Published in final edited form as:

J Pers Assess. 2008 March ; 90(2): 105–109. doi:10.1080/00223890701845104.

A Note on Some Measures of Profile Agreement

Robert R. McCrae

Gerontology Research Center, National Institute on Aging

Abstract

Profile similarity or agreement is increasingly used in personality research and clinical practice and has potential applications in many other fields of psychology. I compared 4 measures of profile agreement—the Pearson r , Cattell's (1949) r_p , McCrae's (1993) r_{pa} , and an intraclass correlation coefficient (double entry), ICC_{DE} —using both broad factor and specific facet profiles. Matched versus mismatched self-ratings/other ratings on the NEO Personality Inventory–3 (McCrae, Costa, & Martin, 2005) were used as criteria. At the factor level, r_{pa} and ICC_{DE} were comparable, and both were superior to r_p in distinguishing matched versus mismatched profiles. At the facet level, ICC_{DE} was superior to the other coefficients. The Pearson r performed better than expected.

Measures of profile agreement quantify the extent to which two profiles match across a range of characteristics. The profiles to be compared might be a set of ability measures administered to one individual on two occasions or ratings from two observers on the personality traits of a single target. Measures of agreement are useful in clinical applications in which they can help to evaluate the accuracy of data about an individual (see Costa & Piedmont, 2003), and they have been used to validate aggregate personality profiles of cultures (McCrae, Terracciano, & 79 Members of the Personality Profiles of Cultures Project, 2005) and assess the accuracy of national character stereotypes (Terracciano et al., 2005). Profile agreement between individual scores and fixed prototypes has been proposed as a guide to personality disorder diagnosis (e.g., Lynam & Widiger, 2001) and might have applicability in employee selection. For all these reasons, it would be useful to identify the most effective measure of profile agreement.

A good index of profile agreement would show high values when the two profiles were in fact similar. However, there are several ways to conceive of agreement between profiles, and it is not clear which of these is optimal. To circumvent this problem, I evaluated several indexes using self-reports and observer ratings of personality traits under the assumption that true similarity will be higher when the target for the two profiles is the same individual rather than two randomly paired individuals. Correlations with this matched/mismatched criterion can be used to evaluate the different indexes. Note that this criterion does not prejudge the form of agreement. Either absolute closeness of scores or similarity of profiles shapes might be most effective in distinguishing matched from mismatched profiles, and different indexes place different emphases on these components. The data I examined here

are from the NEO Personality Inventory–3 (NEO–PI–3; McCrae, Costa, & Martin, 2005), but it seems likely that the results would be generalizable to other instruments.

The most obvious candidate for agreement is the Pearson correlation, but it is generally dismissed on the grounds that it is insensitive to differences in profile elevation (Cattell, 1949; Terracciano & McCrae, 2006): Two profiles may have the same shape and thus a high correlation, but one may consist entirely of high and the other of low scores. Cattell's (1949) alternative, r_p , was based on the sum of squares of the distance between corresponding profile elements, using the following formula:

$$r_p = \frac{2k - \sum d^2}{2k + \sum d^2},$$

where k “is the median for χ^2 on a sample size n [profile elements]” (p. 292). McCrae (1993) noted that Cattell's measure is insensitive to the extremeness of the scores, when, conceptually, differences are less important at very high or very low levels. T scores of 44 and 56 differentiate low from high scorers on the NEO–PI–3, but T scores of 66 and 78 are both interpreted as very high. In Cattell's index, these two distances—both 12—contribute equally to the total agreement. McCrae (1993) proposed an index of profile agreement, I_{pa} , that reflects both the distance between the assessments (d) and the extremeness of their mean (M):

$$I_{pa} = \frac{k + 2\sum M^2 - \sum d^2}{(10k)^{1/2}},$$

where k is the number of profile elements. For example, if a self-report yields standardized Neuroticism (N), Extraversion (E), Openness (O), Agreeableness (A), and Conscientiousness (C) factor scores of 1.2, 0.3, –0.5, 0.1, and 0.9, whereas an observer rating of the same person yields corresponding scores of .6, .7, –1.1, –.3, and .7, then

$$\begin{aligned} k &= 5, \\ \sum M^2 &= (0.9^2 + 0.5^2 + (-0.8)^2 + (-0.1)^2 + 0.8^2) = 2.35, \\ \sum d^2 &= (0.6^2 + 0.4^2 + 0.6^2 + 0.4^2 + 0.2^2) = 1.08, \text{ and} \\ I_{pa} &= 1.22. \end{aligned}$$

This variable is quasi-normally distributed, and, when the profile has more than two elements, can be converted to a coefficient of profile agreement, r_{pa} , by the following formula:

$$r_{pa} = \frac{I_{pa}}{[(k-2) + I_{pa}^2]^{1/2}}.$$

for the example given, $r_{pa} = .58$. McCrae (1993) showed that r_{pa} was superior to r_p in assessing agreement across matched versus mismatched observers.

The final index considered here is an intraclass correlation (double entry), ICC_{DE} , calculated as a Pearson correlation using the double entry method (Griffin & Gonzalez, 1995). Consider profiles of N, E, O, A, and C scores (see Table 1). The usual Pearson's r would correlate self-reports with observer ratings across the 5 rows of data. The double-entry ICC_{DE} is based on 10 rows of data: Each element in the paired profile is entered twice but in reversed order across rows; the entry in the first column is entered in the second column on a new row, and the entry that had been in the second column is placed in the first column on that new row. The Pearson correlations between the two columns yield ICC_{DE} . ICC_{DE} is sensitive to differences in profile elevation as well as shape. If one profile consists entirely of high scores (e.g., Table 1, Self-Reports) and one entirely of low scores (Table 1, Observer Ratings), double entry ensures that in half the cases, high will be paired with low; and in the other half, low will be paired with high scores, resulting in a low or negative correlation, even if the shapes of the two profiles are identical. Table 1 provides an illustration. ICC_{DE} is akin to McGraw and Wong's (1996) one-way random effects model for exact agreement, but their formula, based on analysis of variance, gives slightly different values. For the data in Table 1, their $ICC(1) = -.43$.

In applying all these indexes, the metric of the profile elements is crucial. Both r_p and r_{pa} require z scores, standardized across the people in the sample, or using published norms. Pearson r s and ICC_{DE} s can, of course, be calculated on any scores, but use of raw scores will generally inflate correlations and suggest agreement between unrelated pairs of profiles. If, for example, a profile consists of one scale with a group mean of 10, a second with a mean of 50, and a third with a mean of 100, virtually everyone will show a profile of low, middle, and high and thus, there will be high r s and ICC_{DE} s between any two profiles. This is a documented problem with Q-correlations using the California Adult Q-set (Block, 1961) because some items are more desirable and readily endorsed than others (Ozer & Gjerde, 1989). Q-correlations are still meaningful because variations around the baseline indicate relative degree of agreement, even though the baseline is not zero. It would appear, however, that an index of profile agreement ought to have a near-zero baseline value in cases of randomly paired profiles.¹ A solution is to standardize all profile elements. T scores, z scores, or any other metric can be used as long as it is used consistently across all profile elements.

In an evaluation of the classification accuracy of 13 indexes of profile agreement, Carroll and Field (1974) used simulated data with random errors and concluded that Cattell's (1949) r_p was equal or superior to the other indexes. McCrae (1993) used a more realistic criterion: He assigned a value of 1 to matched pairs of self-reports and ratings by spouses or peers and a value of 0 to mismatched pairs in which the self-reports were compared to observer ratings of a randomly chosen target. The correlation of this dichotomous criterion with r_{pa} was

¹In the combined samples of mismatched profiles ($N = 1,596$; see Methods), the mean values for the four indexes ranged from $-.14$ to $.10$ for factors and from $-.06$ to $.04$ for facets.

consistently, although not markedly, larger than the correlation with r_p . In this article, I use that strategy to compare r , r_p , r_{pa} , and ICC_{DE} as measures of profile agreement.

Many psychological profiles are hierarchically organized (Gustafsson, 2002). For example, the NEO-PI-3 consists of 30 facet scales that define the broad factors of the Five-factor model (FFM). In principle, if the facets have valid specific variance (McCrae & Costa, 1992), the 30-facet profile should contain more information than the 5-factor profile, and profile agreement based on the facets should be more accurate: Distinctions within a domain (e.g., between levels of E1, Warmth, and E3, Assertiveness) might contribute to more detailed agreement. Conversely, the facet scales are less reliable than the domain and factor scores, which might detract from agreement. The statistical rationales for r_p and r_{pa} were based on an analysis of orthogonal profile elements such as NEO-PI-3 factor scores. However, r_p and r_{pa} can be calculated for the 30 correlated facet scales and their accuracy empirically evaluated. Indeed, McCrae (1993) showed that r_{pa} works relatively well when the Revised NEO-PI (NEO-PI-R; Costa & McCrae, 1992) domain scores (the unweighted sums of relevant facets) are analyzed, despite the fact that there are nontrivial correlations between domains. Thus, r_{pa} is likely to be useful in analyzing a variety of profiles. The other profile agreement indexes I consider here, r and ICC_{DE} , are applicable at either hierarchical level but may function better at one level than the other.

Different indexes may also function best in different profile ranges. In particular, one of the peculiarities of r_{pa} is that agreement is limited when the profile elements are near the mean: Perfect agreement on a profile consisting of five z scores of zero results in an r_{pa} of only .38. However, in clinical applications, extreme profiles are common, so indexes that perform best in that range may have particular value. I therefore examine the performance of these four indexes at different levels of profile elevation.

Method

I took data for this analysis from a study of the NEO-PI-3, a revision of the NEO-PI-R in which 37 of the 240 items were changed to increase readability or internal consistency. The NEO-PI-3 consists of 30 eight-item facet scales that define the factors of the FFM: N, E, O, A, and C. Domain scores are the unweighted sum of the 6 relevant facets; factor scores are calculated from factor scoring weights derived from the NEO-PI-R normative sample (Costa & McCrae, 1992). There are two versions of the NEO-PI-3: a self-report Form S and an observer rating Form R. McCrae, Martin, and Costa (2005) reported evidence for the equivalence of the NEO-PI-3 and NEO-PI-R as well as reliability and validity for the former.

To establish norms for the NEO-PI-3, self-reports were gathered from 635 adults aged 21 to 91 (McCrae, Martin, et al., 2005). For 532 of these individuals, observer ratings were also obtained from friends or relatives who rated each other. Self/other correlations across individuals for the five domain scores ranged from .52 to .65. In this study, factor scores were based on T scores derived from combined sex adult NEO-PI-3 norms; self/other correlations for the five factor scores ranged from .56 to .67. As with the NEO-PI-R, factor

scores showed increased convergent as well as discriminant validity compared to the domain scores. All factors and facets were then *z* standardized for all subsequent analyses.

To create a criterion by which to evaluate the accuracy of the profile agreement indexes, an artificial data set, X, was created. The first 532 cases consisted of self-reports and matched observer ratings from McCrae, Martin, et al. (2005); they were coded “1” on the variable Match. The next 532 cases consisted of the same self-reports but with randomly mismatched observer ratings—that is, ratings of a target different from the self-report; they were coded “0” on Match. Note that the ratings for the mismatched pairs have a distribution that is identical to that of the matched pairs because the same cases are used in a different order. To evaluate the generalizability of findings, I repeated the same procedure with different randomizations to form data sets Y and Z. Cross-observer correlations of the five factors in the 1,596 mismatched cases from data sets X, Y, and Z ranged, as intended, from $-.04$ to $.03$.

In each of the three data sets, the extremeness of the self-report profiles was quantified as the sum of the squares of the five *z*-standardized factor scores. By squaring the scores, both low and high factor scores contributed to a high extremeness score. Cases above the median (*Mdn*) value of 3.83 were considered extreme because they had a combination of high and/or low scores greater than half the sample.

Results

Intercorrelations among the eight profile agreement coefficients ranged from $.39$ (for r_{pa} for factors with r_p for facets) to $.97$ (for r and ICC_{DE} for facets), with *Mdn* = $.72$ in the matched sample ($n = 532$). In the three artificial data sets ($ns = 1,064$) in which the range of the agreement statistics is greater, correlations were even higher, ranging from $.60$ to $.99$, with *Mdn* = $.81$. Given these high correlations, the differences among the indexes are likely to be subtle.

Table 2 reports correlations of the profile agreement indexes with the Match criterion. All the correlations are significant, and most are large by Cohen’s criteria, probably because agreement between couples in the original study was quite high. Coefficients are largest in the subsample with extreme scores and smallest in the subsample with nonextreme scores, which is expectable, at least for r_{pa} . Indexes based on the full 30-facet profile are consistently larger than those based on the five factors, perhaps reflecting the additional information contained in a facet-level profile and perhaps reflecting greater reliability from the sheer number of profile elements.

Of particular interest is the comparison of the four different indexes. The statistical significance of differences between these correlated correlations (Meng, Rosenthal, & Rubin, 1992) are reported for Full Sample A in Table 2 (see table footnote). Given the high intercorrelations among the profile agreement coefficients, most of these differences are statistically significant. For example, the correlation for r_p for factor profiles is lower than that for any other coefficient except the Pearson r for factor profiles; the correlation for ICC_{DE} for facet profiles is higher than that for any other coefficient.

Results are consistent across randomizations, and the same pattern is seen for the full sample and the two subsamples. At both factor and facet levels, r_p was inferior to the other three indexes, and ICC_{DE} was as good or better than the others—at least using the criterion of agreement adopted in this study. At the factor level, r_{pa} and ICC_{DE} were equally good, but ICC_{DE} was the better choice at the facet level.

How should profile agreement statistics for individual cases be evaluated? Although it is possible to compute statistical significance for these indices (e.g., McCrae, Terracciano, et al., 2005), that is not likely to be optimal. For example, a Pearson correlation of .88 is required for significance, $p < .05$, with a profile of five elements. In this sample, only 25% of 532 matched targets showed this degree of agreement. That should surely not be interpreted to mean that there is no agreement for the remaining 75% of the sample because in fact, cross-observer correlations for the five factors ranged from .45 to .55, all $ps < .001$, in that 75% subsample. Profile agreement indexes are better interpreted with respect to norms, and, for the NEO-PI-3, these are reported in Table 3. Values reported for r_{pa} for the factor profiles are comparable to those reported by McCrae (1993) for self/spouse agreement on the NEO-PI-R.

Discussion

These results replicate earlier findings that r_{pa} is superior to r_p as a metric of profile agreement (McCrae, 1993) and suggest that its use for estimating agreement at the factor level is justified. Currently, r_{pa} is used as the basis for profile agreement in the computer interpretation of the NEO-PI-R (Costa, McCrae, & Psychological Assessment Resources Staff, 1994), and it is likely to be used in interpretation of the NEO-PI-3. Instruments that assess only the five factors, such as the NEO Five-Factor Inventory (Costa & McCrae, 1992) or the Big Five Inventory (Benet-Martínez & John, 1998), can appropriately use r_{pa} , and it seems likely that r_{pa} will also be applicable to other profiles consisting of a relatively small number of characteristics (e.g., attachment style scores; Brennan, Clark, & Shaver, 1998).

One advantage of r_{pa} over ICC_{DE} is that it is possible to calculate I_{pa} for individual profile elements and thus identify particular instances of disagreement for further investigation (McCrae, Stone, Fagan, & Costa, 1998) or clinical interpretation (Costa & McCrae, 2005).² I_{pa} based on a single profile element may also prove useful in quantifying individual stability in longitudinal studies of personality. Because it is sensitive to both the difference between scores and their mean elevation, it should be superior to individual stability coefficients that consider only the difference (Asendorpf, 1992).

However, for faceted instruments such as the NEO-PI-3 or the Sixteen Personality Factor questionnaire (Conn & Rieke, 1994), ICC_{DE} appears to be a better measure of profile agreement. McCrae, Terracciano, et al. (2005) found that facet-level ICC_{DES} were more informative than r_{pa} s in comparing aggregate personality profiles of cultures, and the present data show that ICC_{DES} across 30 facets more accurately characterize agreement. That is a nonobvious finding because eight-item facet scales are substantially less reliable than factor

²The Q-sort method also provides a convenient way to judge disagreements, as the difference in item scores from two sorts.

scores and might have introduced error rather than additional precision. Specific variance in facet scales may have contributed, and the number of profile elements is certainly part of the explanation: Given equal reliability and validity, 30 independent scales would surely outperform 5 scales. However, the top panel of Table 1 makes it clear that profiles based on five elements are sufficient to distinguish matched from mismatched profiles with considerable success, at least when based on reliable scores.

Perhaps the most surprising finding was the performance of the Pearson r . Despite the theoretical objections to this measure, it was nearly as accurate as ICC_{DE} and superior to r_{pa} at the facet level. Although it is possible to have a large r for two profiles that are very dissimilar (when the profiles are parallel but at different elevations), that appears to be a very rare circumstance. In part, this may be attributable to the diversity of content in the NEO–PI–3, for which the concept of overall elevation is not psychologically meaningful. A profile restricted to a single factor (say, the six facets of E) might show more differences between r and ICC_{DE} because some pairs of raters might have differing ideas about the overall level of E but similar perceptions about the relative levels of the E facets.³

The validity of the Pearson r as a measure of profile agreement demonstrated here suggests that studies in the literature that have employed r (e.g., Biesanz & West, 2000; London & Wohlers, 1991), are trustworthy. However, future research should use ICC_{DE} instead; it is slightly more accurate (see Table 2) and avoids even the possibility of artifacts related to elevation.⁴ As Table 2 shows, ICC_{DE} is generally smaller than r . This is a perhaps healthy reminder that raters can differ in elevation as well as the shape of the profile, and it does not present a problem if agreement is interpreted in terms of norms rather than conventional statistical significance.

These results are directly applicable only to the NEO–PI–3, but any researcher who has matched sets of profiles can repeat the analyses I have reported here to evaluate their generalizability to other instruments. Such studies could also provide normative values for other instruments comparable to those provided in Table 3 for the NEO–PI–3.

In this study, I examined agreement between two different assessments of the same target, but profile agreement statistics are also used to compare assessments to a fixed prototype such as expert consensus ratings of a personality disorder (Lynam & Widiger, 2001). Conceptually appealing as this approach is, it is computationally cumbersome and unlikely to be used by clinicians interpreting personality profiles of individual clients. An alternative is simply to sum the scales identified as salient by the prototype and compare the sum to specified cutoffs. Costa and McCrae (2005) showed that such a scheme worked as well as their r_{pa} measure of personality disorders; and Miller, Bagby, Pilkonis, Reynolds, and Lynam (2005) reported similar results comparing summed scores with an ICC index of profile agreement. When profile agreement must be assessed because there is no fixed standard of comparison, both r_{pa} and ICC_{DE} are useful indexes. However, for comparisons of individuals

³Supplementary analyses correlating Match with r and ICC_{DE} based on the six profile elements for each domain separately showed that for all domains and all samples, ICC_{DE} was larger ($Mdn r = .46$) than r ($Mdn r = .41$).

⁴It is of some interest to note that in Carroll and Field's (1974) simulation study, ICC performed almost as well as r_p and considerably better than r , in correctly classifying profiles.

with established prototypes, summed scores may be preferable to profile agreement measures.

The profile agreement measures I have reviewed here are designed for use with two profiles. In some cases, researchers will have more than two measures—for example, father's ratings, mother's ratings, and child's self-reports (Laidra, Allik, Harro, Merenakk, & Harro, 2006). These profiles could be compared pairwise, but it would also be possible to combine the father's and mother's ratings to get a mean or adjusted mean (McCrae et al., 1998) parental rating profile and compare this to the self-report profile. Alternatively, one could use a standard intraclass correlation with three (or more) judges. It remains for future research to determine if the normative values provided in Table 3 would be appropriate for such applications.

Acknowledgments

Robert R. McCrae is at the Gerontology Research Center, National Institute on Aging, National Institutes of Health, Department of Health and Human Services. Robert R. McCrae receives royalties from the Revised NEO Personality Inventory. This research was supported by the Intramural Research Program of the National Institutes of Health, National Institute on Aging.

References

- Asendorpf JB. Beyond stability: Predicting inter-individual differences in intra-individual change. *European Journal of Personality*. 1992; 6:103–117.
- Benet-Martínez V, John OP. *Los cinco Grandes* across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*. 1998; 75:729–750. [PubMed: 9781409]
- Biesanz JC, West SG. Personality coherence: Moderating self-other profile agreement and profile consensus. *Journal of Personality and Social Psychology*. 2000; 79:425–437. [PubMed: 10981844]
- Block, J. *The Q-sort method in personality assessment and psychiatric research*. Springfield, IL: Thomas; 1961.
- Brennan, KA., Clark, CL., Shaver, PR. Self-report measurement of adult attachment: An integrative overview. In: Simpson, JA., Rholes, WS., editors. *Attachment theory and close relationships*. New York: Guilford; 1998. p. 46–76.
- Carroll RM, Field J. A comparison of the classification accuracy of profile similarity measures. *Multivariate Behavioral Research*. 1974; 9:373–380. [PubMed: 26805485]
- Cattell RB. r_p and other coefficients of pattern similarity. *Psychometrika*. 1949; 14:279–298.
- Conn, SR., Rieke, ML., editors. *16PF fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing; 1994.
- Costa, PT., Jr, McCrae, RR. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources; 1992.
- Costa, PT., Jr, McCrae, RR. A Five-factor model perspective on personality disorders. In: Strack, S., editor. *Handbook of personology and psychopathology*. Hoboken, NJ: Wiley; 2005. p. 257–270.
- Costa, PT., Jr, McCrae, RR., Psychological Assessment Resources Staff. *NEO software system [Computer software]*. Odessa, FL: Psychological Assessment Resources; 1994.
- Costa, PT., Jr, Piedmont, RL. Multivariate assessment: NEO-PI-R profiles of Madeline G. In: Wiggins, JS., editor. *Paradigms of personality assessment*. New York: Guilford; 2003. p. 262–280.
- Griffin D, Gonzalez R. Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin*. 1995; 118:430–439.

- Gustafsson, JE. Measurement from a hierarchical point of view. In: Braun, HI, Jackson, DN., Wiley, DE., editors. *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Lawrence Erlbaum Associates; 2002. p. 73-95.
- Laidra K, Allik J, Harro M, Merenakk L, Harro J. Agreement among adolescents, parents, and teachers on adolescent personality. *Assessment*. 2006; 13:187–196. [PubMed: 16672733]
- London M, Wohlers AJ. Agreement between subordinate and self-ratings in upward feedback. *Personnel Psychology*. 1991; 44:375–390.
- Lynam DR, Widiger TA. Using the Five-factor model to represent the *DSM-IV* personality disorders: An expert consensus approach. *Journal of Abnormal Psychology*. 2001; 110:401–412. [PubMed: 11502083]
- McCrae RR. Agreement of personality profiles across observers. *Multivariate Behavioral Research*. 1993; 28:13–28.
- McCrae RR, Costa PT Jr. Discriminant validity of NEO-PI-R facets. *Educational and Psychological Measurement*. 1992; 52:229–237.
- McCrae RR, Costa PT Jr, Martin TA. The NEO-PI-3: A more readable Revised NEO Personality Inventory. *Journal of Personality Assessment*. 2005; 84:261–270. [PubMed: 15907162]
- McCrae RR, Martin TA, Costa PT Jr. Age trends and age norms for the NEO Personality Inventory–3 in adolescents and adults. *Assessment*. 2005; 12:363–373. [PubMed: 16244117]
- McCrae RR, Stone SV, Fagan PJ, Costa PT Jr. Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality*. 1998; 66:285–313. [PubMed: 9615420]
- McCrae RR, Terracciano A, 79 Members of the Personality Profiles of Cultures Project. Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology*. 2005; 89:407–425. [PubMed: 16248722]
- McGraw KO, Wong SP. Forming inferences about some intraclass correlations. *Psychological Methods*. 1996; 1:30–46.
- Meng XL, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. *Psychological Bulletin*. 1992; 111:172–175.
- Miller JD, Bagby RM, Pilkonis PA, Reynolds SK, Lynam DR. A simplified technique for scoring *DSM-IV* personality disorders with the Five-factor model. *Assessment*. 2005; 12:404–415. [PubMed: 16244121]
- Ozer DJ, Gjerde PF. Patterns of personality consistency and change from childhood through adolescence. *Journal of Personality*. 1989; 57:483–507. [PubMed: 2769564]
- Terracciano A, Abdel-Khalak AM, Àdàm N, Adamovová L, Ahn CK, Ahn HN, et al. National character does not reflect mean personality trait levels in 49 cultures. *Science*. 2005; 310:96–100. [PubMed: 16210536]
- Terracciano A, McCrae RR. How to measure national stereotypes? Response. *Science*. 2006; 311:777–779.

TABLE 1

An example of the double-entry intraclass correlation.

Profile Element	T Score, Single Entry		T Score, Double Entry	
	Self-Report (S)	Observer Rating (R)	Column 1	Column 2
N	50	25	50	25
			25	50
E	55	30	55	30
			30	55
O	60	35	60	35
			35	60
A	65	40	65	40
			40	65
C	70	45	70	45
			45	70
$r_{SR} = 1.00$			$r_{12} = ICC_{DE} = -.52$	

Note. N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness; ICC_{DE} = intraclass correlation (double entry).

TABLE 2

Correlations of profile agreement indexes for factors and facets with the match criterion in three randomizations.

Correlation	Full Sample			Extreme Subsample			Nonextreme Subsample		
	X	Y	Z	X	Y	Z	X	Y	Z
Factor profile									
$r(5)$	.57 _{a,b,c}	.58	.57	.66	.63	.64	.48	.53	.51
$r_f(5)$	.54 _a	.55	.54	.63	.63	.63	.49	.50	.49
$r_{pw}(5)$	.60 _{b,d,e}	.60	.60	.67	.66	.67	.52	.55	.54
ICC _{DE} (5)	.61 _{d,f}	.60	.60	.69	.66	.68	.53	.56	.54
Facet profile									
$r(30)$	.69	.68	.69	.74	.70	.72	.64	.68	.65
$r_f(30)$	.57 _{c,e,f}	.58	.57	.64	.63	.63	.55	.56	.54
$r_{pw}(30)$	.65	.65	.65	.71	.68	.69	.61	.63	.62
ICC _{DE} (30)	.70	.69	.69	.74	.71	.72	.65	.68	.66

Note. ICCDE = intraclass correlation (double entry). X, Y, and Z are artificial data sets created by adding three different random mismatches to the original matched data. Ns = 1,064 for the Full Samples, 534 for the Extreme Subsamples, and 530 for the Nonextreme Subsamples. Number of profile elements is given in parentheses. In the first data column, correlations sharing the same subscript are not significantly different, $p < .05$.

TABLE 3

Distribution of profile agreement indexes between self-reports and observer ratings on the NEO-PI-3 in an adult sample.

Statistic	Factor Profiles				Facet Profiles			
	<i>r</i>	<i>r_p</i>	<i>r_{pa}</i>	ICCD _E	<i>r</i>	<i>r_p</i>	<i>r_{pa}</i>	ICCD _E
M	.60	.50	.53	.46	.49	.37	.41	.43
<i>SD</i>	.36	.29	.24	.36	.24	.23	.19	.25
Minimum	-.83	-.55	-.81	-.75	-.37	-.50	-.65	-.37
Maximum	.99	.96	.96	.98	.91	.90	.89	.90
Percentiles								
1	-.56	-.28	-.41	-.53	-.18	-.29	-.13	-.23.
5	-.18	-.06	.14	-.28	.02	-.06	.09	-.01
7	-.07	.00	.20	-.19	.09	-.02	.13	.04
31	.51	.39	.45	.33	.39	.27	.33	.33
50	.70	.54	.56	.52	.53	.39	.42	.47
69	.85	.67	.67	.68	.63	.50	.51	.57
93	.97	.89	.83	.90	.80	.69	.66	.77
95	.98	.90	.86	.92	.81	.72	.69	.79
99	.99	.95	.91	.97	.89	.81	.80	.88

Note. *N* = 532. NEO-PI-3 = NEO Personality Inventory-3; ICCD_E intraclass correlation (double entry). By the conventions for interpreting profiles of the NEO-PI-R, the 7th, 31st, 69th, and 93rd percentiles divide the sample into *very low*, *low*, *average*, *high*, and *very high* agreement (see McCrae, 1993).