Running head: Comparability of self- and informant-ratings

Do self-reports and informant-ratings measure the same personality constructs?

René Mõttus *
Department of Psychology, University of Edinburgh
Institute of Psychology, University of Tartu, Estonia

Jüri Allik
Institute of Psychology, University of Tartu, Estonia
Estonian Academy of Sciences

Anu Realo
Department of Psychology, University of Warwick
Institute of Psychology, University of Tartu, Estonia

* Corresponding author
7 George Square, EH8 9JZ Edinburgh, UK
Phone: +441316503453
E-mail: rene.mottus@ed.ac.uk

Author's Note: Data collection was supported by the University of Tartu (SP1GVARENG) and by institutional research funding (IUT2-13) from the Estonian Ministry of Education and Science.

Accepted for publication in the European Journal of Personality Assessment on the 11th of October 2018.

Abstract

Personality researchers often supplement or substitute self-reports with ratings from knowledgeable informants, at least implicitly assuming that the same constructs are measured regardless of the source of ratings. However, measurement invariance (MI) of personality constructs across these rating types has rarely been empirically tested. Here, this was done for the Five-Factor Model domains and their 30 facets (N = 3,253). All facets and all domains but Agreeableness met the level of invariance (metric MI) required for comparing the relative standings of individuals across self-reports and informant-ratings, which is what researchers mostly do. However, ten facets and the Agreeableness domain scale failed to achieve the level of invariance (scalar MI) recommended when comparing mean scores. In conclusion, self-reports and informant-ratings appear to measure similar constructs for most research purposes.

Keywords: cross-rater agreement; informants; peer-ratings; measurement invariance; equivalence

Do self-reports and informant-ratings measure the same personality constructs?

Self-reports are by far the most widely used method for collecting data on individuals' personality differences. In fact, scores of self-report scales have been the default operationalization of the very concept of personality, with prominent personality models such as the Five-Factor Model (FFM; McCrae & John, 1992) being mostly developed and refined by analyzing the co-variance structures in self-reports. But a substantial body of research has also relied on alternative sources of information, most notably ratings provided by knowledgeable informants of the targets (i.e., people being rated). Often, informants provide ratings using the same questionnaires as their targets *could* use—and often *do* use—for providing their self-ratings.

Supplementing self-reports with informant-ratings has a number of potential benefits (Vazire, 2006). Using both methods simultaneously allows quantifying the extents to which personality trait scores reflect trait-relevant information as opposed to other effects (e.g., McCrae, 2015). Cross-rater agreement has been used to validate traits as capturing something real about individual differences as opposed to judgment biases (McCrae et al., 2004). Employing informant-ratings can enhance the predictive validity of personality trait scores (Kolar, Funder, & Colvin, 1996). Outcomes (i.e., variables outside personality domain that could be influenced by personality) are sometimes simultaneously predicted from both self- and informant-rated personality scores, expecting convergent findings to strengthen the evidence for the associations (e.g., Čukić et al., 2016). Developmental patterns may be cross-validated across self- and informant-ratings (e.g., Mõttus, Briley, Tucker-Drob et al., 2018). Using informant-ratings has also contributed towards elucidating the structure of personality (Tupes & Christal, 1961; Norman, 1963).

Sometimes, informant-ratings may provide a practically more viable of collecting information about people’s personality than their self-reports—or even the only way. For example, by asking college students to rate a person they knew well, McCrae and colleagues (2005) were able to collect

personality ratings of nearly 12,000 people of various ages from 50 cultures; it would have been difficult, if not impossible, to obtain personality scores for such a demographically diverse sample of people based on their self-ratings. Likewise, young children may be unable to provide valid self-ratings, whereas parental ratings can be used as substitutes.

Substituting self-ratings with informant-ratings or using them in parallel—correlating the scores (e.g., McCrae et al., 2004), comparing findings based on the two methods (e.g., Čukić et al., 2016; Mõttus et al., 2018) or aggregating them (Realo et al., 2015)—may at least implicitly rest on the assumption that the scores reflect the same constructs regardless of who provides the ratings. For example, comparing the FFM trait scores based on self- and informant-ratings typically yields correlations somewhere between .40 and .60 (McCrae et al., 2004). Assuming that the degrees to which these correlations differ from unity reflect general method effects and random error (e.g., McCrae, 2015), is based on the premise that the scales measure FFM traits invariantly in both rating types. But imperfect cross-rater agreement may also result from scales reflecting partly different underlying constructs in self-reports and informant-ratings. Also, finding that self- and informant-rated personality traits correlate differently with an outcome (e.g., Čukić et al., 2016) may not constitute a lack of replication, but result from lack of measurement invariance (MI).

Indeed, there are theoretical reasons to consider the possibility of self- and informant-ratings reflecting somewhat different constructs or, at least, reflecting the same constructs somewhat differently (i.e., lack MI). Accurate personality judgments require relevant (behavioral) cues that are available to, and detected and utilized by, the rater (Funder, 1995). However, the trait-relevant cues may be differently available to the self- and even well-acquainted informant such that, for example, the former has more privileged access to thoughts and feelings, whereas the latter is in a better position to judge external (behavioral) trait manifestations (Vazire, 2010). Likewise, the ratings by the self and informant may be subject to different motivational biases such as self-enhancement

(Vazire, 2010). Some personality indicators (e.g., items) may therefore be more accurately reflective of "true" personality in self-ratings and some in informant-ratings. Moreover, personality ratings may contain veridical components other than personality trait variance *per se* such as identity (readily available to the self) and reputation (readily available to informants; McAbee & Connelly, 2016).

When constructs have been measured using identical scales in different groups (here, rating types), MI can be operationalized as the equality of sets of measurement model parameters across these groups (Meredith, 1993). The most basic form of invariance, *configural* MI, is met when the constituents of a measurement scale (items, facets) define (or load on) the same trait in different rating types, whereas a more stringent form of invariance, *weak* or *metric* MI, implies that these loadings are also equal in size. This ensures that the scale constituents contribute to the operationalization of the construct to the same degree and therefore individuals are more likely to be ranked based on the same construct in both rating types. This level of MI is needed for comparing findings based on the two methods (e.g., Mõttus et al., 2018) and it facilitates the interpretation of cross-rater agreement (e.g., McCrae, 2004): granted metric MI, imperfect agreement suggest that raters have different perceptions or knowledge about target's traits *per se* rather than scales measuring different things in different rating types. For comparing mean construct levels across rating types, *strong* or *scalar* MI is required, which is met when the intercepts of the scale constituents are also equal, in addition to loadings. Without scalar MI, mean estimates may correspond to different levels of scale constituents in different rating types and mean differences are influenced not (only) by underlying traits but (also) by their specific constituents. One can also test for the equality of residual variances of scale constituents (*strict* MI), in addition to loadings and intercepts; strict MI suggests that the scales measures the trait with the same level of reliability in both ratings types.

The degree to which personality scale scores based on self-reports and informant-ratings display MI has rarely been explicitly tested. Perhaps the only study to focus on this question (Olino & Klein, 2015), reported configural, metric and scalar invariance across rating types for all four personality constructs considered: Well-Being, Social Closeness, Stress Reaction and Harm Avoidance. However, perhaps atypically to many currently popular personality assessment instruments, the responses were collected using a binary, yes/no, answer format. Here we test for MI between self-reports and informant ratings across a wide range of personality traits from different levels of personality trait hierarchy: the FFM domains and their 30 facets, as measured with the NEO Personality Inventory-3 (NEO-PI-3; McCrae & Costa, 2010), which relies on a five-point Likert scale response format. Evidence for MI across self- and informant-ratings would provide more confidence in research making use of informant-reports or combining these with self-reports, whereas evidence for poor MI would also have implications for interpreting the findings of such research (e.g., imperfect self-informant agreement).

Method

Sample

Participants constituted a subset of the Estonian Biobank cohort study, a volunteer-based sample of the Estonian resident adult population, recruited by general practitioners, hospital staff and using other means (for details see Leitsalu et al., 2014). Each participant signed an informed consent form. This study uses data from 3,253 cohort members (1,917 women; mean age 46.55 years, standard deviation 17.02, range from 18 to 91) for whom both personality self-reports and informant-ratings were available.

Measure

Participants and their knowledgeable informants (typically a spouse/partner, parent/child or friend) completed the Estonian version of the NEO-PI-3 (McCrae & Costa, 2010). The NEO-PI-3 has 240

items that measure 30 personality facets, which are then grouped into the five FFM domains, each including six facets consisting of eight items. The items were answered on a five-point Likert scale (0 = false/strongly disagree to 4 = true/strongly agree). For cross-rater correlations, see Mõttus and colleagues (2014).

Analytic Strategy

We tested for MI in the NEO-PI-3 scales as they are usually scored—each item only contributing to its intended scale—rather than in multi-dimensional factor models that incorporate cross-loadings and correlations among traits. Although MI in the multi-trait factor models may be of psychometric interest, it is of limited practical value, given that such models tend to fit poorly (Hopwood & Donnellan, 2010) and are less frequently used for actually scoring personality. Each trait was modeled separately.

Separately in self-reports and informant-ratings, we started our mean and co-variance structure analyses by constructing a unidimensional confirmatory factor analysis (CFA) model for each facet and FFM domain such that facets (as latent causes) were defined by the eight items intended to measure them and domains (as latent causes) were defined by their respective six facets; for each domain, we also ran supplementary analyses where the latent domain was directly defined by the 48 items intended to measure it (i.e., disregarding facets). Latent trait variances were fixed at unity. Both items and facet scores were specified as continuous variables (with interval scale) and the models were fit using Robust Maximum Likelihood estimator in the *lavaan* (Rosseel, 2012) package of R statistical software (R Development Core Team, 2017). We did not model item responses as ordered-categorical because real-world applications of the scales are mostly based on sum-scores, which assumes items having interval scales, and because for several items some response options had not been used in either rating type, creating estimation problems. As indices of model fit, we relied on Comparative Fit Index (CFI), Root Mean Square Error of Approximation

(RMSEA) and Standardized Root Mean Square Residual (SRMR) . For CFI, values at least 0.95 are generally desirable, whereas for RMSEA/SRMR values below 0.06/.08 are often considered as indicating good model fit (Hu & Bentler, 1999). Where model fit was not satisfactory due to residual correlations among items, item pairs that required residual correlations were identified in self-reports. This process was iterative, always based on the highest modification index, and as many item pairs were allowed to have correlated residuals as was necessary for models fit to become satisfactory. However, we also report the main results based on models *without any* correlated residuals.

Next, the constructed models were fit in collapsed self- and informant-ratings, with results serving as the baseline for subsequently fitted multi-group CFA models (MGCFA), in which self-reports and informant-ratings were specified as separate groups (latent trait variances were set unity in both rating types). In the first set of MGCFA models (for configural MI), no parameter equality constraints were applied, whereas in the second set residual correlations and in the third set (metric MI) also factor loadings were constrained equal. Scales were flagged for lack of configural MI when their models fit data more poorly than the respective baseline models, as indicated by a drop in CFI (ΔCFI) of at least .01, combined with increases in RMSEA and SRMR of at least .015 and .03, respectively (Cheung & Rensvold, 2002; Chen, 2007). Same fit-index-change-based criteria were applied for testing equality of residual correlations and metric MI, comparing each model with one step less constrained model. In the two subsequent sets of models, intercepts (for scalar MI) and then also residuals variances (for strict MI) were constrained equal across groups (in addition to all previous equality constraints), and again each model was compared to the corresponding previous, less constrained model, with $\Delta CFI \geq .01$, $\Delta RMSEA \geq .015$ and $\Delta SRMR \geq .01$ flagging potential

lack of the respective level of MI (Chen, 2007). Stricter ΔSRMR criteria for scalar and strict MI tests as opposed to other MI tests were proposed by Chen (2007) based on a simulation design.[1]

Results

For both self-reports and informant-ratings, the fit indices (chi-square, CFI, RMSEA, SRMR) for models with and without correlated residuals are reported in Electronic Supplementary Material (ESM). Before allowing for correlated residuals, the indices varied widely. For example, CFIs ranged from .54 to .96 (median .89) in self-reports and from .37 to .96 (median .90) in informant-ratings, whereas RMSEAs ranged from .049 to .159 (median .091) in self-reports and from .043 to .204 (median .093) in informant-ratings. According to SRMRs, only N5: Impulsiveness and O2: Openness to Aesthetics failed to achieve acceptable fit. After allowing for correlated residuals, all model fit indices were satisfactory in self-reports, whereas CFIs ranged from .87 to .98 (median .96) and RMSEAs ranged from .039 to .127 (median 0.058) in informant-ratings (all SRMRs < .06). That is, in a few instances residual correlations that had been identified based on self-reports did not sufficiently improve model fit in informant-ratings according to CFIs and RMSEAs. Fit indices of the supplementary domain-models identified based on their 48 items (see ESM) were very poor (e.g., CFIs ranged from .41 to .65 and from .46 to .67, respectively in self-reports and informant-ratings). This was expected because items of the same facets were designed to have residual correlations. We did not allow correlated residuals for these models because hundreds of them would have been required.

Fit indices for the baseline models (fitted in collapsed self- and informant-ratings without specifying groups) are reported in Table 1. Model fit indices for all MGCFA models as well as ΔRMSEAs and ΔSRMRs are reported in the ESM. There was no strong evidence for any facet or

[1] In order to test whether the adopted ΔCFI, ΔRMSEA and ΔSRMR criteria could have been too strict (falsely signaling potential lack of MI), we reran the steps above having randomly reshuffled the group indicator. There should not have appeared any evidence for the lack of MI and, indeed, none appeared.

domain lacking configural MI, except for the E1: Warmth facet and Agreeableness domains showing ΔCFI = .01 and the latter also showing an increase in RMSEA of .022. Constraining residual correlations equal did not cause deterioration in model fit, except for ΔCFI = .01 for O4: Openness to Actions and O6: Openness to Ideas.

Eighteen facets and all domains but Extraversion were flagged for poor metric MI according to the ΔCFI criterion, although the deterioration of model fit was more than minor (i.e., ΔCFI > .01) only for four facets, O6: Openness to Values (ΔCFI = .03), A3: Altruism (ΔCFI = .04), A4: Compliance (ΔCFI = .03), and C1: Competence (ΔCFI = .02), and the Agreeableness domain (ΔCFI = .03). According to ΔRMSEA, only Agreeableness domain showed a minor lack of metric MI (ΔRMSEA = .018), whereas this was the case for seven facets (N2: Hostility, O1: Openness to Fantasy, A3: Altruism, A5: Modesty, C1: Competence, C2: Order, and C3: Dutifulness) and three domains (Openness, Agreeableness, and Conscientiousness) according to ΔSRMR (it only exceeded .03 for A3: Altruism, Agreeableness and Conscientiousness). Thus, except for all criteria converging on poor metric MI for Agreeableness, they varied in flagging other scales for the lack of this level of MI and in most cases fit deteriorations were minor. Therefore, factor loadings could be considered reasonably equivalent in self-reports and informant-ratings for all scales except, perhaps, for Agreeableness.

However, according to the ΔCFI criterion, all 35 scales were flagged for potentially poor scalar MI. Although for five facets the drop in fit was minor (ΔCFI = .01), evidence for poor scalar MI was particularly noteworthy (ΔCFI > .10) for E1: Warmth, O6: Openness to Values and A3: Altruism and also notable (ΔCFI > .05) for N3: Depression, N5: Impulsiveness, O4: Openness to Actions, A4: Compliance, C2: Order, C3: Dutifulness, C6: Deliberation and Agreeableness. According to the ΔSRMR criterion, too, majority of facets and all domains failed to achieve scalar MI, although ΔSRMRs exceeded .01 for no more than nine facets and were above .02 for only E1: Warmth, O6:

Openness to Ideas, and A3: Altruism. According to the ΔRMSEA criterion, ten facets and two domains failed to achieve scalar MI, with ΔRMSEA exceeding .02 for N3: Depression, E1: Warmth, O4: Openness to Actions, O6: Openness to Values, A1: Trust, A3: Altruism, C2: Order, C3: Dutifulness, C6: Deliberation and Agreeableness and being .02 for N5: Impulsiveness and Conscientiousness. All of these ten facets and Agreeableness had also been flagged by other criteria, suggesting potentially poor scalar MI for them.

Seventeen facets were also flagged for potentially poor strict MI according to the ΔCFI criterion, although the evidence was more than marginal (ΔCFI > .01) for only O1: Openness to Fantasy, N4: Self-Consciousness, O3: Openness to Feelings, O6: Openness to Values, A4: Compliance, A6: Tendermindedness and C1: Competence. The ΔSRMR criterion also flagged 12 of these facets for poor strict MI, as well as A3: Altruism and Neuroticism and Openness domains; however, ΔSRMR exceeded .01 only for O1: Openness to Fantasy and A4: Compliance. The ΔRMSEA did not highlight any scale as lacking strict MI. Overall, thus, most if not all scales showed an acceptable level of strict MI, had it not been for problematic scalar MI for many scales.

In the ESM, MGCFA results based on models allowing *no* residual correlations are also reported. In these analyses, according to the ΔCFI criterion N5: Impulsiveness showed clear lack for configural (ΔCFI = .10) and metric MI (ΔCFI = .28) alongside the other scales identified above as having ΔCFI > .01 for metric MI. Similarly to what was reported above, there was also evidence for a general lack of scalar MI (for 24 facets and all domains ΔCFI > .01), coupled with poor strict MI for a number of facets (for ten facets ΔCFI > .01). According to ΔRMSEA, only N5: Impulsiveness failed the configural and metric MI test, whereas only six facets (N3: Depression, E1: Warmth, O6: Openness to Values, A3: Altruism, C3: Dutifulness and C6: Deliberation) failed to meet the required level of scalar MI and all scales met strict MI. According to the ΔSRMR criterion, all scales achieved configural invariance, four facets (N2: Hostility, O1: Openness to Fantasy, A3: Altruism

and C3: Dutifulness) and Agreeableness and Conscientious domains failed to achieve metric MI, only nine facets and no domain met scalar MI (only for three of them, ΔSRMR > .02) and 13 facets and Neuroticism and Openness domain showed some evidence for poor strict MI (mostly minor). In conclusion, when residual correlations were not allowed and models tended to be grossly mis-specified there was generally similar evidence for (occasionally poor) MI.

Finally, we tested for MI in (very poorly fitting and without residual correlations) domain-models specified based on 48 items rather than facet scores (see ESM for fit indices). According to the ΔCFI criterion, Neuroticism, Openness and Agreeableness marginally failed to achieve metric MI (ΔCFI), all five domains failed to achieve scalar MI (ΔCFI .02 to .05) and all domains but Neuroticism marginally failed to achieve strict MI (ΔCFI = .01). According to other criteria, all domains achieved all levels of MI. It therefore seems possible that such grossly mis-specified models are less sensitive to violations of MI. Such models may become less sensitive to further fit deteriorations.

Discussion

Do FFM domain and facet scales—as embodied in the NEO-PI-3, one of the most comprehensive and widely used personality questionnaires—measure the same constructs in self-reports and informant-ratings? There are theoretical reasons to consider the possibility that they may not, because the self and external raters may have access to different information about the target or they may bias their ratings in different ways (e.g., Vazire, 2010; McAbee & Connelly, 2016). However, the presented evidence suggests that self- and informant-ratings generally measure the same constructs—at least, to the extent that it matters in most applications. In most cases, researchers are interested in comparing the relative standings of individuals based on self-reports and informant ratings, be this by directly correlating the scores based on the two rating types (McCrae et al., 2004) or by comparing some findings based on them (e.g., Čukić et al., 2016; Mõttus et al., 2018). For

such attempts, similarity of factor loadings (metric MI) suffices, and we found this to apply to most scales. Even for the scales that did not meet the metric MI threshold, the discrepancies in loadings appeared minor, being consistently pointed out by different criteria only for the Agreeableness domain scale.

However, as has been noted in other instances (e.g., Mõttus et al., 2015), mean score comparisons of FFM traits and/or their facets may be complicated due to poor scalar MI. When this level of MI is not met, observed mean differences are at least partly not driven by the ostensible latent trait that the scale constituents ought to define, but by the constituents themselves. For example, Allik and colleagues (2010) showed that informants generally ascribed people the same level of Agreeableness as the target themselves did, but a closer look at the findings suggested that this trend was opposite for different facets of the Agreeableness domain. Therefore, an unqualified difference between the two rating types in mean Agreeableness domain scores could have been misleading. Moreover, given our evidence that several facets themselves (especially O6: Openness to Values) lack scalar MI themselves, the patterns reported by Allik and colleagues (2010) could have been even more nuanced than the authors considered at the time—driven by some specific items of the facets. Lack of scalar invariance may also complicate research based on differences between self- and informant-reported personality scores (e.g., McCrae, Mõttus, Hřebíčková, Realo, & Allik, in press).

What is common to the scales that showed evidence for potentially poor MI: Depression, Impulsiveness, Warmth, Openness to Actions and Values, Trust, Altruism, Order, Dutifulness, Deliberation and Agreeableness? Vazire (2010) suggests that self-informant discrepancy in ratings may be moderated by traits' evaluativeness, so one could expect that these scales have more evaluative content than others (see also Mõttus et al., 2014). Indeed, Agreeableness is among the most evaluative domains (alongside Conscientiousness) and Warmth, Altriusm and Dutifulness are

highly evaluative facets (Allik et al., 2010), suggesting that their comparatively poor MI may result from different motivational biases in self- and informant-ratings. The other scales do not stand out as being among the most evaluative (Allik et al., 2010), but they may contain items that vary greatly in evaluativeness. Items of these scales may also vary in their level of observability, which might moderate cross-rater consistency (e.g., Mõttus et al., 2014). General rating biases such as acquiescence and extreme responding are less plausible explanations for why some scales show poorer MI than others.

In conclusion, there appeared an acceptable level of invariance in how the FFM domains and facets were measured in self-reports and informant-ratings, given how data from these different sources are typically used. Informant-ratings provide a valuable complementary, and sometimes alternative, source of information and lack of MI with self-ratings is not something that should prevent personality researchers from relying on this information. s

References

Allik, J., Realo, A., Mõttus, R., Borkenau, P., Kuppens, P., & Hrebícková, M. (2010). How People See Others Is Different From How People See Themselves: A Replicable Pattern Across Cultures. *Journal of Personality and Social Psychology*, *99*, 870–882.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling*, *14*, 464–504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255.

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

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

Hopwood, C. J., & Donnellan, M. B. (2010). How Should the Internal Structure of Personality Inventories Be Evaluated? *Personality and Social Psychology Review*, *14*, 332–346.

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.

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.

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*, *44*, 1137–1147.

McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. *Psychological Review*, *123*, 569–591.

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. (2010). *NEO Inventories professional manual*. Odessa, FL: Psychological Assessment Resources.

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., Mõttus, R., Hřebíčková, Realo, A., & Allik, J. (in press). Source Method Biases as Implicit Personality Theory at the Domain and Facet Levels. *Journal of Personality.*

McCrae, R. R., & Terracciano, A. (2005). Universal Features of Personality Traits From the Observer's Perspective: Data From 50 Cultures. *Journal of Personality and Social Psychology*, *88*, 547–561.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Mõttus, R., Briley, D. A., Zheng, A., Mann, F., Engelhardt, L., Tackett, J., … Tucker-Drob, E. M. (2018). Kids becoming less alike: A behavioral genetic analysis of developmental increases in personality variance from childhood to adolescence. *Journal of Personality and Social Psychology*.

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.

Norman, W. X (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. Journal of Abnormal and Social Psychology, 66, 574-583

Olino, T. M., & Klein, D. N. (2015). Psychometric comparison of self- and informant-reports of personality. *Assessment*, *22*, 655–664.

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

Realo, A., Teras, A., Kööts-Ausmees, L., Esko, T., Metspalu, A., & Allik, J. (2015). The relationship between the Five-Factor Model personality traits and peptic ulcer disease in a large population-based adult sample. *Scandinavian Journal of Psychology*, *56*, 693–699.

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

Tupes, E. C, & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings* (USAF ASD Tech. Rep. No. 61-97). Lackland Air Force Base, TX: U.S. Air Force

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

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*, 281–300.

Table 1. *Baseline model fit indices for FFM domains and their 30 facets.*

	Chi-square	df	CFI	RMSEA	SRMR
N1: Anxiety	380.605	19	.966	.054	0.027
N2: Hostility	354.478	19	.967	.052	0.029
N3: Depression	472.943	19	.945	.061	0.040
N4: Self-Consciousness	392.784	18	.949	.057	0.033
N5: Impulsiveness	475.171	18	.941	.063	0.045
N6: Vulnerability to Stress	511.074	17	.949	.067	0.034
E1: Warmth	366.778	19	.957	.053	0.036
E2: Gregariousness	468.112	19	.963	.060	0.031
E3: Assertiveness	659.791	20	.944	.070	0.036
E4: Activity	637.414	17	.952	.075	0.038
E5: Excitement Seeking	314.738	19	.966	.049	0.029
E6: Positive Emotion	526.318	19	.953	.064	0.036
O1: Openness to Fantasy	694.716	19	.938	.074	0.042
O2: Openness to Aesthetics	520.988	18	.963	.066	0.037
O3: Openness to Feelings	280.304	18	.953	.047	0.029
O4: Openness to Actions	260.174	17	.962	.047	0.028
O5: Openness to Ideas	629.592	18	.948	.072	0.043
O6: Openness to Values	295.797	16	.915	.052	0.037
A1: Trust	317.739	18	.972	.051	0.033
A2: Straightforwardness	669.332	19	.938	.073	0.049
A3: Altruism	407.764	17	.942	.060	0.042
A4: Compliance	282.626	18	.946	.048	0.031
A5: Modesty	356.590	15	.971	.059	0.032
A6: Tendermindedness	205.020	19	.966	.039	0.027
C1: Competence	308.019	18	.960	.050	0.030
C2: Order	452.875	18	.959	.061	0.034
C3: Dutifulness	201.728	19	.968	.039	0.023
C4: Achievement Striving	411.876	18	.960	.058	0.035
C5: Self-Discipline	402.697	17	.962	.059	0.033
C6: Deliberation	452.152	18	.953	.061	0.038
Neuroticism	445.471	7	.970	.098	0.033
Extaversion	266.557	7	.979	.076	0.024
Openness to Experience	135.258	5	.982	.063	0.024
Agreeableness	75.902	6	.991	.042	0.016
Conscientiousness	192.632	6	.987	.069	0.022

NOTE: df = degrees of freedom; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.