JOURNAL OF PERSONALITY ASSESSMENT, *83*(2), 120–127

STATISTICAL DEVELOPMENTS AND APPLICATIONS
SPECIAL SERIES: Personality Measurement

Personality in Proportion: A Bipolar Proportional Scale for Personality Assessments and Its Consequences for Trait Structure

Willem K. B. Hofstee and Jos M. F. Ten Berge
The Heymans Institute
University of Groningen

Trait structures resulting from personality assessments on Likert scales are affected by the additive and multiplicative transformations implied in interval scaling and correlational analysis. The effect comes into view on selecting a plausible alternative scale. To this end, we propose a bipolar bounded scale ranging from –1 to +1 representing an underlying process in which the assessor would review and discount positive and negative behavioral instances of a trait. As an appropriate index of likeness between variables *X* and *Y,* we propose $L_{XY} = \Sigma XY/N$, the average of the raw scores cross products. Using this index, we carried out a raw scores principal component analysis on data consisting of 133 participants who had each been rated by 5 assessors including self on 914 items. Contrary to the Big-Five structure that was found in these data on standard analysis, the results showed a relatively large first principal component F_1 and 2 very small ones, F_2 and F_3. The sizes $L_{FF} = \Sigma F^2/N$, the averages of the squared component scores, were modest to small. It thus appears that the scale, bipolar proportional versus standard, has a profound impact on the size and structure of personality assessments. The dissimilarity remains on analyzing self-ratings rather than averaged (over the 5 assessors) ratings.

We consider personality assessment as a way of communicating about people's traits or qualities. The main parties are the assessors and their audience, consisting of the target persons themselves or third parties. In between, there are investigators represented by instruments for data gathering—questionnaires in a wide sense—and psychometric and statistical procedures for data processing. In this perspective, their task is to facilitate the communication between assessor and audience by asking the right questions, by adequately summarizing the data, removing biases, and the like. We propose that standard procedures for data processing imported from the area of objective tests and measurement and consisting of relative scaling and correlational analysis may not be the most obvious choice from this communicative point of view. We present a bipolar proportional scale as an alternative. We give an empirical demonstration in which the set of alternative procedures is applied. It appears that they have profound implications for prevailing conceptions about the size of individual differences and their structure. We discuss whether these implications are nonetheless acceptable.

REPRESENTING ASSESSMENT DATA

In developing alternative procedures, we refer to the point of view of the assessor (and the audience). Evidently, that perspective represents a construction. We do not claim to be able to look inside the heads of assessors; what we argue is that the construction is recognizable and plausible. We also do not assume that the assessor is necessarily right; we analyze whether the reconstruction is rational. Most notably, we do not assert that relative scaling is wrong. In comparative contexts, for example, in strictly comparative personnel selection, relative models may be preferable over this one.

A Bipolar Scale

The pivot of our assessment model is a conception of the midpoint of a Likert scale (e.g., 3 on a scale ranging from 1 to 5) as a natural zero point ruling out additive transformations. Personality traits or qualities in general are understood in a bipolar fashion: Negations (e.g., unfriendly) do not denote the mere

absence of a trait but its reverse. Opposites have high negative correlations after removal of acquiescence (Hofstee, Ten Berge, & Hendriks, 1998); traits such as unfriendly are rated socially undesirable rather than neutral as would logically be the case with null friendliness. Such negations are thus understood in the manner of a *litotes* (expressing of an affirmative by the negative of its contrary). In the typical case, most people are assessed at the socially desirable side of the midpoint of the scale. Therefore, for persons with a score between the midpoint of the scale and the mean of the distribution, relative scaling would reverse the sign of the assessment. For the others, the shift is less dramatic but still substantial.

One could argue that relative scaling corrects for a bias, namely, socially desirable responding, instead of introducing one. That position, however, is difficult to maintain. First, social desirability is not confined to self-report: On average, all sorts of assessors judge people to be socially desirable, although some are judged more desirable than others, and although assessors differ to some extent. Second, assessors including self agree on the differential social desirability of target individuals, making social desirability a trait (a target person effect) rather than a mere response style (an assessor effect). Thus, an absolute conception of the scale midpoint provides a plausible representation of the assessor's point of view.

We do not presume that personality traits are bipolar, for example, at a behavioral level. Whereas an individual can meaningfully be assessed to be unforgiving—meaning the opposite of forgiving—it may not be so easy to "unforgive" someone whether linguistically or behaviorally. Nonetheless, behavioral and verbal specifications of traits, such as respects others versus looks down on others, appear to show the same bipolar structure as trait adjectives (the example was taken from Hendriks's, 1997, p. 121–122, study, in which these items had high opposite loadings on one and the same factor).

Bounded Scales

Hofstee and Hendriks (1998), on the basis of the preceding argument, proposed the use of scales anchored at the scale midpoint rather than the mean of the distribution. In their procedure, the spread of the scores is set at unity by dividing the deviation scores (from the midpoint instead of the mean) by their standard deviation. The procedure has the advantage of leaving correlational and factor structures untouched. Here, we propose a bounded scale that may be found to come closer to the assessor's perspective.

All data-gathering procedures use bounded scales. One could think of an unbounded scale: Assessors might be instructed to assign any number between plus and minus infinity to a person's friendliness. We do not know of any case in which such scales, whether bipolar or unipolar, have been applied. An obvious reason is that the assessor would have to decide whether John's utter friendliness should be rated at +1,000 or +100,000, which is difficult to do. A related reason is that the ratings of two assessors would be all but incomparable. The problem is not that the number of scale points is infinite: Likert scales may have large numbers of scale points up to infinity if the assessor is instructed to place a mark on a line (a procedure that might have become more popular with automatic computerized scoring). It is unbounded scales that are problematic. Moreover, observed distributions on different traits differ in spread. Relatively neutral traits may span the whole scale, but it would be difficult to find any case in which a person is assessed to be totally or extremely unreliable or murderous. In distributions with small standard deviations, as with clearly desirable or undesirable traits, assessments in the mildly undesirable region may easily become extreme (e.g., minus 3 *SD*s) on transformation. Presumably, that is not what the assessor had in mind; it is definitely not what he or she has said. In other words, the assumption underlying classical standardization—namely, that spreads are arbitrary and carry no meaning—is not fulfilled.

We propose to standardize assessments at the scale end, that is, adopting a bipolar proportion (or percentage) scale running from –1 to +1 (or –100 to +100). Scores X on Likert scales are linearly transformed by taking

$$(X - ½h - ½g)/(½h - ½g), \quad (1)$$

with h the highest possible scale value and g the lowest. In simpler terms, set the scale midpoint at 0, the scale ends at –1 and +1, and interpolate linearly. For example

Rating on 5-point scale	1	2	3	4	5
Transformation	–1	–.5	0	+.5	+1

For another example, any binary scale transforms into [–1, +1].

The proposed convention suggests an underlying assessment model that is well in line with classical notions in personality assessment, namely, Bem and Allen's (1974) summary label interpretation of trait ascription and Buss and Craik's (1983) act frequency approach to personality. According to these notions, the assessor reviews relevant behavioral instances of a trait, for example, cases in which the person has helped others, cheered them up, expressed an interest in them, and the like, when assessing that person's friendliness. In view of the bipolar conception of traits, we add that counterinstances are also relevant to the assessor, for example, unfriendly behaviors such as short-changing others, making them feel uncomfortable, turning one's back on them, and the like.

Consequently, we interpret a score of 4 on a 5-point Likert scale (transformed into +.5 on the bipolar proportion scale with .75 representing the boundary between the scale points +.5 and +1 and .25 the boundary between 0 and +.5) along the following lines: In relevant situations, this individual has shown a clear preponderance of friendly over unfriendly be-

haviors, for example, between 7 versus 1 and 5 versus 3 if the number of situations is 8 so that the proportion of friendly minus the proportion of unfriendly behaviors is at most (7 – 1)/8 = .75 and at least (5 – 3)/8 = .25. Clearly, one should not postulate deliberate and meticulous counts and calculations in the mind of the assessor; all that can be asked for are rough intuitive estimates and a discounting of positive and negative instances subject to all sorts of error. However, the model is at least correct in the sense that it describes what one would like the assessor to do: review and discount concrete instances and counterinstances relevant to the question.

The bipolar proportion model is not restricted to direct trait ratings. At a more specific level, the same estimation process can be postulated. Helping others versus turning one's back on them constitutes a large set of more specific relevant behaviors in relevant situations. For example, turning on the favorite music of a person who is cleaning the room may be found helpful. Even assisting a person crossing the street versus not doing so has subspecifications according to the density of the traffic, the status or demeanor of that person, and so on and so forth.

To summarize the argument thus far, a bipolar proportion scale with a natural zero point in the middle provides a plausible reconstruction of the assessment process. As either additive or multiplicative transformations are arguably inappropriate, our proposal amounts to conceiving of assessment scales as (bipolar) absolute scales. We are aware that the bipolar proportion model may be refined, for example, by allowing differential weights for instances according to their relevance or prototypicality, by taking assessment biases into account such as acquiescent responding and individual differences in that respect (see Hofstee et al., 1998), by establishing the value of intermediate scale points in an empirical manner instead of interpolating linearly, and so on.

Likeness Coefficients

The next problem is choosing a coefficient of association for absolute scales. Hofstee (2002) proposed adopting the most elementary measure of likeness, namely, the averaged cross-product $L_{XY} = \Sigma XY/N$. With scores between –1 and +1, L_{XY} is conveniently bounded within those same limits. As a consequence, $L_{XX} = \Sigma X^2/N$, the extent to which X is like itself, is generally not unity but ≤ 1; $L_{XX} = 1$ only if all ratings are at the extreme scale ends. So L_{XX} functions as an index of saliency: It represents the size of the **X** vector.

The primary argument in favor of L_{XY} is its analogy to $r_{XY} = \Sigma z_X z_Y/N$ for likeness between variables scored on interval scales: Both are averaged cross-products, which is how correlations, interactions, and likenesses in general are represented. For absolute scales, cross-products of the raw scores rather than z scores should be averaged.

Zegers and Ten Berge (1985) presented the identity coefficient $e_{XY} = 2\,\Sigma XY/(\Sigma X^2 + \Sigma Y^2)$, thus $2L_{XY}/(L_{XX} + L_{YY})$, as an appropriate association coefficient for (unbounded) absolute scales. The denominator is needed to keep e_{XY} within bounds. As one might wish to disregard the fact that the bipolar proportion scale happens to be bounded and prefer the identity coefficient, we present a detailed comparison of L_{XY} and e_{XY}.

Unlike r_{XY}, both coefficients are defined at the level of a single pair of observations, permitting sayings such as "In my case, friendliness and socialness do not go together." Both coefficients can thus accommodate the notion of intraindividual structure at a particular point in time, that is, without intraindividual replication. Unlike e_{XY}, however, L_{XY} at the aggregate level is the mean of the individuals' coefficients. L_{XY} is thus perfectly separable in the sense that the likeness coefficient at the individual level stays the same after aggregation; with e_{XY}, the denominator changes on taking other cases into account.

Using L_{XY}, two individuals of equal but moderate friendliness are less alike in that respect than two individuals with pronounced friendliness: More extreme or salient scores obtain higher weights. For example, L_{XY} is higher for two individuals with scores +.2 and +.7 than for two individuals with scores +.2 and +.2. Using e_{XY}, such effects would be corrected on calculating individual coefficients; they would, however, return if the coefficient is calculated at the aggregate level and then split up, as the denominator is a constant at that level.

A dramatic difference arises on comparing the sizes of the two coefficients. In the general case, $L_{XY} < e_{XY}$ as ΣX^2 and ΣY^2 are smaller than N. Only if all scores are +1 or –1, as would automatically be the case with a binary scale transformed into the bipolar proportion scale, would the two coefficients be identical. They would degenerate into a coefficient proposed by Holley and Guilford (1964), among others, which takes the value of the diagonal proportions minus the off-diagonal ones in the fourfold table (proportion agreement p_A minus proportion disagreement, or $2p_A - 1$). In the general case, size or saliency influences the likeness coefficient, whereas the identity coefficient disregards those aspects. Using continuous scales, assessments of ±.5 would be typical rather than extreme ones, so the typical likeness coefficient would be about four times as small as the corresponding identity coefficient. Another consequence is that variables with relatively small sizes L_{XX} will have relatively little impact on the multivariate structure of sets of variables as in factor analysis. The question is how to appreciate these properties of the likeness coefficient.

From the point of view of representing the concept of likeness, the coefficient may be found quite defensible. Take an example in which two individuals are equally but only vaguely at the friendly side of the scale; say their scores are .1, indicating a 55 to 45 preponderance of friendly over unfriendly behaviors. One may envisage a thought experiment in which the two persons would have been observed to behave in a friendly or unfriendly manner in the same situations, creating a 2 × 2 table with 55–45 marginal proportions.

Given behavioral inconsistencies and observer error and the resulting low correlations between situations, the off-diagonal proportions would be sizeable and the overlap would be modest. Conversely, extreme marginals for either or both variables automatically create more overlap and higher likeness coefficients. So to the extent that likeness or association is based on overlap rather than statistical dependence, L_{XY} captures it.

We submit that assessors and their public overwhelmingly opt for overlap rather than statistical dependence in defining likeness or even correlation. We base our prediction on informal classroom experiments that are easily replicated. Present people with fourfold tables with extreme marginal frequencies, for example, a table with 90 and 0 in the diagonal cells, and 5 in both off-diagonals (so that $r_{XY} = -.05$ and $L_{XY} = +.80$), and they will say that the two variables are clearly positively related. It does not even help much if the respondents have followed a course in applied statistics. One may of course object that they are mistaken and that coefficients that reflect overlap are wrong because they need not be zero on statistical independence. However, that argument is not nearly as straightforward as it may look. The layperson's point of view receives support from a classical argument mostly referred to as Meehl's paradox (Meehl & Rosen, 1955).

Take a diagnostic setting (rather than a comparative selection setting) in which clients or students are assessed on some trait or quality with a heavily skewed base rate, say 95% positive and 5% negative according to some reasonable criterion. Take two assessors, one of whom displays a 95–5 selection rate but a maximally negative validity ($r = -.05$, see previously), whereas the other has a 50–50 selection rate but maximally positive ($r = .23$) validity. By the rational standard consisting of proportions of correct diagnoses (90% vs. 55%), the first assessor outperforms the second. The difference is reflected in the values of L_{XY}, which are .80 and .10, respectively. Surely, a random procedure with a 95–5 selection rate would work even better than the first assessor, but that is because the procedure has been fed with prior knowledge: It has been told to take the base rate as its selection rate. A truly random procedure, which would have to draw its own selection rate in some random fashion, would be vastly inferior to the first assessor. In other words, the assessor is credited for choosing the right selection rate apart from relative validity. Similar arguments apply if the scale is continuous rather than dichotomous.

In conclusion, we propose to adopt L_{XY} as the most elementary and appropriate measure of likeness for bipolar proportion scales. The identity coefficient for absolute scales caters to the situation in which variables, for example, tests with different numbers of items, have arbitrary sizes and would be prevented from being perfectly associated because of that. No such provision is necessary with proportions. On the contrary, correction for size appears to detract from the representing of likeness between variables scored on a bipolar proportion scale.

Multivariate Structure

In this context, multivariate analysis is primarily a procedure for summarizing assessment data. On one hand, it is considered good practice not to ask assessors questions at a high level of abstraction but to spell out concepts by means of a questionnaire containing more concrete instances. On the other, the audience would not be served by receiving an answer sheet; the data had better be summarized. The obvious way of summarizing is to take an average (taking signs into account) of the scores on the relevant items. An objection, however, is that some items are more relevant than others so that weighting them would be more appropriate. The objection is of limited practical importance if the number of items is large, as weighted and unweighted sums are very much alike in that case. On the other hand, any gains from weighting are obtained for free with computerized scoring. Secondarily, multivariate analyses have theoretical spin-off: They provide insight into the multivariate structure of personality.

In the absence of external criteria for the differential relevance of questionnaire items, one would weight them according to their likeness to the (unweighted) total score on the relevant items. To keep weighted averages on the same bipolar proportion scale as the item scores, the weights w_j should be divided by the sum of their absolute values $\Sigma|w_j|$. On having assigned weights, however, the total score has been implicitly transmuted into a weighted average, so logic dictates weighting of the items according to their likeness to that weighted average. Consequently, an iterative process should be carried out, which ends on sufficient convergence. In practice, that procedure amounts to calculating the scores on the first principal component of the raw scores data matrix (see Horst, 1965). It maximizes the sum of the squared likenesses of the items with the weighted average. The second principal component does that with respect to the residual scores matrix, and so on. We thus conceive of principal components as weighted averages of variables and of principal component analysis as the way to find optimal weights. Finally, item loadings are calculated as likeness coefficients L_{XF} between item X and principal component F. MATLAB® (MATLAB Inc., 1994) routines for carrying out these procedures are available from Jos M. F. Ten Berge. Using standard programs, one would (a) transform scores onto the bipolar proportion scale, (b) calculate the matrix of average cross-products L_{XY}, (c) find the principal components of L, (d) find the corresponding component weights, (e) divide the weights for each component by the sum of their absolute values, (f) find the component scores, and (g) find the loadings L_{XF} on the principal components.

As an aside, we note that our approach resolves the dispute between proponents of a person centered (Magnusson, 1992) or typological conception of personality structure and the dominant variable-centered conception. Raw scores principal component analysis not only produces both matrices of factor scores and factor loadings, but contrary to standard

factor or component analysis, it is impartial, as it does not standardize scores in one direction (Q-analysis) or another (R-analysis). The bipolar proportion scale meets a classical (Cattell, 1944) plea for an interactive rather than either a normative or an ipsative conception of scores. As a matter of fact, one may conceive of the entries in a matrix of assessment data as likeness coefficients between individuals and variables, decomposable through principal component analysis. This rectangular or off-diagonal matrix of likeness or proximity relations (see Coombs, 1964) has the same principal component scores and loadings as the triangular matrices of likenesses between persons and likenesses between variables, respectively.

Another ramification is $N = 1$ principal component analysis. This novelty may find its way in situations in which the data do not form a matrix as in the common case in which each individual has been rated by a different set of assessors. In the minimal and most radical application, one individual has been rated by two assessors on one variable, say friendliness; the question is how these ratings are weighted in an optimal fashion. Take the following example:

	Assessor 1	Assessor 2	Average
Scores	1.0	–.5	.25

The first approximation to the average score is the unweighted average. To find the next approximation, we find the likenesses L between assessors and average:

Likeness	.25	–.125

We rescale the likenesses into weights with $\Sigma|w| = 1$:

Weights	.667	–.333,

so that the weighted average is now .833.

We could iterate the procedure, but in this elementary case, that would give the same weights, so convergence is immediate. What the example demonstrates, apart from the fundamentals of principal component analysis, is that the weighted average is drawn toward the more extreme assessment. As the likeness coefficient between a rating and the average rating is proportional to the extremeness of that rating, so is the weight. Note also that the second assessor receives a negative weight.

Evidently, analyses with small numbers capitalize heavily on chance. In the following, we present a less radical application in which the number of items is large.

REPRESENTING PERSONALITY

We applied the methodology set out previously to data collected by Hendriks (1997, p. 35, and following) in the process of constructing the Five-Factor Personality Inventory (FFPI; see also Hendriks, Hofstee, & De Raad, 2002). We used the ratings on a 5-point scale of 133 target persons, mostly first-year students of psychology at the University of Groningen, on 914 sentence items (e.g., "Keeps apart from others") by self and four others, forming 133 matrices of 5 × 914 assessments.

Weighting Assessors

All ratings were linearly transformed onto the bipolar proportion scale following Equation 1. For each of the 133 target persons, a 914 × 5 matrix of such scores was available. A raw scores principal component analysis using L as an index of association was applied to each of these 133 matrices to find optimal weights for averaging the five raters in question. The five assessors' weights on the first principal component were divided by the sum of their absolute values. The end result of this operation was a matrix of 133 × 914 averaged assessments of the target participants on the bipolar proportional scale.

The obtained weights are interesting by themselves. In the first place, only 2 of the 133 × 5 assessor's weights were negative (–.01 and –.02), which affirms the high quality of Hendriks' (1997) data. In the second place, self-ratings contributed slightly but significantly less than others' ratings: Whereas the overall average (absolute) weight is .20 as a consequence of the procedure, 84 self-weights were below that figure and only 49 above, $\chi^2(1, N = 133) = 9.21$, $p < .01$. If one accepts a definition of personality (cf. Hofstee, 1994) as the common component in the assessments of the population of relevant judges, these results point to a relative inferiority of self-reports amidst others' assessments. The result should be surprising to those who conceive of self as sharing more relevant information with others than do others amongst each other: Apparently, that pivotal position does not help enough. Theoretically, self-assessments could still have superior external validity; however, real-life criteria also tend to be in the hands of third persons.

Structuring Personality Assessments

The resulting 133 × 914 matrix of target individuals by items was in its turn subjected to raw scores principal component analysis using L as an association index. Five principal components were extracted to facilitate comparisons with the original solution. The sum of the absolute values of the 914 item weights per principal component was set to 1 to obtain component scores on the bipolar proportion scale.

In a standard principal component analysis using z scores and correlations, the first 5 eigenvalues were 210.2, 115.6, 56.6, 44.1, and 31.0. Thus, the first eigenvalue was 1.8 times as high as the second. In the raw-scores analysis, the first 5 eigenvalues of the matrix of L coefficients were 60.56, 9.62, 4.45, 3.40, and 2.56. Here, the first eigenvalue was 6.3 times the second. An additional size index in raw scores principal component analysis is $L_{FF} = \Sigma F^2/N$, the mean of the squared component scores. For the first 5 principal components, the sizes are .0866, .0156, .0081, .0056, and .0045. According to

this index, the first principal component is 5.6 times as big as the second. We emphasize that these high figures are not an artifact in the sense of an automatic consequence of adopting a particular scale. That would be the case with a unipolar scale leading to all-positive L_{XY}s. With a bipolar scale, however, there is no such restraint: If item scores would be symmetrically distributed around the zero point, results from raw scores principal component analysis would not differ much from standard outcomes. The dramatic increase in the relative contribution of the first principal component reflects the dominant role of individuals' social desirabilities in personality assessments, which is partly suppressed in standard analysis.

On interpreting these scores, their size should be kept in mind. Scores between –.25 and +.25 are in the neutral range indicating neither a positive nor a negative likeness between the individual and the principal component. Of the 133 participants, 47 (35%) had neutral scores on all five principal components. Of the remaining scores, none even approached the extreme (> +.75 or < –.75) ranges; only two scores were (slightly) above .5.

Of the 86 participants with nonneutral scores, 79 (92%) had their highest absolute score on the first principal component, which may be interpreted as (un)desirable social behavior; it is mostly negatively defined by items such as takes advantage of others, abuses people's confidence, insults/hurts people, treats people as inferiors, and cuts others to pieces. Most of these items were markers for the negative pole of Factor 2, Mildness, in Hendriks' (1997) study. Note that this is not social (un)desirability in the self-presentational sense but undesirable social behavior in a moral perspective. Of the 79 nonneutral scores on this component, the large majority (75, being 95%) were positive; 4 were negative, ranging from –.27 to –.45. Of all 133 participants, 127 (again 95%) had positive scores on the first principal component. If this sample is at all representative, our Orwellian saying about people's desirabilities has to be qualified somewhat: By far the most people are weakly to mildly socially desirable; a few are weakly to mildly undesirable.

Four participants had nonneutral scores on the second principal component of which three were positive and one negative. A number of 56 (out of 914) items had their highest loading on this component, constituting a self-willed type that takes charge, wants to pull the strings, makes his/her own rules, wants to have it his/her own way, seeks confrontations, and knows how to manipulate a situation. Three participants had nonneutral scores on the third principal component of which two were negative. On this component, only 17 items had their highest loading, such as needlessly worries a lot and is afraid that he/she will do the wrong thing, versus readily overcomes setbacks and can take his/her mind off his/her problems; according to Hendriks' (1997) research, the latter two are marker items for Emotional Stability. On the fourth and fifth principal component, all 133 scores were in the neutral range.

As might be expected given the large size of the first principal component, varimax rotation of the matrix of component scores did not bring about great changes; all diagonal elements in the rotation matrix were above .8. The number of participants with neutral scores on all components rose from 47 to 54. Of the 70 participants with a nonneutral score on the first rotated component, 4 were negative (these belong to the same participants as before rotation). The numbers of participants having their highest absolute score on the second through fifth rotated component were 3, 3, 2, and 1, bringing the total back to 133. Before and after rotation, 3 participants had nonneutral scores on both the first and the second component, so the rotation did also not result in a simpler person structure.

Single-Source Analysis: Self-Assessments

We consider aggregation over assessors to be the standard for personality assessments, if only to compensate for assessor error. However, single-source perceptions of personality may be of interest as such, for example, if self-assessments are to be confronted with other assessors' points of view as in therapeutic or personnel management settings. Evidently, the findings from the aggregated data cannot be automatically generalized to single-source data. We therefore carried out a supplementary analysis on just the 133 self-ratings in the Hendriks (1997) study using the same methodology.

Using standard principal component analysis, the first 5 eigenvalues were 112.3, 84.6, 48.6, 42.4, and 32.0. Therefore, the first eigenvalue was 1.33 times the second. In the raw-scores analysis, the first 5 eigenvalues were 47.47, 14.42, 7.78, 6.50, and 5.51. Therefore, the first eigenvalue was 3.33 times the second. The sizes $L_{FF} = \Sigma F^2/N$ of the first 5 principal components were .0688, .0223, .0135, .0109, and .0095. Thus, according to both criteria, the first principal component was more than three times as large as the second. The relative increase of the first principal component as a result of absolute scaling is less than in the aggregated data but still sizeable.

Of the 133 participants, 71 (53%) had neutral scores on all five principal components. This is far more than the 35% in the aggregated data and reflects the smaller sizes of the principal components in the self-assessments. Still, the component scores ran higher, the three largest scores being between .65 and .70. Thus, any extreme item scores that were present in the self-assessments were not smoothed out as much as happens on aggregation over assessors; but even in these data, extreme (> +.75 or < –.75) component scores did not occur.

Of the 62 participants with nonneutral scores, 44 (71%) had their highest absolute score on the first principal component; for the other components, the numbers were 10, 3, 2, and 2, respectively. The interpretation of the principal components in terms of their highest loading items remained virtually unchanged. Clearly, however, the dominance of the first principal component was somewhat mitigated in the

self-assessments. Of all 133 scores on the first principal component, only 4 (3%) were negative; none of the nonneutral scores were. These figures were in line with the results from the aggregated analysis.

The single-source analysis was based on only 133 assessors, so not all aspects of the results should be expected to generalize to other samples. The main results from the aggregate analysis, namely, absence of principal component scores in the extreme regions and a large relative boost of the first principal component, are replicated in the self-assessments, although the effects are less extreme than in the aggregate analysis.

DISCUSSION

Our procedures render an unromantic turn to the study of personality. Little remains of the shades and subtleties of individual differences in temperament and character. By far the most people appear to be faintly to mildly okay, a few are not, and a handful (in the self-assessments, a sizeable minority) may better be characterized in other terms; that is about all there is to it. It is as if we are reminded of the fact that we share 99% of our genes with primates and almost all with one another. How do these outcomes relate to our educated intuitions about personality?

Size of Individual Differences

Perhaps the most counterintuitive outcome is that extreme scores were not observed: Even on varimax rotation of the component scores matrix, the highest score of all fell short of .55 in the aggregate analysis. Few persons should thus be expected to be more than halfway agreeable, conscientious, and the like. However, don't we all know individuals who seem to be much more extreme than that in one respect or another? The answer is undoubtedly affirmative, for if assessors rate individuals directly on traits, a sizeable fraction of these ratings will be at the most extreme scale points.

However, the contradiction is easily resolved. First, assessors do not agree all that much. On averaging their ratings, regression toward the midpoint of the scale is virtually automatic. It occurs even if extreme assessments receive greater weights as was the case in this analysis. Who is right, the individual assessor or the collective? That depends on the frame of reference. In a poetic or romantic context, truth is subjective; one's individual judgment is all that counts. However, in an intersubjective context, we have to account for the well-established fact that even our own judgments about ourselves and about others had better be subjected to statistical regression. Hendriks' (1997) exemplary data involving five assessors per target are much more relevant and authoritative in this respect than our intuitions, scientifically speaking.

Second, the component scores in this study came about as a weighted average over 914 items or facets of personality-relevant behavior, each of which triggers another aggregation operation at a more specific level. Human behavior is not very consistent. In the supreme autonomy of our private intuitions, we can easily discount the fact that the prototypical extravert is seen sitting in a corner. However, scientifically speaking, we can only acknowledge that a score of .5 corresponding with a 75–25 preponderance of positive over negative instances of a trait is probably a whole lot. On second thought and in view of the additional imperfect assessors' reliability, one may even become distrustful the other way and wonder whether such high scores do not represent capitalizations on chance (which they probably do to some extent).

Personality Structure

Next, the results run counter to the bigness of factors beyond the first principal component. The dominant impression of a highly differentiated, five-dimensional structure (see, e.g., De Raad, 2000; De Raad & Perugini, 2002) came about through the combined use of standard scores and varimax rotation, which spreads their variance over factors. With the bipolar proportion scale, the results are varimax resistant. That is not because the content of the principal components changes much on using an absolute scale. We calculated a standard Big-Five solution (five principal components plus varimax) on Hendriks's (1997) data and carried out a multiple regression for each of these factors using the five raw scores principal components as predictors. The multiple correlations ranged between .983 and .999, indicating excellent coverage of the standard Big-Five space.

This change in perspective is reminiscent of what happened to the concept of intelligence. In the tradition of Thurstone (1938) to Guilford (1967), intelligence was conceived as multifaceted and complex. Toward the end of the 20th century, the hierarchical conception with *g* at the top took over (see, e.g., Herrnstein & Murray, 1994). The intercorrelations of personality items, if scored in socially desirable direction, will be found to be in the same order of magnitude as the intercorrelations of intelligence items. Thus, even on adhering to relative scales and corresponding statistics, a general *p* factor of personality (Hofstee, 2003) seems to deserve serious consideration. As with intelligence, the implication is not that there is nothing more to personality than the first principal component. The implication is that other dimensions are of secondary importance. The overall picture that arises is a positive manifold, bipolar in the case of personality assessment, according to which lower level concepts form an oblique structure: a sort of multidimensional double cone (Peabody & Goldberg, 1989) but with a spatial angle of less than 90°.

As we noted at the beginning of this article, relative scales and statistics are relevant in comparative contexts; therefore, so are well established trait structures such as the Big Five. It thus looks as if we have two conceptions of trait structure,

one relative and the other absolute. Could the true structure lie in between? That question amounts to asking for contexts that have both relative and absolute features. Those situations may well constitute the general case. Even in comparative personnel selection, one would like to make sure that the best candidate is at all fit for the job; in selecting a prizewinner, the jury may decide not to award the prize. On the other hand, the central tendency of a trait's distribution tends to induce anchoring effects and therefore, a relative component. So both strictly absolute and relative scaling may be found to be extreme cases. The development of mixtures of absolute and relative scales is beyond the scope of this article. At this stage, it may suffice to suggest that standard structures based exclusively on relative scaling are not located in the middle of the road but at its edge, the other edge being occupied by the bipolar proportion model.

ACKNOWLEDGMENTS

The content of this article was presented as an invited paper in the symposium on Personality Types Versus Personality Dimensions at the Eleventh European Conference on Personality, July 2002, in Jena, Germany. We have profited a great deal from incisive comments by anonymous reviewers.

REFERENCES

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

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review, 90,* 105–126.

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51,* 292–303.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

De Raad, B. (2000). *The Big Five personality factors: The psychological approach to personality.* Göttingen, Germany: Hogrefe.

De Raad, B., & Perugini, M. (2002). *Big Five assessment.* Göttingen, Germany: Hogrefe.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Hendriks, A. A. J. (1997). *The construction of the Five-Factor Personality Inventory (FFPI).* Unpublished doctoral dissertation, University of Groningen, Groningen, The Netherlands.

Hendriks, A. A. J., Hofstee, W. K. B., & De Raad, B. (2002). The Five-Factor Personality Inventory: Assessing the Big Five by means of brief and concrete statements. In B. de Raad & M. Perugini (Eds.), *Big Five assessment* (pp. 79–108). Göttingen, Germany: Hogrefe.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve.* New York: Simon.

Hofstee, W. K. B. (1994). Who should own the definition of personality? *European Journal of Personality, 8,* 149–162.

Hofstee, W. K. B. (2002). Types *and* variables: Towards a congenial procedure for handling personality data. *European Journal of Personality, 16,* 89–96.

Hofstee, W. K. B. (2003). Structures of personality traits. In I. B. Weiner (Series Ed.) & T. Millon & M. Lerner (Vol. Eds.), *Handbook of psychology: Vol. 5. Personality and social psychology* (pp. 231–254). Hoboken, NJ: Wiley.

Hofstee, W. K. B., & Hendriks, A. A. J. (1998). The use of scores anchored at the scale midpoint in reporting people's traits. *European Journal of Personality, 12,* 219–228.

Hofstee, W. K. B., Ten Berge, J. M. F., & Hendriks, A. A. J. (1998). How to score questionnaires. *Personality and Individual Differences, 25,* 897–909.

Holley, J. W., & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement, 24,* 749–753.

Horst, P. (1965). *Factor analysis of data matrices.* New York: Holt.

Magnusson, D. (1992). Back to the phenomena: Theory, methods, and statistics in psychological research. *European Journal of Personality, 6,* 1–14.

MATLAB, Inc.. (1994). MATLAB (Computer software). Natick, MA: Author.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

Peabody, D., & Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptors. *Journal of Personality and Social Psychology, 57,* 552–567.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 1.* Chicago: University of Chicago Press.

Zegers, F. E., & Ten Berge, J. M. F. (1985). A family of association coefficients for metric scales. *Psychometrika, 50,* 17–24.

Willem K. B. Hofstee
The Heymans Institute
University of Gronigan
Grote Kruisstraat 2–I
9712 TS Groningen
The Netherlands
E-mail: w.k.b.hofstee@ppsw.rug.nl

Received November 4, 2002
Revised October 2, 2003