Swiss Journal of Psychology, 71 (1), 2012, 35–45

Original Communication

Person-Fit to the Five Factor Model of Personality

Jüri Allik[1,2], Anu Realo[1], René Mõttus[1,3], Peter Borkenau[4], Peter Kuppens[5], and Martina Hřebíčková[6]

[1]Department of Psychology, University of Tartu, Tartu, Estonia, [2]Estonian Academy of Sciences, Tallinn Estonia, [3]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK, [4]Psychology Department, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany, [5]Department of Psychology, Katholieke Universiteit Leuven, Leuven, Belgium, [6]Institute of Psychology, Academy of Sciences of the Czech Republic, Prague, Czech Republic

Abstract. The Five Factor Model (FFM), a valid model of interindividual differences in the personality of a group of people, reportedly does not always provide a good fit for the individuals of that group. In addition to intraindividual variation across a considerable period of time, meaningful intraindividual variation can be observed within a single test administration. Two person-fit indices showed that the FFM is an adequate model for 95% of the 1,765 target-judge pairs in four different countries (Belgium, the Czech Republic, Estonia, and Germany): the double-entry intraclass correlation (ICC_{DE}), which indicated that the 30 NEO PI-R scores on scales measuring the same personality trait are more similar and certainly less different than scores measuring different traits, and the individual contribution to the extracted eigenvalues (Z_{eig}). The individual response pattern to the personality questionnaire characterized by the ICC_{DE} and Z_{eig} strongly determined the percentage of explained variance for the group-level factor structure of interindividual differences and the mean self-observer profile agreement. We demonstrate that, if the percentage of variance explained by the first five principal components is high enough, the FFM also provides an adequate fit at the individual level for most people.

Keywords: personality, Five Factor Model, person-fit, cross-cultural comparison, personality profiles

Unlike personality researchers, educational and applied psychologists have been concerned with unusual test-response patterns for some time now. Attempts to systematically identify people with unusual test-response patterns have led to an elaborate methodology for detecting individual aberrations from the common response pattern. A number of person-fit statistics (e.g., the caution index, the norm conformity index, the individual consistency index) have been developed, all of which measure how well statistical models fit at the level of the individual (Karabatsos, 2003; Meijer & Sijtsma, 2001). Most person-fit statistics are based on the item response theory (IRT) and measure an individual's congruence with the general response pattern of a group. In its simplest form, a lack of fit is illustrated by the responses of an examinee who responds correctly to the more difficult items but incorrectly to the easier ones.

Although personality questionnaires are even more vulnerable to potential distortions than intelligence tests, the applications of IRT in personality assessment have been less impressive, mainly because of the multidimensionality of personality instruments and perhaps because the interpretation of a person's score on latent personality variables is more complex than that on mental abilities (Reise & Waller, 1993). Unlike intelligence tests, personality instruments do not have common response patterns since we expect some individuals to score high and others to score low on personality trait measures. Indeed, the assumption that some personality items represent more extreme expressions of a trait than others (the so-called item "difficulty"), and that this pattern of endorsement needs to be identical across individuals, is a premise that has yet to be proven. Researchers have also identified distinct subpopulations within large samples, each of which responds differently to a set of personality items (Egberink, Meijer, & Veldkamp, 2010; Rost, 1990). Moreover, there is evidence that individuals can respond differently to personality items than to ability test items, making IRT application to personality data even more problematic (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). Except for the detection of outliers in multivariate analysis, little research exists that examines how latent variables in factor or structural equation analysis fit at the level of the individual (Reise & Widaman, 1999).

This study approaches the problem of person-fit to theoretical models of personality from a different perspective, one not related to the IRT research tradition. All personality mod-

DOI 10.1024/1421-0185/a000066

els are based on observations that covariations between personality traits are relatively stable. Individuals who describe themselves or whose close acquaintances describe them as talkative are also believed to experience positive emotions frequently, and those who are reported to be ambitious often describe themselves as organized and systematic. These kinds of covariations tend to group around the same five basic themes, which transcend language and culture, hinting that this structure of covariation may be universal (McCrae & Costa, 1997). Although personality psychologists still debate about the correct number of factors, many of them believe that the Five Factor Model (FFM) of personality trait covariation is the most parsimonious description of the basic human tendency to think, feel, and behave in a consistent manner (Goldberg, 1993; McCrae & John, 1992).

However, according to some researchers, personality trait covariation models such as FFM provide information that holds true at the level of groups or populations, but maybe not at the level of the individual (Borsboom, 2005). For example, it has been demonstrated that if a latent factor model fits a given population, it does not necessarily fit each or even any individual at all in that population when intraindividual variation is measured by repeated administration of the same instrument (Borsboom, Mellenbergh, & Van Heerden, 2003; Molenaar & Campbell, 2009). Not all researchers agree with Borsboom and his colleagues' (2003) conclusion that models derived from between-subject variation cannot provide causal explanations for the behavior of individuals (McCrae & Costa, 2008). As noted by an anonymous referee, if you repeatedly administered the same instrument, you might observe the covariation of nothing but noise. Nevertheless, it is regrettable that there is no convincing demonstration of how individual response patterns to personality questionnaires fit or misfit theoretically postulated latent factor models derived from a group-level analysis of interindividual differences. Thus, our main goal is to analyze how individuals contribute to a theoretically postulated latent multifactor model of personality.

Individual's Contribution to Fit Statistics

One of the best solutions to the person-fit problem came from Reise and Widaman (1999), who proposed that an individual's contributions can be calculated with the chi-square (χ^2) statistic, which measures the log likelihood that the observed covariance matrix is reproduced by the statistical model. To this end, they partitioned the model's overall χ^2 value into the subjects' individual contributions such that a relatively large decrease in fit would result if an individual who contributes more to the overall model fit and more to the increase in χ^2 value was removed from the total sample. Consequently, a relatively large drop of the χ^2 value represents individuals whose responses fit the statistical model well. It is interesting that this covariance structure, which is based on the person-fit index, did not correlate strongly with person-fit statistics computed on the basis of various IRT models (Reise & Widaman, 1999). Unfortunately, however, the idea of partitioning a model's fit indicators into individual contributions has not been widely used in personality research. Our intention is to use this largely underexploited idea for the study of person-fit to the FFM by determining individual contributions to the cumulative amount of variance explained by the first five principal components.

Intraclass Correlation (ICC)

Another way to solve the problem of the person-fit to the FFM is to take up Campbell and Fiske's (1959) basic idea about convergent and discriminant validity. Nowadays, the main message of their seminal paper almost seems self-evident: Measures of the same variable using different methods should agree (converge) more than measures of different variables. In an effort to apply their methodological vision, Campbell and Fiske developed the multitrait-multimethod matrix approach. The main idea was to use a factorial combination of traits and measurement methods that would allow one to partition the total variance into separate components that could be identified with traits, methods, and unique error terms. Many current personality measurement models are implementations of this simple methodological principle: They measure several independent factors, each tapped by several parallel subscales. For example, a general tendency toward extraversion is conceptualized in the NEO PI-R as consisting of several specific tendencies such as being assertive, looking for excitement, having warm feelings toward other people, and experiencing positive emotions (Costa & McCrae, 1992). These subscales can be viewed as different methods aimed at measuring various manifestations of the same broad trait – extraversion. Therefore, it is reasonable to expect these subscales to converge more than subscales that measure different traits, for example, openness to experience or agreeableness. In terms of covariation, the correlation between any two subscales that measure the same trait should be higher than that between any two subscales that measure different and theoretically unrelated traits. For example, of the 435 possible pairwise intercorrelations between the NEO PI-R's 30 facets, 75 are among subscales that measure the same basic trait (e.g., N1: Anxiety and N2: Angry Hostility), and the remaining 360 are between subscales measuring different traits (e.g., A1: Trust and C1: Competence). For the matrix of intercorrelations in the North American normative sample of 500 men and 500 women (Costa & McCrae, 1992, Appendix F), the mean correlation between subscales measuring the same trait was .38 and the mean correlation between subscales measuring different traits was, as expected, close to zero: .01. In the Estonian normative sample (Kallasmaa, Allik, Realo, & McCrae, 2000), these mean values were .45 and –.02, respectively. In both cultures, the convergent (within-factor) correlations

were substantially higher than the discriminant (between-factor) correlations.

For obvious reasons, no such matrix of intercorrelations can be calculated for a single subject, unless he or she answers the same questions repeatedly a sufficient number of times. However, repeated measures over extended time intervals are not the only source of within-subject variance. One possible way of compensating for a lack of repeated measures on the same subscales is to regard parallel measures of the same trait as repeated measures of the same general disposition. Following this line of reasoning, an extravert is expected to score high on all or at least most subscales assessing extraversion – E1: Warmth, E2: Gregariousness, E3: Assertiveness, E4: Activity, E5: Excitement Seeking, and E6: Positive Emotions – whereas an emotionally stable person should score low on the majority of subscales measuring neuroticism. Consequently, scores assessing the same trait should be similar and certainly less different than scores measuring different traits.

For each individual response pattern – scores on the 30 NEO PI-R subscales – we can compute the *intraclass correlation* (ICC) by arranging all scores on the same dimension pairwise into two columns, entering each pair twice (N1-N2, N2-N1, N1-N3, N3-N1, . . ., C5-C6, C6-C5) and then computing the correlation between the two columns. However, the most efficient way to calculate the ICC is to use the analysis of variance (ANOVA) framework and split the total variance produced by the 30 NEO PI-R individual subscale scores into two components: within-factor (σ^2_W) and between-factor (σ^2_F) variance (i.e., the variance produced by "error" and the variance produced by differences between the factor means). Any one-way ANOVA program in which the 30 scores are assembled into the 5 respective groups on that factor can be used to compute these values. The ICC is defined as the ratio of variance attributable to the differences between factor means to total variance: ICC = $\sigma^2_F/(\sigma^2_W + \sigma^2_F)$. Thus, the perfect ICC in fact detects individual patterns of response in which the scores of subscales measuring the same factor are maximally similar but their mean levels differ substantially. Obviously, the ICC is not limited to the FFM and can be computed for six (Ashton & Lee, 2010; Ashton et al., 2004) or any other number of personality dimensions. More technically, the ICC demonstrates how well the 30 NEO PI-R individual subscale scores can be reproduced by only five mean values, one for each of the five personality dimensions.

Aim of the Study

This study investigates how well theoretical multifactor models, such as the FFM, derived from aggregated interindividual differences, apply to individual subjects. In addition, we are also interested in the ability of the two person-fit indices to predict the accuracy in personality judgment. If self-descriptions that agree with the FFM do not converge with the observer descriptions of the same people, then we will still be uncertain about the validity of these personality descriptions.

Method

Samples

Belgian Sample

Flemish data were collected from 345 target participants (270 women and 75 men) who were psychology students at the Katholieke Universiteit Leuven and who, as a course requirement, rated their own personality with the Dutch version of the NEO PI-R (Hoekstra, Ormel, & DeFruyt, 1996). They also recruited a well-acquainted person (n = 345; 190 women, 112 men, and 43 who did not specify sex), either a relative or a friend, who rated their personality using the observer report form of the same instrument. The targets' mean age was 18.4 (SD = 3.0) years. The observers' mean age was 29.5 (SD = 13.7) years.

Czech Sample

The Czech sample included 808 targets (329 men, 479 women) recruited in a series of studies (McCrae et al., 2004). They ranged in age from 14–83 years, with a mean age of 35.7 (SD = 14.2) years. Peer ratings were provided by 909 raters (377 men, 532 women) aged 14–83 years (M = 35.8; SD = 14.3 years) who participated in one of two research designs. In the self-other agreement studies (N = 616), each target provided a self-report and was rated by one informant. In the consensus study, 196 targets (85 men and 111 women aged 17–77 years; mean age 36.4, SD = 15.2) provided a self-report and were each rated by three informants. All participants used the Czech version of the NEO PI-R questionnaire (Hřebíčková, 2002).

Estonian Sample

The Estonian data came from two previously published studies. The first sample consisted of 218 Estonian-speaking participants (180 women and 38 men; mean age 22.3 years, SD = 5.2) who completed the NEO PI-R questionnaire accompanied by standard instructions to describe themselves honestly and accurately (Konstabel, Aavik, & Allik, 2006). They were also asked to provide two peer reports (n = 436) from acquaintances, relatives, or close friends. The Estonian version of the NEO PI-R (Kallasmaa et al., 2000) was completed voluntarily; some students studying psychology received extra credit for the fulfillment of their course. The second Estonian sample consisted of 154 participants (53 men and 101 women; mean age 43.9 years, SD = 17.6) who were described by one or two judges (Mõttus, Allik, & Pullman, 2007). The sample of judges (n

= 308) included 203 women, 67 men, and 38 participants who did not report their sex. The judges' mean age of 38.2 (*SD* = 15.9) years. Both targets and judges used the Estonian version of the EE.PIP-NEO (Mõttus, Pullmann, & Allik, 2006), which has a facet structure identical to the NEO PI-R, but was designed to be linguistically simpler, containing shorter and grammatically less complex items.

German Sample

The participants were 304 students (169 women, 134 men, and 1 person who did not report sex) at a German university, only 3 of whom were studying psychology (Borkenau & Zaltauskas, 2009). Their mean age was 23.38 (*SD* = 2.68) years, ranging from 18 to 35 years. They received 45 EUR for their participation and were recruited in 76 groups, each comprising four people who all knew each other well. First, the participants described the three other group members on 30 bipolar adjective scales; these, however, are not relevant to the present study. Next, each four-person group was split into two dyads and the participants all described themselves and the other dyad member on several personality inventories. The participants all described themselves and the other member of their dyad on the German version of the NEO PI-R (Ostendorf & Angleitner, 2004).

Taken together, the total sample for this study included 1,765 target participants (mean age = 29.7 years, *SD* = 13.9; 623 males, 1,141 females and 1 person who did not report sex) who rated their own personality and were rated by one or several observers (see also Allik et al., 2010).

Informed consent in written form was obtained from all participants in this study. Ethics approval for this study was received from the ethics committees of the respective universities.

Normalizing Data

In order to eliminate the confounding effect of culture, all personality scores were normalized within each country after which the mean country value on each trait became 0 and the standard deviation was made equal to one.

Person-Fit Indices

Intraclass Correlation (ICC)

For each self- or observer-rating, the ICC was computed on the basis of 30 normalized individual subscale scores. The ICC is defined as the ratio of variance attributable to the differences between factor means to total variance: ICC = $\sigma^2_F/(\sigma^2_W + \sigma^2_F)$. In the context of the analysis of variance, it is easier to calculate the ICC with the mean squares, MS_W and MS_F. The ICC, or, more precisely, the one-way model ICC(1) according to the classification introduced by McGraw and Wong (1996), can be calculated from the mean squares in the following way: ICC = $(MS_F - MS_W)/[\ MS_F + (K - 1)\ MS_W\ ]$, where K is the number of subscales (in the context of the NEO PI-R, $K = 6$).

However, the ICC is imperfect in that its value varies with arbitrary decisions in which direction variables are coded, for example, whether Neuroticism or Emotional Stability is scored high. In order to get rid of this shortcoming, we computed a double-entry ICC_{DE} by entering all scores twice in original form X and reflected form X_0, which is computed from the original score as $X_0 = 2\ m–X$, where m is the midpoint of the scale (see Cohen, 1969). For normalized scores, the reflection simply means entering the z-scores twice with the opposite signs. Unlike the ICC, the double-entry ICC_{DE} is invariant to the direction in which variables are coded. It may come as a surprise that ICC and ICC_{DE} are usually highly correlated (in this study, $r = .90$) and the values for ICC_{DE} tend to be on average higher than those for ICC although the double-entry method does not require an increase in the number of observations (Cohen, 1969).

It should be noted that Bem and Allen (1974) proposed the ipsatized variance index (IVI) as a criterion of inconsistent responding. The IVI was defined as the ratio of the variability of the subject's responses to the items on a given trait scale to that across the entire set of items included in the questionnaire. Clearly, the IVI is related to the ICC in the sense that it is its inverse value: $ICC = IVI^{-1}$. Another similar measure to the ICC is the (partial) eta squared (η^2), which is used as a standardized measure of effect size in statistical packages such as SPSS (IBM Corporation). The η^2 statistic was also previously used as an index of personality consistency (Campus, 1974).

Individual Contribution to the Extracted Eigenvalues (Zeig)

Reise and Widaman (1999) proposed that an individual's contribution be calculated with the χ^2 statistic, which measures the log likelihood that the observed covariance matrix is reproduced by the statistical model. Following the same logic, we calculated individual contributions using extracted eigenvalues. After a principal component analysis was performed and the ratio of the first five cumulative eigenvalues to the total number of variables was found for the whole sample containing all N subjects, individuals were excluded one by one from the sample and the proportion of explained variance was again calculated for each of them in the reduced N–1 sample. This change corresponds to the contribution of the individual excluded from the total sample. The percentage of explained variance in the whole sample N minus the percentage of explained variance in the reduced N–1 sample, Z_{eig}, was used as a measure of each individual's contribution. All principal component analyses were performed within each country sample.

Results

We started the analysis by computing the ICC_{DE} values for the 1,765 self- and observer-ratings on the basis of their normalized scores on the 30 NEO PI-R scales. The mean values of the ICC_{DE} were .41 and .43, respectively, for self- and observer-ratings. At the same time, the 1,765 individuals' contributions to the extracted eigenvalues, Z_{eig}, were nearly symmetrically distributed around 0 with 57.2% and 55.9% of those who made a positive contribution to self- and observer-ratings, respectively. It is interesting that for both ICC_{DE} and Z_{eig} there was some crossobserver agreement, with the respective correlations being $r(1763) = .20$ and $r(1763) = .18$ ($p < .0001$).

The mean squares for the factors (i.e., between-factor variance) must be at least 2.76 times larger than the mean squares for the error (i.e., within-factor variance) with 4 and 25 degrees of freedom to exceed the critical value at the significance level $p < .05$. There were 1,460 (82.7%) self-ratings and 1,477 (83.7%) observer-ratings whose ICC_{DE} values were statistically significant. Out of 1,765 target-judge pairs, the self- and observer ICC scores only did not simultaneously reach statistical significance in 76 cases (4.31%). This percentage is higher than expected on the basis of the independence assumption (2.82%). However, for 55 (3.12%) of these 76 "poor-fit" participants, self-ratings nevertheless showed significant profile agreement (for Pearson correlation $p < .05$) with observer-ratings, suggesting that the pattern of personality scores may not have been accidental. Put differently, there were only 36 participants out of 1,765 (2.04%) whose self- and observer-rated personality scores did not resemble the expected five-factor pattern and who also failed to demonstrate significant agreement between self- and observer-ratings. Thus, only approximately 2% of our participants from four countries responded to the NEO PI-R items in a manner that deviates from the FFM and also failed to show acceptable self-observer agreement.

Although we know for ICC_{DE}, we do not know at which point the individual's contribution to the extracted eigenvalues (Z_{eig}) becomes significant. Nevertheless, individual ICC_{DE} and Z_{eig} values were significantly correlated: $r(1763) = .64$ and .59 for self- and observer-ratings, respectively ($p < .0001$). This indicates that although ICC_{DE} and Z_{eig} are related, they still characterize two different aspects of personal fit to the FFM.

Since the strict mathematical relationship between the total amounts of variance explained by the FFM and any person-fit indices is complicated, one initially needs to establish this link empirically. This purpose in mind, we divided the whole sample $N = 1{,}765$ into 10 approximately equal-sized groups on the basis of the percentile values of the respective person-fit indicators: ICC and Z_{eig}. Thus, in each group there were 176.5 participants on average. For example, the mean ICC_{DE} values in the lowest groups were $ICC_{DE} = .05$ and .02 for the self- and observer-ratings, respectively; in the highest groups, these values were $ICC_{DE} = .70$ and .71, respectively. Next, we performed a series of principal component analyses to determine the percentage of variance explained by the five components extracted for each of the 10 percentile groups and each person-fit index. Figure 1A and Figure 1B show the percentage of explained variance for each percentile group as a function of the mean

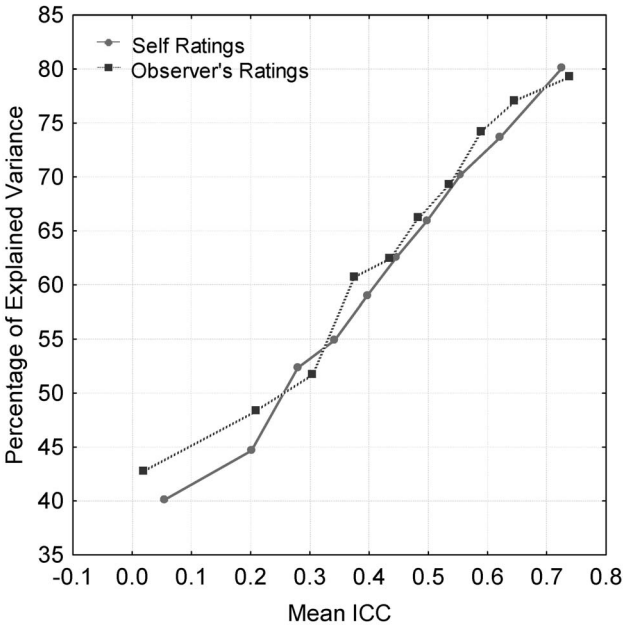

Figure 1A. Percentage of variance explained by the first five principal components as a function of the mean double-entry intraclass correlation (ICC_{DE}).

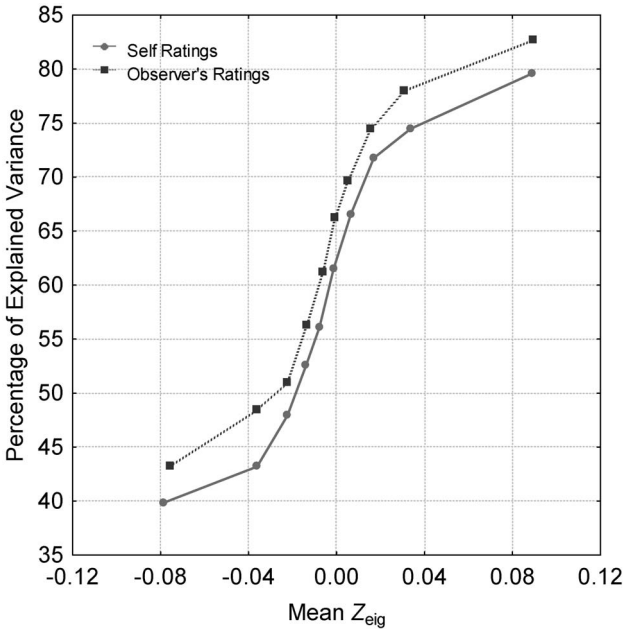

Figure 1B. Percentage of variance explained by the first five principal components as a function of the mean individual contribution to the extracted eigenvalues (Z_{eig}).

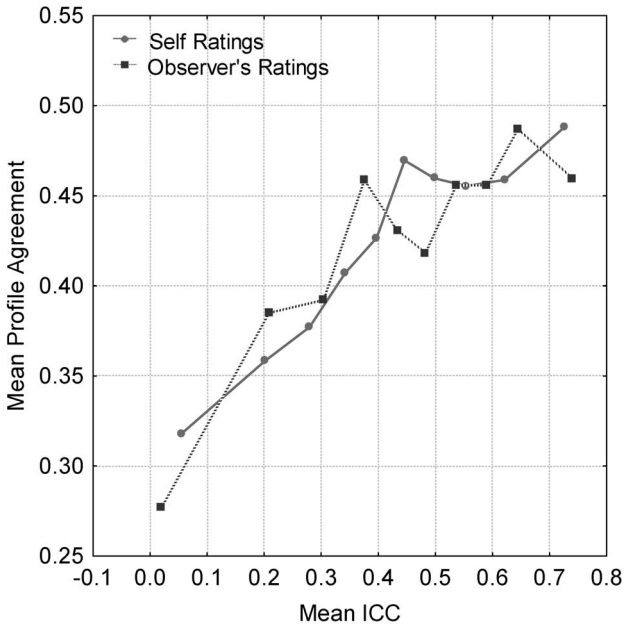

Figure 2A. Mean self-observer profile agreement as a function of the mean double-entry intraclass correlation (ICC_{DE}).

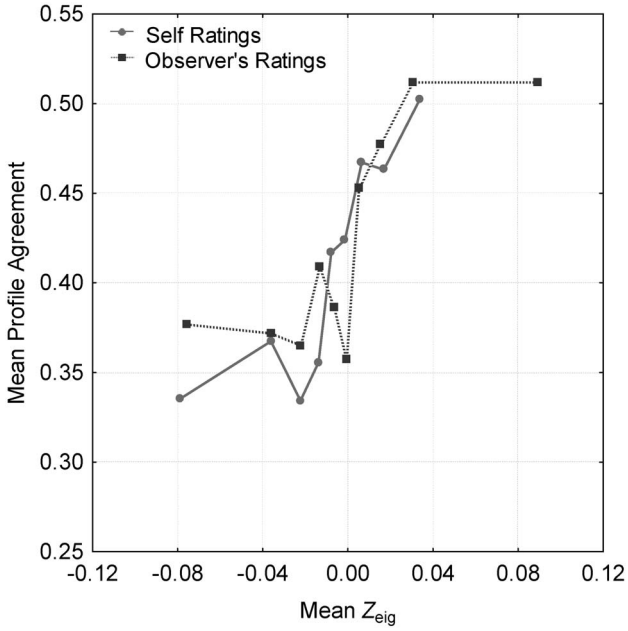

Figure 2B. Mean self-observer profile agreement as a function of the mean individual contribution to the extracted eigenvalues (Z_{eig}).

ICC_{DE} and Z_{eig}, respectively. For both the ICC_{DE} and Z_{eig}, the percentage of explained variance varied systematically with increasing percentile group. For example, in the lowest ICC_{DE} groups, five components explained only 44.0% and 48.2% of variance for self- and observer-ratings, respectively. However, in the highest ICC_{DE} groups, the amount of explained variance was 80.1% and 79.2%, respectively. For the ICC_{DE}, the relationship between percentage of explained variance and decile group mean ICC was almost perfectly linear [$r(8) = .995$ and .985 for self- and observer-ratings, respectively]. Due to the slightly curvilinear relationship, the correlation between percentage of explained variance and the mean Z_{eig} was slightly lower [$r(8) = .931$ and .929 for self- and observer-ratings, respectively]. Not surprisingly, the mean value of each decile group's ICC_{DE} was strongly correlated with the mean individual contribution to the extracted eigenvalues, Z_{eig}. The correlations were highly significant, .98 and .95, respectively, for self- and observer-ratings.

Next, we computed the mean self-observer profile agreement for each ICC_{DE} and Z_{eig} decile group. As shown in Figure 2A and Figure 2B, the mean self-observer agreement increases with the increase in mean ICC_{DE} and Z_{eig}. The relationship between mean profile agreement and ICC_{DE} was again close to linear [$r(8) = .950$ and .890 for self- and observer-ratings, respectively]. The relationship between mean profile agreement and Z_{eig} was only slightly more irregular [$r(8) = .912$ and .787 for self- and observer-ratings, respectively]. It is well established that correlations based on aggregated data are higher than the same correlation computed on individual data (see Epstein, 1983). As expected, this was true for the present data as well. The correlation between profile agreement and the person-fit indices, ICC_{DE} and Z_{eig}, were lower but statistically significant at the level of the individual: $r(1763) = .195$ and $r(1763) = .215$, respectively ($p < .0001$).

To explore the link between ICC and the percentage of explained variance more thoroughly, we ran several simulations. Every simulation was started by filling the 30 individuals by 30 scales matrix with randomly generated integers in the range from 1 to 5. As expected, the ICC values computed on these random values are almost always close to 0. After that, the program randomly selected one element in the matrix and replaced it with another randomly generated integer only if it improved the ICC values in the corresponding row. Repeating this procedure a different number of times generated response patterns with the mean ICC varying in the range from 0 to one. If the procedure was repeated a sufficiently large number of times, the pattern became identical to one with the mean ICC = 1. For each resulting response matrix, we applied principal component analysis to extract the first five principal components. The relationship between mean ICC and percentage of explained variance is shown in Figure 3A. Indeed, the relationship is very close to functional: The mean ICC determines almost precisely the percentage of variance explained by the first five principal components. A simple parabolic function describes 99.5% of the variance. Figure 3B demonstrates the same relationship for a larger sample of hypothetical responses ($N = 100$). Except for the smaller percentage of explained variance for a completely random response pattern (the ICC values close to 0), the relation-

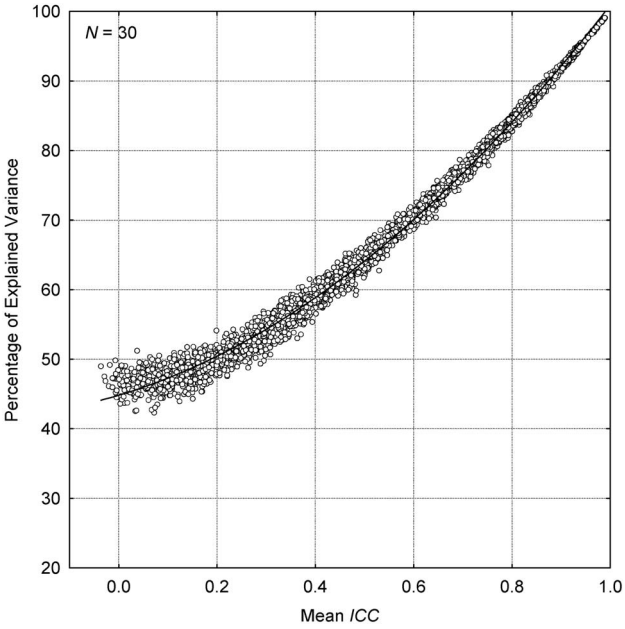

Figure 3A. Percentage of variance explained by the first five principal components as a function of the mean intraclass correlation (ICC) based on more than 4,000 simulations.

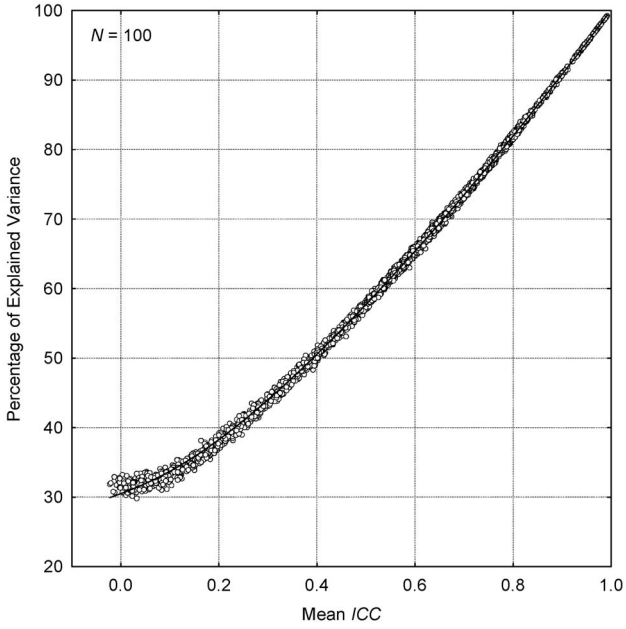

Figure 3B. Percentage of variance explained by the first five principal components as a function of the mean intraclass correlation (ICC) based on more than 4,000 simulations.

ship has exactly the same shape leaving less than 0.1% of variance unexplained. The results of this simulation demonstrates that there is indeed a simple property of the individual response pattern – the mean ICC or how well the 30 subscale scores can be reproduced by only five mean values – which almost precisely predicts how much variance can be explained by the first five principal components. Perhaps for the first time, we found a characteristic of individual test responses which directly determines properties of the model characterizing behavior at the group level.

Although the FFM seems to be an adequate description of personality scores for most of the participants from these four cultures (at least the structure of the instruments designed to reveal the FFM is widely replicable), it is still possible that a relatively small number of atypical constellations of personality traits exist. In addition to individuals and their observers who had very high ICCs and almost perfect agreement with each other (Figure 4A), it is possible to find individuals who demonstrate very high agreement between self- and observer-ratings and yet, at the same time, display very low ICC_{DE} values. Figure 4B demonstrates almost ideal self-observer agreement [$r(28) = .94$] accompanied by near-0 ICC_{DE} values. The reason for these low ICC_{DE} values is obvious: Both the person and her judge assessed her N1: Anxiety and N2: Angry Hostility at about one standard deviation above the sample mean, while she was perceived to be well below the sample mean on the other Neuroticism subscales, especially on N4: Self-Consciousness. In other words, there is consensus in the portrayal of her as an anxious and angry individual who is also fearless and does not often feel shame or embarrassment. Figure 4C shows another atypical profile of an 18-year-old Flemish women and her judge. In spite of very high self-observer agreement [$r(28) = .84$], this woman was assessed to be at about one standard deviation above the sample mean for E1: Warmth and about one standard deviation below the sample mean for E2: Gregariousness. Similarly, she and her judge characterized her as relatively closed to fantasies (O1) and, at the same time, very open to new ideas (O6). From the point of view of the FFM model, these are unexpected combinations of personality traits since affectionate and friendly people usually prefer other people's company and individuals who have an active fantasy life are also ready to re-examine their social, political, and religious values.

Discussion

Even though the FFM and six-factor models have repeatedly been shown to be valid models of aggregate personality in many languages (Allik & McCrae, 2002; Lee & Ashton, 2008), personality psychologists remain split on the issue of the universality of the FFM and any other factorial model. This study demonstrated that, in addition to intraindividual variation across a considerable period of time, it is also possible to observe meaningful individual variation within one test administration. Quite simply, an individual subject fits the FFM or any other model when his or her subscale scores measuring a given general personality trait are relatively similar to, and certainly less different than, the differences between scores on different factors. The intraclass correlation, ICC_{DE}, which measures precisely this property, demonstrated

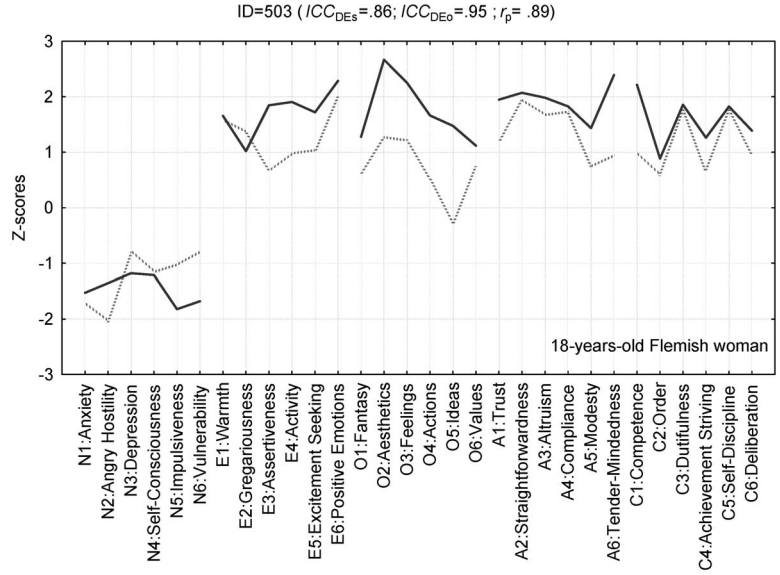

Figure 4A. An example of high self-observer profile agreement (r_p) and high person-fit values of self (ICC_{DEs}) and observer (ICC_{DEo}) ratings.

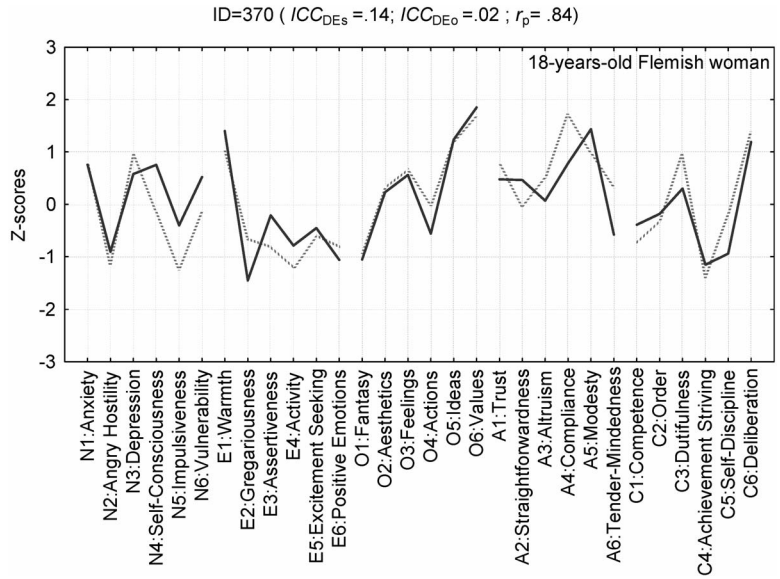

Figure 4B. An example of high self-observer profile agreement. Although agreement between profiles is perfect, the fit to the FFM is completely absent.

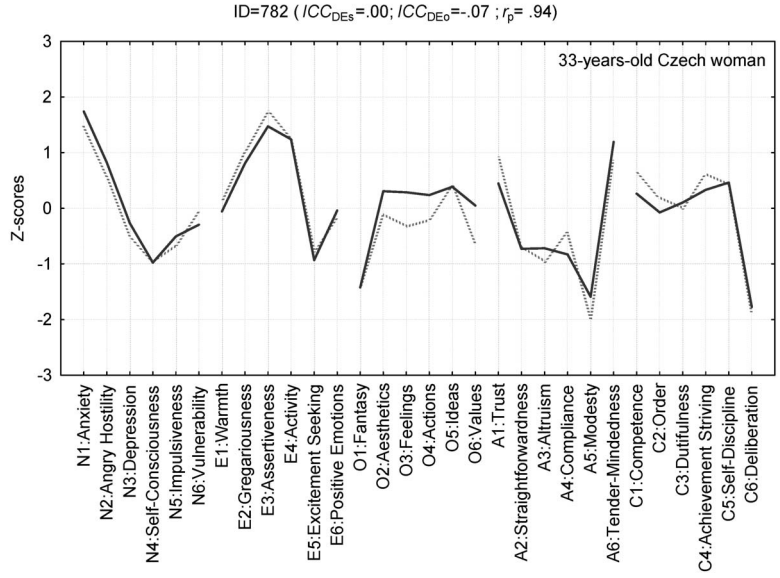

Figure 4C. An example of high self-observer profile agreement and large disparities within factors. The index "s" refers to self- and "o" to observer ratings.

that approximately 95% of all participants had statistically significant ICC_{DE} values either in self- or observer descriptions, indicating that the FFM is an adequate model for most people in at least four rather different countries. Conversely, it also provides a tool for identifying the relatively small group of individuals whose responses deviate from what is in accordance with the standard FFM.

This result also undermines the conviction that variable-centered and person-centered approaches are, if not incompatible, then clearly distinct approaches that represent two different perspectives on human personality (Cervone, 2005). This division has misled some researchers into believing that the structure of covariation established for a group of people tells us almost nothing about the individuals of that group (Borsboom, 2005; Borsboom et al., 2003). However, this was an invalid conclusion as we were indeed able to find a characteristic in the individual pattern of responses which determined, in a very straightforward manner, properties of the group-level factor structure. If a latent factor model fits a given population by explaining a substantial amount of variance, the same model has to fit the majority of individuals of that population. For example, if the FFM explains more than 80% of the variance, the mean ICC will necessarily also be over .70, which leaves a very small fraction for those whose ICC could be below statistical significance. Or, if the mean ICC is above .40, the observed factor structure will necessarily resemble the FFM sufficiently well and the factors will explain at least 60% of the variance. Thus, there is a mathematical relationship between the two fit indices, one of which is applied at the group level and the other at the individual level. Consequently, it is not true that the structure of covariation established between personality traits in a group of people tells us nothing or very little about each subject in the group. The established causal link between within-individual and between-individual levels of description closes the gap that existed between the theoretical explanations that are supposed to apply to individuals and the data that were collected on the basis of between-individual differences. There is good reason to say that the FFM, which has proved to be an adequate description of the basic human personality traits at the group level, has proved to be an equally suitable personality model at the individual level, for most of the people tested. Thus, we can rule out the notion that, even if a given latent factor model fits a population, it does not necessarily fit the majority of the individuals in the population.

For another group of personality researchers, the exposed connection between ICC and group level factor structure may seem trivial and little more than a pedantic replication of what was already known: Good fit at the level of the typical individual follows necessarily from the findings of good fit at the level of the entire group. Is it not so that, when most individuals respond in a manner consistent with the FFM, it is almost inevitable that a nearly perfect structure will emerge in the group-level factor analysis? As strange as it may sound, we are not aware of any other studies exploring possibility of using ICC to test the goodness of fit of a postulated personality model. True, Bem and Allen (1974) proposed using the ipsatized variance index (equivalent to ICC^{-1}) as a criterion of inconsistent responding and Campus (1974) proposed using the eta squared statistic as an index of personality consistency. Besides recognizing that testing the consistency and inconsistency of responses is not the same as testing fit to a theoretical model, researchers have concluded that moderator variable effects in personality are generally small and do not serve to transform weak relationships among personality variables into strong ones (Chaplin, 1991). However, even if it is true that the link between ICC and factor structure is self-evident, one puzzle remains: Why was this rather obvious link between person-fit indices and group-level factor structure missed in previous studies? One possible reason for this oversight is the "error paradigm," which dominated this area of research for several decades (Funder, 1995). Within this paradigm, the person-fit problem was mainly used to single out individuals who deviated from the postulated theoretical model, and not as a measure of the adequacy of the model. Another reason is the focus on one or more personality traits but not the whole structure in general. Finally, person-fit indices have been applied to the response pattern at the level of individual items. As a consequence, unreliable individual answers have resulted in unreliable indicators.

Undeniably, neither ICC_{DE} nor Z_{eig} is an ideal person-fit index. For example, the ICC alone or in combination with Z_{eig} cannot be the only criterion of person-fit for the FFM. Although it is rather unlikely that one would obtain equal scores on all personality dimensions, it is still perfectly compatible with the FFM (Robert R. McCrae, personal communication, November 9, 2001). Nevertheless, the ICC_{DE} would be 0 in the absence of between-factor variance, even if the within-factor variance was relatively small. It is, indeed, counterintuitive that the statistically most probable response pattern – in which all factor means are equal to the population means – represents misfit in terms of the ICC_{DE} but a perfect fit according to IRT and other similar models. Another shortcoming is that the value of ICC_{DE} differs depending on whether it is computed on the basis of raw or normalized scores. The choice between raw and normalized scores is even more perplexed since they are both similarly related to the amount of explained variance. As normalization is a linear transformation, the correlation matrices of raw and normalized traits are identical and consequently must lead to identical percentages of explained variance. However, it is important to notice that, as our study demonstrated, normalization and coding direction of the response scales have much less practical importance than is usually believed.

When it comes to Z_{eig}, we do not have information about how specific it is to the extracted variance in the five-factor solution in particular. It is possible that individuals who have a significant contribution to FFM also have significant contributions to the explained variance in four-, three-, or even one-factor solutions. Thus, we always need an additional judgment or several parallel person-fit indices to increase the reliability of the decisions made about personal fit for a theoretical model. Indeed, there were a minority of cases in which individuals with statistically insignificant ICC_{DE} made

a substantial contribution to the percentage of explained variance of the FFM and vice versa. It is also true that ICC_{DE}, like any person-fit index, is at least partly confounded with response characteristics (e.g., total variance, extreme responding) that are not related to the FFM. Therefore, by combining different but not completely overlapping person-fit indices, it is possible to increase the reliability of identifications of individuals for whom the FFM can be regarded as a sufficiently adequate description of their personality. Nevertheless, only future studies can reveal what percentage of nonfitting individuals can be regarded as acceptable from both a theoretical and a practical standpoint.

The two person-fit indices, ICC_{DE} and Z_{eig}, not only characterized the congruence with the FFM, they also predicted the accuracy of personality judgments. There was less agreement between self- and observer-profiles that alone or both deviated from the canonical five-factor pattern. This seems to provide indirect proof that FFM is indeed a "true" personality description since it is easier to agree on traits that exist rather than produced by imagination.

Among our participants, there was still a sizeable number of individuals who described their personality, or whose personality was described by their acquaintances, in a manner that is incompatible with the FFM according to the ICC_{DE} and Z_{eig}. Nevertheless, due to very high self-observer agreement, these seem to be realistic descriptions of existing combinations of personality traits (e.g., Figures 4B and 4C). These cases, despite being infrequent, suggest that some atypical combinations of personality traits are, in principle, possible. For example, although it is unusual, some people may be anxious and hostile but seldom feel embarrassment. Moreover, in some cases, high levels of friendliness may be associated with a lack of desire to be in the company of others. Unfortunately, this is largely unexplored territory for theorists and practitioners of the FFM. Although the concept of a personality type for a group of people who share a similar combination of personality traits has long been appealing, a continuous dimensional approach has been shown to yield more convincing results (Asendorpf, 2003). Nevertheless, it is still plausible that there are several distinct but internally consistent ways of expressing approximately the same personality disposition.

Implications for Future Research

What are the practical implications of these findings? Even skeptics who disagree with the theoretical rationale that ICC_{DE} is an adequate index of fit have to admit that, at least technically, ICC_{DE} is by far the best predictor of model fit – the proportion of variance explained by five factors. Since it is easy to compute ICC_{DE} on the basis of the individual scores of the NEO PI-R or any other multifactor instrument (besides specialized programs, any one-way ANOVA program computing MS_W and MS_F can be used), we recommend always supplementing the subscale scores with the ICC_{DE} value. Clinical, educational, and other applied psychologists are less interested in individuals whose responses fit the model sufficiently well and more concerned about those who deviate from the expected response pattern. While all IRT-based approaches assume the existence of only one or sometimes a few different dominating response patterns (see Egberink et al., 2010; Rost, 1990), ICC tolerates a great variety of different response profiles. There are numerous response profiles that are all examples of perfect fit to the FFM. On the basis of the mean ICC, it is possible to predict, without actually carrying out factor analysis, what the fit of a given group to the FFM and to determine the likelihood that secondary "wrong" factor loadings will occur. If the ICC_{DE} value happens to be low for a particular individual, it may be a sign of possible deviation from the FFM. We recommend calculating his or her Z_{eig} and if this is also, let's say, among the 10% of the lowest contributions to the extracted eigenvalue, then an independent opinion of an external observer should be obtained. One can only use low ICC_{DE} values to correct clinical or counseling decisions after additional control measures like these have been taken.

Acknowledgments

We are extremely thankful to Kenn Konstabel for writing a Visual Basic program for computing the Z_{eig} values. This project was supported by a grant from the Estonian Ministry of Science and Education (SF0180029s08) to Jüri Allik and by a Primus grant (3-8.2/60) from the European Social Fund to Anu Realo. Writing of this article was supported by a Mobilitas grant from the European Social Fund to René Mõttus (MJD44; via Estonian Science Foundation). Martina Hřebíčková was supported by grant P407/10/2394 from the Grant Agency of the Czech Republic.

References

Allik, J., & McCrae, R. R. (2002). A Five-Factor Theory perspective. In R. R. McCrae & J. Allik (Eds.), *The Five Factor Model of personality across cultures* (pp. 303–322). New York: Kluwer Academic/Plenum.

Allik, J., Realo, A., Mõttus, R., Borkenau, P., Kuppens, P., & Hřebíčková, 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.

Asendorpf, J. B. (2003). Head-to-head comparison of the predictive validity of personality types and dimensions. *European Journal of Personality, 17*, 327–346.

Ashton, M. C., & Lee, K. (2010). Trait and source factors in HEXACO-PI-R self- and observer reports. *European Journal of Personality, 24*, 278–289.

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., . . . De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86*, 356–366.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistency in behavior. *Psychological Review, 81*, 506–520.

Borkenau, P., & Zaltauskas, K. (2009). Effects of self-enhancement on agreement on personality profiles. *European Journal of Personality, 23*, 107–123.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. New York: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–219.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Campus, N. (1974). Transituational consistency as a dimension of personality. *Journal of Personality and Social Psychology, 29*, 593–600.

Cervone, D. (2005). Personality architecture: Within-person structures and processes. *Annual Review of Psychology, 56*, 423–452.

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality, 59*, 143–178.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523–562.

Cohen, J. (1969). r_c: A profile similarity coefficient invariant over variable reflection. *Psychological Bulletin, 71*, 281–284.

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

Egberink, I. J. L., Meijer, R. R., & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality, 44*, 232–244.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51*, 360–392.

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

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26–34.

Hoekstra, H. A., Ormel, J., & DeFruyt, F. (1996). *NEO Persoonlijkheidsvragenlisten NEO-PI-R en NEO-FF-I: Handleigin* [NEO Personality Questionnaire NEO-PI-R and NEO-FF-I: User manual]. Lisse, The Netherlands: Swets & Zeitlinger.

Hřebíčková, M. (2002). Internal consistency of the Czech version of the NEO Personality Inventory (NEO-PI-R). *Ceskoslovenska Psychologie, 46*, 521–535.

Kallasmaa, T., Allik, J., Realo, A., & McCrae, R. R. (2000). The Estonian version of the NEO-PI-R: An examination of universal and culture-specific aspects of the Five-Factor Model. *European Journal of Personality, 14*, 265–278.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277–298.

Konstabel, K., Aavik, T., & Allik, J. (2006). Social desirability and consensual validity of personality traits. *European Journal of Personality, 20*, 549–566.

Lee, K., & Ashton, M. C. (2008). The HEXACO personality factors in the indigenous personality lexicons of English and 11 other languages. *Journal of Personality, 76*, 1001–1053.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509–516.

McCrae, R. R., & Costa, P. T. (2008). Empirical and theoretical status of the five-factor model of personality traits. In G. Boyle, G. Matthews, & D. Saklofske (Eds.), *Sage handbook of personality theory and assessment* (Vol. 1, pp. 273–294). Los Angeles, CA: Sage.

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.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person-fit. *Applied Psychological Measurement, 25*, 107–135.

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science, 18*, 112–117.

Mõttus, R., Allik, J., & Pullman, 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*, 149–157.

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

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65*, 143–151.

Reise, S. P., & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods, 4*, 3–21.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282.

Jüri Allik

Department of Psychology
University of Tartu
Tiigi 78
EE-50410 Tartu
Estonia
juri.allik@ut.ee