Cross-rater agreement on common and specific variance of personality scales and items

Citation for published version:
Mõttus, R, McCrae, RR, Allik, J & Realo, A 2014, 'Cross-rater agreement on common and specific variance of personality scales and items', *Journal of Research in Personality*, vol. 52, pp. 47-54. https://doi.org/10.1016/j.jrp.2014.07.005

Digital Object Identifier (DOI):
10.1016/j.jrp.2014.07.005

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Research in Personality

SHORT TITLE: Common and specific variance in personality

Cross-rater agreement on common and specific variance of personality scales and items

René Mõttus *
Department of Psychology, Centre for Cognitive Ageing and Cognitive Epidemiology,
University of Edinburgh, UK
Department of Psychology, University of Tartu, Estonia

Robert R. McCrae
Baltimore, Maryland

Jüri Allik
Department of Psychology, University of Tartu, Estonia
Estonian Academy of Sciences, Tallinn, Estonia

Anu Realo
Department of Psychology, University of Tartu, Estonia

* Corresponding author:
7 George Square
EH8 9JZ Edinburgh, Scotland, UK
rene.mottus@ed.ac.uk
Tel: +441316503410

Abstract

Using the NEO Personality Inventory-3, we analyzed self/informant agreement on personality traits at three levels that were made statistically independent from each other: domains, facets, and individual items. Cross-rater correlations for the common variance in the five domains ranged from 0.36 to 0.65 (*M* = 0.49), whereas estimates for the specific variance of the 30 facets ranged from 0.40 to 0.73 (*M* = 0.56). Cross-rater correlations of residual variance of individual items ranged from -0.14 to 0.49 (*M* = 0.15; 88% statistically significant at $p < 0.002$). Agreement on common variance was moderately related to item observability and evaluativeness, whereas variance played a larger role. Facets and even single items detect nuances of personality variation that may merit substantive attention.

Keywords: Corss-rater agreement; cross-informant agreement; cross-observer agreement; consensual validity; specific variance; nuances; bi-factor.

> We should be glad that chemists and physicists are not still trying to account for all matter in terms of, say, ten elements that combine into 100 compounds. If there is an *a priori* reason to believe that personal attributes can be portrayed using a scheme simpler than the periodic table, I am unaware of it (Blaney, 1991, p. 62).

Personality traits are defined as pervasive and enduring patterns of thoughts, feelings, and behaviors and are typically assessed by self-reports or by informant ratings. When a single assessment method is employed, important features of traits such as structure, stability, and correlates can be examined, but to an unknown degree, the results may reflect the structure, stability, or correlates of method artifacts. For that reason, cross-rater agreement on personality trait ratings has become one of the central topics and tools of personality research (Kandler, Riemann, Spinath, & Angleitner, 2010). For many traits, it is likely that trait variance shared by different raters is a more valid source of information than single rater-based variance (Kolar, Funder, & Colvin, 1996).

Moreover, a substantial level of cross-rater agreement on particular traits is sometimes interpreted as one type of supporting evidence for the reality of traits as underlying psychological attributes (Funder, 1991; McCrae et al., 2004; McCrae & Costa, 2008). Of course, agreement across raters on observed trait scores is not sufficient to establish their reality, because it might arise from false consensus. It does, however, strengthen the case, because the veridicality of traits provides a plausible and parsimonious explanation of agreement. Certainly, without substantial agreement one cannot speak of personality traits as something that exist outside of our imagination and judgment.

Substantial cross-rater agreement has been shown for all Five-Factor Model (FFM) personality domains. For example, the correlations between self-ratings and ratings of knowledgeable informants typically range between 0.39 and 0.62 (Connolly, Kavanagh, & Viswesvaran, 2007; Connelly & Ones, 2010) and are sometimes even higher (Allik, Realo, Mõttus, & Kuppens, 2010). Typically, only somewhat lower agreement has been shown for the facets of the broad FFM domains (McCrae et al., 2004; Mõttus, Allik, & Pullmann, 2007).

Although less extensively documented, there is also some evidence that the specific variance in facets is agreed upon by different raters. For example, facet scores residualized for the five FFM factors show significant cross-rater correlations (McCrae & Costa, 1992; Costa & McCrae, 2008; Kandler et al., 2010). This is taken to show that facets contain valid specific variance over and above the variance that they share with their respective FFM domains. This has important implications as it may suggest that domain (sum) scores reflect multiple underlying mechanisms in addition to any possible common aetiology, and that facets could provide additional predictive variance to the factor scores. The latter has been empirically demonstrated (e.g., Judge, Rodell, Klinger, Simon, & Crawford, 2013).

Personality at the Item Level

Objective personality assessment relies ultimately on responses to individual items, but items have been less widely researched than scales. There is a substantial literature on the desirability of items (e.g., Edwards, 1957), and occasional efforts to describe other item characteristics (Blaney, 1991; Johnson, 2004). Angleitner, John, and Löhr (1986), for example, looked at surface characteristics, including item length, grammatical complexity, and format, and reported that these characteristics predicted item stability and internal validity. Likewise, Mõttus, Pullmann, and Allik (2006) found that scales with shorter items tended to have higher internal consistencies.

However, personality assessors have tended to assume that the *content* of items is nothing more than the trait they are intended to assess. From the classical trait perspective, the personality-related characteristics reflected in different items in a scale are essentially exchangeable (Bollen & Lennox, 1991); items only reflect the traits of which they are indicators (and error). In practice, trait researchers have long known that assessment is improved (at the cost of internal consistency) by diversifying item content (Cattell, 1973), but this diversity is seldom considered as a source of variance of interest in its own right. But if, as Blaney (1991) surmised, human personality is as complex as organic chemistry, then personality items might express real but subtle distinctions within the construct assessed by the scale as a whole. McCrae (in press) has argued that items correspond to a third level of the trait hierarchy—*nuances*—located beneath domains and facets. Tension, for example, might be a nuance of the anxiety facet of Neuroticism.

If there is substantive content in individual items beyond that of the superordinate trait they assess, it should appear as cross-rater agreement (i.e., consensual validity) on single items, over and above agreement on the variance of their respective FFM domains and facets. If the residual variance is agreed upon by independent raters, it probably reflects something veridical about the person; and if there is something veridical about the person, it may prove to be consequential or otherwise interesting.

Such evidence would suggest that items reflect substantively meaningful narrow personality characteristics with their own underlying mechanisms and predictive utility, at least partially independent from broad domains and facets. If so, items could be addressed accordingly in personality research as a potential source of incremental validity. This possibility is consistent with the observations that items of the same scale sometimes demonstrate different age-group differences (Lucas & Donnellan, 2009), and that individual items may predict external criteria over and above the scores of the scales they belong to. For example, it was reported that the correlation of the Impulsiveness facet of Neuroticism with being overweight was entirely driven by its two items that referred to overeating (Terracciano et al., 2009). It is not known whether items with valid specific variance are exceptional, or whether most items in personality inventories share this property, because item-level associations have been rarely reported in the literature.

In the current study, data from a large sample of Estonians were used to decompose cross-rater agreement into three independent components: agreement on (a) the broad FFM domains, (b) the residual variance of their facets and (c) the residual variance of single items. The major purpose thus was to establish the degree of unique consensual validity at three levels of personality variance. At the item level, this constitutes a test of the hypothesis that items embody distinct nuances of personality.

Moderators of Cross-Observer Agreement

This study also gives an opportunity to explore possible determinants of variation among traits at each level in the magnitude of cross-observer agreement. There is a substantial literature on this topic with regard to the five factors. John and Robins (1993) reported that agreement was highest for traits related to Extraversion and lowest for those related to Agreeableness. Across domains, agreement was higher for easily observable traits, and for evaluatively neutral traits, especially when (as in the present study) agreement was assessed by self/observer correlations. Funder and Dobroth (1987) also reported effects of evaluation and visibility, and concluded that agreement was higher for Extraversion than for Neuroticism because manifestations of the former were more easily observed. More recent studies (Edmonds, Goldberg, Hampson, & Barckley, 2013; Vazire, 2010) have generally supported these conclusions, while showing the influence of different kinds of raters on agreement. To date, however, no studies have examined differential cross-observer agreement for the specific variance in facets or items.

At least two other variables are likely to moderate the degree of agreement. Scales, and particularly items, differ in variance, and agreement is likely to be reduced for traits with a circumscribed variance (Allik, Realo, Mõttus, Esko, et al., 2010). Finally, items vary in the degree to which they are clearly understood by different respondents; long and difficult items may cause problems for some respondents and reduce validity (Angleitner et al., 1986). This is a particularly important consideration in samples (as in the present study) where participants represent a wide educational range. Although the NEO-PI-3 was designed to minimize such problems, there is still a range of variation in reading level across items, and it may affect the magnitude of cross-observer agreement. There are no reading level estimation procedures for Estonian, but item length can serve as a proxy measure for item difficulty.

Therefore, as a corollary to the main focus on the consensual validity of specific variance, the present study will investigate the possible moderating effects of item observability, evaluativeness, length, and variance on cross-informant agreement.

Method

Participants

Participants came from the Estonian Genome Centre (EGC) of the University of Tartu. The EGC is a large database of health, genealogical, and genome data that aims to cover 5% of Estonia’s population (for details see www.biobank.ee). The current EGC cohort includes over 51,000 people and roughly reflects the age, sex, and educational distribution of the Estonian adult population (Letisalu et al., 2014). A subset of the EGC cohort was asked to complete a self-report personality questionnaire (self-ratings). In addition, these participants identified a person who knew them well and and asked this person to complete the same questionnaire about them (informant-ratings). The informants were spouses or partners (51.5%), friends (15.1%), parents (14.6%), (grand)children (6.1%), siblings (6.3%), other relatives (3.2%) or acquaintants (3.2%) of the participants.

After excluding protocols with missing responses in the personality questionnaire, we used data for 2,658 participants (mean age 45.86, standard deviation 17.28, range 18 to 91, for 31 people age was unknown; 1,453 women, for 3 people sex unknown). Of these, 8.0% had basic, 24.5% secondary, 27.5% vocational, and 40.0% higher education. Of the raters, 72.2% were women and their mean age was 42.24 years (standard deviation 16.14, range 11 to 90).

Measures

Personality traits were measured with the Estonian version of the NEO Personality Inventory-3 (NEO-PI-3; De Fruyt, De Bolle, McCrae, Terracciano, & Costa, 2009; McCrae & Costa, 2010). This is a slightly modified and more readable version of the Revised NEO Personality Inventory

(NEO-PI-R; Costa & McCrae, 1992). The NEO-PI-3 has 240 items that measure 30 personality facets which are grouped into the five FFM domains, such that each domain score is a composite of six facet scores. Participants respond on a 5-point Likert scale ranging from *completely disagree* to *completely agree*.

Statistical analyses

Correlations. First, cross-rater (Pearson) correlations between sum-scores of facets and domains were calculated. Second, each facet was residualized for its respective domain score, and cross-rater correlations between the residuals were calculated. Residualization involved calculating residuals from linear regression models, whereby score of each facet scores were predicted by its domain scores. Third, each facet was residualized for all five FFM domains, as was done in Costa and McCrae (2008), and cross-rater correlations between these residuals were obtained again. Finally, cross-rater correlations were calculated for raw item scores and for item scores residualized for the sum-scores of their respective facet. These analyses have a straightforward interpretation: Correlations between residual facet scores address the question of whether facets have any valid specific variance remaining after information from the domains has been taken into account; correlations between residual item scores address the question of whether items have any consensually validated variance after information from their facet scale has been removed.

SEM analyses. Structural equation modelling (SEM) was used to decompose the common variance of domains (shared variance across all items of a domain), facets (shared variance of all items measuring a given facet independently of what they shared with the common variance of the domain) and items (the variance in items not shared with other items of the domain and facet). This procedure allowed for the common and specific variance to be simultaneously separated at all three levels of analysis—something that could not be done with sum-scores, which incorporate both common and specific variance of their constituents.

For each domain, the following bi-factor-type model was fitted (Figure 1). All 48 self-report items defined the self-report score and 48 informant-report items defined the informant-report score of the respective FFM domain; the correlations between these latent traits were taken as the estimates of error-free agreement on the variables underlying the FFM domains. In both types of ratings, all 8 items of each facet also defined the respective facet score. Importantly, the correlations between the facet scores and their respective domain scores were set to zero, resulting in orthogonal domain and facet scores (facets were allowed to correlate among themselves, to allow for lack of local independence among facets; we acknowledge that ultimately any such lack of local independence would have pointed to the need to explicitly model additional tiers in trait hierarchy, but due to computational as well as conceptual complexity we did not specify additional traits between domains and facets). The correlations between the self- and informant-report facet scores were taken as estimates of error-free cross-rater agreement on the specific variance of facets. In each model, corresponding items from self- and informant-reports were also allowed to have residual correlations; these were taken as estimates of cross-rater agreement on the specific variance of items (i.e., the item variance not accounted for by the shared variance of all items of the domain as well as the shared variance of the respective facet).

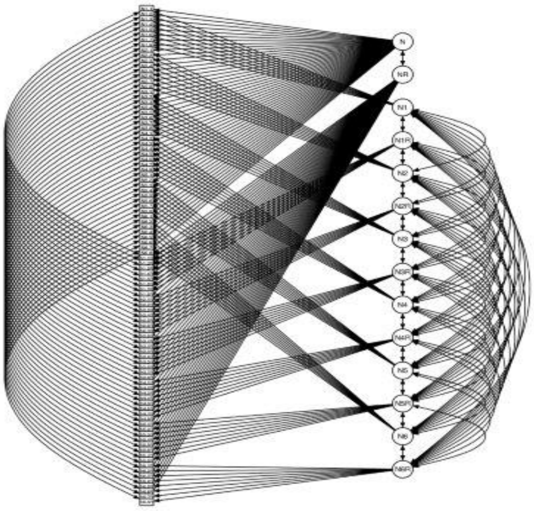

Figure 1. The baseline model (before any tweaks) for Neuroticism. Equivalent baseline models were specified for other FFM traits. “R” at the end of item or trait symbol indicates informant-ratings. Full trait names can be seen in Table 1.

As indices of model fit, we relied on Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA). For the former, values at least 0.95 are generally desirable, whereas for the latter values below 0.06 are sometimes considered as indicating good model fit (Hu & Bentler, 1999). However, as it has been shown that CFI tends to decrease in models with a large number of variables even when the model is correctly specified (Kenny & McCoach, 2003), we relied on a more lenient CFI cut-off (CFI ≥ 0.90) in this instance (each of our models featured 14 latent traits and 192 factor loadings alone).

When very complex models are tested, it is common to find occasional poor fits. When this occurs, two strategies can be followed: One can ignore the fit indices and calculate results using the a priori model, or one can successively modify (“tweak”) the model until acceptable fit is reached, and evaluate the corrected model. We used both approaches, with very similar results, which demonstrated that the results were robust, and not due to excessive tweaking. We will present the corrected model results, and then note at the end the occasional instances in which somewhat different results were obtained when the a priori model was analyzed.

Therefore, when the initial model did not fit data according to our criteria, it was tweaked by first iteratively omitting all non-significant ($p < 0.001$) loadings (starting from the weakest loading) and then by allowing for as many correlated residuals or secondary loadings as was necessary for the models to fit data sufficiently well (the only restriction was that the domains were not allowed to correlate with their facets, as this was a central requirement for decomposing their variance). Correlated residuals and secondary loadings were selected based on modification indices (MI). The tweaking process was iterative in that always only the parameter with the highest MI was set free and the model was then re-fitted to find the next highest MI, until the condition of CFI ≥ 0.90 and RMSEA ≤ 0.06 was met. In the process of model tweaking, no equivalence across self- and informant-ratings was assumed. Maximum Likelihood estimation was used. The models were fitted using 'lavaan' (Rosseel, 2012). For a reader uncomfortable with such post-hoc tweaking, we note that commonly used exploratory factor analysis and factor rotation procedures, for example, are also *prima facie* examples of allowing model parameters to choose their values such that a desired outcome is achieved (e.g., the sum of the variances of the squared loadings is maximized); such procedures simply carry out the orgy of model tweaking behind the scenes.

Finally, differences between facets and their domains in cross-rater correlation were tested for significance by comparing models where the respective estimates were constrained equal to those where the estimates were free (by means of chi-square difference test).

Moderators of Agreement

Nine judges rated the items in terms of their observability and social desirability: four female psychologist (one of them the fourth author of the study), two male psychologist (one of them the first author of the study), and three females with no expertise in psychology (the pattern of results was very similar when the ratings from the first and fourth author of this study were excluded). All were native Estonian speakers.

For observability, the following instruction was used, based on John & Robins (1993): “Some aspects of personality are easy to judge by external observer, whereas some aspects may be judgeable only by people themselves. For each item, please indicate how easy it would be for an external observer to decide if it descibes the person being rated.” The items were rated on a 7-point scale (*very difficult* to *very easy*). The mean of the ratings was calculated for each item. For desirability, the following instruction was used: “The descriptive characteristics of people often contain an evaluative component. Some characteristics are considered very important for gaining social approval, whereas other characteristics are not approved at all. For each item, please indicate how helpful agreeing with it would be for gaining others' approval." The ratings were provided on a 7-point scale (*not helpful at all* to *very helpful*). Evaluativeness of each item was operationalized as the absolute value of its mean desirability rating from scale midpoint (3.5).

Item length was the number of words in each item. Item variance was calculated using both self- and informant-report data; for each item the average of these was used. For facet scales, observability, evaluativeness, and length were the mean of the eight items in the facet; variance was calculated for the facet scale using both self- and informant-report data, taking the average of these.

Results

Means, standard deviations and inter-correlations of the personality scale scores are given in the Supplementary Material 1.

Correlations

Results for domains and facets are reported in the first three data columns of Table 1. For the five domains, the sum-score based cross-rater correlations varied from 0.48 to 0.66 with a mean of 0.56. For the facets, the sum-score based correlations ranged from 0.39 to 0.63 with a mean of 0.49 and standard deviation of 0.07. When facets were residualized for their own domain score, their cross-rater correlations ranged from 0.29 to 0.55, with a mean of 0.40 and a standard deviation of 0.07. When facets were residualized for all five domains, cross-rater correlations were slightly lower, ranging from 0.24 to 0.49, with a mean of 0.34 and a standard deviation of 0.07; these estimates are similar to those of Costa and McCrae (2008), who reported a median of 0.33 (range = 0.18 to 0.58) for cross-rater correlations of residualized facets. Residualizing facets for all domains probably attenuated cross-rater correlations because facets tended to be correlated with multiple domains (commonly referred to as cross-loadings). In either way, residualizing facets for the FFM domains attenuated cross-rater agreement by less than one-third, on average. Thus, facets show substantial consensually-validated specific variance.

Table 1. *Cross-rater agreement.*

	Correlations			
		Residuals controlling for:		
	Raw scores	Own domain	All domains	SEM estimates
Neuroticism	.53			.54
Extraversion	.66			.65
Openness	.62			.36
Agreeableness	.48			.38
Conscientiousness	.53			.50
N1: Anxiety	.51	.37	.30	.55
N2: Hostility	.50	.41	.33	.42[a]
N3: Depression	.50	.36	.32	.55
N4: Self-Consciousness	.44	.38	.31	.53
N5: Impulsiveness	.42	.41	.29	.52
N6: Vulnerability	.44	.29	.24	.51
E1: Warmth	.51	.44	.34	.40[a]
E2: Gregariousness	.60	.45	.43	.65
E3: Assertiveness	.59	.51	.47	.66

E4: Activity	.60	.48	.43	.65
E5: Excitement Seeking	.63	.55	.49	.69
E6: Positive Emotion	.56	.42	.38	.62
O1: Fantasy	.46	.32	.30	.53[a]
O2: Aesthetics	.60	.55	.49	.71[a]
O3: Feelings	.47	.39	.27	.59[a]
O4: Actions	.51	.39	.33	.60[a]
O5: Ideas	.55	.45	.39	.65[a]
O6: Values	.39	.30	.29	.59[a]
A1: Trust	.41	.38	.29	.49[a]
A2: Straightforwardness	.39	.35	.31	.44
A3: Altruism	.42	.34	.29	.49[a]
A4: Compliance	.46	.37	.35	.55[a]
A5: Modesty	.39	.37	.29	.44
A6: Tender-mindedness	.43	.36	.34	.53[a]
C1: Competence	.41	.37	.26	.50
C2: Order	.60	.50	.47	.73[a]
C3: Dutifulness	.44	.35	.27	.51
C4: Achievement Striving	.49	.46	.41	.60[a]

C5: Self-Discipline	.46	.32	.32	.51
C6: Deliberation	.44	.39	.32	.46

Note: SEM estimates = standardized estimates from models decomposing the variances of domains, facets and items. [a]The difference in cross-rater correlations between the facet and its respective domain is significant at $p < 0.002$ (Bonferroni correction: 0.05 / 30).

Before residualizing, cross-rater correlations of individual items ranged from 0.13 to 0.56 with a mean of 0.31 and a standard deviation of 0.08. Cross-rater correlations for the residualized item scores varied from 0.06 to 0.47 with a mean of 0.19 and a standard deviations of 0.07; all correlations were significant at $p < 0.001$. (Almost identical results were obtained when items were corrected for both the facet and the domain they measured: the mean correlation was then 0.18 and the range from 0.05 to 0.47.) Therefore, on average less than half of cross-rater agreement on single items could be accounted for by agreement on the construct purportedly assessed by the item's facet. Although the amount of specific variance in individual items varied, every item showed significant agreement: Item-level specific variance appears to be ubiquitous.

SEM Analyses

None of the bi-factor-type models decomposing cross-rater agreement on the FFM domains, their facets, and items fit data according to our threshold for CFI, but all models met the RMSEA criterion (≤ 0.06). The CFIs were .85 (Neuroticism), 0.85 (Extraversion), 0.84 (Openness), 0.85 (Agreeableness), and 0.84 (Conscientiousness). The models were tweaked until they met the fit criteria, such that a number of non-significant loadings were omitted (loadings of 16, 10, 24, 22, and 15 items, respectively for Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness) and residual correlations (10, 24, 10, 13, and 11, respectively) and secondary loadings (25, 12, 11, 20, and 22, respectively) were allowed. Outprints for all models are given in the Supplementary Material 2 such that all estimated model parameters can be seen and omitted parameters worked out.

The results pertaining to traits are reported in the last column of Table 1. Cross-rater agreement estimates of the residual variances of facets were often of the same magnitude or even larger than the estimates of the FFM domains. Specifically, the estimates for the five domains ranged from 0.36 to 0.65 with a mean of 0.49, whereas the estimates for the residual variance of the 30 facets ranged from 0.40 to 0.73 with a mean of 0.56 and a standard deviation of 0.09. The estimates for the specific variances of 12 facets were significantly (Bonferroni-corrected $p < 0.002$) higher than the estimates for their domain scores (all facets of Openness, 4 facets of Agreeableness and 2 facets of Conscientiousness); the reverse was true for two facets, N2: Hostility and E1: Warmth.

Two findings deserve highlighting. First, cross-rater agreement for Neuroticism, Extraversion, and Conscientiousness was similar whether based on raw scores (that simply summed up items and thus included both shared and specific variance) or latent trait scores (that estimated the error-free variance common to all items of the domain). However, for Openness and Agreeableness the latter type of agreement was lower than the former. This indicates that in these two domains raters agreed relatively more on the specific variance of facets or items than on their shared variance. Second, the specific variances of facets demonstrated higher cross-rater agreement in SEM models as opposed to the analyses based on sum-scores, regardless of whether the sum-score based residuals controlled for only the respective domain of the facet or for all five domains. This may be because the SEM models residualized facets only for the *common variance* of all facets of the same domain, whereas the sum-score residualizing involved both common and the unique variance of each facet. SEM models can yield higher estimates because they adjust associations for measurement error, although this did not happen for the Neuroticism, Extraversion, and Conscientiousness domains.

In the SEM models, cross-rater correlations of the residual variance of individual items ranged from -0.14 to 0.49 with a mean of 0.15 and a standard deviation of 0.08. Therefore, on average about half of cross-rater agreement on single items (*M* = 0.31) could be accounted for by the agreement on the common variance of their domain and the specific variance of the facets of that domain. Most (87.5%) of the residual correlations were statistically significant (Bonferroni-corrected $p < 0.002$); 182 (75.8%) were higher than 0.10, 46 (19.2%) were higher than 0.20, and 14 (5.8%) were higher than 0.30.

Of the 14 items with residual cross-rater correlations higher than 0.30, three were from the E5: Excitement Seeking facet and referred to listening to loud music, watching scary movies and taking part of crowded events, three were from the O2: Aesthetics facet and referred to enjoying visual art, music and ballet, and two were from C3: Dutifulness and referred to paying debts promptly and going to work even when unwell. Two other Openness items referred to trying unusual foods and to the importance of religion, two Extraversion items referred to spending holidays in crowded places and being talkative, and two Conscientiousness items referred to being work-oriented and planning trips carefully. The item with negative residual correlation was from the N5: Impulsiveness facet and referred to controlling emotions. Among the 30 items with non-significant ($p \geq 0.002$) residual cross-rater correlations, three were from both A1: Trust and C6: Deliberation, whereas other facets were represented with fewer items.

It must be noted that neither of the procedures adjusted item residuals for measurement error, which means that cross-rater correlations of item residuals were always attenuated by measurement error. Likewise, some of the item residual correlations could be attenuated by restriction of variance, because item responses often had very skewed distributions. Correlations between the standard deviations of items in self- and informant-ratings and cross-rater agreement on these items (unresidualized or residualized) were all above 0.30 ($p < 0.001$; cf. Allik, Realo, Mõttus, Esko, et al., 2010; see below). Overall, this is likely to mean that cross-rater correlations on the personality characteristics reflected in the residual variances of single items were underestimates.

Of additional note is that the results would have been relatively similar for poorly fitting SEM models (i.e., models without any tweaks), the only notable difference being for E4: Activity, for which the unmodified model would have resulted in cross-informant correlation of 0.40 as opposed to 0.52 in the tweaked model. For the facets, the average cross-informant correlation of residual variances would have been 0.55 based on unmodified models as opposed to 0.56 for the modified models. For the residual variances of items, the average cross-informant correlations would have been 0.17 based on unmodified modes as opposed to 0.15 for the tweaked models. For the five domains, the differences would have been no greater than |0.01|. Hence, the model tweaking had effectively no impact on the main conclusions.

Moderator analyses

We first examined agreement as a function of trait content. Regardless of whether the correlations were based on raw scores or residualized scores, the facets ranked relatively similarly in terms of cross-rater agreement. For example, the agreement estimates based on raw facet scores of the 30 facets correlated at r = .73 with those based on residualized scores (from SEM models). This associations, depicted in Figure 2, shows that Extraversion facets (with the exception of E1: Warmth) as well as O5: Ideas and C2: Order tended to be the most agreed upon, whereas several Agreeableness (e.g., A2: Straightforwardness and A5: Modesty), Conscientiousness (e.g., C1: Competence and C6: Deliberation) and Neuroticism facets (e.g, N4: Self-Consciousness, N5: Impulsiveness and N6: Vulnerability) demonstrated lower agreement. The finding that highest cross-rater agreement tend to be found in the Extraversion domain and the lowest agreement in the Agreeableness domain is consistent with previous research (e.g., Connolly et al., 2007; John & Robins, 1993; McCrae et al., 2004).

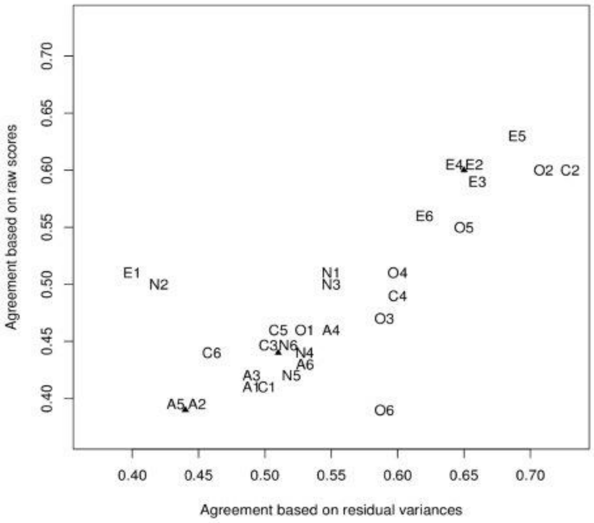

Figure 2. Cross-rater agreement based on raw facets scores and residualized facet scores (from the SEM models). For full facet names see Table 1.

Next, we considered item characteristics. Across the 240 items, average cross-rater agreement (zero-order correlation) on ratings of observability and evaluativeness were .37 and .62, respectively (Cronbach's alphas = .84 and .93). Observability and evaluativeness correlated at r = .29 (p <.001). As Table 2 shows, cross-rater agreement at the level of single items had a moderate positive correlation with observability, a moderate negative correlation with evaluativeness and a fairly strong positive correlation with variance. The associations tended to become weaker when cross-rater agreement on residual variances was concerned; the attenuations were statistically significant in all cases (p < .05). When agreement on raw items scores was simultaneously predicted by observability, evaluativeness and variance in a multiple regression model, the standardized effect sizes were .21, -.21, and .52 (p < .05 for all). For cross-rater agreement on both types of item residuals, only item variance remained a significant predictor in the multiple regression model, whereas other predictors had effect sizes below .10 in all instances. At the level of the facets, a similar pattern was seen, except that correlations with agreement on residualized scores were not significantly lower than correlations with agreement on raw scores.

Table 2. *Correlations of item characteristics with cross-rater agreement.*

	Observability	Evaluativeness	Word count	Variance
		Items (*N* = 240)		
Raw scores	.27***	-.32**	-.02	.61***
Regression-based residuals	.15*	-.17**	.02	.48***
SEM-based residuals	.12	-.09	.05	.37***
		Facets (*N* = 30)		
Raw scores	.49***	-.50**	-.29	.72***
Regression-based residuals	.39*	-.52**	-.10	.70***
SEM-based residuals	.12	-.68***	-.09	.45*

Note: Regression-based residuals refer to items scores being residualized for the respective facet of the item or to the facets being residualized for all FFM domains using multiple regression models. SEM-based residuals refer to cross-rater correlations for unique variances at the respective levels of the bi-factor models. In facet-level correlations, observability, evaluativeness and word count are averages of the items of the respective facets, whereas the variance is that of facet scores.

It thus appears possible that raters agree slightly more highly on more observable and less evaluative items, whereas increased variance plays a more major role. The moderating effects of observability and evaluativeness may be less important for residual variances in items, suggesting that observability and evaluativeness pertain more to the shared (trait-related) than unique variance of specific personality characteristics. However, it is also the case that residual item contain error, which probably attenuates associations with moderators.

Discussion

First, the findings add to the existing literature on cross-rater agreement on the FFM domains. In a large sample of Estonians, there appeared a fairly typical, if somewhat higher, level of cross-rater agreement for both broad FFM domains and their facets (Connolly et al., 2007; McCrae et al., 2004; Connelly & Ones, 2010). Second, the findings add to the scarce literature on cross-rater agreement on the *specific* variance of the FFM facets and show that raters agree to a substantial extent on this part of variance in facet scores that is not accounted for the by the FFM domains. Finally, the results uniquely demonstrate that there tends to be variance in individual items that is not accounted for by either facets or domains but that is nevertheless significantly agreed upon by different raters.

This study appears to be the first in which traits were also modelled as latent variables that reflected the common variance of all items designed to measure the purported traits. Correlations between latent traits are adjusted for measurement error, which is akin to the familiar practice of disattenuating correlations for unreliability. The SEM-based estimates of agreement at the domain level were comparable in magnitude to the typical findings based on sum-scores (Table 1; Connolly et al., 2007; McCrae et al., 2004) for Neuroticism, Extraversion and Conscientiousness, but were somewhat lower for Openness and Agreeableness. People apparently agreed more on the specific variance of the facets of these two domains than on the variance common to all of their manifestations.

Both correlational and SEM analyses concur in showing that NEO Inventory facets contain substantial consensually-validated variance that is not accounted for by their membership in a broader domain. When sum-scores of facets were residualized for the sum-scores of their respective domains as well as other domains, they still showed significant and substantial cross-rater correlations; residualizing facets for the FFM domains attenuated cross-rater agreement by less than third, on average. As for the results based on SEM latent traits, in many cases the estimates of cross-rater agreement tended to be even significantly higher for the specific variance of facets compared to their common variance (i.e., the latent domain scores). To the extent that such cross-trait level comparisons are valid, these findings may strengthen the claim that facet-level traits yield potentially meaningful information about human personality differences over and above the domains they are subsumed under. This is consistent with findings suggesting, for example, different developmental trajectories (Soto & John, 2012) and predictive validities (Tett, Steele, & Beauregard, 2003; Judge et al., 2013) for facets of the same domain.

The finding that estimated correlations for specific variance in facets were often as large or larger than the estimated correlations for the five domains (especially for the Openness and Agreeableness domains) suggests that in some sense observers are often particularly attuned to narrow rather than broad traits, perhaps because the former are more concretely manifested in particular behaviors and thereby more visible and diagnostic of personality differences, or because Openness and Agreeableness are relatively less coherent domains to start with. However, it must be recalled that these SEM based results reflect hypotheticals, pertaining to the theoretical status of *latent* traits and their consensual validity. In the context of real-world assessment that is typically based on sum-scores and neither decomposes variance into common and specific proportions nor adjusts for measurement error, it remains the case that agreement is generally higher on broad domains than on narrow facets, largely because domains (at least in the NEO-PI-3) are assessed by six times as many items as single facets.

Finally, this study was unique in that it also estimated the level of cross-rater agreement on the residual variance of single test items. Most items showed significant cross-rater agreement over and above the variance they shared with other items of the same facet and broad domain; often this level of agreement was rather substantial (e.g., depending on the method of residualizing, about one fifth to nearly one half of residual item correlations were higher than 0.20). Thus, the variance in individual test items that is not shared with other items of the same scale is not merely error variance as would appear from the perspective of classical test theory and related assessment practices. If it were merely error, different raters would not be likely to agree on it.

This finding suggests that single items may be potentially interesting variables in their own right. In scale development, items are normally chosen to represent different manifestations of the trait they assess and they are assumed to be exchangeable (Bollen & Lennox, 1991). However, the differences in nuances of meaning across items of the same trait may correspond to theoretically or practically important ways in which people differ from each other. This possibility is consistent with the findings that items assessing the same trait may show different developmental trajectories (Pullmann, Realo, & Allik, 2009; Lucas & Donnellan, 2009) as well as correlate with external variables over and above trait scores (Terracciano et al., 2009).

Indeed, researchers rarely consider single test items as something worth substantive interest—that is, beyond the role they play in the measurement of purported traits. In discussing the use of scales and subscales, Reise, Moore, and Haviland (2010) wrote that “any two items that are not perfectly correlated potentially have different correlations with external criteria, yet it would be silly to argue that one should investigate external correlates for each item separately” (p. 554). We would agree that in many research applications it is more efficient to focus on domain- and facet-level assessment. But in some cases (e.g., Terracciano et al., 2009), a consideration of individual items may be warranted. As item-level findings have been very rarely reported in the literature, we currently do not know how often considering individual items would confer incremental value over considering traits only; such cases may well appear quite prevalent. The specific characteristics reflected in single items may help to elucidate the personality correlates of external variables or help to understand what drives their associations with broader domains or facets.

Concurring with previous research (e.g., Funder & Dobroth, 1987; John & Robins, 193; Vazire, 2010; Allik et al., 2010), the present study showed that cross-rater agreement may be higher for more visible and less evaluative personality characteristics and for characteristics that have greater variance among people. The finding that the links with observability and evaluativeness were less apparent for item-specific variances suggests that they may pertain more to broader traits than to the unique variance of specific personality characteristics.

Conclusion

At a theoretical level, the present results support the view that the hierarchy of trait-like characteristics in human personality goes deeper than the two layers of domains and facets, at least to a third level of nuances (McCrae, in press). Individual differences in personality traits are as subtle as they are pervasive and enduring.

Authors' Note

The project was supported by the University of Tartu (SP1GVARENG) and research funding from the Estonian Ministry of Science and Education (SF0180029s08 and IUT2-13). Authors are grateful to Andres Metspalu and Tõnu Esko for access to the data. Robert R. McCrae receives royalties from the NEO Inventories.

References

Allik, J., Realo, A., Mõttus, R., Esko, T., Pullat, J., & Metspalu, A. (2010). Variance determines self-observer agreement on the Big Five personality traits. *Journal of Research in Personality*, *44*, 421–426. doi:10.1016/j.jrp.2010.04.005

Allik, J., Realo, A., Mõttus, R., & Kuppens, P. (2010). Generalizability of self–other agreement from one personality trait to another. *Personality and Individual Differences, 48*, 128–132. doi:10.1016/j.paid.2009.09.008

Angleitner, A., John, O. P., & Löhr, F.-J. (1986). It's *what* you ask and *how* you ask it: An itemmetric analysis of personality questionnaires. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaires* (pp. 61-107). Berlin: Springer.

Blaney, P. H. (1991). Not personality scales, personality items. In W. M. Grove & D. Cicchetti (Eds.), *Thinking clearly about psychology. Vol. II: Personality and psychopathology* (pp. 54-71). Minneapolis: University of Minnesota Press.

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

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

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

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment, 15,* 110–117. doi:10.1111/j.1468-2389.2007.00371.x

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

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

De Fruyt, F., De Bolle, M., McCrae, R. R., Terracciano, A., & Costa, P. T., Jr. (2009). Assessing the universal structure of personality in early adolescence: The NEO-PI-R and NEO-PI-3 in 24 cultures. *Assessment*, *16*, 301–311. doi:10.1177/1073191109333760

Edmonds, G. W., Goldberg, L. R., Hampson, S. E., & Barckley, M. (2013). Personality stability from childhood to midlife: Relating teachers' assessments in elementary school to observer- and self-ratings 40 years later. *Journal of Research in Personality, 47,* 505-513. doi: 10.1016/j.jrp.2013.05.003

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.

Funder, D. C. (1991). Global traits: A Neo-Allportian approach to personality. *Psychological Science*, *2,* 31–39. doi:10.1111/j.1467-9280.1991.tb00093.x

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52,* 409-418. doi: 10.1037/0022-3514.52.2.409

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. doi:10.1080/10705519909540118

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61,* 521-551. doi: 10.1111/j.1467-6494.1993.tb00781.x

Johnson, J. A. (2004). The impact of item characteristics on item and scale validity. *Multivariate Behavioral Research*, *39,* 273-302.

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *The Journal of* *applied psychology*, *98*, 875–925. doi:10.1037/a0033901

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

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 10*, 333–351. doi:10.1207/S15328007SEM1003_1

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64*, 311–337. doi:10.1111/j.1467-6494.1996.tb00513.x

Leitsalu, L., Haller, T., Esko, T, Tammesoo, M.-L., Alavere, H., Snieder, H., . . . Metspalu, A. (2014). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology*. doi: doi: 10.1093/ije/dyt268

Lucas, R. E., & Donnellan, M. B. (2009). Age differences in personality: Evidence from a nationally representative Australian sample. *Developmental Psychology*, *45*, 1353–1363. doi:10.1037/a0013914

McCrae, R. R. (in press). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review.*

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

McCrae, R. R, Costa, P. T., Jr., 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., & Costa, P. T. (2008). 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., Jr. (2010). *NEO Inventories professional manual*. Odessa, FL: Psychological Assessment Resources.

Mõttus, R., Allik, J., & Pullmann, H. (2007). Does personality vary across ability levels? A study using self and other ratings. *Journal of Research in Personality*, *41*, 155–170.

Mõttus, R., Pullmann, H., & Allik, J. (2006). Toward more readable Big Five personality inventories. *European Journal of Psychological Assessment*, *22*(3), 149–157. doi: 10.1027/1015-5759.22.3.149

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

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, *92*, 544–559. doi:10.1080/00223891.2010.496477

Rosseel, I. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1–36.

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

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

Tett, R. P., Steele, J. R., & Beauregard, R. S. (2003). Broad and narrow measures on both sides of the personality–job performance relationship. *Journal of Organizational Behavior*, *24,* 335–356. doi:10.1002/job.191

Vazire, S. (2010). Who knows what about a person? the Self-Other Knowledge Assymetry (SOKA) model. *Journal of Personality and Social Psychology, 98,* 281-300. doi: 10.1037/a0017908.