VII

Objective Scoring of Projective Tests

WAYNE H. HOLTZMAN
The University of Texas

EVER SINCE L. K. Frank's first use of the term "projective method" in 1939 (15), there has been a rapid mushrooming of techniques for encouraging an individual to reveal aspects of his personality by the way in which he perceives, organizes, or relates to potentially affect-laden, ambiguous stimuli. Stemming largely from psychoanalytic theory, such projective techniques range all the way from free association in relatively unstructured situations to rather highly structured, formalized devices such as the Thematic Apperception Test. Before considering the problems of quantification and objective scoring, it might be instructive to examine closely the assumptions implicit in the projective method as contrasted to those underlying psychometric tests and measurement theory.

PROJECTIVE COMPARED WITH PSYCHOMETRIC METHODS

Unlike the standardized aptitude test, the projective approach deals with the idiomatic expression of the individual as revealed in the context of his needs, fears, strivings, and ego-defensive behavior. As Frank has so aptly stated, "The essential feature of a projective technique is that it evokes from the subject what is, in various ways, expressive of his private world and personality process." (16, p. 47).

Given any projective technique where the subject is offered a wide latitude in which to reveal himself, the particular sample of responses obtained is assumed to reflect significant aspects of the subject's personality organization, if only the examiner can find the key to its interpretation.

Macfarlane and Tuddenham have pointed out that such an isomorphic assumption concerning the subject's test protocol and his personality leads to three corollaries that are rarely explicit: (a) belief that a protocol is a sufficiently extensive sampling of the subject's personality to warrant formulating judgments about it; (b) belief that the psychological determinants of each and every response are basic and general; and (c) belief that projective tests tap the durable essence of personality equally in different individuals (27, p. 34). Many of the more wary, sophisticated projectivists would argue that none of these three assumptions *necessarily* follows from the basic assumption underlying the projective method—that even the best of projective test protocols is but a tiny fragment of the total personality, fraught with innumerable possibilities for misinterpretation. Nevertheless, in actual practice it is difficult to avoid falling into the dogmatic position of over-interpretation in an attempt to weave together a consistent picture of the personality dynamics presumably reflected by the clinical techniques employed. It can be argued that elaborate, clinical interpretations of personality from projective protocols often reveal more about the personality of the clinician than that of the subject.

In contrast to a projective technique, a psychometric test is based upon the fundamental assumption that an obtained score on the test reflects a hypothetical "true" score which is characteristic of the attribute in question for a given individual under specified testing conditions and at a given moment in time. Any deviation of the obtained score from the true score represents error of measurement which can be assessed provided one is willing to make certain assumptions about the nature of such errors. By defining the true score so that it includes all constant errors of measurement, the discrepancy between obtained and true score becomes a random error component. Since a random event by definition is uncorrelated with any other event, a general theory of measurement can be developed out of which components of error variance can be estimated, both with regard to the concept of reliability and the concept of validity (18).

Contrary to the opinion of some writers (37), such psychometric theory is not necessarily limited to a nomothetic universe where one

is interested in group or inter-individual differences. As Cattell (6) has been quick to point out, one can legitimately utilize psychometric theory for idiographic purposes by considering k different measures on m different occasions for a single person. Nor need psychometric theory be restricted to consideration of one response variable at a time—the oft heard criticism that a psychometric, statistical, or quantitative approach is too atomistic to provide more than a ridiculous caricature of the individual personality. While it is true that most contemporary uses of test scores deal with isolated traits, or at best with linear combinations of several traits, the advent of configural scoring methods (30), the possibilities of profile analysis (19), and other complex, multivariate procedures open new vistas for effective utilization of psychometric theory in the study of the individual personality.

Use of psychometric theory as a basis for assessment of personality commits one to a trait theory of personality. Postulating some sort of "true" score as a hypothetical construct to be inferred from observed scores is tantamount to saying that John Doe has X amount of the trait in question. It is not necessary, however, to think of John's possession of the trait as a "fixed" quantity. An individual's true score remains invariant only so long as the specific testing conditions remain constant and there is no real change in the individual with respect to the trait in question. A primary purpose of test standardization is to minimize constant sources of error that are ordinarily confounded with the inferred true score. Only errors of measurement that are random in nature can be adequately assessed and taken into account by the usual concepts of reliability and validity within contemporary psychometric theory.

Rosenzweig (37) has observed that assessment procedures can be ordered on a continuum depending upon the degree of structuring and control introduced by the assessor. At one extreme are the completely qualitative, unstructured methods of psychoanalysis, free association by a patient in the presence of an analyst. At the other extreme are highly structured paper-and-pencil tests which meet all the standards of psychometric theory. Projective techniques are seen as falling somewhere in between the particular position on the continuum depending upon the degree of standardization and control. In most instances, the projectivist has tried to preserve the qualitative, idiographic essence of the projective method while also searching for ways in which to categorize, quantify, and standardize the response variables underlying test behavior. He would like to have a technique for assessing personality

which covers a wide band of the above continuum with a high degree of power throughout the range. Very few psychologists indeed have completely and consistently refrained from some form of abstraction later leading to quantification.

As soon as an individual decides to classify and enumerate any characteristics of a subject's responses to a projective technique, however crude and elementary the system, he has shifted from a purely projective point of view to a psychometric frame of reference. Such measurement may be quite nominal and only faintly resemble full-blown quantification. Nevertheless he has made the first and most significant step by classification of responses. For example, to classify a given response to an inkblot as a *W* assigns meaning to the response that transcends the idiosyncratic, private world of the subject. Unless one considers such symbols as *W*, *D*, and *d*, mere short-hand devices that have no real meaning beyond calling one's attention to certain aspects of the protocol, the symbols take on nominal characteristics of measurement. Those subjects who use the whole inkblot are seen as one class of individuals (*W*-tendency type), while those who use only a small part of the inkblot for their response are seen as another class (*d*-tendency type).

Such symbols of classification can be considered "signs" depicting specified characteristics abstracted from the raw protocol. More or less elaborate patterns of signs can be derived, either rationally or empirically, which point toward a syndrome or personality attribute to be inferred from the protocol. The pattern of signs may be complex and highly conditional so that predictive statements of the "if A and B but not C, then X" type can be formulated. Or the set of admissible signs may all contribute to some sort of "global" measure like the adjustment score derived from the Rorschach by Munroe's Inspection Technique (32). Such clusters of signs may have some pragmatic value in predicting a criterion, but they have a disjunctive quality or arbitrariness which makes any theoretical interpretation exceedingly difficult. In most instances when a series of responses is classified, some types of response will appear more than once. Counting of such response frequencies is the first step in the construction of a quantitative scoring system. A Rorschach protocol with 10 movement responses would be thought of as indicating a greater tendency to see movement than a record with only two movement responses. Such a statement implies a crude kind of ordinal scale by which people can be ordered according to their degrees of *M*-tendency, provided the total number of responses is controlled.

As one becomes engrossed with the counting of symbols it is very easy to forget the nature of the projective material being classified. In his eagerness to make a given technique meet the demands of both psychometric and projective theory, the psychologist often compromises the two sets of conflicting standards to the point where the technique fails to accomplish either aim. There are some projective devices that should always be treated by qualitative methods of analysis since almost any attempt to abstract quantitative scores will fail to have any meaning. Other projective techniques may be altered sufficiently to yield scores meeting acceptable psychometric standards while at the same time preserving the projective nature of the task. It is too much to expect a technique designed originally as a purely projective method to lend itself to a meaningful kind of quantification without some revision, and in many projective techniques no amount of revision will produce adequate scores in the true psychometric sense.

Frank (16) has divided the projective techniques into five general kinds: constructive, interpretive, constitutive, cathartic, and refractive. The constructive methods consist of those techniques which require the subject to arrange materials into larger configurations or to produce drawings as in the Draw-A-Person Test. The interpretive methods are primarily verbal-associational techniques such as the Thematic Apperception Test. The best known example of a constitutive method is the Rorschach in which the subject must organize relatively amorphous, unstructured inkblots into meaningful concepts. While most projective techniques may stimulate cathartic reactions, some, such as play therapy with dolls, are designed specifically for this purpose. The last of Frank's classes, the refractive method, is based upon the fact that any conventionalized mode of communication—handwriting, gestures, and other forms of expressive movement—may be used as an approach to the individuality of a person.

The above classification serves as a convenient basis for a more detailed discussion of scoring problems and quantifications in the analysis of projective techniques. Since cathartic methods cut across the other procedures, and since the analysis of expressive movement and individual style of communication can be considered as a special topic apart from more conventional projective methods, only the first three of Frank's classes will be discussed. Considerably more attention will be given to the Rorschach and related techniques than to the constructive or interpretive methods, partly because the

Rorschach has been studied longer and more exhaustively than any other projective test and partly because it provides an unusually good illustration of various problems of quantification encountered throughout the projective-psychometric continuum.

Constructive Methods

The way in which a child or adult arranges miniature life toys, draws a figure of a man or woman, or builds mosaics from colored pieces can reveal a great deal about his personality. Generally speaking, however, such creative productions are very difficult to analyze in any objective, quantitative fashion. Most clinicians only use qualitative procedures when dealing with constructive methods. Occasionally the characteristics of a construction may be classified to formalize its description, but inferences regarding personality, whether based upon symbolic interpretations or more direct expressions by the subject, remain at the clinical intuitive level. Of course, rating scales for recording clinical judgment can be employed with such materials, as with any other individual response or style of expression. But it is not difficult to see why quantification in the psychometric sense has failed to prove useful in the analyses of drawings or other creative products, even though the situation may be rather highly structured as in the Bender-Gestalt Test. Usually the construction has to be viewed as a whole or as only a very small number of separate units analogous to test items. The configuration, color, shading, and other characteristics of a drawing are complex, defying quantification in the usual sense. Nevertheless, in some special cases, fairly successful attempts have been made to score objectively certain limited aspects of such productions. Several of these will be briefly discussed.

Drawing a human figure has been employed rather extensively as a projective technique in recent years, largely due to the persistent studies of Karen Machover (28). Working primarily from a psychoanalytic point of view in which the drawing is assumed to reflect the body image of self, Machover and others have developed systems of graphic analysis utilizing a sign approach to the scoring of drawings. For full use of the system, the subject must draw both a man and a woman so that comparisons of self-sex and opposite-sex figures can be made. A good example of this graphic sign method is the scale of figure drawing items which is presumed to measure field-dependency (50). Sets of 40 items for men and 45 for women were constructed by Machover on the basis of a preliminary

analysis. Criterion groups for the initial selection of items consisted of college students with high and low field dependence as measured by a battery of perceptual tests. A total score is obtained by summing the number of signs checked during the detailed analysis of the two figure drawings. Some of the signs are completely objective such as transparency, lack of ears, or hair shaded. Others, like consistency rating and rigidity rating, are subjective and require a clinical judge. For the most part, however, the list of signs is sufficiently objective to merit further study.

Graphic signs have been used with similar success by Pascal and Suttell in the objective scoring of drawings in the Bender-Gestalt Test (34). The test consists of nine geometric forms that are copied by the subject. The number of scorable signs on each design varies from 10 to 13, with seven additional signs dealing with the total configuration of all nine drawings. Each sign is given a numerical weight varying from one to eight. The size of the weight was empirically determined in earlier studies differentiating normals from such groups as psychotics and organics.

A single score is obtained by summing the weights of positive signs, the higher the score the more pathological the record. Although much valuable information may have been sacrificed at the expense of obtaining a single quantitative index, the resulting score has sufficiently high reliability and validity in a variety of situations to prove highly useful as a screening procedure.

A third variation of semi-structured drawing which represents an attempt at objective quantification is the Drawing-Completion Test described by Kinget (23). Eight squares are presented to the subject, each containing small, but suggestive, stimuli such as a dot, a wavy line, or a black square, around which the subject draws whatever he wishes. Kinget has attempted to develop a graphic system with a series of crudely quantitative variables, some based on content analysis and others dealing with style and expressive features of the drawings. A personality profile is constructed by recording signs and then adding them together in more abstract categories, somewhat like the first attempts to quantify the Rorschach. While the rationale behind the scoring system is highly speculative and smacks of arm-chair analysis without adequate empirical support, the method itself is interesting and sufficiently novel to deserve careful study.

Working with spontaneous finger paintings, a construction which has proved very difficult to quantify, Dorken (10) has developed a series of objectively defined rating scales for energy output, affective

range, contact with reality, and clarity. Pictorial norms were used as points of reference to anchor the scales. The variable, Affective Range, illustrates the technique. "Spontaneous" colors, red and yellow, were each assigned scale values of three, blue and green were given values of two each, and the "somber" colors, black and brown, were each scored one. Combination colors were scored in relation to this primary scale. Test-retest reliability ranged from .13 to .84, depending upon the sample and time interval between administrations. By using a series of finger paintings, reasonably adequate summary scores on the four variables defined by Dorken should be possible.

It is significant to note that in each of the above examples of attempts to achieve objective scoring of projective techniques, the degree of quantification is pretty much limited to the complex sign approach in which numerous signs are scored, weighted, and summed to yield some sort of "global" but quantitative, measure which is purported to reflect important dimensions of personality. Ideally, the sign approach should begin with sufficient theoretical rationale to construct a coherent system. After careful operational definition of each sign, the objectivity of scoring should be determined by having at least two trained individuals independently score a large number of protocols. In some instances where several signs have similar rationales in their definition, their consistency should be examined empirically in a study to validate the construct which they theoretically represent (7). In most cases, however, a straight empirical analysis without regard for the construct in question will be undertaken with the practical view in mind of establishing a weighting system that has maximum efficiency for predicting some criterion. In any case, the burden of proof concerning the reliability and objectivity of any proposed scoring system rests with the individual who proposes it.

Interpretive Methods

Assessing personality from the way in which an individual reveals his fantasy life in telling a story or interpreting a scene goes back through centuries of mankind. However, the first notable attempt to develop a projective test for uncovering a person's needs, wishes, and related fantasies by having him tell stories was made by Morgan and Murray in 1935 (31). In the past 20 years, Murray's Thematic Apperception Test (TAT) has become a standard projective technique, second only to the Rorschach in its widespread use both

in the clinic and laboratory. Numerous other interpretive methods —Rosenzweig's Picture-Frustration Study (36), Bellak's Children's Apperception Test (22), and Shneidman's Made-A-Picture-Story Test (43), to mention but a few—stem more or less directly from Murray's pioneering work and attest to the fruitfulness of the basic method.

Interpretive methods range all the way from one end of the projective-psychometric continuum to the other. Representative of the purely projective approach is the standard TAT analyzed entirely in a qualitative manner, focusing upon the content of stories and stylistic aspects of the story telling as illustrated by Stein (44), such analysis draws heavily upon careful deduction and clinical intuition. Only one step removed from this intuitive approach is the more formal kind of qualitative analysis in which various characteristics of each story are classified according to theme expressed, kinds of affect, need categories, and the like. Such qualitative systems tend to vary considerably according to the predilection of the analyst. Representative of the diverse approaches to analysis of TAT protocols is Shneidman's (43) compilation of systems used by 15 different authorities working with the same TAT record.

Several investigators have developed sets of rating scales to be used with the TAT. One of the most extensive systems is Hartman's (21) consisting of five-point scales for 65 categories covering thematic elements, feeling qualities, topics of reference, and more formal characteristics, each of which can be scored for a given story. Total scores are obtained by summing ratings across stories. While such scales utilize the clinical skill of the interpreter, serious difficulties often arise when one is concerned with the objectivity of the scoring. When categories deal with the manifest aspects of a story, independent raters can generally agree at a satisfactory level to insure fair objectivity. But as soon as attention is focused upon covert aspects of the response or upon the personality of the story-teller rather than his production, agreement falls off sharply (46).

The reason for this greater subjectivity when dealing with the personality of the subject is apparent when one examines closely the nature of the factors influencing response to a TAT picture. Holt (22) discusses nine different determinants of the manifest response, ranging from situational context to personal style of the story-teller. The interpreter is faced with the very complex task of weighing the probable influence of each factor before he can arrive at an interpretation of the subject's personality. It is somewhat like having an equation with nine variables, several of which can be

partially discounted while most remain unknown quantities. Several judges will weigh the unknowns quite differently, resulting in widely varying ratings.

This difference between *test-oriented systems* dealing with formal characteristics of the response and *personality-oriented systems* in which the interpreter makes direct inferences concerning the personality of the story-teller is fundamental. The more superficial or concrete the system, the more objective the scoring and the less relevant the derived variables to the personality of the subject. Young (51) developed a set of 23 well-defined traits, such as Anxiety, Dominance, and Need to be Loved, which could be used in rating the personality of the interpreter as well as the subject. Fifteen trained interpreters independently rated 12 TAT stories from seven different individuals, a total of 84 responses, on each of the 23 traits. Ratings on the same 23 traits were obtained for each of the 15 interpreters by a sociometric method. Even though the average agreement among interpreters was fairly high for such personality-oriented variables, differences in the interpreters' ratings proved significantly related to their own personalities, demonstrating the intrinsic subjectivity of such methods of analysis.

Several fairly objective variables dealing with story content seem sufficiently relevant to important aspects of the story-teller's personality to merit special attention. McClelland and his colleagues (26) have carefully developed the personality construct, Achievement Motive, and have demonstrated how it can be reliably scored in TAT stories. The scoring involves simple classifications of response elements by objective criteria that are then summed to yield an overall index of the individual's Need-Achievement score. A number of experimental studies are also cited indicating the validity of the personality construct.

A similar careful derivation of two test-oriented variables of relevance to the story-teller's personality was undertaken by Eron (12). Using well-anchored rating scales, Eron and co-workers developed fairly objective measures of emotional tone and outcome that could be applied to single responses and summed to get an overall score. Both variables have satisfactory inter-scorer reliabilities, .86 for emotional tone and .75 for outcome. Eron is chiefly concerned with the development of norms for TAT themes that can be used to define the general characteristics of each card in terms of the ease with which certain themes are evoked. Such data for the TAT can be roughly thought of as analogous to difficulty level or other item-parameters in aptitude tests. A recent application of Eron's ap-

proach demonstrates how Guttman's scaling method can be employed using normative TAT data to construct a uni-dimensional scale for need-Sex (1).

A final example of an objective approach to the scoring of the TAT is one devised recently by Dana (9). Three fundamental aspects of test behavior—approach to the situation, normality of response, and rarity of response—were used by Dana to define three variables amenable to objective scoring, Perceptual Organization, Perceptual Range, and Perceptual Personalization. Inter-scorer reliability in terms of percentage agreement between independent judges ranged from 76 to 94 for the three scoring categories in a study of 150 TAT stories. The unique aspect of Dana's approach is the fact that these three variables are sufficiently pertinent to a large variety of projective techniques to permit inter-test comparisons for sharpening the validity of the personality constructs involved.

Variations of the sentence completion method provide much more suitable data for psychometric development than the TAT. The technique consists of providing the subject with a list of incomplete sentences to which he responds with whatever completions come to mind. By wise selection of sentence stems, content fairly similar to the thematic apperception methods can be obtained. Of course the response is much more highly structured and discrete from one item to the next than is the case with the TAT. Herein lies the chief virtue of the method with respect to quantification.

Rotter and Willerman (38) developed one of the first sentence completion tests with high objectivity. Designed for large-scale screening purposes in the Army Air Force, their 40-item version yielded a single adjustment score having inter-scorer reliability of .89 and split-half reliability of .85. A refined version of this test designed for college students, the Rotter Incomplete Sentences Blank (39) has an objective scoring manual with reported interscorer reliability of .96 and split-half reliability of .84, unusually high for a projective technique.

Trites and his colleagues (47) developed a military version of the sentence completion method to a high degree of objectivity while at the same time dealing with a number of response-categories rather than just one. A scoring manual was written on the basis of 1038 test protocols which yielded interscorer agreement ranging from .80 to .96 for eight major variables, Conformity, Ego Esteem, Gregariousness, Sexuality Attitudes, Air Force-oriented Motivation, Hostility, Insecurity, and Unscorable Response. Although there

is little direct evidence to support the validity of these variables with respect to the personality constructs implied, in a later factor analysis of inter-item correlations where the items had been scored dichotomously as indicating either a positive or negative attitude with reference to adjustment to flying, Trites (48) obtained four factors which were meaningfully linked to several of the original major variables.

It is instructive to note the characteristics of the sentence completion method which are responsible for achievement of satisfactory psychometric standards. Unlike the TAT, the number of discrete items can be very large, making possible an atomistic treatment of test elements without undue distortion of the technique. Where the TAT has at most 20 pictures, each with an infinite variety of complex responses possible, the sentence completion method has highly structured items for which the variety and extent of responses are relatively limited. The more circumscribed nature of the technique makes possible the development of an objective scoring manual for any variables that may be present in the response. That such psychometric treatment does not necessarily reduce the usefulness of a projective method is demonstrated by the repeatedly high validity obtained for the Rotter Incomplete Sentences Blank in assessing level of personal adjustment (39).

Constitutive Methods

The Rorschach test stands alone among projective techniques in the amount of attention, both clinical and experimental, which it has received during the past twenty years and illustrates problems encountered in scoring responses to constitutive methods. Quantitative analysis of responses to inkblots has ranged all the way from one extreme of the projective-psychometric continuum to the other. Some writers (25, 41) have pointed out how the Rorschach can be dealt with in a purely qualitative manner, emphasizing the dynamic and symbolic nature of the content and leaning heavily upon psychoanalytic theory and the intuitive skills of a clinician. Associations to inkblots are seen as only one step removed from completely free association in the psychoanalytic session. Others (20, 33) have shown how highly structured and completely objective multiple-choice methods can be applied to the study of individual differences in the perception of inkblots. And curiously enough, the same 10 inkblots are used throughout!

To what extent are these various degrees of structuring and quan-

tification based upon sound principles of measurement theory? Does the Rorschach really span the entire projective-psychometric continuum with the high degree of power claimed by some of its proponents?

The most rudimentary form of quantification in the Rorschach is the assigning of symbols to certain kinds of responses which are then looked upon as signs pointing to various personality attributes or nosological classes. An excellent example of such a classification of qualitative signs is the analysis of verbalization described by Rapaport (35), who presents a very careful rationale for the scoring of such pathognomic verbalizations as confabulations, contaminations, confusion, absurd responses, and ideas of reference. Such signs are not additive except in the very crude sense that a number of positive signs in a single record tend to pile up in confirming the diagnosis.

The widely used "formal" scoring methods for the Rorschach represent attempts to measure the perceptual variables implicit in the response. The complex nature of the stimulus permits a wide latitude of location, of determinants, and conceptual content. Once decisions have been made as to what constitutes a discrete response, the number of such responses to a given inkblot or to all 10 Rorschach plates can be determined. Although there are some minor problems encountered in deciding when a verbalization is truly a response for purposes of scoring, one can safely assume that interscorer agreement as to number of responses (R) is quite high regardless of the judge's theoretical position. Similarly, the scoring of location, at least in its gross elements of whole, usual large detail, or small and unusual detail, does not pose serious problems in the attaining of reasonable objectivity. Aside from specialized uses of content such as Elizur's anxiety score (11), the categorizing of concepts into human, animal, and other generic classes is quite straightforward also. The greatest difficulties in achieving scoring objectivity arise in the realm of response-determinants.

Trying to determine those stimulus attributes which are responsible for eliciting a given response amounts to a kind of global psychophysics for which the general laws have yet to be worked out. Although logical in their conception, most scoring systems for determinants involve a number of highly arbitrary decisions, the wisdom of which is highly debatable. The subjectivity of the method, the influence of factors extraneous to the blots such as the examiner-subject interaction (40) and variation in style of inquiry (17) raise troublesome questions concerning the meaning of scores once achieved.

Presumably the inquiry phase of the Rorschach is designed to discover the characteristics of the inkblot which prompted the subject to give a response. The subject is asked by rather vague and indirect questions to introspect, to analyze the perceptual process and report to the examiner what about the blot suggests, for example, "a bloody finger," or "a pretty flower." A helpful subject who senses what the examiner is after may reply by saying, "It's shaped like a man's thumb and is colored red, suggesting blood." More than likely, however, the subject will say, "It just looks like it to me," leaving the examiner about where he started. And even if the subject does mention the color as playing a part in the concept, do we have any way of knowing whether the subject would have reported blood in the absence of color? How do we know it wasn't the combination of form and shading that suggested a bloody thumb? The unfortunate fact is that we simply don't know, although recent studies by Baughman (2) provide a better basis for guessing.

Zubin (52) has recognized this problem and has tried to overcome it by introducing a much more exhaustive inquiry than the usual brief, indirect questioning. In addition to asking many more questions per response, he has experimented with inquiry immediately following the response rather than waiting until all 10 inkblots have been administered. Sixty scales were constructed that could be applied in scoring a single response, provided the inquiry was sufficiently exhaustive. Five scales deal with location, six with the objective attributes of the stimulus, six with determinants or the relative importance of stimulus attributes in the formation of the percept, 14 with interpretation categories such as surface texture or strength of movement, three with organization activity, 15 with content, and 11 with other aspects of the single response such as reaction time and popularity. In addition, there are six scales dealing with variables present in the protocol as a whole. When one stops to think that Rorschach records frequently contain upward of 50 responses, the amount of energy invested in scoring 60 scales on each response is tremendous.

If a sufficient amount of information were available about the objective stimulus attributes and the correlates between these attributes and characteristics of the response, the amount of work required to utilize Zubin's system might be justified. However, the very nature of the complex stimulus confronting the subject in the form of an inkblot defies all but the crudest, global type of description as far as the specific stimulus attributes are concerned. With

respect to the determinants or global psychophysics of the reported percept, even a highly trained introspectionist would be hard put to verbalize accurately the relative importance of various inkblots characteristics in forming the percept. Since the greatest value for the Rorschach is claimed to be the study of psychopathology where the subject's ability to introspect accurately may be seriously impaired, there appears to be little real hope of obtaining the kind of information necessary to use many of the scales Zubin has proposed. Although Zubin's system may not really increase the objectivity of scoring for the Rorschach, since it is comprised largely of five-point scales for recording clinical impression, his exhaustive approach immediately points out the fundamental weaknesses inherent in the standard methods of scoring.

In addition to the fact that objective scoring for most inkblot variables cannot be achieved without the use of arbitrary rules, the standard Rorschach is inherently poor as a psychometric device in some other important respects. Providing the subject with only ten inkblots and then permitting him to give as many or as few responses to each card as he wishes characteristically results in a set of unreliable scores with sharply skewed distributions, the majority of which fail to possess the properties of even rank-order measurements. One record with an R of 20 may be comprised of single responses to the first nine cards and 11 responses to Card X, while another may consist of two responses per card. Any of the usual scores with the possible exception of form level will have quite different meanings in the two contrasting protocols even though the total number of responses is constant. Add to this the difficulties arising when R varies from less than 10 to over 100, and it is easy to see why most quantitative studies involving the standard Rorschach yield confusing or negative results.

In a general review of statistical methods applied to Rorschach scores, Cronbach (8) has considered several ways in which the confounding effect of R upon most other variables can be reduced. (a) Computing percentage ratios of each variable over R; (b) removing the linear effect of R by partial regression techniques; (c) reducing the effect of R by plotting the variable against R and drawing a freehand line fitting the medians of the columns (a crude form of curvilinear partial regression); or (d) dividing the total sample into a number of subgroups that are homogeneous with respect to R before proceeding with any quantitative analysis of other variables. The usual procedure of computing percentage ratios is highly unsatisfactory because of the crude metric qualities of most Ror-

schach variables and the lack of a linear relationship between R and other variables. In a study of 790 cases, Fiske and Baughman (14) demonstrated that the relationships between R and other scoring categories are usually complex and nonlinear. Consequently the usual linear regression methods for removing the confounding effect of R will generally fail. Given a standard free-response Rorschach, the only procedure which has any real promise for controlling R is to form subgroups according to R and analyze each one independently. But even this very inefficient procedure leaves unanswered the serious criticism that two records with identical number of responses may be quite different in meaning due to different patterning of responses across the 10 cards.

Recognizing the serious problems in the interpretation of scores when R is a variable, most clinicians make allowance for R in a crude intuitive way. Buhler (5) goes one step further by trying to structure the test administration so that three to five responses will be given to each blot. Blake and Wilson (4) avoid the problem in part by considering only the first response to each card. However, having only 10 responses from which to obtain scores, many of which occur rather rarely, creates a whole host of new problems in attempting to achieve satisfactory standards of measurement.

Standardization of testing conditions and development of procedures for administering the Rorschach to large groups at a time represents another attempt to achieve more objectivity. Munroe (32), Harrower (20), Sells (42), and others have demonstrated the feasibility of group procedures provided one is willing to sacrifice certain aspects of the more unstructured, personalized individual Rorschach. The usual procedure is to project each inkblot on a large screen for three minutes while the subject writes down his responses in a standard booklet. The number of responses is uncontrolled, the subject is usually given a very simple, direct inquiry concerning the role of shape, color, movement, and texture, and location is indicated by drawing the outline of his percept on a miniature replica of the blot.

Most of the scoring difficulties inherent in the standard Rorschach are aggravated still further by use of such group methods. Where one at least has the opportunity for such things as the recording of verbalizations and individualized inquiry to help clear up scoring problems in the standard Rorschach, the group method deprives the examiner of all but the most superficial cues for scoring determinants, increasing further the arbitrary nature of the system.

If one uses standard paper-and-pencil aptitude tests as a model

to be emulated, the most highly structured, psychometrically sound form of the Rorschach would appear to be a multiple-choice test with sufficiently standard instructions to permit its use with large groups of subjects. Under pressure of screening demands during wartime, Harrower and others (20) developed a multiple-choice version in which the subject chooses from a list of thirty concepts those three which look best to him for the particular blot in question. Fifteen of the 30 available concepts presumably indicate psychopathology while the remainder reflect normality. Harrower's own system of scoring is unusual and unnecessarily complicated. Normal answers are arbitrarily weighted "1" for any concept involving human movement, "2" for any that represent a popular response, "3" and "4" for those which involve color-form integration, and "5" for space responses. The set of abnormal answers is assigned weights varying from "6" to "9" in a similar arbitrary fashion. The total score obtained by summing the weights for the concepts chosen is confused in its meaning because of the arbitrary weighting system.

More recently, O'Reilly developed a simpler multiple-choice form with 12 choices per blot, four from psychotic records, four from neurotic records, and four from normals. The subject is asked to select the two concepts which best describe the inkblot. Answers are weighted on a three-point system with "1" for normal and "3" for psychotic. Almost complete separation of normals from psychotics was achieved in a cross-validation, although the neurotics had only slightly higher total scores than did the normals.

Another interesting, objective approach utilizing the multiple-choice format is the concept evaluation technique developed by McReynolds (29). Using Beck's list of good and poor responses according to form level (3), McReynolds selected 25 good and 25 poor concepts spread throughout the 10 Rorschach plates. The subject is shown the location of the concept and asked to indicate whether or not the inkblot looks like the concept. Generally given after a standard Rorschach as part of the testing-the-limits phase, McReynolds' concept test yields an objective, scorable, reliable, and well-defined measure of the degree to which the subject can discriminate good from poor concepts. One of the main advantages of McReynolds' test is the fact that the number of discrete stimuli (intact areas of inkblots) has been increased from 10 to 50 by breaking up the standard 10 Rorschach plates into smaller components. This point is a highly significant departure from the usual ipsative method of allowing repeated response to the same stimulus

and probably accounts for the satisfactory internal consistency (split-half reliability of .82) that McReynolds obtained.

As Harrower (20) has pointed out, the highly structured multiple-choice versions of the Rorschach are no longer equivalent to the standard individual Rorschach except for the inkblots themselves. One could go a step further and question whether or not tests that have completely fixed response alternatives can even be considered projective techniques. In all respects they appear to be objective tests of perception which may have implications for the measurement of important personality traits. The course of development from an unstructured projective technique to a completely structured objective test is complete.

A New Solution

The fundamental question of how to develop psychometrically sound scoring procedures for responses to inkblots while also preserving the rich qualitative projective material of the Rorschach has been approached from a new point of view at The University of Texas.[1] The major modifications undertaken consist of greatly increasing the number of inkblots while limiting the number of responses per card to one, and extending the variety of stimulus colors, pattern, and shadings used in the original Rorschach materials. From an exploratory study it was concluded that a test containing 45 inkblots, to each of which only one response is given, would be feasible to construct and would probably tap essentially the same variables as the classical Rorschach method. Special efforts might have to be made, however, to develop materials which have high "pulling power" for responses using small details, space, and color and shading attributes to compensate for the tendency to give form-determined wholes as the first response to an inkblot.

Such a test would have several advantages over the standard Rorschach: (a) The number of responses per individual would be relatively constant. (b) Each response would be given to an independent stimulus, avoiding the weaknesses inherent in the Rorschach where all responses are lumped together regardless of whether they are given to the same or different inkblots. (c) Making a fresh start in the production of stimulus materials, especially in

[1] Initial impetus for this research was given the writer by a Faculty Research Fellowship from the Social Science Research Council, Inc., of New York. More recently the research program has been supported by a grant-in-aid from the Hogg Foundation for Mental Health, The University of Texas.

view of recent experimental studies of color, movement, shading, and other factors in inkblot perception, would yield a richer variety of stimuli capable of eliciting much more information than the original 10 Rorschach plates, And finally, (d) A parallel form of the test could easily be constructed from item-analysis data in the experimental phases of test development, and adequate estimates of reliability could be obtained independently for each major variable.

The research to date has borne out all original expectations. Two matched alternate forms, A and B, of the Holtzman Inkblot Test have been developed, each containing 45 inkblots. Two additional blots are common to both forms of the test and appear as practice blots before the others. Instructions to the subject are similar to those used in the standard Rorschach with the exception that the subject is asked to give just the primary response to each card, and a brief, simple inquiry is made after each response where necessary to clarify the location or determinants. Administration of the test is easier than the Rorschach, and the subject generally finds giving only one response per card is a fairly simple task.

Six major variables are scored for each response, while a number of minor variables or qualitative signs are scored when deemed appropriate. The major variables were selected and defined according to the following criteria: (a) The variable had to be one which could be scored for any legitimate response. Variables which only rarely occurred were set aside for the moment. (b) The variable had to be sufficiently objective to permit high scoring agreement among trained individuals. (c) The variable had to show some *a priori* promise of being pertinent to the study of personality through perception. And (d) each variable must be logically independent of the others. Location, Form Appropriateness, Form Definiteness, Color, Shading, and Movement Energy Level were selected for intensive study and provided the basis for item-analyses in the final selection and matching of inkblots for Forms A and B.

Location as a variable was defined strictly in terms of the amount of blot used and the extent to which the natural gestalt of the blot was broken up by the response. A three-point weighting system was adopted with "0" for wholes, "1" for large details, and "2" for small areas, making possible a theoretical range of scores from 0 to 90.

The scoring of color was based entirely upon the apparent primacy or importance of color, including black, gray, and white, as a response-determinant. When the subject named the color in his

response, scoring was relatively simple. On rare occasions, when it was apparent that the response would have been highly improbable without the presence of color, credit for color was given even though never mentioned by the subject. A four-point system similar to the Rorschach was adopted with "0" for completely ignoring color and "3" for use of color as the sole determinant. Total scores for Color have a theoretical range from 0 to 135.

While subtle distinctions in the different uses of shading as a determinant are usually made in the Rorschach, no such differentiations are made in the Holtzman Inkblot Test. As with Color, the scoring of Shading was based solely upon the apparent primacy of shading as a determinant. Because pure shading responses are so rare, only a three-point scoring system was used, yielding a theoretical range from 0 to 90.

The scoring of movement is linked closely to content in most contemporary scoring systems for the Rorschach. Too frequently such practices lead to highly arbitrary convention as to whether or not movement is scored or how it is scored. In the Klopfer system (24), for example, "airplane" and "bat" present difficult problems. Can you be sure the airplane is flying? Even when an airplane does fly, there is no movement of its parts and no movement relative to any frame of reference unless landscape is added. Is "bat" to be scored FM for animal movement while "airplane" is scored Fm for inanimate movement when both concepts are really precision alternatives rather than uniquely different responses? The resulting picture is often highly confusing from a psychometric point of view. The essential character of the movement response is the energy level or dynamic quality of it, rather than the particular content. Leaning heavily upon Zubin (52), Sells (42), and Wilson (49), a five-point scale was adopted varying from "0" for no movement or potential for movement, through static, casual, and dynamic movement to a weight of "4" for violent movement such as whirling or exploding. Movement Energy Level ranges theoretically from "0" to 180.

Different authorities vary in the extent to which concept elaborations and specifications are confounded with the goodness of fit of the concept to the form of the inkblot. In the Holtzman Inkblot Test, Form Definiteness was defined independently of form level in the usual sense and refers solely to the definiteness or specificity of the form of the concept represented in the response, disregarding completely the characteristics of the inkblot. Working independently with a large number of concepts culled from inkblot responses, five psychologists placed them in rank order with the most form-

definite concept at the top. The independent sets of ranked concepts were then merged to yield an overall rank order for the entire list. Cutting points were chosen so that five levels of form definiteness could be distinguished. The resulting set of examples served as a scoring manual, with a weight of "0" for the most indefinite concepts, such as anatomy drawing, squashed bug, or fire, and a weight of "4" for the most definite concepts, such as Indian chief, violin, or knight with a shield. Form Definiteness has a theoretical range from 0 to 180.

Form Appropriateness, the last of the six major variables, is by its very nature a subjective variable, requiring extensive preliminary work to make scoring reasonably objective. And yet, it is this very subjectivity which gives the variable great theoretical importance. Beck (3) recognized the likelihood that goodness of fit of the concept to the form of the inkblot would be closely related to degree of contact with reality and undertook a major study of form level that has proved to be one of the most valuable contributions to the Rorschach. Considerable effort was spent in arriving at acceptable standards for scoring Form Appropriateness. Different responses to each inkblot were listed separately for each location and rated independently by at least three judges. A seven-point scale was used with "0" representing extremely poor fit. Although there was good agreement of judges in most cases, a final judgment for each response was reached only after full discussion in conference. The resulting manual provides a guide to the scoring of Form Appropriateness on a three-point system with zero for unusually poor form and "2" for unusually good form. Form Appropriateness can range theoretically from 0 to 90.

The agreement among independent but well trained scorers for a sample of 46 records proved in general to be very high: product-moment correlations of .99 for Location, Form Definiteness, and Movement Energy Level, .97 for Shading, .95 for Color, and .91 for Form Appropriateness. Good estimates of reliability based upon internal consistency were obtained by using Gulliksen's matched random subtest method (18). Correlations ranged from .80 for Form Appropriateness to .91 for Shading. All six variables proved to be reasonably normal and continuous in distribution. Studies are now underway to determine the correlations between Forms A and B with several time intervals and populations of subjects.

Once the standardization of the Holtzman Inkblot Test is complete, it should be possible to develop specialized multiple-choice versions of test for measuring variables of particular interest. Sey-

mour Fisher and Sidney Cleveland have already had some success in developing a series of multiple-choice items to be used with 40 of Holtzman's inkblots which yields a measure of their Barrier Score (13). The particular inkblots used were selected on the basis of earlier item-analysis data so that each blot would be accompanied by three fairly acceptable choices, one representing a barrier response (such as "a knight in armor"), one representing a penetration response (such as "x-ray"), and one which was neutral (such as "flower"). The subject was asked to check the one he liked most and place a different mark on the one he liked least, leaving the third choice blank. Both the Group Rorschach and the new multiple-choice test were given to 60 college students by Fisher and Cleveland. The correlation between the two sets of Barrier Scores was .64 [2]. This fairly high correlation, coupled with the fact that the distribution of scores on the multiple-choice test was much greater than on the Rorschach and was more normally shaped, suggests that the multiple-choice Barrier Score would be superior to the measure reported earlier by Fisher and Cleveland (13).

Considerable ground has been covered in this analysis of the more common problems encountered in the objective scoring of projective techniques. The very nature of the projective hypothesis, that an individual will reveal something of his private self in the way in which he responds to ambiguous stimuli, has encouraged an almost unbelievably wide range of assessment techniques under the rubric of projective methods. In focussing upon quantitative methods of analysis and their objectivity as measured by reproducibility, a whole host of important problems concerning the meaning of projective responses has been deliberately side-stepped. Concepts of validity and their empirical determination, examiner-subject interactions, variability of response across different populations of subjects have been dealt with only tangentially if at all.

One cannot help but observe that few, if any, of these many projective devices can serve well two masters at the same time, particularly when their original purpose is exploitation of the projective hypothesis in the clinical diagnosis of personality. While not necessarily incompatible, the assumptions and historical biases inherent in the projective approach on the one hand and those in the psychometric approach on the other are at opposite extremes of a continuum defined roughly in terms of the degree of structure and control of the subject's response that is imposed by the method. An

[2] Personal communication from Dr. Sidney E. Cleveland.

unfortunate and bewildering array of inadequate quantification characterizes most projective techniques when there is pressure upon the projectivist to conform to the rigorous statistical standards of psychometric theory without concomitant pressure to revise the technique itself. A major challenge to psychologists interested in the objective assessment of personality is the development of psychometrically sound personality tests from available projective devices, a point made by Thurstone (45) 10 years ago which still stands today.

References

1. Auld, F., Jr., Eron, L. D., and Laffal, J. Application of Guttman's scaling method to the TAT. *Educ. psychol. Measmt.*, 1955, 15, 422-435.
2. Baughman, E. E. A comparative analysis of Rorschach forms with altered stimulus characteristics. *J. proj. Tech.*, 1954, 18, 151-164.
3. Beck, S. J. *Rorschach's test: I. Basic processes.* New York: Grune and Stratton, 1944.
4. Blake, R. R., and Wilson, G. P., Jr. Perceptual selectivity in Rorschach determinants as a function of depressive tendencies. *J. abnorm. soc. Psychol.*, 1950, 45, 459-472.
5. Buhler, C., Buhler, K., and Lefever, D. W. *Rorschach standardization studies.* Published privately by authors, 1948.
6. Cattell, R. B. *Personality: A systematic theoretical and factual study.* New York: McGraw-Hill, 1950.
7. Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, 52, 281-302.
8. Cronbach, L. J. Statistical methods applied to Rorschach scores: A review. *Psychol. Bull.*, 1949, 46, 393-429.
9. Dana, R. H. An application of objective TAT scoring. *J. proj. Tech.*, 1956, 20, 159-163.
10. Dorken, H., Jr. The reliability and validity of spontaneous finger paintings. *J. proj. Tech.*, 1954, 18, 169-182.
11. Elizur, A. Content analysis of the Rorschach with regard to anxiety and hostility. *Rorschach Res. Exch.*, 1949, 13, 247-284.
12. Eron, L. D. Responses of women to the Thematic Apperception Test. *J. consult. Psychol.*, 1953, 17, 269-282.
13. Fisher, S. and Cleveland, S. D. *Body image and personality.* Princeton, N. J.: D. Van Nostrand, 1958.

14. Fiske, D. W. and Baughman, E. E. Relationships between Rorschach scoring categories and the total number of responses. *J. abnorm. soc. Psychol.*, 1953, 48, 25-32.
15. Frank, L. K. Projective methods for the study of personality. *J. Psychol.*, 1939, 8, 389-413.
16. Frank, L. K. *Projective methods.* Springfield, Ill.: C. C. Thomas, 1948.
17. Gibby, R. G. Examiner influence on the Rorschach inquiry. *J. consult. Psychol.*, 1952, 16, 449-455.
18. Gulliksen, H. *Theory of mental tests.* New York: Wiley and Sons, 1950.
19. Haggard, E. A. *Intraclass correlation and analysis of variance.* New York: Dryden Press, 1958.
20. Harrower, M. R. Group techniques for the Rorschach test. In Abt, L. E. and Bellak, L. (eds.) *Projective psychology.* New York: A. A. Knopf, 1950.
21. Hartman, A. A. An experimental examination of the Thematic Apperception Technique in clinical diagnosis. *Psychol. Monogr.*, 1949, 63, No. 303.
22. Holt, R. R. The Thematic Apperception Test. In Anderson, H. H. and Anderson, G. L. (eds.) *An introduction to projective techniques.* New York: Prentice-Hall, 1951.
23. Kinget, G. M. *The Drawing-Completion Test.* New York: Grune & Stratton, 1952.
24. Klopfer, B. and Kelley, D. M. The Rorschach technique. Yonkers-on-Hudson, N. Y.: World Book Company, 1942.
25. Lindner, R. M. The content analysis of the Rorschach protocol. In Abt, L. E. and Bellak, L. (eds.) *Projective psychology.* New York: A. A. Knopf, 1950.
26. McClelland, D., Atkinson, J. W., Clark, R. A., and Lowell, E. L. *The achievement motive.* New York: Appleton-Century-Crofts, 1953.
27. Macfarlane, J. W. and Tuddenham, R. D. Problems in the validation of projective techniques. In Anderson, H. H. and Anderson, G. L. (eds.) *An introduction to projective techniques.* New York: Prentice-Hall, 1951.
28. Machover, K. *Personality projection in the drawing of the human figure.* Springfield, Ill.: C. C. Thomas, 1948.
29. McReynolds, P. Perception of Rorschach concepts as related to personality deviations, *J. abnorm. soc. Psychol.*, 1951, 46, 131-141.
30. Meehl, P. E. Configural scoring. *J. consult. Psychol.*, 1950, 14, 165-171.
31. Morgan, C. D. and Murray, H. A. A method for investigating phantasies: the Thematic Apperception Test. *Arch. Neurol. and Psychiat.*, 1935, 34, 289-306.
32. Munroe, R. L. The inspection technique for the Rorschach protocol.

In Abt, L. E. and Bellak, L. (eds.) *Projective psychology.* New York: A. A. Knopf, 1950.
33. O'Reilly, B. O. The objective Rorschach; a suggested modification of Rorschach technique. *J. clin. Psychol.*, 1956, 12, 27-31.
34. Pascal, G. R. and Suttell, B. J. *The Bender-Gestalt Test.* New York: Grune & Stratton, 1950.
35. Rapaport, D., Gill, M., and Schafter, R. *Diagnostic psychological testing. Vol. II.* Chicago: The Year Book Publishers, 1946.
36. Rosenzweig, S. The picture-association method and its application in a study of reactions to frustration. *J. Pers.*, 1945, 14, 3-23.
37. Rosenzweig, S. Idiodynamics in personality theory with special reference to projective methods. *Psychol. Rev.*, 1951, 58, 213-223.
38. Rotter, J. B. and Willerman, B. The Incomplete Sentence Test as a method of studying personality. *J. consult. Psychol.*, 1947, 11, 43-48.
39. Rotter, J. B. Word association and sentence completion methods. In Anderson, H. H. and Anderson, G. L. (eds.) *An introduction to projective techniques.* New York: Prentice-Hall, 1951.
40. Sarason, S. *The clinical interaction.* New York: Harper & Bros., 1954.
41. Schafer, R. *Psychoanalytic interpretation in Rorschach testing.* New York: Grune and Stratton, 1954.
42. Sells, S. B., Frese, F. J., Jr., and Lancaster, W. H. *Research on the psychiatric selection of flying personnel. II. Progress on development of SAM Group Ink-Blot Test.* Project No. 21-37-002, No. 2, Randolph Field, Texas: USAF School of Aviation Medicine, April, 1952.
43. Shneidman, E. S. *Thematic test analysis.* New York: Grune & Stratton, 1951.
44. Stein, M. I. *The Thematic Apperception Test.* Cambridge, Mass.: Addison-Wesley, 1955.
45. Thurstone, L. L. The Rorschach in psychological science. *J. abnorm. soc. Psychol.*, 1948, 43, 471-475.
46. Tomkins, S. S. *The Thematic Apperception Test.* New York: Grune & Stratton, 1947.
47. Trites, D. K., Holtzman, W. H., Templeton, R. C., and Sells, S. B. *Psychiatric screening of flying personnel: Research on the SAM Sentence Completion Test.* Project No. 21-0202-0007, No. 3, Randolph Field, Texas: USAF School of Aviation Medicine, July, 1953.
48. Trites, D. K. *Psychiatric screening flying personnel: Evaluation of assumptions underlying interpretation of sentence completion tests.* Report No. 55-33. Randolph Field, Texas: USAF School of Aviation Medicine, March, 1955.
49. Wilson, G. P. *Intellectual indicators in the Rorschach test.* Unpubl. doctoral dissertation, The University of Texas, Austin, Texas, 1952.
50. Witkin, H. A., Lewis, H. B., Hertzman, M., Machover, K., Meissner, P. B., and Wapner, S. *Personality through perception.* New York: Harper & Bros., 1954.

51. Young, R. D., Jr. *The effect of the interpreter's personality on the interpretation of TAT protocols.* Unpubl. doctoral dissertation, The University of Texas, Austin, Texas, 1953.
52. Zubin, J. and Eron, L. *Experimental abnormal psychology.* (Preliminary Edition) New York: New York State Psychiatric Institute, 1953.