PSYCHOMETRIKA—VOL. 30, NO. 4
DECEMBER, 1965

MATCHING COMPONENTS OF SELF-REPORT AND PEER-NOMINATION PERSONALITY MEASURES

WARREN T. NORMAN
AND
THAD R. HARSHBARGER

UNIVERSITY OF MICHIGAN

Two common methods of measuring personality variables are examined with respect to the ways in which they control respondent biases. A form of the mental-test theory model is developed which makes explicit differences in score components that result from different methodological constraints imposed by the two methods. From the model, it is possible to specify the operations necessary to provide equivalent information from the two sets of data. Performance of these operations in an empirical test supports the model by producing generally higher correlations between the variables having score components which are more closely matched.

Under one form of the mental-test theory model, an individual's score on a personality test can be considered to be a function of three independent types of components: the true status of the individual on the attribute the test is intended to measure, other systematic sources, and a random error component. In recent years, there has been a growing interest in systematic effects on test scores deriving both from subject characteristics other than the intended attribute and from methods of stimulus presentation, scoring, and the like. Techniques have been developed to minimize the effects of certain of these systematic components. However, efforts to control for one source of systematic bias frequently introduce other, easily overlooked, varieties of contaminants into test scores. The present paper attempts to show that two commonly used forced-response procedures, each employed to control *respondent biases* in its respective class of assessment devices, result in the introduction of distinct kinds of *methodological bias* in the two sets of scores. It then proposes a series of transformations to be applied to the two classes of measures to being them into closer structural alignment and presents the results of an empirical test which supports the analysis and provides an indication of the relative magnitude of the effects attributable to the several components considered.

The Problem

The two kinds of instruments used in the present study were (i) a set of peer-nomination rating scales developed by Cattell and found by Tupes and Christal [5] and by Norman [3] to assess five orthogonal personality factors, and (ii) a forced-choice, self-report test, the Descriptive Adjective Inventory, developed by Norman to assess the same five factors. These devices are described in detail in two previous papers [2, 3].

Peer-nomination ratings had been collected within small groups under instructions to nominate a fixed percentage of peers in each rating category. The result of this forced-nomination procedure was that no between-group differences which might have existed on the attributes were reflected in these data. Self-report data, however, were collected on an individual basis and required of each respondent a forced-choice on each item between adjectives representing different factors. It was therefore possible for group averages on the latter measures to vary.

On the other hand, there was no restriction imposed in collecting the peer-nomination data concerning which people were to be nominated in the several categories of each scale. A given person conceivably could be nominated at, say, the positive pole on all scales by all other members of his group, thus getting maximum scores on all five of the peer-nomination measures. This latter condition is in contrast to that which obtained for the self-report scores. The method of item construction and keying for this self-report inventory led to scores that were partially balanced with respect to one another, so that persons having high scores on one factor tended to have correspondingly lower scores on the remaining factors.

The problem was thus to develop a model to represent the structural components present in each of the two types of measures and to derive from it analysis procedures to place the two sets of data under the same methodological constraints. Increases in correlation as one or another of these corrections was made were taken to indicate that the corresponding score component or components had been at least partially accounted for.

The Model

Let P_{igjk} denote the raw score for person i in group g on factor j assessed on occasion k by use of the peer-rating scales;

$$\begin{aligned} i &= 1, 2, \cdots, N_g \text{ (persons within groups)},\\ g &= 1, 2, \cdots, G \text{ (groups)},\\ j &= 1, 2, \cdots, J \text{ (factors)},\\ k &= 1, 2, \cdots, \infty \text{ (occasions)}. \end{aligned}$$

Let S_{igjk} denote the raw score for person i in group g on factor j assessed on occasion k by use of the self-report inventory.

Let T_{igj} (capital Greek letter "tau") denote the true score of person i in group g on factor j.

Averages will be indicated by replacing the appropriate subscript or subscripts over which summations are taken with a dot (.). Thus

$$\mathrm{T}_{ig.} = \frac{1}{J} \sum_{j=1}^{J} \mathrm{T}_{igj}$$

denotes the "average elevation" of the true scores of person i in group g over the J factors. The letters Γ, Λ, and Ω will be employed with appropriate subscripts to denote the random error components of the several kinds of measures. Upper-case letters will be used for raw-score variables and their components while corresponding lower-case letters will be used for deviation scores (computed about expectations on i and g) and their components.

The self-report inventory scores, S_{igjk}, can be represented in terms of components by

$$S_{igjk} = \mathrm{T}_{igj} - \Delta\mathrm{T}_{ig.} + \Lambda_{igjk} ,$$

where

$\Delta\mathrm{T}_{ig.}$ = the portion of the average elevation, $\mathrm{T}_{ig.}$, removed by the counterbalancing of the test stimuli,

Λ_{igjk} = error component.

Converting to deviation score form we obtain

$$\begin{aligned} s_{igjk} &= S_{igjk} - S_{..jk} \\ &= (\mathrm{T}_{igj} - \Delta\mathrm{T}_{ig.} + \Lambda_{igjk}) - (\mathrm{T}_{..j} - \Delta\mathrm{T}_{...} + \Lambda_{..jk}) \\ &= (\mathrm{T}_{igj} - \mathrm{T}_{..j}) - \Delta(\mathrm{T}_{ig.} - \mathrm{T}_{...}) + (\Lambda_{igjk} - \Lambda_{..jk}). \end{aligned}$$

Adopting the deviation-score notation for components on the right we obtain

(1) $$s_{igjk} = \tau_{igj} - \Delta\tau_{ig.} + \lambda_{igjk} .$$

Similarly the model for the peer-rating scores, P_{igjk}, can be represented by

$$P_{igjk} = \mathrm{T}_{igj} - \mathrm{T}_{.gj} + \Gamma_{igjk} ,$$

where

$\mathrm{T}_{.gj}$ = location parameter for group g on factor j which, by virtue of collecting these data by a forced-nomination procedure within peer groups, does not contribute to these data,

Γ_{igjk} = error component.

Again expressing both sides in deviation score form we obtain

(2) $$p_{igjk} = \tau_{igj} - \tau_{.gj} + \gamma_{igjk} .$$

The first step was to remove from the s_{igjk} the remainder of the elevation component which had not been eliminated by the partial counter-balancing technique empolyed in item construction. This was necessary because the amount $(1 - \Delta)\tau_{ig.}$ could not be matched by corrections applied to the p_{igjk} whereas the quantity $\tau_{ig.}$ could. The correction was performed by subtracting the average elevation of an individual's self-report factor scores from each of his factor scores:

$$
\begin{aligned}
(3) \qquad s'_{igjk} &= s_{igjk} - s_{ig.k} \\
&= \tau_{igj} - \Delta\tau_{ig.} + \lambda_{igjk} - \tau_{ig.} + \Delta\tau_{ig.} - \lambda_{ig.k} \\
&= (\tau_{igj} - \tau_{ig.}) + (\lambda_{igjk} - \lambda_{ig.k}).
\end{aligned}
$$

The second step was to remove the elevation component from the p_{igjk} to match more closely the s_{igjk}. That is,

$$
p'_{igjk} = p_{igjk} - p_{ig.k} .
$$

The model for p'_{igjk} thus becomes

$$
\begin{aligned}
(4) \qquad p'_{igjk} &= \tau_{igj} - \tau_{.gj} + \gamma_{igjk} - \tau_{ig.} + \tau_{.g.} - \gamma_{ig.k} \\
&= (\tau_{igj} - \tau_{ig.} - \tau_{.gj} + \tau_{.g.}) + (\gamma_{igjk} - \gamma_{ig.k}).
\end{aligned}
$$

The third step was to remove differences between groups from the self-report scores. The differences between the mean self-report score for each group and the mean for the total sample was computed for each factor, and this group-difference score was subtracted from the score of each member of the group on that factor. Thus self-report data were impoverished of between group differences, matching the constraint of the peer-nomination data. That is,

$$
s''_{igjk} = s'_{igjk} - (s'_{.gjk} - s'_{..jk}).
$$

The model for s''_{igjk} thus becomes

$$
\begin{aligned}
(5) \qquad s''_{igjk} &= (\tau_{igj} - \tau_{ig.}) + (\lambda_{igjk} - \lambda_{ig.k}) - (\tau_{.gj} - \tau_{.g.}) - (\lambda_{.gjk} - \lambda_{.g.k}) \\
&= (\tau_{igj} - \tau_{ig.} - \tau_{.gj} + \tau_{.g.}) + (\lambda_{igjk} - \lambda_{ig.k} - \lambda_{.gjk} + \lambda_{.g.k}),
\end{aligned}
$$

since $s'_{..jk}$ is a constant for all subjects in all groups.

Matching the two kinds of measures on the group location parameters was also accomplished in a second way; by augmenting the peer-nomination scores using a direct estimate of these parameters derived from group ratings.

Let $H_{.gjk}$ denote the raw score of group g on factor j assessed on occasion k by use of the group-rating scales. (The dot in place of the subscript i reflects our assumption that these ratings, though collected with direct reference to the various groups, in fact represent the average of the group members on the factor.) Thus, in terms of components of the model

$$
H_{.gjk} = \mathrm{T}_{.gj} + \Omega_{.gjk} ,
$$

where

$$\Omega_{.gik} = \text{error component.}$$

Expressed in terms of deviations this variable becomes

$$h_{.gik} = \tau_{.gi} + \omega_{.gik} \; . \tag{6}$$

These estimates of group location were added to the original peer-nomination scores, i.e.,

$$p''_{igjk} = p_{igjk} + h_{.gik} \; ,$$

which in terms of the structural components of p and h becomes

$$\begin{aligned} p''_{igjk} &= \tau_{igj} - \tau_{.gj} + \gamma_{igjk} + \tau_{.gj} + \omega_{.gjk} \\ &= \tau_{igj} + \gamma_{igjk} + \omega_{.gjk} \; . \end{aligned} \tag{7}$$

Augmenting the p' scores is most appropriately done by using group location estimates from which elevation components have first been removed. Thus

$$\begin{aligned} h'_{.gjk} &= h_{.gjk} - h_{.g.k} \\ &= \tau_{.gj} + \omega_{.gjk} - (\tau_{.g.} + \omega_{.g.k}) \\ &= (\tau_{.gj} - \tau_{.g.}) + (\omega_{.gjk} - \omega_{.g.k}), \end{aligned} \tag{8}$$

and

$$\begin{aligned} p'''_{igjk} &= p'_{igjk} + h'_{.gjk} \\ &= [(\tau_{igj} - \tau_{ig.} - \tau_{.gj} + \tau_{.g.}) + (\gamma_{igjk} - \gamma_{ig.k})] \\ &\qquad\qquad + [(\tau_{.gj} - \tau_{.g.}) + (\omega_{.gjk} - \omega_{.g.k})] \\ &= (\tau_{igj} - \tau_{ig.}) + (\gamma_{igjk} - \gamma_{ig.k}) + (\omega_{.gjk} - \omega_{.g.k}). \end{aligned} \tag{9}$$

The comparisons listed in Table 1 were carried out in a preliminary test of the model. A rough ordering of the correlations between various pairs of variables was constructed, based on the following assumptions. (i) Unmatched non-error score components attenuate correlations. (ii) Since differences among the $\tau_{ig.}$ reflect inter-individual, within-group variability while differences among the $\tau_{.gj}$ reflect variation among group means, the attenuation due to unmatched $\tau_{ig.}$ should be greater than that caused by unmatched $\tau_{.gj}$. (iii) Since the component $\tau_{.g.}$ is itself in turn a component of both $\tau_{ig.}$ and $\tau_{.gj}$, an unmatched componentjof the form $(\tau_{ig.} - \tau_{.g.})$ has an attenuating effect less than one of the form $\tau_{ig.}$ alone.

On this basis, one would predict correlations that increase in size as one reads from top to bottom in Table 1; except for the last two rows, between which no prediction can be made on the above basis.

TABLE 1

Comparisons Made In Order of Increasing Predicted Correlation

Correlation	Unmatched Non-Error Components
$r_{ps'}$	$\tau_{.gj}, \tau_{ig.}$
$r_{p''s'}$	$\tau_{ig.}$
$r_{ps''}$	$\tau_{ig.} - \tau_{.g.}$
$r_{p's'}$	$\tau_{.gj} - \tau_{.g.}$
$r_{p'''s'}$	None
$r_{p's''}$	None

TABLE 2

Reliabilities of Fraternity Ratings

Scale	Scale Reliabilities	Factor	Factor Reliabilities
1	.869	I	.949
2	.558		
3	.891		
4	.849		
5	.570	II	.838
6	.413		
7	.510		
8	.509		
9	.848	III	.946
10	.815		
11	.905		
12	.784		
13	.879	IV	.959
14	.852		
15	.856		
16	.848		
17	.798	V	.928
18	.780		
19	.849		
20	.424		

Procedure

Self-report and peer-nomination data on individual subjects were available from an earlier study [2]. For the present purposes it was necessary to gather independent assessments of status on the five factors for the *groups* from which those subjects had been drawn. The subjects in the original study had been recruited from 16 fraternities having chapters on the University of Michigan campus.

Accordingly, two officers or other upper-classmen familiar with fraternities at the University of Michigan were selected from each fraternity represented in the sample and asked to rate the other fraternities in the sample directly, using the same forced-nomination forms with which the peer-nomination data had been collected. A split-raters reliability coefficient was computed for each scale, one rater from each fraternity being randomly assigned to each half-sample. Rating-scale scores were then computed for each fraternity for each half-sample, and the resultant sixteen pairs of ratings were correlated for each rating scale. These scale reliabilities were then combined via Mosier's equation ([1], p. 393). Results are shown in Table 2.

House-nomination factor scores, $H_{.gik}$, were computed by adding together, for each house on each factor, the rating scale scores received by that house on the four rating scales representing that factor. This scoring technique was the same used to obtain factor scores, P_{igik} , for individuals.

It should be noted, however, that house-nomination data were collected over two years after peer-nomination and self-report data, and are therefore more or less appropriate depending on the degree to which they reflected stable differences among the several groups tested.

The final sample for study consisted of eleven groups from nine different fraternities, for which all the necessary information was available. Total sample size was $N = 108$.

Analysis and Results

For these 108 cases, self-report and peer-nomination data were scored on the basic variables, S and P, for each of five factors. These data were then successively subjected to each of the transformations suggested by the model above.

Both peer-nomination and house-nomination scores had been collected using the same instruments. This procedures generated variances for the two kinds of scores that were of comparable magnitude. However, since house-nomination scores were estimates of differences in group means, it was thought that their variance should differ from that of peer-nomination scores by a factor of approximately $1/N_g$. Hence, transformations involving house-nomination scores were carried out by dividing each H score by the square root of average group size and adding this resultant quantity to the individual factor score for each member of the group.

Results are shown in Table 3 with the predicted degree of relationship increasing as one reads down the table. These results should be compared with the predictions of Table 1.

TABLE 3

Concurrent Validities of Peer-Nomination and Self-Report Variables

Correlation	Factors I	II	III	IV	V
$r_{ps'}$	45	17	35	24	38
$r_{p''s'}$	51	15*	28*	27*	39
$r_{ps''}$	49*	19	30*	24*	38*
$r_{p's'}$	51	22	42	41	38*
$r_{p'''s'}$	51	25	44	42	34*
$r_{p's''}$	56	25	37*	43	37*

*The asterisk indicates a correlation that is less than at least one correlation above it in the same column. Decimal points are omitted.

It is evident that the corrections resulted in generally increasing correlations for all factors save factor V. With eleven exceptions noted by asterisks, each correlation coefficient is equal to or greater than all those above it, i.e., differs in the predicted direction. Since both the predicted order of these relationships and the several transformations are all based entirely on a priori considerations, the results presented in Table 2 in no sense reflect the effects of "fitting" the several corrections empirically.

The effects of augmenting the peer-nomination data with the house-rating data in order to match the group differences present in the self-report data can be seen by comparing $r_{p''s'}$ to $r_{ps'}$ and $r_{p'''s'}$ to $r_{p's'}$. Correlations increase for six out of the ten cases and are equal in one other instance. The effects of removing the group location components from the self-report data can be seen by comparing $r_{ps''}$ to $r_{ps'}$ and $r_{p's''}$ to $r_{p's'}$. Correlations increase in five out of ten cases and remain the same in two of the remaining instances.

The effects of removing the elevation component from peer-nomination data can be seen by comparing $r_{p's'}$ to $r_{ps'}$ and $r_{p's''}$ to $r_{ps''}$. In eight out of ten cases, the correlation increases, and in one case the correlation is the same.

Conclusions

In the present research, two kinds of personality measures were considered. In the self-report data an attempt was made to minimize desirability and to maximize intra-individual, inter-trait discrimination by presenting response alternatives in a forced-choice, pair-comparison format. This, however, led to a set of data in which the sum of the factor scores was constrained (unrealistically, we assume) to be approximately the same for everyone.

Peer-nomination data were collected by a within-group, forced-nomination procedure. This prevented a massing of ratings of group members at the desirable pole on each of the scales and thus forced sharper discriminations to be made by each rater but in turn produced scores which precluded meaningful comparisons of individuals belonging to different groups.

A model was developed which represented scores derived from the two kinds of data in terms of true score, systematic bias, and random error components. On the basis of the model it was possible to specify what operations on the data would be necessary to produce equivalent information in terms of the various components present in each case. Performance of the specified operations yielded transforms of the original variables whose score components were matched to a greater or lesser degree. An empirical test yielded generally higher correlations between pairs of variables whose score components were more closely matched than for pairs with less congruity in their structural components.

While it was possible to derive from the peer-nomination data unbiased estimates, p''_{igik}, of true score status, τ_{igi}, it was not possible to do so from the self-report data. Hence the correlations, $r_{p''s'}$, reported in line 2 of Table 3, are probably still spuriously low owing to this deficiency in the s' measures. Correlations below this line of Table 3 involve measures from both sets of data that, while more closely matched in the non-error components that are present, involve in each case only a part of the full τ_{igi} component. Hence these values, as estimates of the saturation of the full τ_{igi} component in the respective measures, are also biased and probably, although not necessarily, in the attenuated direction.

It seems to us to be rather firmly established on the basis of both prior research findings and the arguments and data presented here that both respondent biases and certain features of the assessment process are each capable of producing systematic distortions in measures of personality attributes. Distortions of the former kind would appear, on the basis of previous findings, to be most effectively minimized by use of forced-choice and forced-nomination procedures, but at the cost of introducing distortions of the latter sort. It may, however, be possible to minimize distortions of both kinds by a two-stage approach. Forced-choice and forced-nomination data collection methods can be employed to minimize the effects of respondent biases. The methodological artifacts which these procedures introduce into the resultant score can then be largely eliminated by data analysis techniques of the sort discussed above. This two-stage approach would seem to be necessary since respondent biases of the sort considered here apparently cannot be successfully eliminated by the direct application of statistical corrections to free-response measures.

REFERENCES

[1] Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

[2] Norman, W. T. Personality measurement, faking, and detection: An assessment method for use in personnel selection. *J. appl. Psychol.*, 1963, 47, 225–241.

[3] Norman, W. T. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer-nomination personality ratings. *J. abnorm. soc. Psychol.*, 1963, **66**, 574–583.
[4] Tupes, E. C. and Christal, R. E. Stability of personality trait rating factors obtained under diverse conditions. Lackland Air Force Base, Texas: Personnel Laboratory, Wright Air Development Center, May, 1958. (WADC-TN-61, ASTIA Document No. AD-151 041.)
[5] Tupes, E. C. and Christal, R. E. Recurrent personality factors based on trait ratings. Lackland Air Force Base, Texas: Personnel Laboratory, Aeronautical Systems Division, May, 1961. (ASD-TR-61-97, ASTIA Document No. AD-267 778.)

Manuscript received 11/2/64
Revised manuscript received 3/17/65