THE RELIABILITY OF PROJECTIVE TECHNIQUES: REVIEW OF THE LITERATURE

BY

ARTHUR R. JENSEN [1]

Institute of Psychiatry (Maudsley Hospital)
University of London

Logically, though not historically, projective techniques must be regarded as a part of the general field of psychological measurement. Projective testing and research may therefore legitimately be studied within the theoretical and methodological framework of psychological measurement, always keeping in mind, of course, the problems that are peculiar to projective techniques.

Largely as a result of the historically separate development of projective techniques outside the traditions of psychological measurement, this prominent type of psychological testing has practically ignored a central concept of psychological test theory, namely, *reliability*.

Reliability is seldom mentioned in the projective literature. Occasionally one finds estimates of the reliability of a particular technique; but more often, where reliability is mentioned at all, it is merely to acknowledge the existence of the problem. The problem itself is seldom taken seriously. The general attitude is well expressed in Symonds' statement that " . . . the concept of reliability loses its importance with a projective technique whose purpose is not so much to measure as to describe . . . The accuracy and the consistency which are expected from tests are not looked for in projective techniques" (47, p. 44). Frank, in his well-known monograph Projective Methods (22, Ch. V) builds up an elaborate, though fallacious, argument that the psychometric concepts of reliability and validity do not apply to projective tests. In her overview of thirty years of Rorschach research, Hertz (27) does not mention reliability. Cronbach (12), in his review of statistical methods in Rorschach research, reports fallacious methods of estimating reliability among some of the few studies aimed at this problem. Cronbach was led to conclude: "No entirely suitable method for estimating Rorschach reliability now exists. Studies in this

[1] United States Public Health Service Research Fellow of the National Institute of Mental Health.

area are much needed" (12, p. 426). Perhaps the fullest treatment of the reliability problems of the Rorschach, as well as of projective techniques in general, is to be found in the review by Ainsworth (2). However, she hardly more than suggests some of the difficulties and shortcomings in this special area, and offers little that can help to improve the situation.

The purpose of the present article is to fill some of these gaps in projective methodology by elucidating the meaning and necessity of reliability estimates in projective research and reviewing the past contributions on reliability. In another paper (29) methods are presented for estimating reliability in the face of the difficulties imposed by conditions more or less peculiar to projective tests.

Meaning of Reliability

Reliability is the degree of accuracy of a test in measuring whatever it measures. Statistically defined, reliability is the proportion of non-error variance in test scores. It is important to realize that a test always has some degree of reliability, varying anywhere from 0 to 1.00 as estimated by the reliability coefficient, whether it is formally estimated or not. Some amount of error variance exists in all test scores, and this error, even if it were impossible to estimate it statistically, would enter into any use made of the test. In short, a test's reliability logically has nothing to do with the degree of difficulty of estimating the reliability; the effective reliability of a test would not be at all affected even if a quantitative estimate were totally impossible. Fortunately, it is nearly always possible to make a "best estimate" of a test's reliability, and this is true also of projective techniques.

The concept of reliability and the methods of estimating it have always occupied a prominent place in psychometric theory. Few would argue with Guilford's statement that "Reliability is the minimum information one should have concerning a test" (24, p. 388). There are numerous reasons why it is useful to have an estimate of a test's reliability, but all of them stem from the primary goal in psychological testing, which is to maximize discriminations. Reliability is a necessary, though not sufficient, condition for effective discrimination. An estimate of reliability is one index of the degree to which the test can discriminate, and is thereby also a measure of the test's *potential* effectiveness. The test's *actual* effectiveness depends, of course, upon its validity, that is, the degree to which it measures what it purports to measure.

One of the characteristics of scientific investigation is reproducibility. The reproducibility of a psychological study is a function of the reliability

of the techniques employed. Without a substantial degree of reliability cross validation is not worth while; and where cross validation is not worth while, the original study itself represents only wasted effort. If adequate reliability estimates had been taken into account in the early stages of many a research with projective techniques, the investigators could have been saved their misspent efforts. Low reliability may be largely responsible for the fact pointed out by Cronbach that "In most studies, correlations nearly vanish when a Rorschach prediction formula is tried on a new sample" (12, p. 402). In the research use of any technique one must ask what are the chances that another investigator using the same technique at another time or place will achieve comparable results. This implies an estimate of reliability. Furthermore, if research findings are to be applied in practical situations, reliability becomes even more crucial.

It next becomes necessary to discuss the degree of reliability required for a given purpose. In research aimed at testing the null hypothesis high reliability, while desirable, is not always essential, and tests with very low reliability coefficients (.30 — .50) may still be of value. As long as the errors of measurement are random and there remains a portion of true score variance, a test is capable of serving a purpose in research. But the lower the test's reliability, the larger in the number of cases needed to make true discriminations; and the test of low reliability, while having perhaps proved a theoretical point by making use of large samples, cannot be said to be of any value in making discriminative statements about individual cases. Thus the individual or clinical use of a test, as it is called, is generally said to require a very substantial degree of reliability, that is, a reliability coefficient of .90 or better. Reliabilities in this range are the rule for cognitive tests, but few if any personality tests achieve so high a reliability, most of them falling in the range of .70 to .90.

Even for certain theoretical purposes it is doubtful if reliabilities much below this range can be useful. For example, since the highest possible correlation between any two tests is the square root of the product of their reliabilities, low reliability is anathema to factor analysis. Guilford has mentioned .60 as the lowest test reliability to be tolerated in a factor analysis (24, p. 532). It should also be pointed out that the most frequent use of projective techniques in research is not in testing hypotheses of theoretical interest only, but in demonstrating various clinical uses of a particular technique. Thus a high degree of reliability remains important.

The clinical applicability of research findings with projective techniques

can be judged partly by the standard error of measurement ($SE_{meas.}$)[2], which is a function of the reliability of the scores, diagnostic signs, or whatever the case may be. Indeed for most purposes the best clue to the adequacy of a test's reliability is the relationship of the standard error of measurement to the range of scores or categories in the sample in which the test is being used. From the $SE_{meas.}$ and the standard error of the difference between two obtained scores ($SE_{meas.}\sqrt{2}$), one can get some idea of the number of discriminations that can be made among the scores in one's sample. The degree of reliability below which a test should be discarded is a matter of practical economy and cannot be decided in general terms. The important thing is to have some knowledge of the probable degree of accuracy involved in using a particular test.

Objections to Reliability Estimates on Projective Tests

So many different objections have been made to attempts to estimate the reliability of projective techniques that one is almost left with the impression that reliability criteria are entirely out of the question in this realm of psychological testing. One objection concerns the way in which projective scores, for example Rorschach M, K, FC, etc., or content, e.g. Human, Animal, Anatomy, etc., or various qualitative aspects of verbal style, are used by the clinician in interpreting a projective protocol. It is claimed that these separate elements are important only as they are viewed all together as a gestalt or total configuration. But this fact in no way lessens the need for reliability; a configuration is actually a set of

---

[2] The standard error of estimate may be preferable to the standard error of measurement and should be used when retest reliability is known and one wishes to estimate the probable limits within wich an individual's score will fall on a retest. Since repeatability is an important factor in a scientific study, the standard error of estimate would seem be a more satisfactory estimate of the test's potential discriminatory power. The standard error of measurement is used in determining the significance of an individual's score on one occasion or in determining the significance of the differences between the scores of two or more individuals on the same test. The standard error of measurement tells us the probable limits within which the "true" score lies. The standard error of estimate tells us the probable limits within wich the score on retest lies.

$$SE_{measurement} = s\sqrt{1 - r_{tt}}$$

$$SE_{estimate_{12}} = s_1\sqrt{1 - r_{12}^2}$$

where s = standard deviation of test scores
1,2 = test administered on occassions 1 and 2
r_{tt} = split-half or retest reliability.

differences between various quantities or classifiable qualities. The reliability of these quantities or classifications, therefore, obviously affects the reliability of the whole configuration of which they are a part.

Along these same lines is the argument that a projective technique is not truly a test or measuring instrument at all, but is more like an X-ray plate, in that it gives one a picture rather than a measurement. But X-rays, too, are used for making discriminations and the discriminations must have reliability. The reliability of the discriminations in the case of an X-ray would be a function of the degree of clarity and the amount of artifact in the X-ray picture, and would be determined by the consistency with which different judges could classify the X-ray according to the presence or absence of certain signs. Projective protocols are often quite similar to X-ray plates in this respect.

The split-half method of estimating reliability has also been strongly criticized (2, 40). With respect to the Rorschach, for example, Ainsworth has stated that "Because of the small number of 'items' represented by the ten blots and because of the fact that the standard series of blots was originally selected because the blots differed markedly in 'stimulus quality', it would seem reasonable to reject the split-half method as unreasonable" (2, p. 442). She concludes that "The trend of opinion is currently away from the split-half method" (2, p. 443). Actually it is not at all unreasonable to apply the split-half method to the Rorschach, the TAT, or most other projective techniques. Provided a proper method of analysis is used, the split-half reliability of projective test "scores" can provide valuable information. It is, of course, a statistically different type of information than is provided by test-retest reliability. The split-half estimate tells us something of the internal consistency of the test with respect to the "scores" derived from it. If, for example, it is claimed that the Rorschach F+ response has generally the same psychological meaning irrespective of the particular blot which elicits it, then a split-half estimate of the reliability (internal consistency) of F+, provided a suitable statistical method is employed, is no less justifiable than determining the split-half reliability of an MMPI scale or a subtest of the Wechsler-Bellevue.

Piotrowski, while objecting to split-half estimates of reliability, has recommended retest estimates (40). The two types of reliability, as has already been pointed out, are actually measures of two different properties of the test — internal consistency and stability in time. Retest reliability is important if it is claimed that the test reflects relatively stable, enduring personality characteristics. In both theory and practice this is certainly

claimed for projective techniques. Therefore estimates of retest reliability are of great importance. Yet there have been objections also to retest estimates of reliability, most of which can be summarized as follows:

1. A retest is psychologically not the same experience as the initial test.

2. The subject may remember his initial responses and therefore retesting is only a measure of recall and not of reliability.

3. Projective techniques are so sensitive to the slightest changes in the subject, reflecting transient moods, etc., that a lack of correspondence between test and retest is said to be due to genuine changes in the subject.

Those who discount low reliabilities for this reason also claim, however, that the particular technique can be given repeatedly to the same subject and still be valid. If retest reliability is meaningless, then it can be argued that the particular day or hour on which a subject is tested becomes all-important; and the test results would have little generality beyond this brief time span. The generality of a research study would, therefore, also be seriously in question. Furthermore, the practice of using projective techniques for the measurement of personality change, such as in evaluating the effects of psychotherapy or of drugs, must rest on the assumption of substantial retest reliability. If a test has negligible retest reliability, then it is hard to see how it can be used legitimately to measure specially induced changes in personality. The first two objections listed above can be answered only on the basis of empirical evidence. Methods of estimating retest reliability that tend to minimize the drawbacks suggested above are offered in another paper (29).

It may be argued that reliability is an unimportant concern if the test in question has demonstrable validity, since a valid test must also be reliable. This is of course true. It is possible to make certain inferences about the reliability of a test on the basis of its validity; but it is also possible to do the reverse. If a test can be shown to have negligible reliability, then it also has negligible validity, the highest possible validity coefficient being the square root of the reliability coefficient, a validity which would obtain only under exceptional conditions. In many cases, because of sampling and criterion problems, it is much more difficult to establish validity than reliability. The investigator can sometimes be saved the trouble of a difficult validity study simply by making a careful estimate of the test's reliability. Knowing this, along with some idea of the reliability of the criterion to be used in his validity study, he can estimate the probable outcome of the validity study. It could be shown at the outset that the test is

too unreliable to be valid. This procedure can of course be only of negative value as regards the test's validity. If the highest validity coefficient that could be expected under these conditions proves to be too low for whatever use is proposed for the test, then the validity study would not have to be undertaken.

Another point that is sometimes made concerns the distinction between the hypothesis-generating and hypothesis-testing functions of a clinical instrument. It is assumed that a high degree of reliability and other such paraphernalia of statistical rigor are demanded of hypothesis-testing techniques but are unimportant when a technique is used only as a means of suggesting clinically relevant psychological hypotheses about a patient. Projective tests are usually regarded as being of the hypothesis-generating variety. That is, they do not attempt to prove a point but only to turn up likely clues (hypotheses) as to what the point might be. Even though this may be the case — and in practice it nearly always is — it still does not excuse the projective test from the requirement of reliability. A test that generates just any hypothesis is worthless. The generating of hypotheses is again essentially a discriminatory process; hence the importance of reliability. If a test is to be of value in generating hypotheses, it must generate reliably.

Special Problems of Projective Techniques

There are certain reliability problems which are more or less peculiar to projective techniques. One of these is the fact that, in general, projective techniques do not have a standardized method of administration, scoring, or interpretation. In both research and clinical practice there is far less variation in the administration, scoring, and interpretation of, say, the Wechsler-Bellevue or the MMPI than of any projective technique. An objective standardized test is used with so little variation that a new reliability estimate is not called for every time it is used in a research study. We usually know its approximate reliability from past studies and can assume that it will be affected more by differences in the populations being tested than by differences among the investigators using the test. In surveying the research literature on projective techniques, on the other hand, one finds an enormous variation in the ways particular techniques are used. Because there is no standard procedure in projective testing, reliability estimates are especially important. This applies to both repeat reliability and scoring reliability. It is possible to score something reliably which is so ephemeral as to have no repeat reliability. Ideally such relia-

bility estimates should be reported in every research study using a projective technique.

Projective techniques are as yet unvalidated for many of the purposes to which they lay claim. The use of a projective technique as a criterion instrument for testing hypotheses about personality often seems highly questionable. At the outset all we can be sure that the device is measuring is what is explicitly built into the scoring system, such as K, M, FC or *n* Aggression, *n* Affiliation, *n* Nurturance, etc. The reliability of the scores is not to be taken for granted, but must be explicitly estimated and reported along with whatever research findings are based on them.

In projective testing there is also a longer and far less explicit chain of operations between the basic datum of the test protocol and the clinician's interpretation than is the case with so-called objective tests. Often qualitative, stylistic factors in the protocol, not accounted for in the formal scoring system, enter into the interpretive end-result. As projective techniques become more highly developed along scientific lines these important though seemingly more elusive factors will be subjected to more objective evaluation or scoring. Attempts in this direction have already been made (41, 42, 52). Thus more and more of the projective material will lend itself to adequate reliability estimates. In this realm the problem of *scoring* reliability takes on added importance because of the highly subjective nature of the more subtle qualities of projective productions.

Projective material is similar to the essay-type examination in that the scoring is not completely objective, as it is in a scale like the MMPI. In objective tests scoring reliability is no problem; perfect reliability can be achieved by any clerk whose scoring is free from careless errors. The objective test thus has only one source of error variance, which is estimated by the tests correlation with itself, either in terms of internal consistency or of repeat stability. Projective tests, on the other hand, have an additional source of error. It is estimated by some measure of the degree of agreement between two or more judges in scoring or interpreting the same test protocol. These two sources of error (i.e. test error and scoring error) are logically independent of each other and may therefore be estimated separately. These problems are treated in the paper on Methodology (29).

The Reliability of Various Projective Techniques

Since the aims of this and a following paper (29) are mainly theoretical and methodological, an exhaustive review of the evidence on the reliability of projective techniques is not attempted. However, a number of the most

representative studies of the reliability of specific techniques are presented in order to provide some idea of the reliability one can expect with various techniques. The estimates of reliability reported in the literature can be useful to anyone planning a research with projective techniques.

Because of the great lack of uniformity in scoring methods, reliability coefficients for single test factors, such as F+, C, M, etc., cannot be regarded from one study to another as being simply different reliability estimates of the same factor. For this reason the reliabilities are reported here in only the most general terms. The reader is referred to the original studies for more specific data.

Rorschach

Split-half reliability

Although the split-half method of estimating reliability has met with objections from Rorschachers, split-half estimates have generally been higher than test-retest estimates. As Eysenck has pointed out, "When high split-half reliabilities are in fact obtained, we must conclude that the basis for objection to split-half reliability is invalid" (18, p. 164).

The first attempts to determine the split-half reliabilities of Rorschach scores were those of Vernon (50, 51) and Hertz (26). Vernon split the cards I, III, V, VI, X *versus* II, IV, VII, VIII, IX. His *Ss* were 90 college students. For eight Rorschach scores he obtained an average reliability coefficient of .54 (corrected by Spearman-Brown formula), with a range from .33 (F+%) to .91 (R). Vernon regarded these reliabilities as too low for individual prediction and concluded " . . . if the test is to be regarded as a test at all, or if it claims any objective validity, it must in the future be modified in such a way that the reliabilities of the chief categories of response may achieve a level of at least .70 to .80" (51).

Hertz (26) split the cards odd *versus* even. Her *Ss* were 100 junior high school students. For twenty Rorschach scores she obtained an average reliability coefficient of .83 with a range from .67 (P%) to .97 (Anatomy %). On the basis of the available information it is impossible to account for the great discrepancy between the reliabilities found by Vernon and Hertz. The lesson we can learn from this striking discrepancy is that we cannot speak of *the* reliability of the Rorschach or of any one score; the reliabilities found in one study apparently cannot be generalized at all precisely to another study carried out in a slightly different manner. This emphasizes the need for reliability estimates in every study which introduces any change in procedure or conditions. Reliability varies partly

as a function of the method of scoring, of the training of the scorers, and of the particular sample of *S*s used.

The split-half reliability, not of formal Rorschach scores, but of an "Index of Pathological Thinking" (see 42, 52), based on the content and stylistic aspects of the Rorschach response, was determined by Pope and Jensen (41) by splitting the test by taking odd *versus* even *responses* rather than cards. The corrected split-half reliability coefficient was .52. The *S*s were 41 schizophrenic patients.

The Harrower-Erikson Multiple Choice Rorschach is an objective test and hence there is no scoring error. Eysenck (16) found the split-half reliability of the MCR to be .64. In order to improve the reliability he adopted a different method of administering the test. Instead of having a multiple-choice the *S*s were required to rank 9 alternatives — four typical normal and five typical neurotic responses offered for each blot. The test was given under these conditions to 300 neurotics and the split-half reliability was raised to .84. Since scoring error does not enter the picture, this is probably one of the highest split-half reliabilities ever obtained with any projective test. Thus the objectification of administration and scoring of projective tests would seem to be promising, at least as regards reliability.

Test-retest reliability

Most of the estimates of retest reliability are hardly comparable for the same reasons mentioned in connection with split-half reliability, in addition to the important fact that the time interval between test and retest varies markedly from one study to another.

Hertz (26) measured retest reliability by a method which confounds retest with split-half, which are two different properties of a test and are therefore quite meaningless when so combined. Hertz split the cards odd *versus* even and tested 145 *S*s on the odd and even cards in two sessions separated by a two weeks interval. The "reliabilities" of only six scores were reported for three groups. The average corrected coefficients are W% .83, D% .80, DS% .76, M% .61, C% .74, F% .77. These figures represent neither a split-half nor a retest estimate of reliability but an amalgam of the two. The figures would be more instructive if we also had split-half estimates based on the whole test taken in one sitting by a comparable group of *S*s. Then by means of an analysis of variance we could assess the effect of the two weeks interval between test and retest.

The time interval between test and retest apparently affects the reliability of some scores more than others. Swift (46), using preschool children

reports retest reliabilities of the formal Rorschach scores ranging from .59 to .83 with a two weeks interval between test and retest. With a ten months interval the reliabilities ranged from .18 to .53. Altus and Thompson (1) determined the retest reliability (six weeks interval) in 100 college students on 14 different Rorschach indices of intelligence based on formal scores. The mean reliability was .63, with a range of .13 to .93. A study by Blanton and Landsman (7) specifically addressed to the matter of retest reliability of the Group Rorschach does not report any reliability coefficients at all, but concludes that low reliabilities were to be expected in a college population anyway. Actually the contrary has been found to be true for most tests and questionnaires.

A study by Baker and Creager (4) similar to that of Eysenck (16) referred to previously, again suggests that an objectively scorable Rorschach is more likely to have adequate reliability. Thirty-two college students were required to rate sample Rorschach responses on a five-point scale for "goodness". Split-half reliabilities for the determinant categories ranged from .74 to .97, with a median of .86. The retest reliabilities (two weeks interval) ranged from .51 to .90, with a median of .78. It is interesting to note that when the usual scorer error is eliminated reliabilities compare favorably with those of other objective tests of personality.

An attempt to overcome the objection that retest reliabilities are inflated by the *S*'s memory of his first performance was made by Kelly *et al.* (30) by administering the Rorschach before and after a single electro-shock reportedly sufficient to wipe out any memory of the initial performance. The results obtained from 12 patients show that R (number of responses) shifted as much as 50 percent between test and retest and that the absolute values of other scores shifted also. It was presumed that the electro-shock has no effect on test performance. These results can hardly be interpreted. The authors made no attempt at statistical treatment of the data, probably because of the small number of cases.

Fosberg (21) attempted to determine the retest reliability from 26 cases in the literature where two or more Rorschachs were given the same persons. The reliabilities he obtained are spuriously high because of the fact that his method of determining reliability is fallacious. He correlated two sets of scores for each person; that is, pairs of values such as $W_1 - W_2$. $D_1 - D_2$, etc. were entered into the correlation chart. Since the various scores differ in magnitude for all *S*s in general (e.g. D is generally greater than W, M greater than m, F greater than C, etc.) one would certainly expect to find with Fosberg's method that fairly high

correlations obtain even between protocols paired at random. The conclusions of another study by Fosberg (20) are based on a similar fallacious method of analysis.

Parallel Forms

Test-retest reliability can also be estimated by the correlation between parallel forms of a test given to the same *Ss* at different times. This method gets around the objection that retest reliabilities with the same test may be inflated because of memory. The three essential criteria for a parallel form are that it have the same mean and standard deviation as the original test and that it correlate as highly with the original test as the original test correlates with itself. If these conditions are satisfied, the forms are considered parallel and may be used for estimating retest reliability.

There is a parallel form of the Rorschach which appears to meet these criteria — the so-called Behn-Rorschach. A study by Meadows (see 18, pp. 164-165) shows that for 35 scoring categories the means and standard deviations do not differ significantly in normal and abnormal groups of *Ss*. The correlations between the Rorschach and the Behn on 35 different scores are presented by Eysenck (18, p. 166). For the normal group ($N = 100$) the correlations average .41, with a range from —.05 to .83. For the abnormal group ($N = 96$) they average .52, with a range from —.10 to .86.

From these results one might be tempted to argue that the Behn does not satisfy the third criterion for a parallel test, that of correlating as highly with the original test as the latter correlates with itself. This would be a valid argument except for the fact that the evidence is to the contrary. A study by Eichler (15) shows that the correlation between Behn and Rorschach is not significantly different from the correlation of the Rorschach with itself. Thirty-five *Ss* divided into three groups were treated as follows:

Group I. Behn followed by Rorschach after 20 days.
Group II. Rorschach followed by Behn after 21 days.
Group III. Rorschach followed by Rorschach after 21 days.

The protocols were scored on 12 major categories. The mean correlations were: I. Behn *vs.* Rorschach .65; II. Rorschach *vs.* Behn .56; III. Rorschach *vs.* Rorschach .68. None of the correlations is significantly different from the others. Thus it appears that the Behn may legitimately be used as a parallel form of the Rorschach, at least as far as the formal scores are concerned.

Reliability of Scoring and Interpretation

Even though inter-scorer reliability is the most easily estimated type of reliability and also the easiest to control, it is rarely reported in the projective test literature. The apparent lack of concern about scoring reliability is seen also in the fact that reports of retest reliability seldom, if ever, specify whether the test and retest protocols were administered or scored by the same examiner or by different examiners. Thus the proportion of error that the scoring itself contributes to the estimate of retest reliability usually remains undetermined. In spite of its neglect, scoring reliability remains important and it should be required as the minimum evidence in any projective research that the study is replicable and therefore possibly of scientific value. Its importance for the clinical application of a technique is obvious.

Scoring reliabilities span a wide range. For formal scores reliability is generally good, provided the scorers use the same methods and have practiced together. Scoring reliability is greatly affected not only by the similarity in experience of the scorers but also by the population in which the protocols are obtained. For example, on a Rorschach "Index of Pathological Thinking" Watkins and Stauffacher (52) obtained scoring reliabilities of .04, .47, and .91 for normal, neurotic, and psychotic groups respectively. The scoring reliability of a similar Index was found by Pope and Jensen (41) to be .85 for the protocols of schizophrenics.

Klopfer (32, p. 20) has stated that "Important though reliability may be, it is secondary to interpretation." One might incorrectly conclude that reliability applies only to the scores and not to the interpretation based on them. If the scores are merely secondary to interpretation, then the reliability of the interpretation becomes more important than that of the scores. It may be stated as a general rule that the most crucial reliability is that of the end product of the test. This may be a single score, a profile of scores, or a verbal description based on a global evaluation of the protocol.

The writer has found no study reported in the literature in which reliability of interpretation was higher than that of the basic scoring of the protocols. In fact, it is in the reliability of interpretation that the Rorschach, as well as other projective techniques, falls down most drastically. A general example of this is found in the Kelly and Fiske study (31, p. 63) in which each of four raters rated each of four projective protocols (Rorschach, Bender-Gestalt, TAT, and Sentence Completion) on six personality variables for 20 *Ss*. The median inter-rater correlations

for each of the techniques was: Rorschach .44, Bender-Gestalt .35, TAT .28, Sentence Completion .27.

Korner and Westwood (33) had three clinical psychologists qualified in the Rorschach sort the protocols of 96 Freshmen college students into three categories for level of personality adjustment. The average correlation between the three judges was .31. The same procedure was applied to drawings of the human figure, yielding an average correlation between the judges of .65. The higher reliability of the Figure Drawing indicates that it is easier for a number of judges to make the same global judgment on the basis of a single drawing than on the basis of an entire Rorschach protocol. Probably the more material that must be integrated in making a subjective global judgment, the lower will be the reliability of the judgment.

Probably because of differences in the population sampled, Grant *et al.* (23) obtained higher inter-rater agreement than did Korner and Westwood on adjustment ratings based on the Rorschach. Grant *et al.* had three expert Rorschachers rate the protocols of 146 boys and girls on a four-point scale for general psychological adjustment. The average intercorrelation between the three Rorschachers was .67. Two other raters who were comparatively unknowledgeable about the Rorschach (a social worker who had only been tested with the Rorschach and a psychology student who had taken a semester's course in Rorschach) achieved about as high a degree of agreement with the "experts" as the latter agreed among themselves. The authors concluded: "Apparently several months of training and years of experience in the application of a complicated scoring system have contributed little to the agreement among the "experts" when the discriminations called for are no finer than a four-point scale" (23, p. 7).

Datel and Gengerelli (14) found that when 27 Rorschachers were required to match personality interpretations written by each other on the basis of the protocols of six *Ss* (presented for matching in sets of six) there were more mismatchings than correct matchings. Of the total of 324 discrete matchings, 148 were correct and 176 incorrect. The 18 *Ss* from whom the protocols were obtained were selected so as to differ greatly from one another.

A study by Lisansky (35) also makes possible a comparison of the reliabilities of scoring and interpretation. She asked six highly qualified Rorschachers to rate 40 *Ss* on 10 personality items which they agreed could be confidently assessed from the Rorschach protocol. To make the experiment similar to clinical conditions the Rorschachers were provided

also with an abstract of each patient's history. The degree of agreement between the Rorschachers was measured by the phi coefficient. The average phi was .33 [3]. For reliability of *scoring* the protocols, the average phi was .64. Six other clinical psychologists who answered the 10 personality questions on the basis of the patient's case history alone showed an average phi of .31, that is, a reliability not significantly different from the Rorschachers who used the case history in addition to the Rorschach. The ten personality items were specially selected as being the kinds of questions which the Rorschach, and not particularly the case history abstract, is supposed to be able to answer.

Thematic Apperception Test

Symonds' observation in 1949 that "Many of those who have written about the TAT are quite willing to forego an attempt to determine its reliability." (47, p. 44) still holds true. While there are now over 700 publications on the TAT, only a few provide us with any evidence concerning its reliability.

Split-half Reliability

There is a widespread misconception in the TAT literature that split-half or internal consistency reliability is either impossible, incorrect, or meaningless in the case of the TAT. Actually internal consistency reliability, when properly determined, is as justifiable with the TAT as with any other test in which various elements are additive. The conditions necessary for this kind of reliability to be meaningful seem to be consistent with TAT theory and practice. Only if it is maintained that each single picture in the TAT must be regarded as a unique "test" can we disregard internal consistency reliability. A story involving murder elicited by card 13 MF is said to represent "aggression" and a story involving murder elicited by card 18 BM is also said to represent "aggression". And so with all the *themas* evoked by the TAT.

It should be emphasized that any determination of internal consistency reliability which is based on only one division into halves of the TAT cards is methodologically unsound. The various cards differ markedly in their "pulling-power" for different themas. For this reason reliabilities based on all possible split-halves will span a much wider range than

[3] The phi coefficient is not directly comparable to Pearson's r. A phi of $+.33$ is significant between the .01 and .025 level, wich would correspond to an r between .35 and .39 when $N = 40$.

would be the case with most tests. What is needed is a reliability coefficient representing, in effect, the average of the reliabilities based on all possible divisions into halves.

The formula for determining such a reliability coefficient is given by Gulliksen (25, Formula 11 [4], p. 224), and has been used for the TAT by Child *et al.* (10). Ten of the major Murray TAT variables (e.g., Achievement, Aggression, Autonomy, etc.) were scored for their presence or absence, along with the presence or absence of anxiety concerning the particular theme, in the TAT stories of 183 college students. The Achievement theme was scored quantitatively according to the method used by Mc Clelland in his study of the achievement motive (36). The internal consistency reliabilities of the various themes ranged from —.07 to +.34 with a mean of .13; reliabilities of the "anxiety about the theme" ranged from —.21 to +.44, with a mean of .02. The reliability of the quantitative rating of *n* Achievement ranged from .32 to .43 depending upon various methods of estimation. The low internal consistency reliability cannot be attributed to low scoring reliability, for the latter was quite high. The same judge rescored 25 of the protocols after a six months interval and obtained a correlation of .89 between the two sets of scores. It seems clear from these findings that the TAT themes elicited by different TAT pictures cannot be regarded in an additive way. Apparently there are as many different kinds of "Aggression," "Autonomy," etc. as there are TAT cards. A *S*'s response to one card is no basis for prediction of response on other cards. The authors conclude that "The internal consistency of our group TAT for many of our variables is so low as to suggest that the scores for individuals can represent nothing but random

[4] This is the Kuder-Richardson "Formula 20". The validity of this formula for the purpose discussed here may be questioned because of the one assumption made in its derivation, viz., that all "items" in the test have approximately the same degree of difficulty, which in the case of the TAT would mean that all cards have the same degree of "pulling power" for the various scored themas, an assumption that is obviously untrue. Therefore we must ask what are the effects of applying the Kuder-Richardson Formula 20 when the assumption of "equal item difficulty" does not hold. Brogden (8) investigated this question empirically and found that even extreme differences in "item difficulty" had very little influence on the reliability coefficient as determined by the K–R formula. The greatest deviations from the ideal conditions never caused a reduction in the reliability coefficient as large as its standard error, so the reductions may be regarded as insignificant. Since the ideal conditions are never met in the case of the TAT (or hardly any other test for that matter), it is probably most scrupulous to regard the reliability as determined by the K–R formula as a lowerbound estimate. Still it is probably the best estimate possible of the internal consistency reliability of the TAT.

variation, and of course to suggest that the same may in some instances be true of previous research where similar measures have been used without any attempt to measure internal consistency" (10, p. 104). The authors also constructed a 200 item personality questionnaire covering the same Murray variables (10 items for each of 10 variables and 10 items for anxiety in relation to each of the variables). The split-half reliabilities of these scales, in contrast to the TAT scores, were quite high, ranging from .39 to .91, with a mean of .73. While the TAT scores showed no correlation with certain validity criteria, the questionnaire scales did show significant positive correlations.

By using specially selected pictures it may be possible to achieve higher internal consistency for a particular variable. McClelland (36) used a special set of pictures to elicit achievement related themes. He reports that the split-half reliability for this set of pictures was "over .70".

Auld *et al.* (3) obtained an internal consistency reliability coefficient (as estimated by the Kuder-Richardson Formula 14) of a special Sex Scale on the TAT of .43 ($N = 100$).

Another appropriate means of determining split-half reliability that is not subject to the usual objections is to divide the set of TAT cards into halves and rank-order the *S*s' scores on each variable within each half of the test. A rank order correlation [5] may then be obtained between the two halves of the test on each variable. The differences in "pulling power" of the various TAT pictures for different themes, which has been the main basis of objection to split-half reliability, is eliminated by this method. What is important for reliability is a consistent ordering of *S*s on each theme in two sets of protocols, such as may be formed by an odd-even split of the TAT. The reason for using rankings rather than raw scores is to ensure one of the conditions necessary for determining split-half reliability: that the two halves be parallel as regards their standard deviations. Any split-half sets of raw scores on the TAT are unlikely to have even approximately equal standard deviations; and the skewness of raw scores on TAT variables is usually marked, thus prohibiting the use of the Pearson *r*. Hence the need to reduce the raw scores to a rank order.

Lindzey and Herman (34) determined the split-half reliability of six variables based on eight TAT cards (1, 4, 10, *versus* 2, 5, 13MF, 15) in a college sample ($N = 148$). Each of six TAT variables was rated on a five-point scale and the two sets of scores were correlated (Pearson *r*)

[5] Though the rank order correlation, rho, is identical to the Pearson *r* for ranked data, rho has a larger standard error than *r*.

for each variable. The authors optimistically boosted the obtained *rs* by the Spearman-Brown formula to the magnitude they would theoretically assume if the entire 20 TAT cards were used. This is actually an abuse of the Spearman-Brown formula, the use of which rests upon certain assumptions (25, p. 64-67) which are untenable in the present case. The reliabilities thus obtained are almost certainly gross over-estimates. The correlations between the two halves of the TAT were as follows (*r* corrected by the S-B formula is given in parentheses): *n* Achievement .19 (.54), *n* Aggression .29 (.67), *n* Sex .45 (.80), *n* Abasement .28 (.66), *n* Nurturance .12 (.41), Narcism .20 (.56).

Test-Retest Reliability

This form of reliability is important to the extent that a test purports to tap the more or less stable, underlying structure of personality. Since this claim is made in varying degrees for the TAT, an estimation of its retest reliability is essential. Because there is no parallel form of the TAT, retest reliability is difficult to estimate satisfactorily. Memory for TAT pictures and the stories they elicit probably plays a greater part in retest than is the case with the Rorschach. This kind of "contamination". makes it difficult to evaluate the retest reliabilities that have been reported.

Tomkins (49) tested and retested 45 women at various intervals from two to ten months. The average retest reliabilities of the Murray need-press variables were as follows: after 2 months .80, 6 months .60, 10 months .50. These figures begin to suggest the curve of forgetting and one might wonder if these coefficients represent the reliability of measuring the personality factors presumably tapped by the TAT or simply the reliability of the *S*'s memory. Subsequent studies have shown that *S*s are able to remember their TAT stories even after a lapse of months. Lindzey and Herman (34) tried to get around this problem by requiring *S*'s to make up a different story on retest if they recalled their first story. By eliminating the first story there still was left a theoretically unlimited number of possibilities for the story to be told on the second administration of the test. The retest was given after a two months interval to 20 *S*s, using, unfortunately, only four TAT cards. Retest reliabilities (tetrachoric *r*) for 17 scored variables ranged from .00 to .94 with a mean of .51. Only three of the 17 reliability coefficients were significantly larger than zero. The authors boosted the reliabilities by using the Spearman-Brown formula and theoretically increasing the length of the test by a factor of 5, as if all 20 TAT cards had been used. This procedure increased the total number of significant (at .05 level) reliabilities to six. It should be

pointed out, however, that the use of the Spearman-Brown formula in this case is not only improper, but tests of significance cannot be legitimately applied to S-B corrected *rs*.

Other studies have yielded similar results. Auld *et al.* (3) found the retest reliability (one month interval) on a Sex scale to be .13, which is not statistically significant. However, the *Ss* in this experiment spent the entire interval between test and retest in a submarine, an experience which might well have affected the balance of "needs and preses" in the personality picture.

McClelland (37) reports a retest reliability (1 week interval) for his quantitative scoring of *n* Achievement of .22. He adds that 72.5 per cent of the *Ss* were on the same side of the mean on retest. It is interesting that in McClelland's study the TAT correlated more highly with a number of objective behavior tests than its own reliability, i.e., its correlation with itself. Evidently the reliability of the objective tests was markedly higher than that of the TAT, since the highest true correlation possible between two tests is the square root of the product of their reliabilities.

Scoring Reliability

The correlation obtained between the ratings of different judges (usually two) on the same set of protocols, that is, the scoring reliability, sets the upper limit for the reliability of the test. Actually the effective reliability of a test is always lower than its scoring reliability, and in the case of the TAT this is especially true. Thus the scoring reliability alone, which is by far the most frequently reported form of reliability in TAT research, can lead to over-optimism in judging the tests' effectiveness. The writer combed the TAT literature for estimates of scoring reliability based on sound methods and presented in the form of the product-movement correlation coefficient so as to be strictly comparable to the usual estimates of test reliability. The average of 15 such reliability estimates is .77, with a range from .54 to .91. (For the details of these studies see 3, 10, 34, 38, 44, 49.) On the whole these are not very high reliabilities for scoring. Gulliksen (25, p. 212) states that the scoring reliability of essay examinations should be above .90 and he considers reliabilities below .80 as unacceptable.

More than half the reports of scoring reliability of the TAT are given in the form of percentage agreement between raters. Most of the figures range between 60 and 90 per cent (11, 13, 38, 49). Percentage agreement

between judges in an unsatisfactory way of reporting reliability, for it tells us much less than we wish to know from a reliability coefficient.

Fine (19, p. 307) presents a formula [6] for the scoring reliability of the TAT which should never be used. It is so incorrect as to suggest that it might be a misprint (the reliabilities he reports could not have been arrived at by his formula) but still there is no way in which the elements of the formula can be rearranged to yield anything resembling a reliability coefficient. If the numerator and denominator in the formula are reversed, it simply gives a kind of percentage agreement. (However, it still would not be the correct formula for percentage agreement, which is the number of actual agreements divided by the number of possible agreements and not simply the number of observations.) In any case, Fine is incorrect in stating that his "empirical r" is "closely comparable to the phi coefficient."

Other Projective Techniques

Most of the other projective techniques have been subjected to much less research effort than the Rorschach and TAT and consequently there is scant evidence on their reliability. The few reliabilities that have been reported are quite similar to those for the Rorschach and TAT.

Sentence Completion Test

Rohde (43) scored 33 variables each rated on a scale of 1 to 10, for 64 incomplete sentences given to 50 *S*s. There was an average agreement among four scorers of 78 per cent. Test-retest reliabilities for different groups averaged .79.

Bender-Gestalt Test

This test, as it is generally used in the diagnosis of brain damage, is not a projective technique but a performance test. However, the expressive aspects of the *S*'s drawings are occasionally used by clinicians as projective material. There are no estimates of the Bender's reliability when used in this fashion. Pascal and Suttell (39) have developed a quantitative scoring system for psychiatric diagnosis, excluding brain damage. The scoring scheme consists of 105 types of deviations from the stimulus, the deviations being weighted according to their discriminating power. The authors report that, with practice, scoring reliability can reach .90. Test-retest reliability (24 hours interval) was .71. Suttell (45), using the

---

[6] $$r = \frac{\text{Total number of observations}}{\text{Number of agreements} \times 2}$$

Bender-Gestalt on children (N = 40), obtained a scoring reliability of .91, retest reliabilities of .96 after 24 hours and .74 after 6 months. The higher reliabilities found by Suttell are probably a result of the greater spread of scores among children than among the adult *S*s in the Pascal and Suttell study.

Picture-Frustration Test

Rosenzweig's *P-F* Test has found more applications in research than in clinical practice. The internal consistency reliability for six scoring categories is reported to range from .10 to .58, a finding which undermines the additive use of the scores obtained from each picture. Test-retest reliabilities (2 to 7 month intervals) for the major *P-F* scoring categories range from .34 to .71 for the Adult Form and from .26 to .69 for the Children's Form (9, p. 211). Bernard (6) retested 105 adults after intervals of 3 to 9 months and obtained reliabilities ranging from .45 to .73 for the six scoring categories and the Group Conformity Rating derived from the *P-F* test. The long intervals between test and retest in all these estimates probably create an unfair impression of the *P-F* test's reliability. But this can be decided only in relation to the interpretation that is made of the factors the *P-F* test purports to measure.

EXTRA-TEST FACTORS AFFECTING RELIABILITY

A number of factors that are not strictly a part of the particular test or its scoring system affect the empirical estimates of the test's reliability. The number and magnitude of such factors is greater in the case of projective techniques than in the case of objective tests.

The most important of such factors for which there is substantial evidence are (a) examiner differences and differences in conditions under which the projective protocols are obtained, (b) differences in "ability" of scorers or interpreters of the protocols, (c) the range of the personality factors measured by the test in a particular sample, i.e., what is usually referred to as "range of talent," (d) the quantity of projective material obtained from the *S* (e.g. number of Rorschach responses, length of TAT stories, etc.).

The reliability reported for any technique must be viewed in relation to these factors. Often their combined effects are so intangible and difficult to assess that very little generality can be attached to the reliabilities obtained in any particular study. It is often an unwarranted assumption that a particular technique will have the same reliability from one study

to another when different examiners, different populations, etc. are involved. The evidence points to the conclusion that the reliability of a projective technique has little generality and is rather specific to the particular conditions under which the technique is used. Consequently estimates of reliability are to be desired in every study employing a projective technique.

Examiner Differences

A projective test is clearly not the same test in the hands of different examiners or under different conditions of administration. This fact has perhaps the most drastic implications as regards the reliability of projective techniques. Even if scoring, internal consistency, and retest reliabilities all were perfect for scores obtained by one examiner, the reliability of the test can still be low when it is used by a number of different examiners because of the fact that different examiners obtain different results from the same subject. This was demonstrated in a study by Baughman (5). The Rorschach psychographs obtained by 15 examiners from 633 randomly selected *S*s were analyzed in order ta determine whether different examiners secured different distributions of scores. It was found that the examiners differed significantly (below .05 level) in 16 of the 22 Rorschach scoring categories; 12 of these differences were significant beyond the .001 level. The results obtained by some examiners were comparable while others were markedly deviant, suggesting that some examiners may be considered interchangeable while others may not.

A study by Thornton and Guilford (48) on the reliability of Rorschach's so-called *Erlebnistypus* score (ratio of human movement responses (M) to color responses (sum C)) shows the effect on reliability of only slight differences in the method of administrering the test. Two groups, each of 95 college students, were given the Rorschach under somewhat differing conditions. The corrected splithalf reliabilities of M and C for the first group were .92 and .94 respectively; for the second group they were .77 and .65. The differences between the reliabilities in the two groups are highly significant and clearly show that reliabilities obtained under one set of conditions may not be generalized with confidence to somewhat differing conditions.

Differences in Interpretative Ability

The reliability of interpretations made from projective materials undoubtedly varies acording to the "ability" of the clinician using the technique. The term "ability" as used here is not intended to imply validity. It

refers rather to the fact that agreement is more easily achieved between certain pairs of clinicians than between others. This matter becomes especially important in using the more subtle aspects of projective protocols.

Interpretations of projective material are based not only on the formal scores given in the manuals and textbooks but also on the expressive elements of language, content, and style. A study by Hunt *et al.* (28) on these aspects of Wechsler-Bellevue responses thus has considerable bearing on projective testing. Sixty clinical psychologists ranked ten schizophrenic response items to the Wechsler-Bellevue for degree of pathology. The average rank order of each item was found as well as each clinician's correlation with this average rank order. These correlations, which are actually a measure of reliability, ranged from .02 to .93 for the 60 judges. When the 60 judges were randomly divided into three groups of 29 each, the average correlations for the groups were .19, .51, and .26. This finding strongly suggests there is likely to be similar variability in the reliability of interpretation of projective material from one group of clinicians to another. One wonders how much importance can be attached to any report of the degree of agreement between only two judges in scoring or interpreting a set of protocols, as is the usual practice when any attempt is made to estimate reliability. It appears quite probable that knowing the degree of interpretative agreement between one pair of judges does not permit us to infer with confidence the degree of agreement between another pair of judges.

"Range of Talent"

The reliability of a test may differ significantly from one population to another. The chief cause is differences in the so-called "range of talent" or range of scores represented in the population. A test shown to have high reliability in a very heterogeneous population, such as a random sample of the general population, may have very low reliability in a population that is more homogeneous with respect to scores on the test. In evaluating the reliability of a test, one must therefore take note of the test characteristics of the sample used in estimating the reliability, particularly the range and standard deviation of the test scores. If the test is to be used in a population in which the range of scores is more restricted than in the sample on which the reliability was determined, then one can expect a lowering of the reliability, and hence also of the discriminating power, of the test.

An example of differences in reliability from one sample to another is the study of Powers and Hamlin (42) in which the reliabilities of a Rorschach Index of Patholigical Thinking were .04, .47, and .91 for normal, neurotic,

and psychotic groups respectively. These reliabilities are roughly proportional to the range of scores in each of the three groups. Other factors, such as intelligence, stability, self-consistency, etc. also affect reliability. Eysenck, for example, has found that both split-half and retest reliabilities of objective tests were consistently lower for neurotics than for normals (17, p. 121). Reliabilities would probably be still lower for psychotics on these tests.

Length of Protocol

The length of a projective protocol is a function of the *S* and the examiner, and is not built into the test itself, as in the case of objective tests. This has important implications for reliability. Vernon (25) found that the reliability of Rorschach scores is directly related to R, the number of responses in the protocol. Reliability was much higher in cases where R>30 than when R<30. R is known to vary significantly not only from one *S* to another, but also from one examiner to another (12, p. 410). Cronbach concludes that it is improper to determine reliability on a group of protocols differing greatly in length. The protocols should be grouped on the basis of R (e.g., R of 10—15; 16—25; 26—30; etc.) and the reliability determined separately for each group. If the reliability is based on protocols of all lengths, the standard errors of measurement for any particular protocol would be relatively meaningless. It may be, for example, that a particular Rorschach score when based on protocols of R>30 has a sufficiently small standard error to be discriminative value, while the same score for protocols of R<30 may have so large a standard error as to be worthless. The Spearman-Brown formula comes to mind as a possible means of estimating the reliability of protocols of length *n*R from the known reliability of protocols of length R. This would be at best a rough estimate, since the not very tenable assumption would have to be made that if more responses had been elicited from the *S*, they would be "parallel" to the ones already obtained. One is not even on safe ground when the S-B formula is used for estimating reliability from longer to shorter protocols, in other words, estimating, in effect, the reliability of protocols that are shorter than the ones in which the original reliability coefficient is based. The Spearman-Brown formula is an unsatisfactory substitute for the grouping method suggested by Cronbach. Another solution recommended by Cronbach (12) is to obtain the same number of Rorschach responses from every *S*, so as to eliminate the differential effects of R on reliability. How much violence this procedure would do to the Rorschach would have to be determined empirically.

The problem of protocol length becomes even more ambiguous and difficult in the case of the TAT. There are great individual differences in story length, and the reliabilities of long and short protocols may very well differ significantly. However, the writer knows of no study of the TAT that throws any light on this matter.

The Standard Error of the Reliability Coefficient

The foregoing discussion raises the puzzling question: What is the standard error of the reliability coefficient of a projective test? Like any statistic, the reliability coefficient is subject to errors of sampling. In the case of objective tests, there is little problem in determining the standard error of the reliability coefficient, the general formula for which is

$$SE_{r_{tt}} = \frac{1 - r_{tt}^2}{\sqrt{N - 1}}$$

The odds are about 2 to 1 that the true population value of r_{tt} does not fall outside the range of $\pm$ 1 SE. If the reliability of an objective test (e.g., the MMPI) is based on the scores of a group of *S*s who can be assumed to represent a random sample of some population, then the standard error of the reliability coefficient has a definite meaning. It tells us the probable limits within which the reliabilities obtained from other random samples from the same population will fall.

But this meaning of the standard error of the reliability coefficient does not hold true in the case of projective techniques as they are commonly used. The reason is that there is not just a sampling error for subjects, but also for examiners, scorers, and conditions of administration. Therefore the actual standard error of the reliability of a projective test is bound to be greater than that of an objective test, but just how much greater it is not possibly to say. It is known that a wide range of reliabilities can be obtained from the same *S*s when they are tested by different examiners and from the same set of protocols when scored by different raters. It must be concluded that the standard error, and hence the significance and generality of the reliability coefficient of projective tests is indeterminate in all cases where ordinary correlational methods are used for estimating reliability. (The writer has not encountered in the projective literature a single reliability coefficient the standard error of which was not indeterminate). In another paper (29) a method is presented for obtaining confidence limits for the reliability coefficients of projective techniques.

SUMMARY

The theory and the main facts concerning the reliability of projective techniques have been discussed from the viewpoint of psychometric theory, with emphasis on

the problems more or less peculiar to projective techniques. While there has been unfortunately little concern with reliability in this rather special field of psychological testing, projective techniques, whether used in the clinic or in psychological research, cannot be exempted from demonstrating the fundamental requirement of any measuring instrument, viz., reliability. Projective techniques have remained outside the pale of psychometric theory partly as the result of misconceptions concerning their nature. They have been incorrectly regarded as not actually being tests yielding measurements but rather as means of obtaining a picture, a sort of X-ray, of the personality, or as a means of generating hypotheses about underlying personality dynamics. It has been argued in the present article that these functions attributed to projective techniques are fundamentally no different than the functions of any other kind of psychological test and hence are subject to the same statistical requirements. The nature of projective techniques, however, poses special methodological problems.

The research evidence on the reliability of the major projective techniques (Rorschach, TAT, Sentence-Completion, Bender-Gestalt, Rosenzweig P-F test) was reviewed with reference to five types of reliability: scoring, internal consistency, test-retest, parallel forms, and reliability of interpretation. No general conclusion concerning reliability is possible even with respect to any particular technique. The reported reliabilities are usually lower than is considered acceptable in the case of objective tests, although when projective devices are modified in the direction of greater objectivity of administration, scoring, and interpretation, their reliability compares favorably with that of the best objective measures of personality. Few of the reliability studies reported in the projective literature are methodologically sound or adequate for their stated purpose. The very wide range of reliability coefficients reported for a single technique is so great as to suggest that the reliability of a projective test has little generality.

A number of factors affect reliability, all of which are regarded as sources of error variance which must be taken into account in estimating the reliability of the technique. The main sources of error variance are due to examiner differences, differences among scorers and interpreters, the "range of talent" in the particular sample with respect to the variables being measured, and the differences in quantity of projective material elicited from different subjects. Because of these factors the meaning of the standard error of the reliability coefficient of a projective test is highly ambiguous and usually indeterminate, with the result that the reliability coefficient itself is statistically meaningless. The most satisfactory solution to this problem is to eliminate or reduce as many of the sources of error variance as possible by making the administration, scoring, and interpretation of projective tests more standardized and objective.

References

1. Altus, W. D., & Thompson, Grace M., The Rorschach as a measure of intelligence. *J. consult. Psychol.*, 1949, *13*, 341–347.
2. Ainsworth, Mary, "Problems of validation." In: Klopfer, B., Ainsworth, M. D., Klopfer, W. G., & Holt, R. R., *Developments in the Rorschach technique.* Vol. I. *Technique and theory.* Yonkers, N.Y.: World Book, 1954.

3. Auld, F., Eron, L. D., & Laffal, J., Application of Guttman's scaling method to the TAT. *J. educ. psychol. Measmt.*, 1955, *15*, 422–435.
4. Baker, L. M., & Creager, J. A., Rating scale technique applied to Rorschach responses. *J. clin. Psychol.*, 1954, *10*, 373–375.
5. Baughman, E. E., Rorschach scores as a function of examiner differences. *J. proj. Tech.*, 1951, *15*, 243–249.
6. Bernard, J., The Rosenzweig Picture-Frustration study. I. Norms, reliability and statistical evaluation. II. Interpretation. *J. Psychol.*, 1949, *28*, 325–343.
7. Blanton, R., & Landsman, T., The retest reliability of the Group Rorschach and some relationships to the MMPI. *J. consult. Psychol.*, 1952, *16*, 265–267.
8. Brogden, H. E., The effect of bias due to difficulty factors in product-moment item intercorrelations on the accuracy of estimation of reliability. *Educ. psychol. Measmt.*, 1946, *6*, 517–520.
9. Brower, D., & Abt, L. (Eds.), *Progress in clinical psychology*. Vol. I. New York: Grune & Stratton, 1952.
10. Child, J. L., Frank, Kitty F., & Storm, T., Self-ratings and TAT: Their relations to each other and to childhood background. *J. Personality*, 1956, *25*, 96–114.
11. Combs, A. W., The validity and reliability of interpretation from autobiography and the Thematic Apperception Test. *J. clin. Psychol.*, 1946, *3*, 240–247.
12. Cronbach, L. J., Statistical methods applied to Rorschach scores: a review. *Psychol. Bull.*, 1949, *46*, 393–429.
13. Dana, R. H., Clinical diagnosis and objective TAT scoring. *J. abnorm. soc. Psychol.*, 1955, *50*, 19–25.
14. Datel, W. E., & Gengerelli, J. A., Reliability of Rorschach interpretations. *J. proj. Tech.*, 1955, *19*, 372–281.
15. Eichler, R. M., A comparison of the Rorschach and Behn–Rorschach inkblot tests. *J. consult. Psychol.*, 1951, *15*, 185–189.
16. Eysenck, H. J., A comparative study of four screening tests for neurotics. *Psychol. Bull.* 1945, *42*, 659–662.
17. ————, *Dimensions of personality*. London: Routledge & Kegan Paul, 1947.
18. ————, *The scientific study of personality*. London: Routledge & Kegan Paul, 1952.
19. Fine, R. A scoring scheme for the TAT and other verbal projective techniques. *J. proj. Tech.*, 1955, *19*, 306–309.
20. Fosberg, I. A., Rorschach reactions under varied instructions. *Rorschach Res. Exch.*, 1938, *3*, 12–31.
21. ————, An experimental study of the reliability of the Rorschach technique. *Rorschach Res. Exch.*, 1941, *5*, 72–84.
22. Frank, L. K., *Projective methods*. Springfield, Ill.: Charles C. Thomas, 1948.
23. Grant, Q., Ives, V., & Ranzoni, J. H., Reliability and validity of judges' ratings of adjustment on the Rorschach. *Psychol. Monogr.*, 1952, *66*, No. 2. (Whole No. 334.)
24. Guilford, J. P., *Psychometric methods*. Second edition. New York: McGraw Hill, 1954.

25. Gulliksen, H., *Theory of mental tests.* New York: Wiley, 1950.
26. Hertz, Marguerite R., The reliability of the Rorschach inkblot test. *J.appl. Psychol.*, 1934, *18*, 461–477.
27. ————, "The Rorschach thirty years after." In Brower, D., & Abt, L. (Eds.), *Progress in clinical psychology*, Vol. I. New York: Grune & Stratton, 1952.
28. Hunt, W. A., Arnhoff, F. N., & Cotton, J. W., Reliability, chance, and fantasy in inter-judge agreement among clinicians. *J. clin. Psychol.*, 1954, *10*, 294–296.
29. Jensen, A. R., The reliability of projective techniques: Methodology. (To appear.)
30. Kelley, D. M., Margulies, H., & Barrera, S. E., The stability of the Rorschach method as demonstrated in electric convulsive therapy cases. *Rorschach Res. Exch.*, 1941, *5*, 35–43.
31. Kelly, E. L., & Fiske, D. W., *The prediction of performance in clinical psychology.* Ann Arbor, Michigan: University Michigan Press, 1951.
32. Klopfer, B., Ainsworth, M. D., Klopfer, W. G., & Holt, R. R., *Developments in the technique.* Vol. I. *Technique and theory.* Yonkers, New York: World Book, 1954.
33. Korner, I. N., & Westwood, D., Inter-rater agreement in judging student adjustment from projective tests. *J. clin. Psychol.*, 1955, *11*, 167-170.
34. Lindzey, G., & Herman, P. S., Thematic Apperception Test: a note on reliability and situational validity. *J. proj. Tech.*, 1955, *19*, 36–42.
35. Lisansky, E. S., The inter-examiner reliability of the Rorschach test. *J. proj. Tech.*, 1956, *20*, 310–317.
36. McClelland, D. C., *The achievement motive.* New York: Appleton-Century-Crofts, 1954.
37. ————, (Ed.), *Studies in motivation.* New York: Appleton-Century-Crofts, 1955.
38. Mayman, M., & Kutner, B., Reliability in analyzing Thematic Apperception Test stories. *J. abnorm. soc. Psychol.*, 1947, *42*, 365–368.
39. Pascal, G. R., & Suttell, B. J., *The Bender–Gestalt test: quantification and validity for adults.* New York: Grune & Stratton, 1954.
40. Piotrowski, Z., The reliability of Rorschach's Erlebnistypus. *J. abnorm. soc. Psychol.*, 1937, *32*, 439–445.
41. Pope, B., & Jensen, A. R., The Rorschach as an index of pathological thinking. *J. proj. Tech.*, 1957, *21*, 59–62.
42. Powers, W. T., & Hamlin, R. M., Relationship between diagnostic category and deviant verbalizations on the Rorschach. *J. consult. Psychol.*, 1955, *14*, 120–124.
43. Rhode, Amanda R., Explorations in personality by the sentence completion method. *J. appl. Psychol.*, 1946, *30*, 169–181.
44. Sanford, R. N., Physique, personality, and scholarship. *Mongr. of the Society for Res. in Child Developm.*, Vol. VIII, No. 1, Serial No. 34. Washington, D. C.: Society for Research in Child Development, National Research Council, 1943.
45. Suttell, B. J., The development of visual-motor performance in children as estimated by the Bender-Gestalt test. Unpublished M.A. thesis,

Univ. of Pittsburgh, 1951. Reported in Brower, D., & Abt., L. E., *Progress in clinical psychology*. Vol. I, Sec.1, pp. 185–189. New York: Grune & Stratton, 1952.

46. Swift, Joan W., Reliability of Rorschach scoring categories with pre-school children. *Child Developm.*, 1944, *15*, 207–216.
47. Symonds, P. M., *Adolescent fantasy. An investigation of the picture-story method of personality study*. New York: Columbia Univ. Press, 1949.
48. Thornton, G. R., & Guilford, J. P., The reliability and meaning of Erlebnistypus scores in the Rorschach test. *J. abnorm. soc. Psychol.*, 1936, *31*, 324–330.
49. Tomkins, S. S., *The Thematic Apperception Test: The theory and technique of interpretation*. New York: Grune & Stratton, 1947.
50. Vernon, P. E., The Rorschach inkblot test. *Brit. J. med. Psychol.*, 1933, *13*, 89–118.
51. ————, The Rorschach inkblot test. *Brit. J. med. Psychol.*, 1933, *13*, 179-200.
52. Watkins, J., & Stauffacher, J., An index of pathological thinking. *J. proj. Tech.*, 1952, *16*, 276–286.