# The Future of Diagnostic Testing in Clinical Psychology

William A. Hunt
*Northwestern University*

Originally published in JCLP, 2, 311–317 (1946).  J Clin Psychol 56: 341–347, 2000.

Psychological testing has firmly established itself as a diagnostic procedure in clinical practice. The tests employed and the diagnostic uses to which they are put are well known and it is not proposed to review them here. Rather let us face the fact that diagnostic testing is at present in a state of relative stagnation. Anyone surveying the tremendous development of clinical psychology during the last ten years, and the major importance that psychological testing assumes in such clinical practice, cannot help but be struck by the small amount of progress we have made in developing our psychological tests as diagnostic instruments. Our advances have been in the expansion of physical facilities, in the extension of clinical services, rather than in the improvement of our existing diagnostic techniques and the discovery of new ones. This paper will offer a possible explanation of our lack of progress and suggest certain lines of attack upon our problems that may help to move us out of the present doldrums.

To do this it is necessary to return to fundamentals. My suggestions will stem from two basic premises concerning the fundamental nature of testing:

1. The main contribution of the psychological test is that it offers an opportunity of sampling a subject's behavior in a standard situation.
2. The main contribution of the individual test (as opposed to the group test) is that it offers the tester an opportunity personally to observe such behavior as it takes place.

It follows from the first premise that the primary datum offered by the psychological test is the subject's raw behavior in the test situation. The mathematical symbols into which this behavior can be translated are secondary instruments of convenience and should not be allowed to conceal the primary datum, the actual behavior. That our math-

This paper is based upon some remarks delivered before the Chicago Psychology Club and the University of Minnesota chapter of Psi Chi.

ematical measures, ranging from the simple use of numerical units in scoring to the use of symbolic measures such as the mental age and the intelligence quotient, do obscure the richness of the behavioral data upon which they are based would be admitted by any psychologist, but the point needs constant reemphasis.

Let me illustrate it by quoting some actual answers to two questions on the Information sub-test of the Wechsler-Bellevue Intelligence Scale. In response to the question, "How far is it from Paris to New York?" a subject may answer "About 3,000 miles"; but I have had another subject say "Unfortunately I cannot be as exact as I would like to. No, I don't know exactly. For an approximation—about 3,000 miles. Sorry I can't answer more definitely." Both these answers are correct and count the same in the scoring system, with the numerical symbols concealing the diagnostic richness of the second answer. In response to the question, "Where is Egypt?" a subject may answer "In South America"; but I have had a schizophrenic answer "In a manner of speaking it may be said to be in an oasis—plenty surrounded by sand." Both answers are wrong and in scoring are represented by the same symbol, zero. Not only is the pathological significance of the second answer lost, but I would submit that a real difference in intelligence is overlooked.

A further example may be offered from one of Kent's brief tests of Mathematical Reasoning. One of the questions is "If 8 boys club together and pay *2* dollars for the use of a room, how much should each pay?" Any answer other than 25c is scored as incorrect, but a careful examination of the various "wrong" answers shows some interesting differences among them. The most frequent responses for those mental deficients who attempted this question were 4 and 16. Apparently the mental deficients were able to isolate the necessary mathematical elements of the problem (8 and *2*) and also to comprehend vaguely that something more "complex" than addition or subtraction was called for. Unable to divide 2 by 8, they fell back either upon *multiplying* 2 by 8 or upon *dividing* 8 by 2. On the other hand, a group of malingerers tended to answer 23c or 27c with a range of answers through the twenties. They grasped the fundamental procedure but selected an error in calculation as their response, something the mentally deficient did not show. The examination of test responses on this question thus enables us not only to subject the reasoning processes of the mental deficient to further analysis but also to differentiate true mental deficiency from malingering[(2)]. All this would have been lost had we not carried our examination back beyond the numerical scores to the original test behavior.

Many individual clinicians do not overlook such data. They are not content to base their judgment upon the mere test score or profile of scores but carry their interpretation back to the subject's original performance. This is done somewhat shamefacedly, and is referred to apologetically as the exercise of "clinical judgment" or even more apologetically as "clinical intuition." This is not intuition in the mystical sense. It is the same sort of intellectual process of judgment that ensues when a psychologist considers a test score in the light of the known validity and reliability of the test used before making an interpretation, and in many cases the mathematical data upon which such an interpretation is based are no more reliable than the observational data upon which we base our clinical "intuitions."

Our standard test manuals, however, give little space to any discussion of the quality of test responses and their interpretive significance. The Wechsler-Bellevue manual[(4)] devotes only 22 pages to criteria for scoring. The Terman-Merrill manual[(3)] is much better, but both manuals limit their treatment of test responses to the problem of translating the response into the particular numerical symbols used in their respective scoring systems. Nor is the professional literature more helpful. Our journals are filled with articles on the mathematical treatment of test scores, but only rarely does one find any discussion of actual test behavior and its significance. Behavior as such, seems to be

viewed as a necessary evil, justified only by the fact that it will yield us a numerical symbol with which we can then embark upon a flight of mathematical abstraction.

There is no necessary antithesis between the observation of the subject's test behavior and the expression of this behavior in convenient numerical symbols which lend themselves to statistical manipulation. The two approaches are complementary. Actually some return to the observation of test behavior is a necessary precursor to further numerical "objectification." In a recent paper on "An Analysis of the Concept of Clinical Intuition," Cofer[1] has attempted to identify some of the actual test behaviors upon which our clinical judgments are based. Once such behaviors are identified they can be translated into numerical symbols. If they prove valid diagnostic indicators we then extend our objective scoring system to include them.

We can use the Kent item mentioned above as an illustration. At present the response "25c" contributes one point to a score for intelligence. Any other answer contributes nothing. An examination of these "other" answers, or errors, however, reveals reliable differences between the mistakes made by the truly mentally deficient and those made by malingerers. It is then possible to extend our scoring system, and say that any answer not 25 but within the range of the twenties will also count one point on a scale for malingering.

The purpose of the present paper, however, is not merely to encourage clinical psychologists to look beyond test scores to the underlying test behaviors, nor to suggest the stimulation and new research ideas that might result from such contact with the raw materials of clinical diagnosis. The value of such an approach is accepted in psychology. Rather we would call attention to two consequences which follow logically from the acceptance of this view and which have implications for the development of our diagnostic techniques.

Since we are all agreed upon the diagnostic richness of actual test behaviors, and since most of us, however apologetic we may be in practice, do use such behavior as a basis for our clinical judgments, let us face this fact in the development of new tests. Individual test items as well as types of sub-tests differ in the amount of such clinical material that they offer. Let us rework our present tests and throw out those items that do not offer it. In constructing new tests let us select items that are deliberately chosen not only to allow a numerical measure but also to provide the subject with an opportunity for revelatory clinical behavior even though it goes beyond the present potentialities of "objectification." Without losing the objective efficiency of our present tests we can increase their clinical utility by such deliberate selection of diagnostically rich test items.

If we are to encourage the use of such test material, we must face the fact that only trained clinicians can use it. We need as psychological testers trained observers and interpreters with a wealth of clinical experience behind them, not untrained cashiers to operate an automatic scoring cash register. These last may be left to the field of group testing which is frankly committed to the limitations of the exclusively objective approach. Group testing is essentially nomothetic. In the individual test, however, we can add to the nomothetic all the flexibility of the idiographic approach.

This leads us to the second premise stated at the beginning of our paper—that the main contribution of the individual test (as opposed to the group test) is that it offers the tester an opportunity personally to observe the subject's test behavior as it takes place. There are other contributions. A wider range of test materials can be presented. Moreover, the standard conditions assumed in the group testing situation can more definitely be assured. The opportunity for the observation of test behavior, however, remains the primary value of the individual technique.

This second premise has an implicit corollary that is often overlooked. It is that in the individual testing situation the tester is expected to contribute to raising the level of

prediction. This is not to say that the tester in the individual test situation *does* succeed in raising the level of prediction. A poor tester may even lower it. The fact remains, however, that his participation is based either implicitly or overtly on the belief that his presence will increase the efficiency of the testing. In this connection we might cite the practice of some clinical psychologists in administering group tests on an individual basis. Such a contribution might come about through the wider range of materials that can be presented, or through the extra care which is possible in administering the test and the extra attention which can be given in assuring the desired standard conditions. It may also come about through the deliberate intervention of the tester in changing some of the "standard" conditions to fit same special requirement of the subject or of the testing situation, as well as through the clinical interpretation of the resulting score or the addition of a further clinical judgment based Upon the observation of test behavior which is not amenable to translation into standard scoring measures. Such procedures may be frowned upon officially, but they are used by most practicing clinical psychologists, who seem willing both to accept them and to attempt their justification. Unless the tester in the individual testing method is expected to make some such contribution toward better prediction it is difficult to justify his participation.

At the risk of being accused of making an artificial or erroneous distinction between group testing and individual testing, I should like to bring out what seems to me to be an underlying and tacit, though seldom consciously realized, difference in fundamental philosophy between the two approaches. The accepted goal of both is perfect prediction. I would submit, however, that in general, group testing is used in situations where a certain amount of error is acceptable, and that psychologists using group tests operate acquiescently and even contentedly in many situations in which the test prediction is far below the level of perfection. On the other hand, it seems to me that the individual tester never openly accepts the margin of error inherent in the test, but always strives consciously and definitely toward perfect prediction. I am not saying that the individual test actually does come closer to perfection, nor denying the value of group tests and their necessary use in many situations, but merely suggesting a difference in the fundamental motivation of the people using them. My point is that group tests are used with full acceptance and understanding of the test error involved, whereas the individual test is used in an attempt to lessen the inherent test error, with the clinician striving through personal supervision and the addition of clinical interpretation to achieve better prediction than can be attained with the mass methods of group testing. Whether or not he is successful is another question.

This difference in philosophy between the group test and the individual test was evident in their military uses. In general group tests were used in selection procedures where the manpower pool from which the selectees were drawn was large and where failure on the test did not unduly stigmatize the individual. Thus in the Navy, group tests were used in classification for selecting candidates for the various trade schools, specialized services where the manpower reserve from which the candidates were taken was sufficiently large so that the loss through test error of some potentially acceptable men was not serious, and where men rejected by the test were not lost to the Navy but were passed on to some other branch of the service. Moreover, while the failure to make a certain trade school may have seemed important to the recruit involved, his failure to do so did not entail any serious social stigma. The same was true of the use of group tests in selecting, aviators. As opposed to this, individual tests plus clinical interpretation were used in the neuropsychiatric examination where the manpower reserve being tapped (the military manpower potential of the country as a whole) was low, where rejection meant the loss of the man to the military services, and his return to society with the social stigma attached to a discharge for psychiatric reasons. The same general trend is reflected in

civilian practice where group tests are used for such things as selecting insurance salesmen, admission to college, selecting students for special classes, etc., while the individual test with a clinical interpretation is relied upon in cases such as commitment to an institution, where the social consequences for the individual are particularly severe.

We have already advocated the frank and open acceptance of the importance of the clinical psychologist in the individual testing situation and the admission as a valid clinical instrument of his use of clinical intuition or, as I would rather phrase it, professional judgment. Such acceptance, however, cannot be based upon faith, hope, and professional charity toward one's clinical colleagues. It must be based upon a sound body of scientific evidence. It will be necessary to consider the clinician objectively as a testing instrument and to submit him to the same objective processes of validation that we would use in evaluating any test. The validity and reliability of his judgments are as open to experimental verification as are the validity and reliability of our tests, and such verification must be carried out.

The objection is often raised to the clinical approach that, while many clinicians can make valid clinical judgments, many cannot, i.e., that there are bad as well as good clinicians. The same thing is true of psychological tests. There are bad tests as well as good tests. When we discover that a test is bad, we do not hesitate to discard it. Just so, when we find that a clinician is "bad," or cannot make valid professional judgments, he, too, should be discarded, or limited in his activities to those fields where the value of his contribution can be demonstrated. Such evaluation of clinicians has been lacking in the past, nor will it be easily installed in the future, although the present interest of the American Psychological Association in certification for applied psychologists is a hopeful sign.

In evaluating clinical performance we must be careful to avoid committing the "isomorphic" error that has marked our previous thinking when we have assumed that there is a direct, one-to-one correlation between the performance of the individual clinician and the excellence of the training program which he has undergone. We have limited our critical inspection to a survey of the thoroughness of the curriculum, excellence of the teaching staff, and breadth of clinical experience available in those institutions which offer training programs, assuming that a good training program assured a good clinician. Unfortunately this is not always so, and it will be necessary to supplement our evaluation of training programs by a further professional examination of the clinicians they produce. The evaluation we are suggesting, however, goes well beyond the original selection of clinical workers for the field. We would propose a continuous evaluation of clinical performance as an integral part of administrative practice in any clinic.

Such evaluation may be difficult to obtain on an individual clinician working independently, but is relatively easy if the clinician is functioning as one of a team in an organized clinic. In this latter case it is easy to have an individual's judgment checked by a colleague or by further testing. In most clinics the patient is usually seen by more than one professional worker and the record will contain test scores and case history material which offers a further check. Adequate follow-up material on each case could also be obtained. In fact, if an adequate system of clinical records is established, the checking of each worker's efficiency becomes merely a matter of clinical bookkeeping which can be done with little extra work, and provides a continuous, running evaluation of clinical performance for any staff, member.

Such a system was in practice at the Psychiatric Unit at the Newport, R. I., Naval Training Station during the last war whenever the exigencies of the war emergency left time for its use. It was thus possible to make a direct comparison of the relative efficiency of the psychiatric interview administered by the staff personnel and the group paper-and-

pencil test as neuropsychiatric selection procedures[5]. It developed that both procedures were about equal in their detection rate for potentially unfit personnel but that the false-positive rate (number of fit recruits falsely identified as unfit) was much higher for the test. As a result, the paper-and-pencil test was used as a preliminary coarse screen to select men for subsequent psychiatric interview, and it was possible to cut down the number of personnel engaged in interviewing by two-thirds without loss in the efficiency of the screening procedure. It also developed that there were large individual differences in the ability to handle the brief psychiatric interviewing technique, with the result that specialization was introduced with some members of the staff being assigned to interviews and others being given the task of working up cases on the ward where the pace was more leisurely and a more detailed, painstaking investigation of each case was necessary. Moreover, differences in the ability to handle certain types of case were discovered. Some men were good at handling psychopathic personalities; others were not, but might excel with schizophrenics, epileptics, or homosexuals. These differences were taken into account in assigning cases, and more efficient teamwork was the result.

On the testing side, differences were revealed in the ability to handle the abbreviated intelligence testing techniques used to supplement the original screening interview. Some men were very proficient with these, others did better with the longer tests such as the Wechsler-Bellevue scale. Some clinicians were excellent with the Rorschach test, some with the Minnesota Multiphasic, while others produced adequate judgments of personality structure as a secondary product of administering individual intelligence tests. The demonstration of such individual differences in clinical performance made an efficient allotment of duties possible. The continuous nature of the check provided by such clinical bookkeeping even made possible the detection of the "staleness" and operational fatigue which developed inevitably in a group which was being driven dangerously close at times to the limits of physical capacity. It was a compliment, not merely to the professional caliber of the staff at Newport but to the professional motivation and integrity of psychiatry and psychology as a whole, that such evaluation procedures were actively welcomed and willingly participated in by the individual staff members.

What we are suggesting here is the application of the principles of applied psychology to the field of clinical practice itself. Psychology has long been a leading proponent of efficient personnel procedures in industry and has produced many important studies in the field of motor skills and the efficient organization of work habits. It is fitting now that it turn its attention on itself, and there is no better field in which to begin than clinical psychology. Efficiency engineering is as appropriate in the clinic as it is in industry, and if the volume of future clinical practice turns out to be anywhere near our present estimates it will be not only appropriate but necessary.

## Summary

This paper opened with the observation that diagnostic testing in clinical psychology was in a state of developmental quiescence with little evidence at present of any very effective solution of its many problems. After considering certain premises concerning the basic nature of testing, some suggestions were made for progress within the field. These suggestions may be summarized as follows:

1. As clinical psychologists we should pay more attention to the subject's raw behavior in the actual testing situation and not concentrate exclusively on the resulting numerical scores.

2. We should rework our tests to obtain items which will yield diagnostically rich observable material as well as convenient numerical measures.
3. We should accept the importance of the clinician as a contributing element in the test situation.
4. We should consider the individual clinician as a clinical instrument, and study and evaluate his performance exactly as we study and evaluate a test.

## References

1. Cofer, C. N. An analysis of the concept of clinical intuition. Paper given at the University of Maryland Conference on Military Contributions to Methodology in Applied Psychology.
2. Hunt, W. A. and Older, H. J. Detection of malingering through psychometric tests. Nav. Med. Bull., Wash., 1943, 41, 1318–1323.
3. Terman, L. M. and Merrill, M. A. Measuring intelligence. Boston: Houghton Mifflin Company, 1937.
4. Wechsler, D. The measurement of adult intelligence. (3rd. Ed.) Baltimore: Williams and Wilkins, 1944.
5. Wittson, C. L. and Hunt, W. A. Three years of naval selection—a retrospect. War Med., 1945, 7, 218–221.