OBJECTIVE EVALUATION OF PERSONALITY TESTS

JOSEPH ZUBIN, P.H. D.,* NEW YORK, N. Y.

Progress in scientific use of tests has always lagged behind advances in clinical practice. This held true of intelligence tests in the last generation and now holds true even more of personality tests. This lag, however, is not limited to the clinical field. Intuitive insights nearly always precede scientific confirmation, just as shock troops always precede the regular army in establishing beach heads. Nevertheless, the status of a science is judged not by its still unestablished intuitions or theoretical formulations, but by the consolidated gains of its technicians, in the same way that final victory in war is attained not by the initial attack, but by mopping up activities and the establishment of law and order in a conquered territory. Regardless of the brilliance of the original intuitive hunch, empirical verification is the foundation stone on which science rests. Bleuler and Ziehen and others before them had evaluated intelligence in patients long before Binet, but if we had nothing more than Bleuler's case histories to lean on, evaluation of intelligence today would be severely handicapped.

The popularity of personality tests in the clinical field is one of the outstanding characteristics of the American scene. Personality tests of the projective variety—Rorschach, TAT, and so forth—are now standard laboratory tests in most clinics, are administered routinely in pre- and posttreatment procedures in psychotherapy as well as soma-therapy. It should be borne in mind, however, that the Rorschach test was not received as warmly everywhere. In prewar Germany (1921) when Roemer tried to introduce it at the meeting of the Association of German Psychologists, it was summarily dismissed by William Stern on the grounds that the test was not sufficiently scientific, and as a result it played no important role in Germany until the present American occupation.

More recently, Manfred Bleuler, who learned the test from Hermann Rorschach himself, in describing the attitude towards the test in Switzerland states that, according to the literature, the test enthusiasts seem to outnumber their opponents; clinicians, however, carry on their work without benefit of tests, either because they consider them superfluous, or because they are unacquainted with them, or because they reject them as potential sources of error.

In our own country, too, the verdict regarding projective tests is far from unanimous. Laying aside the objection arising from the misuse of the test there are some fundamental difficulties which the test is facing. When one examines a Rorschach protocol and interpretation dispassionately, two types of material emerge: (1) material that reflects psychopathological behavior directly and (2) material that must undergo interpretation before its psychopathological meaning is attained. One can readily point to a dozen well-known psychopathological characteristics of schizophrenics that the test reflects directly: rejection tendency, confabulation, bizarre thinking, incongruities, flights of fancy, perseverations, stereotypy, confusion, compulsive behavior, poverty of ideas, vagueness, rigidity, and loss of autocriticism. A good clinician uses these criteria in the diagnosis of schizophrenia, and their direct reflection in the test is of clinical value since it affords a systematic framework for their detection. The elements that Rorschach introduced over and above the already well-known psychopathological characteristics: form, color, shading, movement, and so forth, have unfortunately never been scientifically validated. The alteration of these determinants by perseveration or confabulation, etc., can perhaps be accepted as a manifestation of schizophrenia. But the basic significance of these determinants (that color under certain circumstances reflects emotion, or shading depression) rests solidly on intuition but not on science.
Despite the widespread use of the Rorschach test there is hardly one good-sized study dealing with the empirical verification of the usefulness of the test, its reliability and validity. To be sure, there are the clinical testimonials, the blind analyses, the matching experiments, and the comparison of "signs" and similar techniques devised to demonstrate that the test is useful. Nevertheless, these approaches can at best be only auxiliary rather than basic in proving the worth of the test. The fact that a blind analysis was successful does not reveal why and how it was successful. Besides, we usually hear only the successful cases. The unsuccessful apparently go unreported. Matching of personality sketches with test performances is again too global to permit further analyses of the basis of the matching. Furthermore, it is difficult to determine where the test results leave off and the personality of the examiner enters into the results. A considerable number of statistically controlled studies have turned out negative and still a greater number of individual case studies turn out to be either negative or noncontributory to clinical insight. Even in confirmatory results one is never sure of the degree of conscious or unconscious collusion between test results and final clinical judgment. In view of these difficulties what can the scientist do about determining the basic worth of personality tests— their limitations as well as their virtues?

Let us first classify the major types of personality tests. There are three varieties: personality inventories, projective techniques, and expressive movement techniques.

Personality inventories were the first to attract attention in the last few decades. The usefulness of these personality inventories as screening procedures for the war effort is now a matter of history. Such success cannot but help to renew faith in psychoneurotic inventories even in the hearts of those who witnessed the debacle of these same inventories after World War I. Is there any better prospect for their success now? In order to answer this question adequately one must distinguish between the use of inventories for screening purposes and their use for diagnostic purposes. It should be remembered that only as screens have these inventories proved successful. In order to render them useful for diagnostic purposes their construction and evaluation must be seriously modified.

One outstanding shortcoming in the present-day construction of these tests is the method of item validation. By eliminating all the items that fail to differentiate the normals from the abnormals, the questionnaire becomes an inventory of liabilities only, while the assets of an individual are not given as great an opportunity to emerge in the total score. Even the most adjusted personality may show a few quirks, and a test aimed at revealing quirks only will not be as useful for individual diagnosis as a test that permits both the positive and the negative elements of the personality to appear in the final constellation. Methods of pattern analysis for scoring inventories in which liabilities as well as assets are included must eventually replace present-day scoring, if the inventory is to survive as a diagnostic aid, and not remain a subclinical tool.

One outgrowth of the use of psychoneurotic inventories has been the introduction of rating scales for evaluating the status of patients for diagnostic and prognostic purposes as well as an index of change following applications of therapeutic methods. In the diagnostic field, Father Moore (3), Phyllis Wittman (4), and others have applied rating scales to mental patients with notable success. Father Moore was able to demonstrate that the well-known nosological psychiatric categories could be rediscovered and placed on a surer footing by means of factor analysis. More recently the method has been applied by Peters (5), Malamud (6), Wittenborn (7), and others to evaluate preoperative and postoperative characteristics of patients and their changes. Helpful as these techniques are, it is to be regretted that they depend upon rating of items that are not clearly defined in the hands of personnel not always adequately trained.

The application of statistical techniques to data of this nature is bound to yield a harvest of ill-founded results. It would be highly desirable to develop a basic-English rating scale, as Peters has done, that could be applied to the behavior of mental patients. It might perhaps be well to borrow a page our
in order to render ic purposes their must be seriously

coming in the of these tests is ion. By eliminating to differentiate ormal, the questions of liabilities an individual are trinity to emerge he most adjusted to quirks, and a rks only will not diagnosis as a test dve and the negonality to appear methods of patron nies in which hare included must lawyer scoring, if the a diagnostic aid, tool.

The use of psycho
en the producing the status and prognostic x of change fol
spefic method, other Moore's(3), others have ap
apal patients with score was able to own nosological be rediscovered by means of unity the method's(5), Malamud theers to evaluate attive character
gerous. Helpful v to be regretted v of vains that he hands of per

The second major area in which considerable activity is now taking place is the area of validation. Here, on the one hand, we have studies of childhood and adolescence when personality is developing and degree of maturation can serve as a criterion, and on the other hand, pre- and posttherapy studies of patients undergoing psychotherapy, psychosurgery, and other somatic and nonsomatic therapies where rapid and recognizable alterations of personality can serve as a criterion.

The third of these areas deals with the search for a rationale for personality tests. Two trends are noted:

The first of these deals with the analysis of the relation between perception and personality. This in general is an attempt to provide a link between such tests as the Rorschach and other projective techniques and the underlying postulate that perception in some way reflects underlying personality.
movement of the eyeball and the musculature of the eye are probably also involved in the percepts, may be described as consisting of three processes or elements. The final percept will depend on (1) the complex of external stimuli, the retinal image, or what Koffka has called the proximate stimulus; (2) the Gestalt disposition, that is the tendency to structure the complex of external stimuli into figure ground relationships, whole vs. parts, etc.; and (3) stimulation of the cortex from other sources at the time that the retinal image under discussion impinges upon the cortex. On this third level, anxiety, fears, needs, mental sets including inhibitory processes, facilitating processes, higher and lower levels of mentation, etc., become incorporated into the final percept, and in this way personality variables modify perception.

The scoring systems of present-day projective techniques are notoriously in need of greater objectivity and specificity. At least two directions may be traced in the objectification of these variables. First is the attempt to preserve the present tools and provide scales for quantifying and objectifying the clinically useful methods of scoring now in vogue. Second is the attempt to resolve the complex tests into their components to the end that the simpler tasks could be scored more directly and more objectively.

These objective scales were developed as follows. After surveying the field of Rorschach scoring, signs, and qualitative evaluations it was found that 50 to 60 different dimensions had been utilized by clinicians from time to time. A score (9) was prepared for rating each dimension on a continuum from zero to 4. These scores were anchored with examples at each end of the scale. Whenever quantification was not possible, simple categorization was utilized. By providing an objective framework for cataloguing and classifying Rorschach responses, for example, the work of interpretation was not made less dependent on intuition but this intuition could now play against a solid framework of objective facts rather than against a fragile, nebulous framework of clinically scored Rorschach factors. This does not remove the need for many years of experience and much exposure to Rorschach methodology, but it does provide the clinician as well as the research worker with a series of scores on recognizable variables that can then be manipulated either atomistically or in a molar fashion in patterns, to determine the type of personality that gave rise to the particular record.

In applying this method, every response is scored on each one of the 50 scales. Thus the response (Bat) may have a rating of 4 on the W scale, 2 on the D scale, 4 on the F scale, etc. The mean for each scale is computed as a first step in the evaluation of the record. Then patterns of ratings are tabulated until the most significant findings of the record become available. At present the system is used for research only, but as the method develops and the number of scales is reduced and the clinical significance of the findings is established, it may become a useful clinical tool.

Similar scales have been provided for the evaluation of expressive techniques such as handwriting specimens (10) and for evaluating performance on the Bender-Visual-Motor Gestalt Test (11), Mosaic Test, Word Association Test, Incomplete Sentence, TAT, etc. An incidental advantage of such rating scales stems from the fact that they provide a framework for the intercorrelation of results obtained from various personality tests. Most of these scaling devices are at present, however, research tools only, since to utilize them regularly in the clinic would involve many hours of scoring and evaluation.

In order to answer the everyday demands of the clinic a new type of attack on these tests had to be provided. This attack consists of reducing the complex stimuli of each of the major techniques into its components. In the Rorschach Test, for example, the approach is made from the point of view of the factors in the cards that account for the particular contours of the percept, i.e., what combination of contrasting effects accounts for the particular figures that emerge in the percept. There are the following possibilities: contours have been regarded as due to a sudden change in chiaroscuro or brilliance gradients; secondly, they may be produced by the juxtaposition of two contrasted colors or hues; thirdly, they may be produced by black-white contrasts, that is, by uniform black against uniform white. They may also emerge from gradients in saturation. By providing special tests in which the contours are directly attributed to the particular contrasts equally each of these factors influence on the en Movement Blot's purpose of eliciting producing contour chiaroscuro effect in the gradient fashion a "color" prepared for measure and chiaroscuro is either from the from chiaroscuro areas or within the one of these blot prevalence of colors with the basic factors. The final tests are still in only two or three cases, namely the cut-out, and the Still to be prepared cards. Each sheet to yield to patients and not mental gradients.

A third approach, more object through the utilization of the well known and form in a well-defined exposure time to study the test itself by tachistoscopic factors in the mental field.

The tachistoscopy a new approach to.

The usual response to the word is the most examined, and this approach the centers upon the response word itself that it association, while may be simply in
are directly attributable to each of these particular contrasts we can study experimentally each of these factors to note their influence on the emerging percept. The Levy Movement Blots (9) have been made for the purpose of eliciting movement responses by producing contours that emerge from sudden chiascuro effects, that is, sudden alteration in the gradient of brilliancy. In a similar fashion a "color cut-out" test has been prepared for measuring responsiveness to color and chiascuro in which the contours emerge either from the juxtaposition of colors or from chiascuro effects within black or white areas or within the colors themselves. Each of these tests is rated according to the prevalence of contours and in accordance with the basic factors that produced the contours. The final percept is rated according to the usual determinants. Thus far these tests are still in the experimental stage and only two or three of them have been prepared, namely the movement blots, the color cut-outs, and the black-white contour cards. Still to be prepared are viva cards and texture cards. Each of them, however, does seem to yield interesting differentials between patients and normals as well as developmental gradients in children.

A third approach to the problem of creating more objective and more specific tests is through tachistoscopic exposure. This utilizes the well-known principle of ambiguity and formlessness can be created even in a well-defined visual form by shortening the exposure time. Earlier investigators have tried to study the visual perceptual process itself by tachistoscopic exposure. But personality factors in visual perception may also reveal themselves in tachistoscopic studies. The tachistoscopic method is exemplified by a new approach to the word association technique.

In the usual word-association approach the response of the individual to the stimulus-word is the most important item to be studied, examined, and classified. In contrast with this approach the new tachistoscopic method centers upon the stimulus-word rather than upon the response-word. It is the stimulus-word itself that sets off the complex-bound association, while the response-word itself may be simply incidental and is of value only insofar as it may reveal blocking, clang associations, or other aspects of behavior that are tell-tale of an emotional response to the stimulus-word. Stimulus-words, selected either from the case history or from a general list of words that are calculated to give rise to emotionally toned associations, are presented tachistoscopically at very short exposure times (1/100 to 1/2 of a second) and the following objective measures are utilized in the evaluation of the response: First, the exposure time required for correct recognition of the stimulus-word is noted. Secondly, the type of misrecognition that occurs in the pre-solution stage before the final correct recognition becomes available is noted. Much of this work has been done by Bartlett (12) and more recently by Brunner and Postman (13) and in the recent Columbia Greystone study (14). It seems quite promising as a new approach to the understanding of the perceptual process in relation to personality.

Despite progress in objectifying diagnostic methods, very little progress has been made in understanding the utilization of these methods in the final diagnostic conclusion. No amount of objectification can replace the necessary intuition still required for combining the syndrome of test and interview results into a final diagnosis of the illness. The process of intuition itself is still beyond our ken. To be sure, psychoanalysis sometimes releases the tensions of the examiner to the point where he can permit his intuitive skill to operate more effectively, but this does not help in identifying the intuitive process. Anecdotal reports on the importance of the personality of the examiner are quite prevalent.

For example, in a recent screening of candidates for a graduate school all the tests as well as examiners save one failed to detect a psychopathic trend in one of the candidates. The one examiner who felt very strongly that the candidate should not be accepted described later how he arrived at his decision. This examiner is a very short man of slight build whereas all the other examiners were at least average or above in height and weight. He reported that psychopathic individuals who refrain from blustering and showing their hidden aggression before the physically more impressive examiners are so totally unimpressed by his size that they give vent to these trends quite openly. That this clinician's judgment was correct was borne out by the subsequent behavior of the candidate.
Perhaps a beginning can be made in the analysis of intuition by examining the examiner himself. If the standard error of the examiner can be determined, the standard error of the test will be more predictable.

The validation of projective techniques is proceeding along three major lines. The older method of contrasting groups of patients and normals and of different categories of patients or of normals is still the most prevalent method, but it suffers from the fact that the contrasted groups usually contain overlapping boundaries. Furthermore, group studies cannot serve as a basis for understanding the individual. An interesting example of such studies is afforded by the attempts to determine the relationship between the M response and creativity. Such a series of studies has been conducted by Grace O’Neil under the general guidance of Dr. David M. Levy and the present author. Graduate students rated as creative by their major professors were contrasted with other graduate students whose creativity was rated as rather low but who nevertheless were of the same intellectual level as the others. This was done in the department of English and mathematical statistics. Neither the Rorschach M nor the Levy Movement Blot score differentiated between the “creative” and “noncreative.” Rust(15) found similarly that the movement response to the Movement Blots failed to differentiate between artistically creative and noncreative children, and Dorothy Park at Meredith College obtained similarly negative results in comparing creative and noncreative art students. Our criteria of creativity may be poor, but thus far all attempts at catching this elusive relationship between M and creativity have failed. Corroborative evidence for the absence of such a relationship is afforded by the studies of Anne Roe(16).

In the development method, the growth of a given function such as the M response with age is studied. In this manner it has been determined that M response seems to reach its full growth by age 8, and that responsiveness to shading follows after the growth of the color response, and grows differentially in the two sexes(9).

The third approach to validation stems from pre- and posttherapy evaluations of patients. The widespread use of tests for selection and screening of patients for psychotherapy and somatotherapy and the posttreatment testing of such patients has provided an opportunity for testing out the claims of many instruments. Only one example can be cited in the time available. In the recent Columbia-Greystone Study of Frontal Lobe Ablations(17) pre- and postoperative studies were made of patients on some 35 tests. Nearly all the psychological tests failed to show any decrements as a result of the operation. In general, the performance of patients improved in proportion to the drop in their anxiety. The standard scoring of the Rorschach Test, however, failed to show any correspondence with drop in anxiety except with regard to reaction time, which decreased. A psychometric evaluation of the results by means of the objective rating scales described earlier in this paper yielded interesting statistically significant differences. Three pairs of patients were selected, each pair consisting of one individual who decreased in anxiety and one who increased in anxiety after operation. The judgment of loss and gain in anxiety was based on psychological interviewing by means of anchored scaling devices and on the judgment of the psychiatrist. Only those patients in whom the two criteria concurred were selected. The results indicated that perception of movement of whatever variety, regardless of whether it was accompanied by empathy, correlated positively with anxiety, rising when the anxiety level rose and dropping when the anxiety level fell. The degree of tentativeness or insecurity in giving responses also correlated positively with anxiety. The following variables showed only a unilateral relationship to anxiety levels, declining with a decline in anxiety but showing no corresponding rise with rise in anxiety level: sensitivity to chiaroscuro, anatomical responses, perception of animate objects, perception of objects with texture, and degree of self-reference. The following variables also showed a unilateral but negative relationship with anxiety, showing increases as anxiety fell—accuracy of form perception and degree of congruity of the response. The statistical significance of these differences could be readily established since each patient could be analyzed as a separate sample and the significance of the difference for each patient determined. Only consistent changes were reported.

Another recent attempt was begun by the Clinician (18) inverted factor analysis of data. In the recent synthesis, the clinician selected as a raw inductive descriptor emerged with in types who further factored toward the type. By giving evaluate in term found that the in cards portraying a striking type disliked most the theme, overexcitement, emotionality.

Perhaps the group research with pet absence of a "fit" fitting the jigsaw. Scientific progress accumulation of provide hypotheses or or resort to me for remembering particular situations. But before hypotheses be developed we need a law is unthinkableness, volume, pressure, or of gravitation out the concepts, and energy that are emerging and observations clinical experiments it is perhaps concepts. Those in the field of obj (1) self, (2) intr (3) conformity.

A new concept as a result of the
patients for psychotherapy and the post-middle-aged patients has pro-
for the selection among them with the exception of one case, the only one ex-
tremely time available. In a recent paper, with Dr. B. Moore in abnormal psychology but more recently Cattell and Stephenson have continued the approach.

In the recent symposium on Statistics for the Clinician (18) Stephenson reported an in-
verted factor analysis (Q technique) of traits selected as a random sample from Jung's inductive descriptions of personality. He emerged with introverted and extraverted types who furthermore had very definite at-
titudes toward the extremes of their own type. By giving them the Szondi cards to evaluate in terms of likes and dislikes he found that the introverts disliked most the cards portraying apathy and shut-in qualities to a striking degree, while the extraverts disliked most those cards depicting obvious mania, overexcitement, and exaggerated emotionality.

Perhaps the greatest deficiency in clinical research with personality tests today is the absence of a "theoretical framework" for fitting the jigsaw pieces of available facts. Scientific progress does not consist of the accumulation of facts alone. We must provide hypotheses on which to hang our facts, or resort to mechanical mnemonic systems for remembering these facts and the particular situations in which they hold true. Before hypotheses, theories, and laws can be developed we must have concepts. Boyle's law is unthinkable without the concepts of volume, pressure, and temperature. The theory of gravitation would be impossible without the concepts of mass, velocity, acceleration, and energy. What are the concepts that are emerging from the welter of data and observations in case history studies and clinical experimental investigations? At present it is perhaps foolhardy to suggest such concepts. Those that seem to be most useful in the field of objective personality tests are: (1) self, (2) intra-individual variability, and (3) conformity.

A new concept that seems to be emerging as a result of the work in somatic therapy is the concept of emotional charge. Thus lobotomized patients do not lose their compulsive trends, nor their memory of having been compulsive but merely feel less of an urge to carry them out. Schizophrenic patients still hear their hallucinatory voices after operation but they are less urgent and less overpowering and finally fade out because of competing stimulation, whereas formerly the hallucinatory phenomena took precedence. Patients suffering from intractable pain still feel their pain but it is no longer intractable and they can bear it. ECT patients do not lose their traumatic memories but have a raised threshold of the feeling of familiarity for them. Just how this new series of facts is to be conceptualized and integrated is still a question of the future.

Techniques that can evaluate the emotional charge that a person possesses for a given situation may help in understanding the results of projective techniques.

There is need today for a catalogue of the various concepts now in vogue and their relative usefulness in personality research. Perhaps an examination of these concepts on a factorial or logical basis may reveal their structure and potentialities.

**SUMMARY**

The final judgment that the scientist must perform arrive at is that no objective evaluation can now be made of personality tests because they have not yet attained the status of tests yielding specifically defined scores. As techniques for aiding in clinical judgment, they have proved their worth. As independent tests they are found wanting.

The reason for their failing is that no two responses are ever sufficiently identical to be classified as equivalent. In order to evaluate responses some abstract dimensions must be provided. But insufficient theoretical frameworks have thus far been provided for abstracting the concrete response. The present scoring systems are too concrete, too close to the original response, and not sufficiently abstracted to yield a measurable dimension of behavior. When the personality tests yield abstract scorable dimensions of the variety that the physicist finds when he abstracts from a given concrete object its weight, temperature, volume, etc., we shall be able to
make more headway. To those who say that no such measures will ever be possible, let me point out this allocution: The ancient Babylonians and Egyptians in the dawn of history had no thermometers but they did evaluate temperature concretely in terms of the heat of fire, and the heat of summer and cold of winter. They had a scale ranging from the heat of fire, through the hottest day of summer to the coldest day of winter. This scale obtained until the thermometer was born. History does not record the reaction of the populace to the first thermometer, but I can conjure for you the following complaints about it: "You can't measure such an imponderable characteristic as temperature—it is too global, too diffuse, too all-encompassing to yield its secret to that mercury stick. Besides, yesterday I perspired freely, today is relatively cool, but that thermometer registers the same reading." Had the thermometer makers discarded their instrument at that point, science would have suffered a severe loss. Only by persevering with the thermometer did we discover the other factors that go into the global concept of subjective warmth.

For this reason, scientists are providing rating scales for catching the essence of the concrete responses and for classifying these essences along proper dimensions. The virtue of these scales is that they reveal not only what types of responsiveness the patient exhibited, but also which type he failed to exhibit. Another solution that scientists are seeking is to reduce the personality techniques to simpler structure, and to utilize more specific direction for obtaining measurable performance.

Whether such approaches will eventually elevate projective techniques to the status of personality tests is still debatable. Some evidence, however, has been provided to show that scaling of variables and simplifications of test material lead to more precisely definable relationships with clinically observable variables. Whether these tests, even in their higher level of development, can replace clinical judgment, interviews, or examination is highly doubtful. But that they can be of greater helpfulness and of greater dependability is much to be hoped.

**BIBLIOGRAPHY**


**USES OF THE TAT**

The choice of "projective" topic for the theoretical meeting of The American Psychiatric Association is another heuristic sign, of multiple interest and purpose in the field of psychiatry and psychology. It seems that the other side of the two professors is secure enough and may give the bumptious young psychiatrist an opportunity to speak up and be heard.

The choice of this topic seems to me, a mountaineer, to be like psychiatrists for investigating the psychological sort, in a mountain of important researches of the sort.

I would be not a little surprised, if I will, after all, have a chance test with which my name is linked, if this test were soon to reveal more brains than mine. The suggestion for the TAT cannot be a success in abnormal psychology of Mrs. Cecilia T. Rice, in the first phases of its development. The picture selection and the administration and interpretation by Mrs. Christiana D. T. Sanford, Tomkins, Brook-Port, Stein, Rosenzweig for more than character.

This afternoon, with all the present-day wise men shall assume the role of sharpen the argument, and it is the assured proposition of myself—particularly the condemned psychiatrist—should be a part of administering a TAT.

My first reason for sharing with you this presentation is that I am now a TAT examiner with selfish prejudice.  

1. Read at the 100th annual meeting of the American Psychiatric Association, May 1940.