INTERRELATIONSHIPS AMONG PERSONALITY SCALE PARAMETERS: ITEM RESPONSE STABILITY AND SCALE RELIABILITY

RICHARD R. JONES
Oregon Research Institute

AND

LEWIS R. GOLDBERG
University of Oregon and Oregon Research Institute

For the assessment of individual differences in hypothetically stable personality traits, it is desirable to develop scales which will maximize both scale score variance and score retest stability. Since scales are composed of individual items, it is reasonable to expect that characteristics of scales (e.g., score variance and score stability) should be related to characteristics of their items. The cogency of this statement for scale construction is seen in the standard practice of eliminating nonvalid items from scales to increase scale validity. In a similar manner, score retest stability should be related to the stability of responses to the scale's items.

The objective of developing scales to maximize both score variance and score stability suggests two criteria for the selection of items: (a) good items should elicit responses which maximize individual differences, and (b) good items should elicit responses which are stable over time. In general, items with large variances and high retest stability should tend to maximize score variance and score stability respectively. The simultaneous satisfaction of these two

---

1 This report is adapted from a doctoral dissertation carried out by the first author under the direction of the second author. The study was supported by Grant #G-25123 (GS-429) from the National Science Foundation to Lewis R. Goldberg at Oregon Research Institute. Data analyses were carried out at the Western Data Processing Center and the Health Science Computing Facility at the University of California at Los Angeles.
item selection criteria, however, highlights a psychometric paradox (e.g., Goldberg, 1963a), namely that responses to high variance items (i.e., moderate endorsement items) are typically unstable, while responses to low variance items (i.e., items with extreme endorsement percentages) tend to be stable from occasion to occasion. This paradoxical finding from itemmetric research seems to imply that between-individual score variance may have to be sacrificed to obtain within-individual score stability, or vice-versa—unless it can be shown that the parallel between item variance and item stability on the one hand, and score variance and score stability on the other, is not as straightforward as common sense would suggest. The present study was designed to investigate empirically the relationships between scale parameters (such as scale stability, homogeneity, and variance) and various measures of average item stability.

**Procedure**

Ninety-five male and 108 female undergraduates in a General Psychology course at the University of Oregon were administered the California Psychological Inventory (CPI) in class, followed two weeks later by the Minnesota Multiphasic Personality Inventory (MMPI); four weeks after the first administration, each inventory was again administered in class to the same students (Goldberg, 1963b; Goldberg and Rorer, 1963; Goldberg and Rorer, 1964). Scale parameters* for each of 199 MMPI and CPI scales* were computed separately for the male and for the female samples. The scale parameters included: (a) test-retest score stability (the product-moment correlation between scores across the two administrations), (b) scale homogeneity (Kuder-Richardson Formula 20) for the first administration, (c) score variance for the first admin-

---

*For purposes of this study, a scale parameter is broadly defined as any measure based on subjects' responses to a set of items comprising a scale. For a discussion of the item properties from which some of the scale parameters were constructed, see Goldberg and Rorer (1963) and Wiggins and Goldberg (1965).

*The scales include 76 of the most frequently cited MMPI scales (Dahlstrom and Welsh, 1960), 35 nonoverlapping MMPI content scales (Wiggins, 1969), 35 other nonoverlapping MMPI scales (Adams and Horn, 1965), the 18 standard CPI scales, plus 35 additional stylistic CPI scales (e.g., Lovell, 1964; Wiggins and Lovell, 1965). While a considerable amount of item overlap occurs among the total set of 199 scales, the analyses were carried out separately for each set of nonoverlapping scales—with no change in the results.
istration, (d) the number of items in the scale, and (e–i) the mean value of five item properties over the set of items in the scale. The five mean item properties were: (e) mean item variance, (f) mean Stable (the percentage of subjects responding consistently to the item on retest), (g) mean Phi, (h) mean Phi/Phimax, and (i) mean Ambdex (Goldberg's [1963a] index of item ambiguity, which essentially corrects Stable for item endorsement extremeness).

The intercorrelations among these scale parameters, over the 199 scales, were computed separately for the sample of 95 males and for the sample of 108 females. For ease of presentation, the signs of the correlations between mean Ambdex and the other scale parameters were reversed, so that high mean Ambdex will imply high item response stability. Thus, mean Stable, mean Phi, mean Phi/Phimax and mean Ambdex, as measures of item response stability should intercorrelate positively.

For all pairs of variables, four indices of association were computed: (a) Pearson product-moment coefficients for continuous data; (b) Pearson product-moment coefficients for grouped data (the values of each variable having been grouped into stems); (c) and (d) two correlation ratios (etases) for rows and columns, respectively. Computation of the latter three statistics permitted two F-tests for curvilinearity (row-eta vs. grouped r, and column-eta vs. grouped r), each of which was tested for significance. Etas presented in this paper are always significantly greater ($p < .01$) than their corresponding grouped r. Correlations (r) listed are always based on ungrouped data.

**Results and Discussion**

The results of this study will be presented in two sections. First, the relationships among the scale parameters will be presented, and second, the implications of the psychometric paradox for score reliability will be discussed.

**Intercorrelations among Scale Parameters**

Table 1 presents the means, standard deviations, and intercorrelations among the nine scale parameters for the two samples. While the signs of many of these correlations are predictable on the basis
<table>
<thead>
<tr>
<th>Scale Stability</th>
<th>Scale Homogeneity</th>
<th>Number of Items</th>
<th>Score Variance</th>
<th>Mean Item Variance</th>
<th>Mean Stable</th>
<th>Mean Phi</th>
<th>Mean Phi/Phimax</th>
<th>Mean Ambdex</th>
<th>Males (N=95) Mean</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale Stability</td>
<td>74**</td>
<td>41**</td>
<td>57**</td>
<td>46**</td>
<td>-35**</td>
<td>45**</td>
<td>23**</td>
<td>32**</td>
<td>.74</td>
<td>.10</td>
</tr>
<tr>
<td>Scale Homogeneity</td>
<td>79**</td>
<td>49**</td>
<td>67**</td>
<td>23**</td>
<td>-29**</td>
<td>05</td>
<td>00</td>
<td>-12</td>
<td>.53</td>
<td>.22</td>
</tr>
<tr>
<td>Number of Items</td>
<td>50**</td>
<td>50**</td>
<td>72**</td>
<td>08</td>
<td>-03</td>
<td>-02</td>
<td>-03</td>
<td>-06</td>
<td>28.67</td>
<td>18.77</td>
</tr>
<tr>
<td>Score Variance</td>
<td>58**</td>
<td>66**</td>
<td>75**</td>
<td>34**</td>
<td>-37**</td>
<td>12</td>
<td>-01</td>
<td>-06</td>
<td>14.10</td>
<td>15.76</td>
</tr>
<tr>
<td>Mean Item Variance</td>
<td>24**</td>
<td>11</td>
<td>08</td>
<td>31**</td>
<td>-92**</td>
<td>66**</td>
<td>19**</td>
<td>27**</td>
<td>.16</td>
<td>.05</td>
</tr>
<tr>
<td>Mean Stable</td>
<td>-10</td>
<td>-14**</td>
<td>-05</td>
<td>-32**</td>
<td>-92**</td>
<td>-37**</td>
<td>06</td>
<td>08</td>
<td>.86</td>
<td>.04</td>
</tr>
<tr>
<td>Mean Phi</td>
<td>31**</td>
<td>14**</td>
<td>12</td>
<td>21**</td>
<td>68**</td>
<td>-47**</td>
<td>63**</td>
<td>85**</td>
<td>.62</td>
<td>.08</td>
</tr>
<tr>
<td>Mean Phi/Phimax</td>
<td>19**</td>
<td>02</td>
<td>-02</td>
<td>-05</td>
<td>00</td>
<td>21**</td>
<td>45**</td>
<td>68**</td>
<td>.61</td>
<td>.08</td>
</tr>
<tr>
<td>Mean Ambdex</td>
<td>31**</td>
<td>18**</td>
<td>14**</td>
<td>18**</td>
<td>54**</td>
<td>-33**</td>
<td>96**</td>
<td>43**</td>
<td>.55</td>
<td>.07</td>
</tr>
<tr>
<td>Females Mean</td>
<td>.74</td>
<td>.49</td>
<td>28.67</td>
<td>12.54</td>
<td>.16</td>
<td>.86</td>
<td>.46</td>
<td>.60</td>
<td>.64</td>
<td>.09</td>
</tr>
<tr>
<td>(N=108) σ</td>
<td>.10</td>
<td>.22</td>
<td>18.77</td>
<td>13.88</td>
<td>.04</td>
<td>.04</td>
<td>.10</td>
<td>.08</td>
<td>.09</td>
<td></td>
</tr>
</tbody>
</table>

Note.—Correlations from the male sample are presented above the diagonal, and those for the female sample below the diagonal. For both samples, the correlations are based on an N of 199 scales. Decimal points are omitted.

* p < .05.

** p < .01.
of algebraic interdependence between variables, test theoretic predictions, or simple logic, an empirical estimation of the magnitude of these relationships is of considerable interest.

The correlations between the two types of scale reliability estimates, scale stability and scale homogeneity are positive ($r = .74$ and .79) as would be expected, although these coefficients are not sufficiently large to permit the conclusion that these two reliability estimates are interchangeable measures of scale reliability. As Cattell (1964) among others has pointed out, a scale's internal consistency may differ considerably from its temporal stability, either due to design in test construction, or due to the character of the trait which the scale measures.

A second set of expected relationships are those between the number of items in the scale and each of the two reliability estimates. Perhaps the best known conclusion from classical test theory (e.g., Guillicsen, 1950) is that scale reliability may be increased by lengthening the scale, and the Spearman-Brown formula provides a means of estimating the resultant reliability for increases in scale length. However, it does not follow from this test theoretic formulation that, for a set of different scales, long scales necessarily will be more reliable than short scales. In the case of a single scale, the underlying rationale for expecting increased reliability by lengthening the scale is that the augmented scale measures a greater proportion of true score variance than the shorter form of the same scale. For the case of many different scales, as in the present study, longer scales will show higher reliability than shorter scales only insofar as the longer scales measure a greater proportion of true score variance than the shorter scales. Since a relatively short scale may measure a greater proportion of true variance than a longer scale, an empirical measure of the relationship between scale length and scale reliability may be less than perfect. In fact, the present data show that these correlations are considerably less than unity; the correlations between number of items and scale stability were .41 and .50, while the correlations between number of items and scale homogeneity were .49 and .50.

A third set of expected relationships are those between the two types of scale reliability on the one hand and score variance and mean item variance on the other. Guilford (1954) concluded that for scales with less than perfect intercorrelations among items, a
concentration of items with high variances will tend to enhance scale reliability. For the scales in the present study, the correlations between mean item variance and scale stability were .46 and .24, while the corresponding correlations with scale homogeneity were .23 and .11. Although these coefficients were in the expected direction, their magnitude suggests that mean item variance is only moderately predictive of either type of scale reliability. Score variance, on the other hand, was more strongly associated with scale reliability, with correlations of .57 and .58 between score variance and scale stability, and .67 and .66 between score variance and scale homogeneity. These coefficients, although artifically inflated due to algebraic interdependence between the variables, illustrate the well-known conclusion that large score variance is, in general, a necessary but not sufficient condition for high score reliability.

A fourth set of relationships involves those among the number of items in the scale, score variance, and mean item variance. Number of items and score variance were positively related \( (r = .72 \text{ and } .75) \), illustrating the constraint placed on score dispersions by the number of items on which the scores are based. The correlations between score variance and mean item variance should be positive, but low, due to the greater dependence of score variance on a scale's inter-item covariances than item variances. Score variance is simply the sum of the elements in a scale's inter-item variance-covariance matrix which contains \( m \) (number of items) variance terms and \( m(m - 1) \) covariance terms. The relatively small contribution to score variance from item variances compared to covariances is empirically demonstrated by the low correlation \( (r = .34 \text{ and } .31) \) between score variance and mean item variance. Finally, as might be expected, the number of items in a scale and its mean item variance were unrelated \( (r = .08 \text{ in both samples}) \).

Turning now to the relationships among the four mean item stability measures, some unexpected findings emerge. If all of these variables are similar measures of response stability, their intercorrelations should be positive. However, the correlations between mean Stable and mean Phi were negative \( (r = -.37 \text{ and } -.47) \), a function of the fact that Phi is attenuated for extreme endorsement items (which, in turn, tend to be high Stable items). Since the numerical value of Phi is attenuated by item extremeness while simultaneously being increased by increases in response stability, the negative correlations between mean Stable and mean Phi suggest
that the former constraint is more powerful than the latter. When
mean item variance is partialled out, the partial correlations be-
tween mean Stable and mean Phi rise to .75 and .50.

The linear correlations between mean Stable and mean Ambdex
(reflected) are negligible in this study, though non-linearity ($\eta =
.37$ and .42) is evident in these relationships. These relationships
are similar to those obtained in itemmetric studies (e.g., Goldberg,
1965; Wiggins and Goldberg, 1965) when many extreme items are
included in the item pool. When extreme items are progressively
eliminated, however, the relationship between Stable and Ambdex
increases dramatically. In the present study, when mean item var-
ance is partialled out, the partial correlations between mean Stable
and mean Ambdex (reflected) were .85 and .48.

Perhaps the most significant set of relationships illustrated in
Table 1 are those between the four mean item stability measures
and the five other scale parameters. Here the focus of interest is on
the extent to which scale reliability and score variance can be pre-
dicted from the properties of the average item in the scale. As Table
1 indicates, scale retest stability was positively associated with
mean Phi ($r = .45$ and .31), mean Phi/Phimax ($r = .23$ and .19),
and mean Ambdex (reflected) ($r = .32$ and .31). These correlations
are in the expected direction, indicating that the stability of the
average item in a scale is correlated with the stability of the scale.

Unexpected, however, are the negative correlations between scale
stability and mean Stable ($r = -.35$ and -.10)! This unusual
finding is explained by the high negative correlations between mean
Stable and mean item variance ($r = -.92$ in both samples), and
the moderate positive correlations between mean item variance and
scale stability ($r = .46$ and .24). Mean Stable primarily reflects
mean item variance rather than mean item stability, and conse-
quently it is negatively correlated with scale stability. If one were
to use Stable as an item selection criterion, the result would be to
lower the over-all stability of the final scale (the psychometric
paradox)! The resolution of this paradox will be discussed later.

The last findings of interest in Table 1 are the generally low cor-
rrelations between the four mean item stability measures and scale
homogeneity. The few significant correlations that do occur reflect
primarily the correlation of each of the four mean stability mea-
ures with mean item variance, which is moderately correlated with
scale homogeneity.
Partial Correlations

The zero-order correlations presented in Table 1 suggest that mean item variance functions as a suppressor variable in the relationships between scale stability and the various measures of mean item stability. In this section, the interrelationships among mean item stability, mean item and scale variance, and scale stability will be evaluated more completely. Table 2 presents the zero-order correlations among these variables, and then presents partial correlations to evaluate the influence of mean item variance and score variance on these relationships. The top three rows of Table 2 show the zero-order correlations of scale stability, mean item variance, and score variance with mean Stable, mean Phi, and Mean Ambdex (from Table 1). The lower three rows of Table 2 show the effect on these correlations when mean item variance and score variance are partialled out, first separately, then jointly.

**TABLE 2**
Partial Correlations among Scale Parameters

<table>
<thead>
<tr>
<th></th>
<th>Mean Stable</th>
<th></th>
<th>Mean Phi</th>
<th></th>
<th>Mean Ambdex</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>Zero-order Correlations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>With scale stability</td>
<td>-.35**</td>
<td>-.10</td>
<td>.45**</td>
<td>.31**</td>
<td>.32**</td>
<td>.31**</td>
</tr>
<tr>
<td>With mean item variance</td>
<td>-.92**</td>
<td>-.92**</td>
<td>.66**</td>
<td>.68**</td>
<td>.27**</td>
<td>.54**</td>
</tr>
<tr>
<td>With score variance</td>
<td>-.37**</td>
<td>-.32**</td>
<td>.12</td>
<td>.21**</td>
<td>-.06</td>
<td>.18**</td>
</tr>
<tr>
<td>First-order partials</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale stability—with mean item variance partialed out</td>
<td>.19*</td>
<td>.28**</td>
<td>.22**</td>
<td>.21**</td>
<td>.23**</td>
<td>.22**</td>
</tr>
<tr>
<td>Scale stability—with score variance partialed out</td>
<td>-.18*</td>
<td>.11</td>
<td>.47**</td>
<td>.24**</td>
<td>.43**</td>
<td>.25**</td>
</tr>
<tr>
<td>Second-order partials</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale stability—with mean item variance and score variance partialed out</td>
<td>.33**</td>
<td>.41**</td>
<td>.34**</td>
<td>.25**</td>
<td>.36**</td>
<td>.25**</td>
</tr>
</tbody>
</table>

* *p < .05.
** *p < .01.

As indicated earlier, the zero-order correlations shown in Table 2 indicate that mean Phi and mean Ambdex are, in general, positively associated with both scale stability and mean item variance, but that mean Stable is negatively related to the same variables. These
findings suggest that the psychometric paradox (i.e., responses to high variance items tend to be unstable) holds when mean Stable is used as a measure of average item response stability, but not when mean Phi or mean Ambdex is used.

Turning now to the first-order partial correlations with mean item variance partialled out, the correlations between mean Stable and scale stability are .19 and .28, similar to the partial correlations between scale stability and either mean Phi or mean Ambdex. While the relationships between scale stability and mean Stable, mean Phi, and mean Ambdex are all influenced by mean item variance, the suppression effect is greatest for mean Stable due to the high negative correlation between mean item variance and mean Stable ($r = -.32$ in both samples). Note that mean Stable, with mean item variance partialled out, is now predictive of scale stability in the expected direction (positive), and that the first-order partial correlations between scale stability and both mean Phi and mean Ambdex, although still in the expected direction, are lower than their corresponding zero-order correlations. With mean item variance partialled out, all three mean item stability measures correlate approximately equally with scale stability.

When one examines the first-order partial correlations with score variance partialled out, one can see that the influence of score variance is noticeably less than that for mean item variance. In fact, for the male sample, the first-order partial correlations between scale stability and both mean Phi and mean Ambdex increase slightly over their zero-order correlations. These slight changes stem from the lack of correlation between score variance on the one hand and both mean Phi and mean Ambdex on the other.

The bottom row in Table 2 shows the second-order partial correlations between each of the three mean item stability measures and scale stability, with both mean item variance and score variance partialled out. The differences among these correlations are small, leading to the conclusion that mean Stable, mean Phi, and mean Ambdex are equally predictive of scale stability when the influences of mean item variance and score variance are partialled out. What is even more interesting is the fact that these coefficients are not larger, since they reflect the amount of variance in scale stability which is predictable from knowledge of item stability. Since numerous patterns of item responses can produce the same
score on the scale, a given subject could change responses to several items on retest without changing his score. Consequently, any correlation between measures of item response change and score change would be less than unity; indeed, if the present data are at all representative, that correlation is only about .30!

Conclusions

This study investigated the implications of a psychometric paradox for considerations of scale score reliability. The paradox, a finding from itemmetric studies, is that items which show high response stability over time tend to be low variance items. Since both within-individual score stability and across-individual score variability are desirable scale characteristics, the psychometric paradox could present a problem, if maximizing score reliability is an important goal of scale construction. The following conclusions appear to be supported by the data.

1. The psychometric paradox appears to hold true when mean Stable is the item stability measure employed. That is, high mean Stable scales tend to show low mean item variance and low scale variance (and consequently tend to have low scale stability).

2. However, with mean Phi or mean Ambdex as a measure of item stability, positive correlations with mean item variance are obtained—suggesting that the psychometric paradox is dependent on the statistic used to measure item stability.

3. Because of the relative independence of score variance with either mean Phi or mean Ambdex, it should be possible to obtain simultaneously large score variability and high item stability in future personality scales.

REFERENCES


Goldberg, L. R. A Model of Item Ambiguity in Personality Assessment. Educational and Psychological Measurement, 1963, 23, 467–492. (a)


