100%, 97% and 74% of items displayed significant rank-order stability, cross-rater agreement and heritability, respectively, whereas there was evidence for significant shared environmental influences for 13% of item residuals (see Table 6 and Figure 1). Of the 240 item residuals, 173 (72%) simultaneously had FDR-corrected significant cross-rater agreement, rank-order stability and heritability estimates (134 item residuals, or 56%, had all three properties Bonferroni-corrected significant at the same time).

The five most nuanced items referred to trying new foods, liking distressing movies, liking vacations with the crowds, liking roller-coasters and liking attending games. The five least nuanced items referred to needing a lot of help, believing that most people can be trusted, trying to be respectful of others, having poor judgment in difficult situations, and getting easily dishearted and giving up. The five items with the highest heritability in their unique variance referred to fulfilling civic duties, being sometimes entranced in music, trying new foods, liking roller-coasters and liking distressing movies. The five items with the strongest evidence for shared environmental influences referred to being curious about many thing, believing in the value of behaving honestly, wanting to get ahead, blaming oneself and liking action.

Meta-analysis of associations with age, gender and BMI

We also meta-analyzed associations with age, gender and BMI across the current samples and the original German sample. About 22% and 21% of item residuals had Bonferroni-corrected significant associations with age and gender, respectively. However, when we applied FDR correction on the 480 associations (240 for age and 240 for gender), 43% and 44% of associations of item residuals with age and gender were significant, respectively. Among the items with the strongest residual-associations with higher age were those referring to being sickened by others, being interested in patterns, not liking roller coasters, not wanting to get ahead, not acting impromptu, not liking jobs that require working with others, not having liberal principles, misplacing things and not liking distressing movies ($r \geq .10$, $p < 10^{-12}$). Among the items with the strongest residual-associations with being a female were those that referred to liking expressive dance, liking redecorating, being easily frightened, not finding music fascinating and not liking distressing films (Cohen's $d > .20$, $p < 10^{-15}$).

Residuals of 14 items (6%) had Bonferroni-corrected significant associations with BMI, whereas 46 items (19%) had FDR-corrected significant associations with BMI. The strongest association pertaining to items' unique variance was the one with overeating favorite foods ($r = .16$, $p < 10^{-16}$). Among the remaining associations with item residuals, higher BMI was linked with not having a fast-moving life, not being emotionally sensitive to environments, not being liberal in moral principles, being unable to resist carvings, having too much of a good thing, eating excessively, being expected to take lead and being riled by others ($r = .05$ to $.07$, $p < .0001$). All five *a priori* hypothesized item residual-BMI associations put forward by Mõttus et al (2017) were supported at $p < .01$.

The Manhattan plots of *p*-values of age-, gender- and BMI-associations are depicted in Figure 1; for the former two, highly significant associations pertain to items from a range of facets of all FFM domains, whereas for BMI several of the significant associations pertain to the unique variance in N5: Impulsiveness items. Although individual effect sizes were mostly small, the sheer number of highly significant residual associations reveals the degree of nuanced age and gender differences in personality, and the nuancedness of personality-BMI associations. These findings attest to the potential utility of nuances for mapping demographic variability in personality and prediction of life outcomes.

All meta-analytic estimates for each items' raw and residual scores are reported in the Online Supplemental Materials. We cannot reproduce the items, but we have provided short indications of their meanings (note that he meanings of all items are given in the direction required by their respective facet rather than the original direction of reverse-keyed items, so the direction of the effects corresponds to the *given* meanings of items and not their original keying). Furthermore, anyone with access to the manual of the NEO-PI-R (Costa & McCrae, 1992) will be able to match the exact wording of particular items and their estimates. In the Online Supplemental Materials, we also added the Estonian data described in Mõttus and colleagues (2014, 2017) to the meta-analyses; the number of item residuals being significantly associated with age, gender and BMI was substantially higher in these larger meta-analyses. For example, 13% of item residuals had Bonferroni-corrected significant associations with BMI (39% of the associations were FDR-corrected significant) and all of the five *a priori* hypotheses were supported at $p \leq .001$.

Variance

Allik and colleagues (2010) reported that scales with more variance demonstrated higher cross-rater agreement. Could inter-item differences in their nuancedness—degree to which they reflect unique but substantive variance—result from some of them simply having larger variance than others? To test this, we calculated the standard deviations of raw item scores in each sample of this study and the original study (Mõttus et al., 2017) and averaged the estimates across the samples,

and then repeated the same for item residuals. Indeed, items that had more variance to start with (i.e., in raw scores) tended to have more of it after residualizing, $\rho = .84$ ($p < .001$), and they also tended to have higher meta-analytic estimates of cross-rater agreement, rank-order stability and heritability in both raw scores and residuals ($\rho = .37$ to $\rho = .52$, $p < .001$). However, inter-item differences in variance were unlikely to be the only cause for why items systematically differed in their nuancedness, because the correlations between cross-rater agreement, rank-order stability and heritability remained substantial ($\rho = .32$ to $\rho = .56$, $p < .001$) in both raw and residual item scores when controlling for items' standard deviations (either in raw or residual scores).

Disattenuating for unreliability

Cross-rater agreement, rank-order stability, heritability and shared environmental influence estimates pertaining to single items were attenuated (compared to estimates pertaining to aggregate traits) by the elevated ratio of error variance to total variance; this is especially true for items' residual variance, because this is the level of variance where the random error is pushed into.

In an attempt to correct for this, we estimated the reliability of raw and residual item scores by employing the myPersonality dataset (Kosinski et al., 2015), in which a sufficiently large number of participants had completed a 100-item FFM measure (Goldberg at al., 2006) multiple times. We calculated correlations for the raw scores and residuals (controlling for all FFM traits; the measure had no facets, therefore the item residual scores are likely to contain some facet-level variance) for each item across five retesting intervals: 1 day, 2 to 7 days, 8 to 14 days, 15 to 21, and 22 days to a month. The medians (across the 100 items) of the respective retest correlations were .70, .68, .66, .65, and .63 for raw item scores and .57, .52, .50, .48, and .46 for item residuals ($N = 831$ to 3,271). The correlations pertaining to retesting intervals of only one or a few days may have been inflated by participants still remembering their choices to the questionnaire items rather than contemplating on their characteristics anew, whereas correlations pertaining to longer intervals (e.g., a month) may have, in principle and in part, reflected substantive changes in personality. We think that correlations resulting from a retesting interval of 8 to 14 days may provide useful approximations for items' reliabilities, hence we estimate the reliabilities of a typical item at .66 and .50, respectively for raw and residual scores.

Correcting the median meta-analytic estimates of cross-rater agreement (Table 6) for unreliability yielded median cross-rater agreement estimates of .28/.66 = .42 for raw item scores and .12/.50 = .24 for item residual scores. Correcting the median meta-analytic estimates of rank-order stability (over about 12 years) yielded median stability estimates of .41/.66 = .62 for raw item scores and .24/.50 = .48 for item residual scores. The correction of median meta-analytic heritability estimates for reliability resulted in $h^2 = .28/.66 = .42$ and $h^2 = .14/.50 = .28$, respectively for raw and residual item scores. The median estimates for shared environmental influences were 0, so we did not correct these, but we note that these may remain underestimates. It should be noted, however, that the correction of the estimates pertaining to item residuals may have been an under-correction, because the retest correlations of item residuals could have been inflated by facet-level variance. And it should also be noted that items are likely to differ in their reliability, whereas we only corrected median effect sizes for a single point-estimate of unreliability.

Discussion

The present findings clearly replicate the key findings presented by Mõttus and colleagues (2017): Single personality test items often reflect specific and unique personality characteristics with trait-like properties of rank-order stability, cross-rater agreement, and heritability, even when the variance of a wide range of personality traits has been removed from these items. Most median estimates found here are comparable to those of Mõttus and colleagues, and even the rankings of the 240 items in terms of corresponding item properties—cross-rater agreement, rank-order

stability, and heritability estimates—tended to be similar. This level of replicability is even more impressive considering that the current study relied on a set of independent samples from different cultures, sometimes tested in different languages. The original findings were based on a German sample, whereas the present findings were based on Americans, Australians, Canadians, Czechs, Danish, and Japanese.

We also meta-analyzed the current estimates and those reported by Mõttus and colleagues (2017) and found that the unique variance of all items displayed significant rank-order stability over an average of about 12 years, whereas the unique variance of the majority of items also demonstrated significant cross-rater agreement and heritability. Moreover, these three item properties tracked with each other, suggesting that the more items tend to contain (uniquely) stable variance in how individuals differ in personality, the more observable valid and heritable those individual differences tend to be. However, in a relatively small number of cases the unique variance may be more subject to influences from the environment that individuals who are raised together share.

In addition to the replicability of cross-rater agreement, rank-order stability and heritability, the associations of items' unique variance with age, gender and BMI generally replicated across samples. Moreover, items' unique variance allowed for the prediction of BMI even when the weights for the prediction model were based on participants from different countries. Collectively, these findings are consistent with there being a specific level of personality characteristics, nuances, which can be represented by single test items. The nuances extend the personality trait hierarchy below facets and may serve as its most basic units. The nuances can contribute to a more refined mapping of individual differences and prediction of outcomes.

Against the odds

In order to properly interpret these findings, it should be noted that there were at least three reasons why they should *not* have emerged. First, the ratio of measurement error to total variance is likely to be large in item residuals. Second, item scores are recorded in a 5-point scale that artificially constrains variance and, moreover, the distributions of the scores of many NEO-PI-R items are highly skewed with respondents rarely using some response options (Mõttus et al., 2015). The third and perhaps the most important reason is that the items of questionnaires such as the NEO-PI-R have been written and selected with the aim of maximizing their common (facet-level or domain-level) variance rather than their single item-level variance, as the degree of common variance (internal consistency) is typically taken as evidence for scales' reliability. In other words, questionnaire items are specifically designed *not* to capture unique variance (i.e., nuances). And yet most of them do. It is possible that nuances are under-represented among the items of existing questionnaires and in more diverse item pools the trait-like properties of nuances may appear even more pervasive. Also note that when corrected for estimated reliability, typical cross-rater agreement, rank-order stability and heritability estimates of items and their unique variances increased substantially.

Do most items constitute unique nuances?

Most of the 240 items showed evidence for some signal over and above the 30 NEO-PI-R facets, either in terms of significant residual cross-rater agreement, rank-order stability, heritability or, in majority of cases, all of them at the same time. Does this mean that most of the items constitute nuances of their own, conveying unique information and having distinct etiology, or are many of them in fact redundant—captured by the variance of other nuances? We propose that this question requires further research and multiple lines of evidence. For example, a research programme that systematically examines the unique predictive value of questionnaire items for criteria beyond personality (i.e., controlling for the predictive contributions of all other items) can

help. To the extent that particular items tend to be uniquely predictive of at least some criteria, it is plausible that they do constitute unique nuances—they show evidence for some autonomy in how they intersect with phenomena beyond personality. As a working hypothesis, we can speculate that items with stronger unique cross-rater agreement, stability and heritability are more likely to be among those uniquely predictive of criteria, if only because they may contain more unique signal.

Another potentially useful type of evidence could come from examining the partial correlations among large numbers of items such as those of the NEO-PI-R (i.e., correlations controlling for all other correlation). Relatively strong partial correlations between particular items could be suggestive of redundancies. Naturally, redundant items could be either bundled or some of them could be dropped. Further, items that tend to have strong partial correlations with many other items (i.e., are high on centrality; Costantini et al., 2015) could represent the most information-rich nuances, whereas the least connected "fringe" items may have relevance for specific criteria but might generally not be useful for understanding the functioning of personality or might simply not contain much unique signal. As a working hypothesis, we can speculate that among the central, well-connected personality items tend to be those with stronger cross-rater agreement, stability and heritability. Also, genetically informative data (e.g., twins) could be useful for decomposing the (partial) correlations among sets of items into genetic and environmental components; for example, very strong genetic correlations among items could point to genetic common etiology.

If a subset of items tends to be more useful for many purposes than the remaining items, this suggest that researchers with limited resources could get most of the benefit of a long questionnaire administering only the most useful subset of its items.

On stability

The retest interval was nearly three times longer in the current study than in Mõttus and colleagues (2017) and the rank-order stability estimates were correspondingly lower. This is consistent with previous research showing that rank-order or individual-level stability wanes with time (Fraley & Roberts, 2005), although this waning tendency appears to wane itself with longer testing intervals (Anusic & Schimmack, 2016; Terracciano et al., 2006). However, notably the reduction of rank-order stability was comparable in raw item scores and their residuals, suggesting that the stability of nuances does not decay more than the stability of broader traits that the nuances make up.

On heritability

The rankings of items in terms of cross-rater agreement and rank-order stability were somewhat more similar between the present findings and those of Mõttus and colleagues (2017) than the rankings of heritability estimates. This may be because heritability estimates are less reliable than rank-order stability and cross-rater agreement estimates, but also because the specific factors that influence the kinds of specific behavioral, cognitive, affective, and motivational characteristics that single test items capture—that is, genetic backdrop, environmental factors, or their interplay—may genuinely vary across samples and cultural contexts. However, our findings do imply an appreciable level of cross-study/sample consistency in how items ranked in their levels of heritability (as well as age and gender differences), and the average heritability estimates were very similar.

How should we interpret the finding that even very specific behavioral, cognitive, affective, and motivational tendencies that items and their unique variances capture often reflect genetic influences? We cannot be sure at this point, but theories on the etiology of personality traits need to heed this finding. It may seem implausible that there are genetic variants exclusively for liking roller-coasters or taking civic duties seriously, for example. As one possibility, these findings are consistent with the idea that genetic influences on personality are non-specific (Mõttus, Realo,

Vainik, Allik, & Esko, 2017). That is, the genetic influences may not affect particular personality traits *per se*, but act as a general "genetic pull" on the basis of which processes that generally act at the phenotypic level to maximize person-environment fit carve out what would appear as traits (Turkheimer et al., 2014). If so, personality traits as such may be phenotypic phenomena—adaptations between individuals with certain genetic proclivities and their environments.

Another possibility is that same gene variants may be related to different trait constructs at different levels of the (fuzzy) trait hierarchy within a specific group (e.g., family or culture) of individuals at a particular time or age. This would explain genetic correlations between and familial or cultural resemblance within different features of individuals. Similarly, different gene variants may be associated with the same expressions of traits between two individuals at a phenotypic level, just as different possible solutions for one and the same problem. The latter would explain why it is so difficult to identify specific gene variants for trait expressions at population level. One possible way to tackle those challenges could be the analyses of gene variants at very basic (nuanced) levels of behavioral, cognitive, affective, and motivational tendencies.

On effects of growing up together

It would not have been unreasonable to hypothesize that the effects of the environment that children of the same family share are somewhat stronger for specific behavioral, cognitive and affective tendencies captured in items' unique variance that they have typically been found for broader and ostensibly more "deep-routed" traits (see Kandler et al., 2014). Why shouldn't childhood experiences contribute to liking roller-coasters or enjoying music? And yet in most cases the influences of twins' shared environments appeared modest. This finding should be at least heeded by theories that ascribe substantial developmental roles to childhood experiences. Perhaps the systematic and long-lasting influences of childhood environment are indeed very small—even for the most specific habits? Or perhaps perhaps even familial influences are not actually shared by children growing up together such that (genetically non-identical) siblings evoke differential responses from their environments (e.g., parents), based on their own partly genetically influenced proclivities? Of course, influences shared by family members on nuances may be stronger in childhood, which our study did not cover, but even then these effects are unlikely to be lasting, because they were small in adulthood.

On universality

The findings that age and gender differences in the FFM traits are replicable across a variety of cultural contexts and that there are also replicable patterns of cross-rater agreement on the FFM traits are taken as evidence for the human-universality and deep-rootedness of these traits (Allik, Realo, & McCrae, 2013). These findings suggest that the FFM traits are real attributes of human, not just sociocultural constructions. To the extent that these properties also appear cross-culturally replicable, in their typical and relative magnitudes, for single personality test items and even for the unique variance in these, suggests that the specific personality characteristics that these items and their unique variance represent may also be human universals—and that they are likely to describe real attributes of humans similarly to the broader trait constructs such as those of the FFM.

Limitations and future directions

In addition to BMI, Mõttus and colleagues (2017) attempted to estimate the predictive validity of nuances by linking items' unique variance with subjective well-being, conservatism, and interests in various life domains, although they acknowledged that such attempts should be part of a more systematic research program. For the latter reason as well as for the lack of an appropriate combination of data for these specific outcomes, we did not attempt to replicate these findings here. Systematic attempts to address items' unique predictive value have been started elsewhere (Mõttus,

Bates, Condon, Mroczek and Revelle, in preparation; Seeboth and Mõttus, in press). For example, using a large sample of adults, Seeboth and Mõttus (in press) predicted 40 diverse outcomes from both the FFM domains and items, training and validating the models in independent subsamples to avoid artificially inflated predictive power due to model over-fitting. They found that items tended to outperform domains in most cases and the degree of their incremental predictive value did not depend on outcomes' breadth. Moreover, the predictive power of domains was conflated with (and inflated by) items' unique predictive value, whereas residualizing items for domains only marginally decreased their overall predictive power. However, these studies only constitute the very start for the research program into items' (or nuances') incremental predictive value. It may well turn out that there are systematic regularities in where and how nuances provide incremental predictive value and that often they do not add any value at all. It may also turn out that in many cases the items uniquely related to particular outcomes are so similar to these outcomes in content that the associations are rather uninformative. It should be note, however, that Seeboth and Mõttus (in press) found that, for the range of outcomes that they considered, item-outcome content overlap could generally not be responsible for items' incremental predictive value because for most outcomes there were no semantically similar items.

The current study was a replication, using exactly the same personality questionnaire as the original study. The properties of nuances appear robust as far as this particular pool of them is concerned. However, the robustness of these properties should eventually be tested using alternative operationalizations of personality, such as the HEXACO personality trait hierarchy (Ashton & Lee, 2007). It may be that the FFM domains and facets, as measured in the NEO-PI-R, may not cover all broader personality traits and therefore the unique variance of their items could in fact be accounted for by other traits.

Finally, we can envisage some criticism inherent to nuanced-based research. This may pertain to lower reliability of single items, apparently lower parsimony of the results (even if this may be a reflection of reality) and their greater dependence of the particular instruments being used in particular studies. These are valid concerns, but if nuances are real in having robust trait-like properties, then possible inconveniences related to studying them should not put researchers off. For example, lower reliability implies the need to improve items' measurement properties and work with large samples, and working with high-dimensional data implies the need to develop representations of personality that allow us to extract useful regularities from multi-variable association patterns. For example, geneticists working with millions of genetic variants can represent their associations parsimoniously by collating effects into polygenic scores (Dudbridge, 2013).

Conclusion

The present findings reinforce the possibility that the most basic building blocks of the personality hierarchy are nuances—specific personality traits currently represented by single items of personality measures. Even the unique variance in these most specific personality characteristics show trait-like properties of cross-rater agreement, rank-order stability, and genetic influences (and sometimes environmental influences shared by individuals raised together). Moreover, these specific characteristics can provide leverage for mapping demographic variability in personality and the prediction of life outcomes such as BMI. This suggests that how individuals differ in their personality may be highly multi-dimensional in the descriptive sense and possibly as diverse in terms of etiology—and yet aspects of this multi-dimensionality may appear rather universal across time and space. We may choose to ignore this multi-dimensionality, or we may start developing descriptive and explanatory models of personality that can accommodate this.

References

Allport, G. W. (1931). What is a trait of personality? *The Journal of Abnormal and Social Psychology*, *25*, 368–372.

Ando, J., Suzuki, A., Yamagata, S., Kijima, N., Maekawa, H., Ono, Y., & Jang, K. L. (2004). Genetic and Environmental Structure of Cloninger's Temperament and Character Dimensions. *Journal of Personality Disorders*, *18*, 379–393.

Anusic, I., & Schimmack, U. (2016). Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations. *Journal of Personality and Social Psychology*, *110*, 766-781.

Ashton, M. C. & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*, 150–166.

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 (Methodological)*, *57*, 289–300.

Corr, P. J. (2009). The reinforcement sensitivity theory of personality. In P. J. Corr & G. Matthews (Eds.), *The Cambridge handbook of personality psychology.* (pp. 347–376). New York, NY US: Cambridge University Press.

Costa, P. T., & 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.

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13–29.

Deary, I. J. (2009). The trait approach to personality. In P. J. Corr & G. Matthews (Eds.), *The Cambridge handbook of personality psychology.* (pp. 89–109). New York, NY US: Cambridge University Press.

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.

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.

Distel, M. A., Trull, T. J., Derom, C. A., Thiery, E. W., Grimmer, M. A., Martin, N. G., Willemsen, G., & Boomsma, D. I. (2008). Heritability of borderline personality disorder features is similar across three countries. *Psychological Medicine 38,* 1219-1229.

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLOS Genet*, *9*, e1003348. doi:10.1371/journal.pgen.1003348

Eaves, L., Eysenck, H. (1975). Genetic and Environmental Components of Inconsistency and U nrepeatability in Twins' Responses to a Neuroticism Questionnaire. *Behavior Genetics, 6,* 145-160.

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: a dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review*, *112*, 60–74.

Funder, D. C. (1991). Global Traits: A Neo-Allportian Approach to Personality. *Psychological Science*, *2*, 31–39.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*, 652–670.

Gillespie, N. A., Henders, A. K., Davenport T. A., Hermens, D. F., Wright, M. J., Martin, N. G., & Hickie, I. B. (2013). The Brisbane Longitudinal Twin Study: Pathways to Cannabis Use, Abuse, and Dependence Project-Current Status, Preliminary Results, and Future Directions. *Twin Research and Human Genetics, 16,* 21-33.

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

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*, 84–96.

Hill, W. G., Goddard, M. E., & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLOS Genetics, 4,* e1000008.

Hřebíčková, M. (2002). Vnitřní konzistence české verze NEO osobnostního inventáře (NEO PI-R) [Internal consistency of the Czech version of the NEO Personality Inventory (NEO-PI-R)]. *Československa psychologie, 46,* 521-535.

Jang, K. L., Livesley, W. J., & Vernon, P. A. (2002). The aetiology of personality function: The University of British Columbia Twin Project. *Twin Research*, *5*, 342–346.

Jang, K. L., Taylor, S., & Livesley, W. J. (2006). The University of British Columbia Twin Project: Personality is something and personality does something. *Twin Research and Human Genetics*, *9*, 739-742.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), Handbook of personality: *Theory and research* (3rd ed., pp. 114–158). New York, NY: Guilford Press.

Johnson, J. A. (1997). Units of analysis for the description and explanation of personality. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology*. London: Academic Press.

Kandler, C., Kornadt, A. E., Hagemeyer, B., & Neyer, F. J. (2015). Patterns and sources of personality development in old age. *Journal of Personality and Social Psychology*, *109*, 175–191.

Kandler, C., Zimmermann, J., & McAdams, D. P. (2014). Core and surface characteristics for the description anthory of personality differences and development. *European Journal of Personality, 28*, 231–243.

Kosinski, M., Matz, S., Gosling, S., Popov, V. & Stillwell, D. (2015). Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines. *American Psychologist, 70,* 543-556.

Lo, M.-T., Hinds, D. A., Tung, J. Y., Franz, C., Fan, C.-C., Wang, Y., … Chen, C.-H. (2017). Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nature Genetics*, *49*, 152–156.

Makransky, G., Mortensen, E. L., & Glas, C. A. W. (2013). Improving Personality Facet Scores With Multidimensional Computer Adaptive Testing: An Illustration With the Neo Pi-R. *Assessment*, *20*, 3–13.

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.

McCrae, R. R. (1982). Consensual validation of personality traits: Evidence from self-reports and ratings. *Journal of Personality and Social Psychology*, *43*, 293–303.

McCrae, R. R. (2015). A More Nuanced View of Reliability: Specificity in the Trait Hierarchy. *Personality and Social Psychology Review*, *19*, 97–112.

McCrae, R. R., & Costa, P. T. (2008b). Empirical and theoretical status of the five-factor model of personality traits. In B. Boyle, G. Matthews, & D. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment: Volume 1 — Personality theories and models* (pp. 273–295). London: SAGE.

McCrae, R. R., & Costa, P. T. (2008a). The five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research (3rd ed.).* (pp. 159–181). New York, NY US: Guilford Press.

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

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*, 175–215.

McCrae, R. R., Scally, M., Terracciano, A., Abecasis, G. R., & Costa, P. T. (2010). An alternative to the search for single polymorphisms: Toward molecular personality scales for the five-factor model. *Journal of Personality and Social Psychology, 99,* 1014–1024.

McCrae, R. R., Yamagata, S., Jang, K. L., Riemann, R., Ando, J., Ono, Y., ... Spinath, F. M. (2008). Substance and artifact in the higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *95*, 442–455.

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology*, *55*, 1-22.

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.

Mischel, W., Shoda, Y., & Mendoza-Denton, R. (2002). Situation-behavior profiles as a locus of consistency in personality. *Current Directions in Psychological Science*, *11*, 50–54.

Mõttus, R. (2016). Towards more rigorous personality trait–outcome research. *European Journal of Personality*, *30*, 292–303.

Mõttus, R., Bates, T. C., Condon, D. M., Mroczek, D., & Revelle, W. (under review). Your personality data can do more: Items provide leverage for explaining the variance and co-variance of life outcomes. Retrieved from psyarxiv.com/4q9gv

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

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology, 112,* 474–490.

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.

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*, e0119667.

Mõttus, R., Realo, A., Vainik, U., Allik, J., & Esko, T. (2017). Educational attainment and personality are genetically intertwined. *Psychological Science, 28*, 1631-1639.

R Development Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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.

Schousboe, K., Visscher, P. M., Erbas, B., Kyvik, K. O., Hopper, J. L., Henriksen, J. E., … Sørensen, T. I. A. (2004). Twin study of genetic and environmental influences on adult body size, shape, and composition. *International Journal of Obesity and Related Disorders; Hampshire, 28,* 39–48.

Seeboth, A., & Mõttus, R. (in press). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate prediction and mapping of life outcomes. *European Journal of Personality.*

Skovdahl Hansen, H. & Mortensen E. L. (2004). Dokumentation for den danske udgave af NEO PI-R og NEO PI-R Kort Version [Documentation for the Danish version of NEO PI-R and NEO PI-R Short Version]. In: P. T. Costa Jr. & R. R. McCrae (eds.): *NEO PI-R. Manual -*

klinisk [NEO PI-R. Manual -clinical] (pp.53-86). København, Denmark: Psykologisk Forlag A/S.

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

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.

Turkheimer, E., Pettersson, E., & Horn, E. E. (2014). A phenotypic null hypothesis for the genetics of personality. *Annual Review of Psychology*, *65*, 515–540.

Vinkhuyzen, A. A. E., Pedersen, N. L., Yang, J., Lee, S. H., Magnusson, P. K. E., Iacono, W. G., . . . Wray, N. R. (2012). Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. *Translational Psychiatry*, *2*, e102.

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*, 2190–2191.

Yoshimura, K. (1998). Reliability and validity of a Japanese version of the NEO Five-Factor Inventory (NEO-FFI) : A population-based survey in Aomori prefecture. *Japanese Journal of Stress Sciences*, *13*, 45–53.

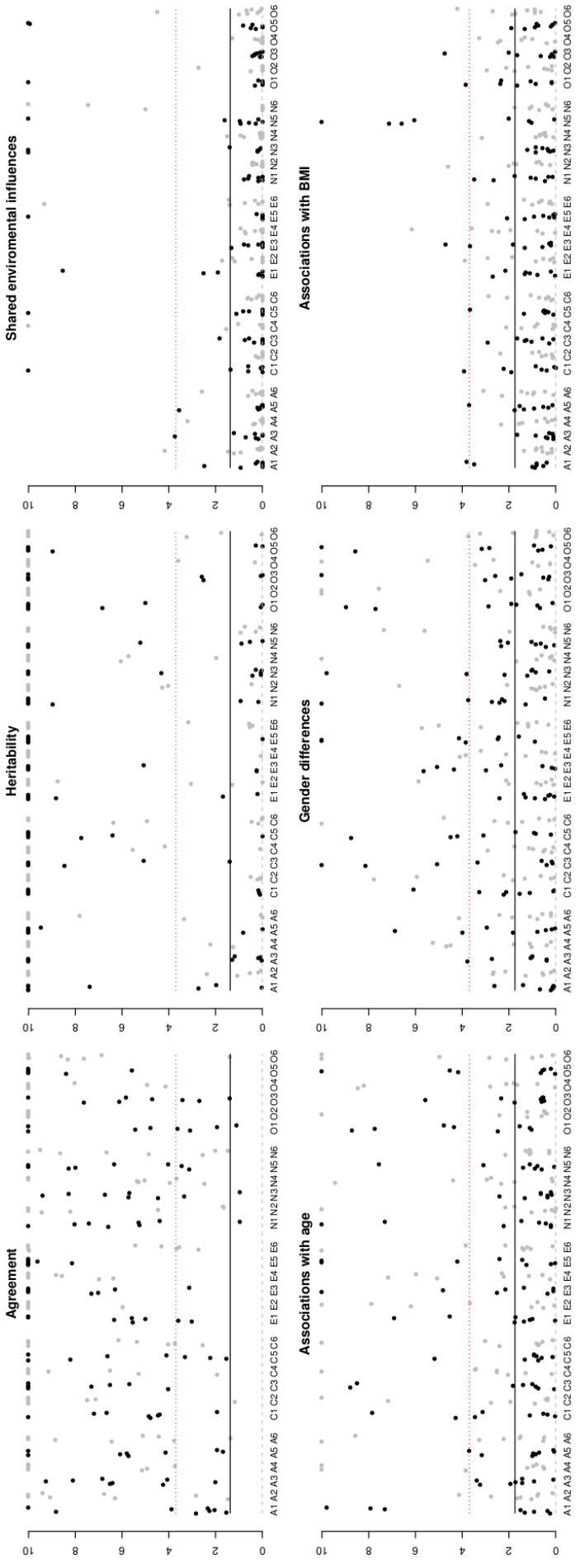
Agreement
Heritability
Shared enviromental influences
Associations with age
Gender differences
Associations with BMI
A1 A2 A3 A4 A5 A6 C1 C2 C3 C4 C5 C6 E1 E2 E3 E4 E5 E6 N1 N2 N3 N4 N5 N6 O1 O2 O3 O4 O5 O6
0 2 4 6 8 10

Figure 1. -Log10 p-values (on vertical axes) of the meta-analytic cross-rater correlations of item residuals (controlling for all facets), their meta-analytic heritability and shared environmental influences, and their meta-analytic associations with age, gender and BMI. Values are capped at 10 (i.e., $p < 10^{-10}$). *P*-values are grouped along the horizontal axes by facets and facets are grouped by the Five-Factor Model domains (A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Openness). Solid lines indicate FDR-corrected significance, whereas the red dotted lines indicate Bonferroni-corrected significance threshold. *P*-values for rank-order stability are not show as they were all Bonferroni-corrected significant (mostly $p < 10^{-10}$).

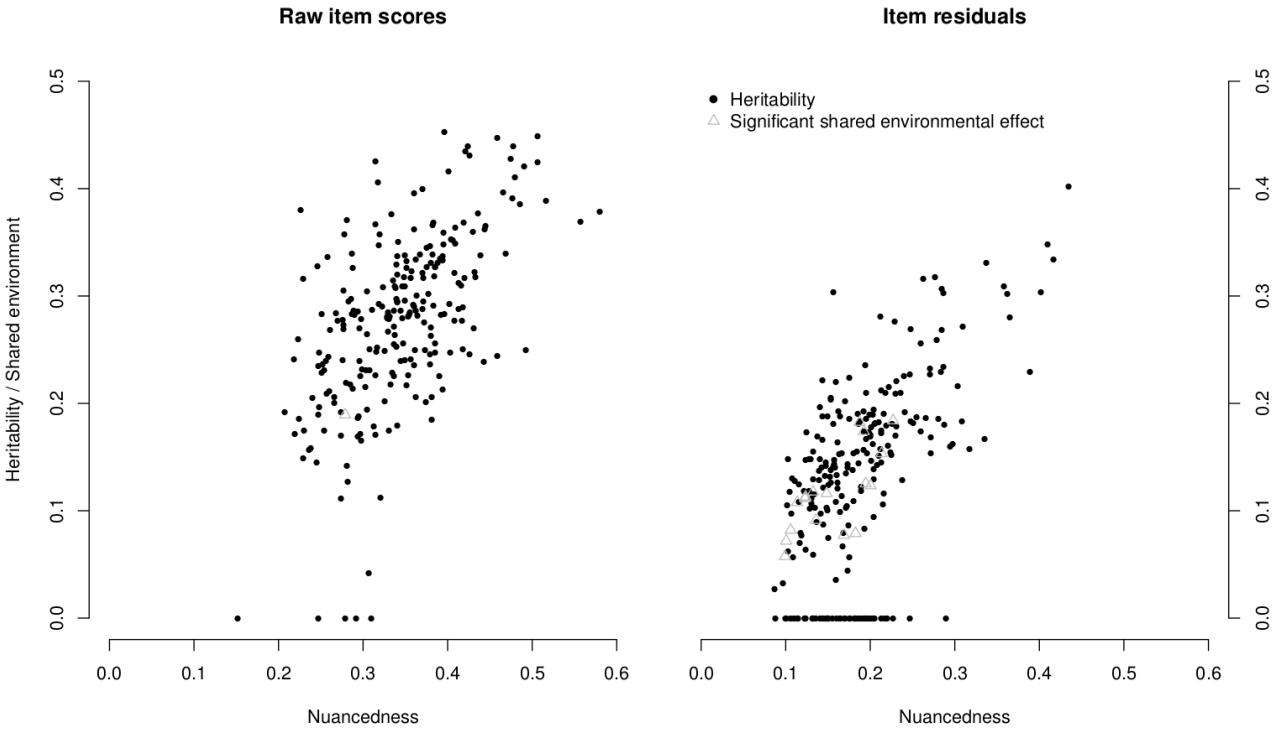

Figure 2. Items' "nuancedness" (average of cross-rater agreement and rank-order stability) and level of heritability/shared environmental influences (only significant shared environment effects are shown).