THE RELIABILITY OF CLINICIANS' JUDGMENTS:
A MULTITRAIT-MULTIMETHOD APPROACH

LEWIS R. GOLDBERG AND CHARLES E. WERTS

University of Oregon and Oregon Research Institute National Merit Scholarship Corporation

4 experienced clinical psychologists independently ranked each of 4 equated samples of 10 patients on 1 of 4 traits (Social Adjustment, Ego Strength, Intelligence, and Dependency), using 1 of 4 data sources (MMPI, Rorschach, Wechsler, and a Vocational History). A $4 \times 4$ Latin square design insured that the usual sources of judgment confounding were absent from this study. The findings indicate quite clearly that the judgments of 1 clinician working from 1 data source bear no systematic relationship to those of another clinician working from another data source, even though both judges are ranking the same patients on the same trait. On the other hand, judgments of diverse traits from the same data source do tend to be related.

Studies of the accuracy of inferences made by experienced clinical psychologists from personality test data have, in general, indicated (a) little validity for these inferences on an absolute basis (e.g., Goldberg, 1959; Holtzman & Sells, 1954; Kelly & Fiske, 1951), and (b) no relative validity for these inferences over rather simple actuarial procedures (e.g., Goldberg, 1965; Gough, 1962; Meehl, 1954; Oskamp, 1962). While early reactions to this literature included some denials of the conclusiveness of the available evidence (e.g., Holt, 1958; Hutt, 1956), later reactions have included some attempts to understand those aspects of the judgmental process which might logically lead to inferential errors (e.g., Cline & Richards, 1962; Hammond, Hursch, & Todd, 1964; Richards, 1963).

The bulk of the empirical studies of clinical judgments have focused on their validity, while relatively few studies have investigated the reliability of these judgments. General psychometric theory, however, posits a strong constraint imposed by the reliability of a measure on its potential validity. Extending the logic of test theory to include clinical psychologists as diagnostic instruments (e.g., Hunt, 1959), one might reasonably inquire whether clinical judgments typically are (a) relatively reliable, but often simply misaligned with reality, or (b) so lacking in reliability that their relationships with any stable external criterion would by necessity be severely attenuated.

Recently, the concept of reliability, itself, has been undergoing a searching reevaluation (e.g., Fiske, 1963). The classic dichotomy between internal consistency (homogeneity) and retest stability has been fused in the recent work of Cronbach and his associates (e.g., Cronbach & Gleser, 1964; Cronbach, Rajaratnam, & Gleser, 1963), while simultaneously being expanded in the writings of Cattell and his associates (e.g., Cattell, 1964; Cattell & Tsujioka, 1964). Both schools of thought, however, recognize generalizability as the common element in all definitions of reliability. Within the domain of clinical judgments, inferential reliability includes, as specific components, generalizability (a) over time, for the same judges using the same data (stability), (b) over judges, for the same data from the same occasion (consensus), and (c) over data sources, administered on the same occasion and interpreted by the same judge (convergence). Each of these three "reliabilities," in turn, might be ex-

---

1 The present report includes a reanalysis and summary by the first author of data originally collected by the second author (Werts, 1960). The authors wish to thank the four clinical judges for contributing their time so generously to this project; Daniel N. Weiner for providing the test protocols; and Dean Peabody, Leonard Rorer, Kenneth Howard, and Robert Nichols for their critical reading of a preliminary draft of this article. The preparation of this report was supported, in part, by Grant MH 04439 from the National Institutes of Health, United States Public Health Service.
pected to have significant interactions with the characteristics (traits) being evaluated, on the one hand, and the reference populations (targets), on the other. Clearly, then, investigation of the reliability of clinicians’ judgments is complicated by covariation across at least five parameters (time, judges, data sources, traits, and targets), and no study of the reliability of clinical inferences is ever likely to provide definitive conclusions.

One type of solution to the general problem of simultaneously displaying covariation across a number of parameters has been offered by Campbell and Fiske (1959). While their multitrait-multimethod matrix was originally proposed as a means of estimating the convergent and discriminant validity of tests within the general model of construct validation (e.g., Campbell, 1960; Cronbach & Meehl, 1955), the suitability of this method for studying clinicians’ judgments was quickly noted by Meehl:

In order to place any confidence in either the theoretical constructs we employ in discussing patients, or in the instrument-interpreter combinations we use to assess them, studies of convergent and discriminative validity must be carried out. The Campbell-Fiske multitrait-multimethod matrix, or the multiperson-multimethod variant of it, should be useful for this purpose [1960; pp. 25–26].

Two clinical judgment studies using variants of the Campbell-Fiske methodology have been published. Howard (1962) had seven clinical psychologists rank-order 10 needs for each of 10 patients from Rorschach, Thematic Apperception Test (TAT), and Sentence Completion protocols. While Howard did not publish his correlation matrix, a chi-square analysis led him to conclude that the average intratest correlation between judges ($r = .19$) indicated some significant, though very slight, consensus in these judgments. Intrajudge convergence between tests, however, averaged only .13, and the average of the interjudge agreement correlations between tests was only .05.

In a later study, Howard (1963) asked his seven clinicians to rate the same 10 patients on five traits, using the same three projective tests. Again, the correlation matrix was not reported, but an analysis of variance indicated that the average interjudge agreement correlations between tests were not significant for any trait. On the other hand, intrajudge-intertest agreement (convergence) and intratest-interjudge agreement (consensus) were both significantly higher for less inferential traits (e.g., verbal fluency, productivity) than for more inferential traits (e.g., psychotherapy prognosis, adjustment).

While both of Howard’s studies explored significant aspects of inferential reliability, certain methodological problems, applicable to any single study, demand further research in this domain. For example, since Howard utilized only one group of 10 patients, the generalizability of his findings to different target populations is still open to question. Moreover, clinicians might legitimately argue that the task of ranking 10 needs (e.g., “harmavoidance,” “sucorance,” “blamavoidance”) ipsatively for each patient (Howard, 1962) falls somewhat outside their usual clinical responsibilities, thereby questioning the generality of these findings to more commonly used traits. The present study included a larger sample of targets (replications across four different patient samples) as well as a set of traits more commonly employed in clinical settings.

Perhaps an even more important methodological issue arises from the relatively common practice of asking a judge to rate the same target, either on the same trait using different data sources, or on different traits using the same data source. In the former case, any idiosyncratic and non-test-specific habit of the target (e.g., use of particular words or phrases, general verbal fluency, etc.) should spuriously increase convergence. In the latter case, any general evaluative impression of the target (e.g., a “halo” effect) might systematically distort trait discrimination. The design of the present study insures that judges, traits, targets, and data sources are all completely unconfounded, and, therefore, these findings should provide an unusual comparison with other studies of inferential reliability.

**PROCEDURE**

Four experienced clinical psychologists independently ranked each of four sets of 10 neuro-
psychiatric patients on one of four traits, using one of four data sources. The intercorrelations among rankings were pooled across the four samples to form an average multitrait-multimethod judgmental matrix. The design, which is illustrated in Table 1, is explained in greater detail below.

Judges

The four judges used in this study were practicing clinical psychologists, all of whom had considerable postgraduate clinical experience and were familiar with the population of patients from which the experimental sample was drawn. The four judges received their PhD degrees, respectively, from the University of Iowa, Northwestern University, the University of Minnesota, and Stanford University.

Traits

Four of the most frequently used constructs in the diagnostic reports of clinical psychologists—Social Adjustment (SA), Ego Strength (ES), Intelligence (IQ), and Dependency (D)—were utilized in this study. In order to ascertain how these concepts were actually being used by experienced clinical psychologists, no attempt was made to further define these constructs for the judges.

Patient Samples

Of the several thousand outpatients who had been tested between the years 1947 and 1950 at the Veterans Administration Mental Hygiene Clinic at Fort Snelling, Minnesota, four groups of 10 each were carefully selected so as to (a) equate the four samples on all relevant traits while (b) maximizing trait variation within each sample. Each patient sample consisted of males who varied (a) in social adjustment, from relatively normal to grossly psychotic; (b) in intelligence, from dull-normal to superior (as measured by Wechsler-Bellevue scores); and (c) in vocational adjustment, from unemployed to presently employed in a well-paying and responsible position which had been held for many years.

Data Sources

The four sources of data were intended to include information commonly available to clinical psychologists carrying out psychological evaluations (e.g., Sundberg, 1961) as well as to represent four rather different types of clinical assessment procedures. An intelligence test (the Wechsler-Bellevue), a projective test (the Rorschach), a structured personality inventory (the MMPI), and a vocational history were selected for study.

Experimental Design

The experimental design of this study might best be visualized as a cube whose three dimensions are tests, judges, and traits. There were four such cubes utilized in this study, each containing a different set of 10 patients. Alternatively, the design may be thought of as a $4 \times 4$ Latin square, replicated four times. Each judge ranked four different sets of 10 protocols each, ranking each of the sets (a) on a different trait and (b) by means of a different data source. Thus, no judge ranked any set of protocols on more than one trait, and no judge ranked more than one trait from any single set of protocols.

Table 1 summarizes the experimental design for this study. Each judge first ranked, on one of the four traits, a packet of 40 test protocols. The 40 protocols, each from a different patient, were composed of four sets of 10, each set comprising a different data source (e.g., Judge 1 first received the 10 MMPIs from Sample A, the 10 Wechsler's from Sample B, and the 10 Vocational Histories from Sample D, and he ranked each set on Dependency; he then received, in turn, the three other packets, each to be judged on a different trait).

Statistical Analysis

Within each of the four patient samples, product-moment correlations were computed among all pairs of rankings. These correlations were converted to $Z$s, averaged across the four samples, and the averages then reconverted to correlation coefficients. The signs of the correlations for the trait of Dependency were reversed, so that all correlations reflect rankings in the socially desirable direction.

Results

Table 2 presents the average multitrait-multimethod correlation matrix. The correla-
<table>
<thead>
<tr>
<th></th>
<th>Social adjustment (SA)</th>
<th></th>
<th>Ego strength (ES)</th>
<th></th>
<th>Intelligence (IQ)</th>
<th></th>
<th>Dependency* (D)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MMPI</td>
<td>ROR</td>
<td>W-B</td>
<td>VOC</td>
<td>MMPI</td>
<td>ROR</td>
<td>W-B</td>
<td>VOC</td>
</tr>
<tr>
<td>MMPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROR</td>
<td>.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W-B</td>
<td>.22</td>
<td>-.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOC</td>
<td>.16</td>
<td>-.15</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROR</td>
<td>.90</td>
<td>.08</td>
<td>.24</td>
<td>-.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W-B</td>
<td>.02</td>
<td>-.01</td>
<td>.66</td>
<td>.11</td>
<td>.06</td>
<td>-.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOC</td>
<td>-.01</td>
<td>-.21</td>
<td>.57</td>
<td>.76</td>
<td>-.11</td>
<td>-.31</td>
<td>.23</td>
<td></td>
</tr>
<tr>
<td>MMPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROR</td>
<td>.13</td>
<td>.05</td>
<td>.15</td>
<td>-.38</td>
<td>.11</td>
<td>.03</td>
<td>.10</td>
<td>-.24</td>
</tr>
<tr>
<td>W-B</td>
<td>.37</td>
<td>.34</td>
<td>.21</td>
<td>-.19</td>
<td>.26</td>
<td>.39</td>
<td>.36</td>
<td>.06</td>
</tr>
<tr>
<td>VOC</td>
<td>.09</td>
<td>.02</td>
<td>.54</td>
<td>.10</td>
<td>.11</td>
<td>.39</td>
<td>.77</td>
<td>-.03</td>
</tr>
<tr>
<td>MMPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROR</td>
<td>.90</td>
<td>.08</td>
<td>.24</td>
<td>-.24</td>
<td>.64</td>
<td>.02</td>
<td>-.04</td>
<td>-.08</td>
</tr>
<tr>
<td>W-B</td>
<td>.12</td>
<td>-.03</td>
<td>.45</td>
<td>.08</td>
<td>.15</td>
<td>.21</td>
<td>.33</td>
<td>.02</td>
</tr>
<tr>
<td>VOC</td>
<td>.11</td>
<td>-.01</td>
<td>.08</td>
<td>-.01</td>
<td>.07</td>
<td>.10</td>
<td>.12</td>
<td>-.03</td>
</tr>
<tr>
<td>MMPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROR</td>
<td>.02</td>
<td>.16</td>
<td>.09</td>
<td>.47</td>
<td>-.25</td>
<td>-.14</td>
<td>.12</td>
<td>.65</td>
</tr>
<tr>
<td>W-B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Signs of the correlations for Dependency have been reversed.
tions between judgments of the same trait made from different data sources (e.g., the montrait-heteromethod correlations) are presented in the triangles just below the main diagonal of the table. There are six such montrait-heteromethod correlations available for each trait: (a) MMPI versus Rorschach, (b) MMPI versus Wechsler, (c) MMPI versus Vocational History, (d) Rorschach versus Wechsler, (e) Rorschach versus Vocational History, and (f) Wechsler versus Vocational History. These correlations indicate the extent to which the clinicians agreed in their judgments of the same patients on the same trait when the judgments were made from different data sources. Since there were four traits under study, there are 24 montrait-heteromethod correlations. If the clinicians tended to agree in their judgments of the same trait evaluated by different data sources, the montrait-heteromethod correlations would be high.

The circled values in Table 2 are the correlations between judgments of different traits made from the same data source (heteromethod-monmethod correlations). There are six such heteromethod-monmethod correlations for each data source: (a) Social Adjustment versus Ego Strength, (b) Social Adjustment versus Intelligence, (c) Social Adjustment versus Dependency, (d) Ego Strength versus Intelligence, (e) Ego Strength versus Dependency, and (f) Intelligence versus Dependency. Since there were four data sources, there are 24 such heteromethod-monmethod correlations presented in Table 2. The size of these correlations indicates the extent to which judgments from the same data sources tend to be correlated across different traits; these correlations reflect the effects of "method variance" (Campbell & Fiske, 1959).

The remaining values presented in Table 2 are the correlations between judgments of different traits made from different data sources (heteromethod-heteromethod correlations). There are 72 such values, reflecting the strength of association between traits, with method variance removed.

Inspection of the montrait-heteromethod correlations in Table 2 reveals evidence of very low convergence. These correlations ranged from −.31 (Ego Strength: Rorschach versus Vocational History) to .64 (Social Adjustment: MMPI versus Rorschach). On the other hand, there is evidence on considerable method variance, the heteromethod-monmethod correlations ranging from −.03 (Rorschach: Social Adjustment versus Dependency) to .80 (MMPI: Social Adjustment versus Ego Strength).

A summary of these findings is presented in Table 3. The top half of Table 3 summarizes the evidence for interjudge convergence. The first column presents the average montrait-heteromethod correlations for each trait (the average of the elements in each of the off-diagonal triangles in Table 2). These average correlations would be high if the clinicians were agreeing with each other. The second column presents the average heteromethod correlations (the average of the noncircled elements from the rectangles in Table 2). These values, which are free from method variance, should be considerably lower than those in the first column, if convergence is to be established. The third column presents the average heteromethod correlations (the average of all of the values from the heteromethod rectangles in Table 2). These values, inflated by some method variance, should ideally be low compared to the montrait-heteromethod correlations if interjudge
convergence—rather than method variance—was the predominant judgmental factor involved.

Of the four traits, only Intelligence yielded an average coefficient of convergence numerically higher (\( r = .25 \)) than the average heterotrait correlations (\( r = .16 \)). While this tiny degree of convergence for a trait as significant as Intelligence seems incredible, the results for the other three traits were even worse! Social Adjustment yielded an average coefficient of convergence of .17, exactly equal to its average heterotrait correlation. The corresponding averages were a mere .03 for Dependency (against a heterotrait \( r \) of .14) and -.01 for Ego Strength (against a heterotrait \( r \) of .18). In general, the monotrait-heteromethod correlations averaged only .11, while the heterotrait correlations averaged .16.

The bottom half of Table 3 presents another summary of the correlations in Table 2, now rearranged by data sources. The first column lists the average heterotrait-monomethod correlations (the circled elements in Table 2), for each data source. Note that the average monomethod correlations (indexes of method variance) were all greater than the average heteromethod correlations. The average monomethod correlation for the MMPI was .45, against a heteromethod \( r \) of .06. For the Wechsler-Bellevue, the average monomethod correlation was .46, against a heteromethod \( r \) of .15, and for the Vocational History the corresponding values were .43, against .03. For the Rorschach, however, the monomethod correlations were small (\( r = .15 \)), against a heteromethod \( r \) of .10.

### Table 4

#### Test-Retest Stability Coefficients for One Judge

<table>
<thead>
<tr>
<th>Traits</th>
<th>MMPI</th>
<th>Vocational history</th>
<th>Rorschach</th>
<th>Wechsler-Bellevue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social adjustment</td>
<td>.88**</td>
<td>.58*</td>
<td>.46</td>
<td>.09</td>
</tr>
<tr>
<td>Ego strength</td>
<td>1.00**</td>
<td>.93**</td>
<td>.78**</td>
<td>.69*</td>
</tr>
<tr>
<td>Dependency</td>
<td>.78**</td>
<td>.77**</td>
<td>.53</td>
<td>-.12</td>
</tr>
<tr>
<td>Intelligence</td>
<td>.10</td>
<td>.54</td>
<td>.72**</td>
<td>1.00**</td>
</tr>
</tbody>
</table>

Note.—Each coefficient represents a rank-correlation (\( r \)) between the original rankings and those carried out 2 months later.

* \( p < .05 \)

** \( p < .01 \)

### Table 5

#### Rank Correlations between Intelligence Rankings and Wechsler-Bellevue IQs

<table>
<thead>
<tr>
<th>Test</th>
<th>Judge</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMPI</td>
<td>.21</td>
<td>-.71*</td>
</tr>
<tr>
<td>Vocational history</td>
<td>.36</td>
<td>.50</td>
</tr>
<tr>
<td>Rorschach</td>
<td>-.33</td>
<td>-.66*</td>
</tr>
<tr>
<td>Wechsler-Bellevue</td>
<td>1.00**</td>
<td>1.00**</td>
</tr>
</tbody>
</table>

* \( p < .05 \)

** \( p < .01 \)

The findings reported in Table 3 indicate that for each of these four traits, the judgments of one experienced clinician working from one data source bear no relationship to the judgments of another clinician working from another data source. However, Table 3 indicates that clinicians' judgments do tend to covary when they are using the same data source but focusing on different traits. Are these enigmatic findings simply a function of the perceived difficulty of the judgmental task? If so, the judgments made by individual clinicians, if made on repeated occasions, should be quite unstable.

Table 4 presents some evidence relating to this possibility. One judge was asked to carry out the identical ranking procedures 2 months after his first attempt, and the rank correlations (\( r \)) for this judge are presented in Table 4. On the average, the test-retest reliability of this judge was .61, indicating considerable stability for his judgments and making less tenable the hypothesis that the task was perceived as an impossibly difficult one. For this judge, judgments of Ego Strength from either the MMPI or the Vocational History and judgments of Intelligence from the Wechsler-Bellevue were remarkably stable. On the other hand, his judgments of either Social Adjustment or Dependency from the Wechsler and Intelligence from the MMPI were extremely unstable.

While the present study focused primarily on the convergence of clinical judgments, some data are available which bear on their validity. Since the clinicians were asked to judge Intelligence from all four data sources, it was possible to relate these judgments to Intelligence as measured by the Wechsler-
Bellevue. Table 5 presents the results of this analysis. Of the 12 rank correlations listed in Table 5 not involving the Wechsler, only 5 were positive and significantly above zero, and 3 of these 5 came from one judge. It appears as if one judge was able to do this task quite accurately, while the other three judges were not. Since in any particular setting the “best” judge is typically not known, the accuracy of the average judge becomes important. For this task, however, the average judge’s inferences, while positive, did not differ significantly from zero.

**DISCUSSION**

In a classic study of the reliability of clinical judgments, Little and Sheidman (1959) reported the results of a comprehensive analysis of diverse kinds of inferences about 12 subjects made by expert clinicians from each of five kinds of data sources—Rorschach, TAT, MMPI, Make-a-Picture-Story (MAPS), and Anamnesis. From their findings, these investigators extrapolated that:

For the tests used, agreement on diagnoses will be only slightly better than chance, judgments of maladjustment will be skewed toward the pathological, agreement with psychiatrists as to personality dynamics will be modest, and the clinician's reliability in all these areas will leave much to be desired [Little & Sheidman, 1959, p. 27].

The findings from the present study confirm Little and Sheidman’s predictions. These results, averaged across four samples of patients, clearly indicate that an experienced clinician’s judgments from one data source do not correlate with another clinician’s judgments from another data source, even though both clinicians are diagnosing the very same patient on—ostensibly—the very same trait! Moreover, this finding occurs regardless of whether the trait being rated is a highly global one (e.g., Social Adjustment, Ego Strength) or a more specific one (e.g., Dependency). These findings become all the more striking when one realizes that each of the four patient samples utilized in this study was selected to span as great a range of the traits as possible, thus eliminating the possibility that a lack of convergence could stem from a restriction of trait range. Moreover, the ranking procedure used in this study in-
ments of Dependency should not be made from an intelligence test, nor judgments of Intelligence from the MMPI. While it is not uncommon for such multitrait-monomethod judgments to be made in clinical practice, the present study should certainly serve to question the wisdom of this activity.

REFERENCES


GOLDBERG, L. R. Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. Psychological Monographs, 1965, 79(9, Whole No. 602).


HUTT, M. L. Actuarial and clinical approaches to psychodiagnosis. Psychological Reports, 1956, 2, 413–419.


LITTLE, K. B., & SHENDE, E. S. Congruencies among interpretations of psychological test and anamnestic data. Psychological Monographs, 1959, 73(6, Whole No. 476).


(Received April 30, 1965)