*Canadian Journal of Psychology* (1959), *13*, 102-128. #047

# SOME RUMINATIONS ON THE VALIDATION OF CLINICAL PROCEDURES[1]

PAUL E. MEEHL

*University of Minnesota*

It is becoming almost a cliché to say that "clinical psychology is in a state of ferment," a remark which is ambiguous as to whether the "ferment" is a healthy or pathological condition. Dr. E. Lowell Kelly finds upon follow-up that about 40 per cent of the young clinicians who were studied in the early days of the Veterans' Administration training programme now state that they would not go into clinical psychology if they had it to do over again (personal communication). In recent textbooks, such as Garfield's, one can detect a note of apology or defensiveness which was not apparent even a decade ago (**13**, pp. vi, 28, 88, 97, 101, 109, 116, 152, 168, 451, and *passim*). No doubt economic and sociological factors, having little to do with the substance of clinical psychology, contribute in some measure to this state of mind within the profession. But I believe that there are also deeper reasons, involving the perception by many clinicians of the sad state of the science and art which we are trying to practise (**17**). The main function of the clinical psychologist is psychodiagnosis; and the statistics indicate that, while the proportion of his time spent in this activity has tended to decrease in favour of therapy, it nevertheless continues to occupy the largest part of his working day. Psychodiagnosis was the original basis upon which the profession became accepted as ancillary to psychiatry, and it is still thought of in most quarters as our distinctive contribution to the handling of a patient. One is therefore disturbed to note the alacrity with which many psychologists move out of psychodiagnosis when it becomes feasible for them to do so. I want to suggest that this is only partly because of the even higher valence of competing activities, and that it springs also from an awareness, often vague and warded off, that our diagnostic instruments are not very powerful. In this paper I want to devote myself entirely to this problem, and specifically to problems of validity in the area broadly labeled "personality assessment."

I have chosen the word "ruminations" in my title. It helps from time to time for us to go back to the beginning and to formulate just what we are

---

[1] Invitational Address to the Canadian Psychological Association's Convention at Edmonton, Alberta, June 12, 1958.

trying to do. I shall have to make some points which are perhaps obvious, but in the interest of logical completeness I trust that the reader will bear with me. In speaking about validity and validation, I shall employ the terminology proposed by the APA committee on test standards, making the fourfold distinction between predictive, concurrent, content, and construct validity. (**1**, see also **6**.)

The practical uses of tests can be conveniently divided into three broad functions: *formal diagnosis* (the attachment of a nosological label); *prognosis* (including "spontaneous" recoverability, therapy-stayability, recidivism, response to therapy, indications for one kind of treatment rather than another); and *personality assessment* other than diagnosis or prognosis. This last function may be divided, somewhat arbitrarily, into *phenotypic* and *genotypic* characterization, the former referring to what we would ordinarily call the descriptive or surface features of the patient's behaviour, including his social impact; and the latter covering personality structure and dynamics, and basic parameters of a constitutional sort (for example, anxiety-threshold). Taking this classification of test functions as our framework, let us look at each one, asking the two questions: "Why do we want to know this?" and "How good are we at finding it out?"

Consider first the problem of formal psychiatric diagnosis. This is a matter upon which people often have strong feelings, and I should tell you at the outset that I have some prejudices. I consider that there are such things as disease entities in functional psychiatry, and I do not think that Kraepelin was as mistaken as some of my psychological contemporaries seem to think. It is my belief, for example, that there is a *disease,* schizophrenia, fundamentally of an organic nature, and probably of largely constitutional aetiology. I would explain the viability of the Kraepelinian nomenclature by the hypothesis that there is a considerable amount of truth contained in the system; and that, therefore, the practical implications associated with these labels are still sufficiently great, especially when compared with the predictive power of competing concepts, that even the most anti-nosological clinician finds himself worrying about whether a patient whom he has been treating as an obsessional character "is really a schizophrenic."

The fundamental argument for the utility of formal diagnosis can be put either causally or statistically, but it amounts to the same kind of thing one would say in defending formal diagnosis in organic medicine. One holds that there is a sufficient amount of aetiological and prognostic homogeneity among patients belonging to a given diagnostic group, so that the assignment of a patient to this group has probability implications which it is clinically unsound to ignore.

There are three commonly advanced objections to a nosological orientation in assessment, each of which is based upon an important bit of truth but which, as it appears to me, have been used in a somewhat careless fashion. It is first pointed out that there are studies indicating a low agreement among psychiatrists in the attachment of formal diagnostic labels. I do not find these studies very illuminating (**2**, **34**, **38**). If you are accustomed to asserting that "It is well known that formal psychiatric diagnoses are completely unreliable," I urge you to re-read these studies with a critical set as to whether they establish that thesis. The only study of the reliability of formal psychiatric diagnosis which approximates an adequate design is that of Schmidt and Fonda (**48**); and the results of this study are remarkably encouraging with regard to the reliability of psychiatric diagnosis. As these authors point out, some have inferred unreliability of formal diagnosis from unreliable assessment of other behavioural dimensions. Certainly our knowledge of this question is insufficient and much more research is needed.

I suppose that we are all likely to be more impressed by our personal experience than by what someone else reports when the published reports are not in good agreement and there is insufficient information to indicate precisely why they come to divergent results. For example, it is often said that the concept "psychopathic personality" is a wastebasket category that does not tell us anything about the patient. I know that many clinicians have used the category carelessly, and it is obvious that one who uses this term as an approximate equivalent to saying that the patient gets in trouble with the law is not doing anything very profound or useful by attaching a nosological label. I, on the other hand, consider the asocial psychopath (or, in the revised nomenclature, the sociopath) to be a very special breed of cat, readily recognized, and constituting only a small minority of all individuals who are in trouble because of what is socially defined as delinquent behaviour (in this connection see **31**, **50**). I consider it practically important to distinguish (*a*) a person who becomes legally delinquent because he is an "unlucky" sociopath, that is, got caught; (*b*) one who becomes delinquent because he is an acting-out neurotic; and (*c*) a psychiatrically normal person who learned the wrong cultural values from his family and neighbourhood environment.

Being interested in the sociopath, I have attempted to develop diagnostic skills in identifying this type of patient, and some years ago I ran a series on myself to check whether I was actually as good at it as I had begun to believe. I attempted to identify cases "at sight," that is, by observing their behaviour in walking down the hall or sitting in the hospital lounge, without conversing with the patient but snatching brief samples of verbal behaviour and

expressive movements, sometimes for a matter of a few seconds and never for more than five minutes. In the majority of cases I had no verbal behaviour at all. In the course of a year, I spotted 13 patients, as "psychopathic personality, asocial amoral type"; accepting staff diagnosis *or* an MMPI profile of psychopathic configuration as a disjunctive criterion, I was "correct" in 12 of the 13. This does not, of course, tell us anything about my false negative rate; but it does indicate that if I think a patient is a psychopath, there is reason to think I am correct. Now if I were interested in examining the "reliability" of the *concept* of the psychopathic personality, I should want to have clinicians like myself making the judgments.

Imagine, if you will, a psychologist trained to disbelieve in nosological categories and never alerted to those fascinating minor signs (lack of normal social fear, or what I call "animal grace," a certain intense, restless look about the eyes, or a score of other cues); suppose a study shows that such a psychologist tends not to agree with me, or that we both show low agreement with some second-year psychiatric resident whose experience with the concept has been limited to an hour lecture stressing the legal delinquency and "immaturity" (whatever that means) of the psychopath. What importance does such a finding have?

This matter of diagnostic skill involves a question of methodological presuppositions that is of crucial importance in interpreting studies of diagnostic agreement. The psychologist, with his tendency to an operational (**20**) or "pure intervening variable" type of analysis (**32**, **47**) and from his long tradition of psychometric thinking in which reliability constrains validity, is tempted to infer directly from a finding that people disagree on a diagnostic label that a nosological entity has no objective reality. This is a philosophical mistake, and furthermore, it is one which would not conceivably be made by one trained in medical habits of thinking. When we move from the question of whether a certain sign or symptom should be given a high weight to the quite different question whether a certain disease entity has reality and is worth working hard to identify, disagreement between observers is (quite properly) conceived by physicians as *diagnostic error.* Neurological diagnoses by local physicians in outstate Minnesota are confirmed only approximately 75 per cent of the time by biopsy, exploratory surgery, or autopsy at the University of Minnesota Hospitals. The medical man does not infer from this result that the received system of neurological disease entities is unsound; rather he infers that physicians make diagnostic mistakes.

Furthermore, it is not even assumed that all of these mistakes could be eliminated by an improvement in diagnostic skill. One of the most highly skilled internists in Minneapolis (**43**) published a statistical analysis of his

own diagnoses over a period of 28 years based on patients who had come to autopsy. Imposing very stringent conditions upon himself (such as classifying a diagnostic error as eliminable if evidence could have been elicited by sufficient re-examination), he nevertheless found that 29 per cent of his diagnoses were errors which could not in principle have been eliminated because they fell in the category of "no evidence; symptoms or signs not obtained." How is this possible? Because not only are there diseases which are *difficult* to diagnose; there are individual cases which are for all practical purposes *impossible* to diagnose so long as our evidence is confined to the clinical and historical material.

Presumably anyone who takes psychiatric nosology seriously believes that schizophrenia (like paresis, or an early astrocytoma in a neurologically silent area) is an *inner state,* and that the correct attachment of a diagnostic label involves a probability transition from what we see on the outside to what is objectively present on the inside. The less that is known about the nature of a given disease, or the less emphasis a certain diagnostician gives to the identification of that disease, the more diagnostic errors we can expect will be made. That some psychiatrists are not very clever in spotting pseudoneurotic schizophrenia is no more evidence against the reality of this condition as a clinical entity than the fact that in 1850, long prior to the clinching demonstration of the luetic origin of paresis by Noguchi and Moore, even competent neurologists were commonly diagnosing other conditions, both functional and organic, as "general paralysis of the insane." By 1913 the luetic aetiology was widely accepted, and hence such facts as a history of chancre, secondary stage symptoms, positive spinal Wassermann, and the like were being given a high indicator weight in making the diagnosis (**27**). Yet the entity could not properly be *defined* by this (probable) aetiology; and those clinicians who remained still unconvinced were assigning no weight to the above-mentioned indicators. This must inevitably have led to diagnostic errors even by very able diagnosticians. It is impossible for diagnostic activity and research thinking to be suspended during the period—frequently long—that syndrome description constitutes our only direct knowledge of the disorder (**33**).

A second argument advanced against nosology is that it puts people in a pigeon-hole. I have never been able to understand this argument since whenever one uses *any* nomothetic language to characterize a human being one is, to that extent, putting him in a pigeon-hole (or locating him at a point in conceptual space); and, of course, every case of carcinoma of the liver is "unique" too. That some old-fashioned diagnosticians, untrained in psychodynamics, use diagnostic labels as a substitute for understanding the

patient is not an unknown occurrence, but what can one say in response to this except *abusus non tollit usum*? We cannot afford to decide about the merits of a conceptual scheme on the grounds that people use it wrongly.

A derivative of this argument is that diagnostic categories are not dynamics, and do not really tell us anything about what is wrong with the patient. There is some truth in this complaint, but again the same complaint could be advanced with regard to an organic disease concept at any stage in the development of the conception of it prior to the elucidation of its pathology and aetiology.

There is some confusion within our profession about the relation between content or dynamics and taxonomic categories. Many seem to think that when we elucidate the content, drives, and defences with which a patient is deeply involved, we have thereby explained why he is ill. But in what sense is this true? When we learn something about the inner life of a psychiatric patient, we find that he is concerned with aggression, sex, pride, dependence, and the like, that is, the familiar collection of human needs and fears. Schizophrenics are people, and if you are clever enough to find out what is going on inside a schizophrenic's head, you should not be surprised that these goings-on involve his self-image and his human relationships rather than, say, the weather. The demonstration that patients have psychodynamics, that they suffer with them, and that they deal with them ineffectively, does *not* necessarily tell us what is the matter with them, that is, why they are patients.

One is reminded in this connection of what happened when, after several years of clinicians busily over-interpreting "pathological" material in the TAT stories of schizophrenic patients. Dr. Leonard Eron took the pains to make a normative investigation and discovered that most of the features which had been so construed occurred equally or more often in a population of healthy college students (**10**).

There is no contradiction between classifying a patient as belonging to a certain taxonomic group and attempting concurrently to understand his motivations and his defences. Even if a certain major mental disease were found to be of organic or genetic origin, it would not be necessary to abandon any well-established psychodynamic interpretations. Let me give you an analogy. Suppose that there existed a colour-oriented culture in which a large part of social, economic, and sexual behaviour was dependent upon precise colour-discriminations. In such a culture, a child who makes errors in colour behaviour will be teased by his peer group, will be rejected by an over-anxious parent who cannot tolerate the idea of having produced an inferior or deviant child, and so on. One who was unfortunate enough to inherit the gene

for colour blindness might develop a colour neurosis. He might be found as an adult on the couch of a colour therapist, where he would produce a great deal of material which would be historically relevant and which would give us a picture of the particular pattern of his current colour dynamics. But none of this answers the question, "What is fundamentally the matter with these people?," that is, what do all such patients have in common? What they have in common, of course, is that defective gene on the X-chromosome; and this, while it does not provide a *sufficient* condition for a colour neurosis in such a culture, does provide the *necessary* condition. It is in this sense that a nosologist in that culture could legitimately argue that "colour neuroticism" is an inherited disease.

I think that none of these commonly heard objections is a scientifically valid reason for repudiating formal diagnosis, and that we must consider the value of the present diagnostic categories on their merits, on their relevance to the practical problems of clinical decision-making. One difficulty is that we do not have available for the validation of our instruments an analogue of the pathologist's report. It makes sense in organic medicine to say that the patient was actually suffering from disease X even though there was no evidence for it at the time of the clinical examination, so that the best clinician in the world could not have made a correct diagnosis on the data presented prior to autopsy. We have nothing in clinical psychology which bears close resemblance to the clinicopathological conference in organic medicine. Our closest analogue to pathology is "structure" and psychodynamics, and our closest analogue to the internist's concept of aetiology is a composite of constitution and learning history. If we had a satisfactory taxonomy of either constitution or learning history, we would be able to define what we meant by saying that a given patient is a schizophrenic. A well-established historical agent would suffice for this purpose, and Freud, for example, made an attempt at this in the early days (before he had realized how much of his patients' anamnesis was fantasy) by identifying the obsessional neurosis with a history of active and pleasurable erotic pre-pubescent activity, and hysteria with a history of passive and largely unpleasurable erotic experience (**12**).

Since anyone who takes formal diagnosis as a significant part of the psychologist's task must be thinking in terms of construct validity, (**1**, **6**), he should have at least a vague sketch of the structure and aetiology of the disorders about which he speaks diagnostically. I do not think that it is appropriate to ask for an operational definition. My own view is that theoretical constructs are defined "implicitly" by the entire network of hypothesized laws concerning them; in the early stages of understanding a taxonomic

concept, such as a disease, this network of laws is what we are trying to discover. Of course, when a clinician says, "I think this patient is really a latent schizophrenic," he should be able to give us *some* kind of picture of what he means by this statement. It could, however, be rather vague and still sufficient to justify itself at this stage of our knowledge. He might say:

> I mean that the patient has inherited an organic structural anomaly of the proprioceptive integration system of his brain, and also a radical deficiency in the central reinforcement centres (or, to use Rado's language, a deficiency in his "hedonic capacity"). The combination of these proprioceptive and hedonic defects leads in turn to developmental disturbances in the body image and in social identification; the result at the psychological level being a pervasive disturbance in the cognitive functions of the ego. It is this defective ego-organization that is responsible for the primary associative disturbance set forth as the fundamental symptom of schizophrenia by Bleuler. The other symptoms of this disease, which may or may not be present, I would conceive as Bleuler does, and therefore my conception of the disorder is perhaps wider than is modal for American clinicians. By "pseudoneurotic schizophrenia" I would mean a patient with schizophrenia whose failure to demonstrate the accessory symptoms (and whose lower quantitative amount of even the primary symptoms) leads to his being readily misdiagnosed. Pseudoneurotic schizophrenia is just schizophrenia that is likely to go unrecognized.

Such a sketch is, to my mind, sufficient to justify the use of the schizophrenia concept at the present state of our knowledge. It is not very tight, and it is not intellectually satisfying. On the other hand, when combined with the set of indicators provided by Bleuler (**3**), Hoch and Polatin (**21**), and others, it is not much worse than the concept of general paresis as understood during most of the nineteenth century following Bayle's description in 1822. In this connection it is sometimes therapeutic for psychologists to familiarize themselves with the logicians' contributions to the methodological problems of so called "open concepts," "open texture," and "vagueness" (**18**, **19**, **23**, **41**, **49**, **57**, **60**). Even a slight acquaintance with the history of the more advanced sciences gives one a more realistic perspective on the relation of "operational" indicators to theoretical constructs during the early stages of a construct's evolution. (See, for example, **39**, **45**, **46**, **56**.)

The formal nosological label makes a claim about an inner structure or state; therefore, the concurrent validity of a test against our psychiatrist as criterion is not an end in itself, but rather is one piece in the pattern of evidence which is relevant to establishing the *construct* validity of *both* the test and the psychiatrist. If I really accept the psychiatric diagnosis as *"the* criterion," what am I doing with my test anyway? If I want to know what the psychiatrist is going to call patient Jones whom he has just finished interviewing, the obvious way to find out is to leave my own little cubicle

with its Rorschach and Multiphasic materials and walk down the hall to ask the psychiatrist what he is going to call the patient. This is a ludicrous way of portraying the enterprise, but the only thing which saves it from really being this way is that implicitly we reject concurrent validity with the psychiatrist's diagnosis as criterion, having instead some kind of construct validity in the back of our minds. The phrase "the criterion" is misleading. Because of the whole network of association surrounding the term "criterion," I would myself prefer to abandon it in such contexts, substituting the term "indicator." The impact of a patient upon a psychiatrist (or upon anyone else, for that matter) is one of a *family of indicators of unknown relative weights*; when we carry out a "validation" study on a new test, we are asking whether or not the test belongs to this family.

Note that the uncertainty of the link between nosology and symptom (or test) is a two-way affair. Knowing the formal diagnosis we cannot infer with certainty the presence of a given symptom or the result of a given test; conversely, given the result on a test, or the presence of a certain symptom, we cannot infer with certainty the nosology. (There are rare exceptions to this, such as thought-disorder occurring in the presence of an unclouded sensorium and without agitation, which I would myself consider pathognomonic of schizophrenia.) This uncertainty is found also in organic medicine, where there are very few pathognomonic symptoms and very few diseases which invariably show any given symptom. An extreme (but not unusual) example is the prevalence of those sub-clinical infections which are responsible for immunizing us as adults, but which were *so* "sub"-clinical that they were only manifested by a mild malaise and possibly a little fever, symptoms which, singly or jointly, do not enable us to identify one among literally hundreds of diagnostic possibilities.

One "statistical" advantage contributed by a taxonomy even when it is operating wholly at the descriptive or syndrome level is so obvious that it is easy to miss; I suspect that the viability of the traditional nosological rubrics, which could not be well defended upon aetiological grounds at present, is largely due to this contribution. When the indicators of membership in the class comprise a long list, none of which is either necessary or sufficient for the class membership, the descriptive information which is conveyed by the taxonomic name has a "statistical-disjunctive" character. That is, when we say that a patient belongs to category X, we are at least claiming that he displays indicators *a* or *b* or *c* with probability *p* (and separate probabilities $p_a$, $p_b$, and $p_c$). This may not seem very valuable, but considering how long it would take to convey to a second clinician the entire list of behaviour dispositions whose probability of being present is materially altered by

placing a patient in category X, we see that from the standpoint of sheer economy even a moderately good taxonomic system does something for us. More important in the long run is the fact that only a huge clinical team, with a tremendous amount of money to spend on a large number of patients over a long period of time, could hope to discover and confirm all $\frac{N(N-1)}{2}$ of the pair-wise correlations among the family of $N$ indicators that relate to the concept, to say nothing of the higher-order configural effects (**22**) that will arise in any such material. The research literature can yield cumulative knowledge and improvement of clinical practice in different settings by virtue of the fact that in one hospital an investigator, working with limited means, is able to show that patients diagnosed as schizophrenic tend to perform in a special way on a proverbs test; while another investigator in another hospital is showing that male patients diagnosed as schizophrenic have a high probability of reacting adversely to sexually attractive female therapists. Imagine a set of one hundred indicator variables and one hundred output variables; we would have to deal with ten thousand pair-wise correlations if we were to study these in one grand research project. The advantages in communicative economy and in cumulating research knowledge cannot, of course, be provided by a descriptive taxonomy which lacks intrinsic merit (that is, the syndrome does not objectively exist with even a moderate degree of internal tightness), or which, while intrinsically meritorious, is applied in an unskillful manner.

Let us turn now to our second main use of tests—prognosis. Sometimes the forecasting of future behaviour is valuable even if no special treatment is contemplated, because part of the responsibility of many clinical installations is to advise other agencies or persons, such as a court, as to the probabilities. But the main purpose of predictive statements is the assistance they give us in making decisions about how to treat a patient. Predictive statements of the form "If you treat the patient so-and-so, the odds are 8:2 that such-and-such will happen," will be with us for a very long time. As more knowledge about behavioural disorders is accumulated, we can expect a progressive refinement and differentiation of techniques; their differential impact will thereupon become greater, so that the seriousness of a mistake will be correspondingly increased. Furthermore, even if—as I consider highly unlikely but as we know some therapists are betting—it is discovered that for all patients the same kind of treatment is optimal, it is easily demonstrated from the statistics of mental illness, together with the most sanguine predictions as to the training of skilled professional personnel, that there will not be adequate staff to provide even moderately intensive treatment for any but a minority of

patients during the professional lifetime of anybody at present alive. So we can say with confidence that the decision to treat or not to treat will be a decision which clinicians are still going to be making when all of us have retired from the scene. As I read the published evidence, our forecasting abilities with current tests are not what you could call distinguished (see, for example, **61**).

In connection with this problem of prognosis, let me hark back a moment to our discussion of formal nosology. One repeatedly hears clinicians state that they make prognostic decisions, not on the basis of a formal diagnosis but on their assessment of the individual's structure and dynamics. Where is the evidence that we can do this? So far as I am aware there is as much evidence indicating that one can predict the subsequent course of an illness from diagnostic categories (**16**) (or from crude life-history statistics) as there is that one can predict the course of an illness or the response to therapy from any of the psychological tests available. I should like to offer a challenge to any clinician who thinks that he can cite a consistent body of published evidence to the contrary.

In order to employ dynamic constructs to arrive at predictions, it would be necessary to meet two conditions. In the first place, we must have a sound theory about the determinative variables. Secondly, we must be in possession of an adequate technology for making measurements of those variables. As any undergraduate major in physics or chemistry knows, in order to predict the subsequent course of a physical system, it is necessary both to understand the laws which the system obeys and to have an accurate knowledge of the initial and boundary conditions of the system. Since clinical psychology is nowhere near meeting *either* of these two requirements, it must necessarily be poor at making predictions which are mediated by dynamic constructs. It is a dogma of our profession that we predict what people will do by understanding them individually, and this sounds so plausible and humanitarian that to be critical of it is like criticizing Mothers' Day. I can only reiterate that neither theoretical considerations nor the data available in the literature lend strong support to this idea in practice.

Let us turn to the third clinical task which the psychologist attempts to solve by the use of his tests, that of "personality assessment." Phenotypic characterization of a person includes the attribution of the ordinary clinical terms involving a minimal amount of inference, such as "patient hallucinates" or "patient has obsessional trends"; trait names from common English, such as the adjectives found in the lists published by Cattell (**5**, p. 219) or Gough (**14**); and, increasingly important in current research, characterizations in the form of a single sentence or a short paragraph of the type employed by

Stephenson (**53**), the Chicago Counseling Center (**44**), Block (**4**), and others. (Example: "The patient characteristically tries to stretch limits and see how much he can get away with.") A logical analysis of the nature of these phenotypic trait attributions is a formidable task although a very fascinating one. I am not entirely satisfied with any account which I have seen, or have been able to devise for myself. Perhaps not too much violence is done to the truth if we say that these are all in the nature of dispositional statements, the evidence for which consists of some kind of sampling, usually not representative, of a large and vaguely specified domain of episodes from the narrative that constitutes a person's life. It is complicated by the fact that even if we attempt to stay away from theoretical inferences, almost any single episode is susceptible of multiple classification under different families of atomic dispositions constituting a descriptive trait. The fact that the evidence for a trait attribution represents only a sample of the concrete episodes that exemplify atomic dispositions introduces an inferential element into such trait attributions, even though the trait name is intended to perform a purely summarizing rather than a theoretical function (**6**, pp. 292-3).

Phenotypic characterization presents a special problem which differentiates it from the functions of diagnosis and prognosis in the establishment of validity. Since it involves concurrent validity, its pragmatic justification is rather more obscure. Suppose we have a descriptive trait, say, "uncooperative with hospital personnel," an item which is not uncommon in various rating scales and clinical Q-pools in current use in the United States. Why administer an MMPI in order to guess, with imperfect confidence, whether or not the patient is being currently judged as uncooperative by the occupational therapist, the nursing supervisor, and the resident in charge of his case? This is even a more fruitless activity than our earlier example of using a test to guess the diagnosis given by the psychiatrist. From the theoretical point of view, the obvious reply is that the sampling of the domain of the patient's dispositions which is made by these staff members is likely to be deficient, both in regard to its *qualitative* diversity and representativeness as seen within the several contexts in which they interact with the patient, and *quantitatively* (simply from the statistical standpoint of size) during the initial portion of a patient's stay in the hospital. This reply leads to a suggestion concerning the design of studies which are concerned with phenotypic assessment from tests. Such designs should provide a "criterion" which is considerably superior in reliability to that which would routinely be available in the clinic on the basis of the ordinary contacts. If it is concurrent validity in which we are really interested (upon closer examination this often turns out

not to be the case), there is little point in administering a time-consuming test and applying the brains of a trained psychologist in order to predict the verbal behaviour of the psychiatric aid or the nurse. If it is our intention to develop and validate an instrument which will order or classify patients as to phenotypic features which are *not* reliably assessed by these persons in their ordinary contacts with the patient, then we need a design which will enable us to show that we have actually achieved this result.

As to the power of our tests in the phenotypic characterization of an individual, the available evidence is not very impressive when we put the practical question in terms of the *increment in valid and semantically clear information transmitted.* (See, for example, the studies by Kostlan (**25**), Dailey (**8**), Winch and More (**58**), Kelly and Fiske (**24**), Davenport (**9**), Sines (**51**), and Soskin (**52**).)

The question of concurrent validity in the phenotypic domain can be put at any one of four levels, in order of increasing practical importance. It is surprising to find that research on concurrent validity has been confined almost wholly to the first of these four levels. The weakest form of the validation question is, "How accurate are the semantically clear statements which can be reliably derived from the test?" It is a remarkable social phenomenon that we still do not know the answer to this question with respect to the most widely used clinical instruments. I do not see how anyone who examines his own clinical practice critically and who is acquainted with the research data could fail to make at least the admission that the power of our current techniques is seriously in doubt.

A somewhat more demanding question, which incorporates the preceding, would be: "To what extent does the test enable us to make, reliably, accurate statements which we cannot *concurrently* and *readily* (that is, at low effort and cost) obtain from clinical personnel routinely observing the patient *who will normally be doing so anyway* (that is, whose observations and judgments we will not administratively eliminate by the introduction of the test)?" In the preceding discussion regarding diagnosis and concurrent validity I oversimplified so grossly as to be a bit misleading. "How the staff rates" cannot be equated with "What the staff sees," which cannot in turn be equated with "What the patient does in the clinic"; and that, in turn, is not the equivalent of "What the patient does." If a patient beats his wife and does not tell his therapist about it, and the wife does not tell the social worker, the behaviour domain has been incompletely sampled by those making the ratings; they might *conclude* that he had beaten his wife, and this conclusion, while it is an inference, is still a conclusion regarding the

phenotype. We cannot, of course, classify a certain concept as "theoretical" merely on the grounds that we have to make an inference in order to decide about a concrete instance of its application. This is a sampling problem, and therefore mainly (although not wholly) a matter of the time required to accumulate a sufficiently extensive sample. On the other hand, in our sampling of the patient's behavioural dispositions in the usual clinical context, it is not wholly a numerical deficiency in accumulation of episodes, because the sample which we obtain arises from a population of episodes that is in itself systematically biased. That is, the population of episodes which can be expected to come to our attention in the long run is itself a non-representative sub-population of all the behavioural events which constitute the complete narration of the patient's life.

A very stimulating paper is that of Kostlan (**25**). There are elements of artificiality in his procedure (of which he is fully aware) and these elements will no doubt be stressed by those clinicians who are determined to resist the introduction of adverse evidence. Nevertheless, his procedure was an ingenious compromise between the necessity of maintaining a close semblance to the actual clinical process, and a determination to quantify the incremental validity of tests. What he did, in a word, was to begin with a battery of data such as were routinely available in his own clinical setting and with which his clinicians were thoroughly familiar, consisting of a Rorschach, an MMPI, a sentence completion test, and a social case history. He then systematically varied the information available to his clinicians by eliminating one of these four sources at a time, arguing that the power of a device is probably studied better by showing the effect of its *subtraction* from the total mass of information than by studying it alone. The clinicians were required to make a judgment, from the sets of data presented to them, on each of 283 items which had been culled from a population of 1,000 statements found in the psychological reports written by this staff. The most striking finding was that on the basis of all three of these widely used tests his clinicians could make no more accurate inferences than they could make utilizing the Barnum effect (**35**, **8**, **11**, **52**, **54**, **55**) when the all-important social history was deleted from their pool of data. A further fact, not stressed by Kostlan in his published report (but see **25** and **26**), is that the absolute magnitude of incremental information, even when the results are statistically significant, is not impressive. For example, clinicians knowing only the age, marital status, occupation, education, and source of referral of a patient (that is, relying essentially upon Barnum effect for their ability to make correct statements) yield an average of about 63 per cent correct statements about the patient. If they have the Rorschach, Multiphasic, and Sentence Completion

tests *but are deprived of the social case history,* this combined psychometric battery results in almost exactly the same percentage of correct judgments. On the other hand, if we consider their success in making inferences based on the social history together with the Sentence Completion test and the MMPI (that is, eliminating only the Rorschach, which made no contribution) we find them making 72 per cent correct inferences (my calculations from his Table 3), that is, a mere 9 per cent increment.

A thesis just completed at the University of Minnesota by Dr. Lloyd K. Sines is consistent with Kostlan's findings (**51**). Taking a Q-sort of the patient's therapist as his criterion. Sines investigated the contribution by a four-page biographical sheet, an MMPI profile, a Rorschach (administered by the clinician making the test-based judgments), and a diagnostic interview by this clinician. He determined the increment in Q-correlation with the criterion (therapist sort) when each of these four sources of information was inserted at different places in the sequence of progressively added information. The contribution of either of the two psychological tests, or both jointly, was small (and, in fact, knowledge of the Rorschach tended to exert an adverse effect upon the clinician's accuracy). For some patients, the application of a stereotype personality description based upon actuarial experience in this particular clinic provided a more accurate description of the patient than the clinician's judgment based upon any, or all, of the available tests, history, and interview data!

A third level of validation demand, in which we become really tough on ourselves, takes the form: "If there are kinds of clear non-trivial statements which can be reliably derived from the test, which are accurate, and which are not concurrently and readily obtainable by other means routinely available, *how much earlier in time* does the test enable us to make them?" It might be the case that we can make accurate statements from our tests at a time in the assessment sequence when equally trustworthy non-psychometric data have not accumulated sufficiently to make such judgments, but from the practical point of view there is still a need to know just how "advanced" this advance information is. So far as I know, there are no published investigations which deal with this question.

A final and most demanding way of putting the question, which is ultimately the practically significant one by which the contribution of our techniques must be judged, is the following: "If the test enables us to make reliably, clear, differentiating statements which are accurate and which we cannot readily make from routinely available clinical bases of judgment; and if this additional information is not rapidly picked up from other sources

during the course of continued clinical study of the patient; in what way, *and to what extent,* does this incremental advance information help us in treating the patient?" One might have a clear-cut positive answer to the first three questions and be seriously in error if he concluded therefrom that his tests were paying off in practice. On this fourth question, there is also no published empirical evidence.

In the absence of any data I would like to speculate briefly on this one. Suppose that a decision is made to undertake the intensive psychotherapy of a patient. A set of statements, either of a dichotomous variety or involving some kind of intensity dimension or probability-of-correctness, is available to the psychotherapist on the basis of psychological test results. How does the therapist make use of this knowledge? It is well known that competent therapists disagree markedly with regard to this matter, and plausible arguments on both sides have been presented. Presumably the value of such information will depend upon the kind of psychotherapy which is being practised; therapists of the Rogerian persuasion are inclined to believe that this kind of advanced knowledge is of no use; in fact they prefer to avoid exposure to it. Even in a more cognitively oriented or interpretative type of treatment, it may be argued that by the time the therapeutic interaction has brought forth sufficient material for interpretation and working-through to be of benefit to the patient, the amount of evidential support for a construction will be vastly greater than the therapist could reasonably expect to get from a psychological test report. It does not help the patient that there is "truth" regarding him in the therapist's head; since there is going to be a lot of time spent before the patient comes around to seeing it himself, and since this time will have to be spent regardless of what the therapist knows, perhaps there is no advantage in his knowing something by the second interview rather than by the seventh. On the other side, it may be argued that any type of therapy which involves even a moderate amount of selective attention and probing by the therapist does present moment-to-moment decision problems (for example, how hard to press, when to conclude that something is a blind alley, what leads to pick-up) so that advance information from psychometrics can set the therapist's switches and decrease the probability of making mistakes or wasting time. It seems to me that the armchair arguments pro and con in this respect are pretty evenly balanced, and we must await the outcome of empirical studies.

One rather disconcerting finding which I have recently come upon is the rapidity with which psychotherapists arrive at a stable perception of the patient which does not undergo much change as a result of subsequent

contacts. I was interested in this matter of how early in the game the psychological test results enable us to say what the therapist *will be saying later on.* In our current research at Minnesota we are employing a Q-pool of 183 essentially "phenotypic" items drawn from a variety of sources. We are also using a "genotypic" pool of 113 items which consists of such material as the Murray needs, the major defence mechanisms, and various other kinds of structural-dynamic content. I was hoping to show that as the therapist leams more and more about his patient, his Q-correlation with the Q-description of the patient based upon blind analysis of the MMPI profile would steadily rise; furthermore, it is of interest to know whether there are **sub**-domains of this pool, such as mild and well-concealed paranoid trends, with respect to which the MMPI is highly sensitive early in the game. (From my own therapeutic work, I have the impression that a low Pa score has almost no value as an exclusion test, but that any patient, however non-psychotic he may be, who has a marked *elevation* on this scale will, sooner or later, present me with dramatic corroborating evidence.) However, I can see already that I have presented the test with an extraordinarily difficult task, because the Q-sorts of these therapists stabilize so rapidly. The therapists Q-described their patients after the first therapeutic hour, again after the second, then after the fourth, eighth, sixteenth, and twenty-fourth contact. If one plots the Q-correlation between each sorting and the sorting after twenty-four hours of treatment (or between each sorting and a pooled sorting; or between each sorting and the next successive sorting), one finds that by the end of the second or fourth hour, the coefficients with subsequent hours are pushing the sort-resort reliabilities. The convergence of the therapist's perception of his patient is somewhat faster in the phenotypic than in the genotypic pool, but even in the latter his conception of the patient's underlying structure, defence mechanisms, need-variable pattern, and so on seems to crystallize very rapidly. Even before examining the MMPI side of my data, I can say with considerable assurance that it will be impossible for the test to "prove" itself by getting ahead, and staying ahead, of the therapist to a significant extent. Of course, we are here accepting the psychotherapist's assessment as one which does converge to the objective truth about the patient in the long run, and this may not be true for all sub-domains of the Q-pool. The extent to which this rapid convergence to a stable perception represents invalid premature "freezing" is unknown (but see **7**).

Personality characterization at the genotypic level will undoubtedly prove to be the most difficult test function to evaluate. A genotypic formulation, even when it is relatively inexplicit, seems to provide a kind of background which sets the therapist's switches as he listens to the patient's

discourse. What things he will be alert to notice, how he will construe them, what he will say and when, and even the manner in which he says it, are all presumably influenced by this complicated and partly unconscious set of perceptions and expectancies. Process research in psychotherapy is as yet in such a primitive state that one hardly knows even how to begin thinking about experiments which would inform us as to the pragmatic payoff of having advanced information, at various degrees of confidence, regarding specific features of the genotype. Even if it can be demonstrated that the therapist's perception of the patient tends with time to converge to that provided in advance by the test findings, this will never be more than a statistical convergence; therefore, in exchange for correctly raising the probability that one sub-set of statements is true of the patient, we will always be paying the price of expecting confirmation of some other unspecified sub-set which is erroneous.

Let me illustrate the problem by a grossly oversimplified example. Suppose that prior to either testing or interviewing, a dichotomously treated attribute has a base-rate probability of .60 in our particular clinic population. Suppose further that it requires an average of five therapeutic interviews before the therapist can reach a confidence of .80 with regard to the presence of this attribute. Suppose finally that a test battery yields this same confidence at the conclusion of diagnostic study (that is, before the therapy begins). During the five intervening hours, the therapist is presumably fluctuating in his assessment of this attribute between these two probability values, and his interview behaviour (as well as his inner cognitive processes) are being influenced by his knowledge of the test results. Perhaps because of this setting of his switches he is able to achieve a confidence around the .8 mark by the end of the fourth session, that is, two hours earlier than he would have been able to do without the test. Meanwhile, he has been concurrently proceeding in the same way with respect to a second attribute; but, unknown to him, in the present case the test is giving him misinformation about that attribute (which will happen in one patient out of five on our assumptions). It is impossible to say from our knowledge of the cognitive processes of interpretive psychotherapists, or from what we know of the impact of the therapeutic interaction upon the patient, whether a net gain in the efficacy of treatment will have been achieved thereby. The difficulties in unscrambling these intricate chains of cumulative, divergent (**29**), and interactive causation are enormous.

I suspect that the present status of process research in psychotherapy does not make this type of investigation feasible. Alternatively, we shift to "outcome" research. Abandoning an effort to understand the fine causal

details of the interaction between patient and therapist, we confine ourselves to the crude question, "Are the outcomes of psychotherapy influenced favourably, on the average, by making advance information from a psychometric assessment available to the therapist?" Granting the variability of patients and therapists, and the likely interaction between these two factors and the chosen therapeutic mode, it seems feasible to carry out factorial-design research in which this question might be answered with some degree of assurance. When so much of the clinical psychologist's time is expended in the effort to arrive at a psychodynamic formulation of the patient through the integration of psychological test data, to the point that in some out-patient settings the total number of hours spent on this activity is approximately equal to the median number of hours of subsequent therapeutic contact, I believe that we should undertake research of this kind without delay.

Whatever the future may bring with regard to the pragmatic utility of the genotypic information provided by psychometrics, I am inclined to agree with Jane Loevinger's view that tests should be constructed in a framework of a well-confirmed psychological theory and with attention devoted primarily to construct validity. In her recent monograph (**28**), Dr. Loevinger has suggested that it is inconsistent to lay stress on construct validity and meanwhile adopt the "blind, empirical, fact-to-fact" orientation I have expressed (**35**, **36**). I do not feel that the cookbook approach is as incompatible with a dedication to long-term research aimed at construct validity as Dr. Loevinger believes. The future use of psychological tests, if they are to become more powerful than they are at present, demands, as Loevinger points out, cross-situational power. It would be economically wasteful to have clinicians in each of the hundreds of private and public clinical facilities deriving equations, actuarial tables, or descriptive cookbooks upon each of the various clinical populations. I would also agree with Loevinger that such cross-situational power is intimately tied to construct validity, and that the construction of a useful cookbook does not, in general, contribute appreciably to the development of a powerful theoretical science of chemistry.

On the other hand, there is room for legitimate disagreement, among those who share this basic construct-validity orientation, on an important interim question. If the development of construct-valid instruments which will perform with a high degree of invariance over different clinical populations hinges upon the elaboration of an adequate psychological theory concerning the domain of behaviour to be measured, then the rate of development of such instruments has a limit set upon it by the rate of development of our psychodynamic understanding. I personally am not impressed with the

state of psychological theory in the personality domain, and I do not expect the edifice of personality constructs to be a very imposing one for a long time yet. Meanwhile, clinical time is being expended in the attempt to characterize patients by methods which make an inefficient use of even that modest amount of valid information with which our present psychometric techniques provide us.

The number of distinct attributes commonly viewed by clinicians as worth assessing is actually rather limited. The total number of distinguishable decision problems with which the psychiatric team is routinely confronted is remarkably small (see, for example, **8**). It is not possible to say, upon present evidence, what are the practical limits upon the validity generalization of configural mathematical functions set up on large samples with respect to these decision classes. It is possible that the general *form* of such configural functions, and even the parameters, can be generalized over rather wide families of clinical populations, with each clinical administrator making correction of cutting scores or reassigning probabilities in the light of his local base-rates (**37**). One could tolerate a considerable amount of shrinkage in validity upon moving to a similar but non-identical clinical population without bringing the efficiency of an empirical cookbook down to the low level of efficiency manifested by clinicians who are attempting to arrive at such decisions on an impressionistic basis from the same body of psychometric and life history evidence. Halbower, for instance, showed that moving from an out-patient to an in-patient veteran population, while it resulted in considerable loss in the descriptive power of a cookbook based upon MMPI profile patterns, nevertheless maintained a statistically significant (and a practically important) edge over the Multiphasic reading powers even of clinicians who were working with the kind of population to which validity was being generalized (**15**). One of the things we ought to be trying is the joint utilization, in one function or table, of the most predictive kinds of life history data *together with* our tests. Some of the shrinkage in transition to allied but different clinical populations might be taken care of by the inclusion of a few rather simple and objective facts about the patient such as age, education, social class, referral source, percentage of service-connected disability, and the like.

Hence, I agree with Dr. Loevinger's emphasis upon the long-term importance of constructing tests which will be conceptually embedded in the network of psychological theory, and therefore superior in cross-situational power; in the meantime we do not have such tests, and there is some reason to think that in making daily clinical decisions a standard set of decision problems and trait attributions can be constructed. Such empirical research

(readily within present limitations of personnel and theory) could result in the near future in cookbook methods which would include approximate stipulations as to those parametric modifications necessary for the main classes of clinical populations and for base rates, whether known or crudely estimated, in any given installation. I do not see anything statistically unfeasible about this, and I shall therefore continue to press for a serious prosecution of this line until somebody presents me with more convincing evidence than I have thus far seen that the clinical judge, or the team meeting, or the whole staff conference, is able somehow to surmount the limitations imposed by the inefficiency of the human mind in combining multiple variables in complex ways.

As for the long-term goal of developing construct-valid tests, maybe our ideas about the necessary research are insufficiently grandiose. Perhaps the kind of integrated psychometric-and-theory network which is being sought is not likely to be built up by the accumulation of a large number of minor studies. If we were trying to make a structured test scale, for instance, which would assess those aspects of a patient's phenomenology that are indicators of a fundamentally schizadaptive makeup, we would be carrying on an uphill fight against nature if we accepted as our criterion the rating of a second-year psychiatry resident on a seven-step "latent schizophrenia" variable! I would not myself be tempted to undertake the construction of an MMPI key for latent schizophrenic tendency unless I had the assurance that the classification or ordering of the patient population would be based upon a multiple attack taking account of all of the lines of evidence which would bear upon such an assessment in the light of my crude theory of the disease. *The desirability of a "criterion" considerably superior to what is routinely available clinically applies to the development of construct-valid genotypic measures even more than to criterion-oriented contexts.* Between such a hypothetical inner variable or state as "schizophrenic disposition," and almost any namable aspect of overt behaviour, there is interpolated quite a collection of nuisance variables. In order to come to a decision regarding, for example, a certain sub-set of cases which are apparently "test misses" (or which throw sub-sets of items in the wrong direction and hence provide evidence that those items should be modified or eliminated) one has to have a sufficiently good assessment of the relevant nuisance variables to satisfy himself that the apparent test or item miss is a miss in actuality.

This brings me to what I have often thought of as the curse of clinical psychology as a scientific enterprise. There are some kinds of psychological test construction or validation in which it suffices to know a very little bit

about each person, provided a large number of persons are involved (for example, in certain types of industrial, educational, or military screening contexts). At the other extreme, one thinks of the work of Freud, in which the most important process was the learning of a very great deal about a small number of individuals. When we come to the construction and validation of tests where, as is likely always to be true in clinical work, higher-order configurations of multi-variable instruments are involved, we need to know a great deal about each individual in order to come to a conclusion about what the test or item should show regarding his genotype. However, in order to get statistical stability for our weights and to establish the reality of complex patterning trends suggested by our data, we need to have a sizable sample of individuals under study. So that where some kinds of psychological work require us to know only a little bit about a large number of persons, and other kinds of work require us to know a very great deal about a few persons, construct validation of tests of the sort that Loevinger is talking about will probably require that we know a great deal, and at a fairly intensive or "dynamic" level, about a large number of persons. You will note that this is not a reflection of some defect of our methods or lack of zeal in their application but arises, so to speak, from the nature of things. I do not myself see any easy solution to this problem.

I am sure that by now you are convinced of the complete appropriateness of my title. I am aware that the over-all tenor of my remarks could be described as somewhat on the discouraged side. But we believe in psychotherapy that one of the phases through which most patients have to pass is the painful one between the working through of pathogenic defences and the reconstitution of the self-image upon a more insightful basis. The clinical psychologist should remind himself that medical diagnostic techniques frequently have only a modest degree of reliability and validity. I have, for instance, recently read a paper written by three nationally known roentgenologists on the descriptive classification of pulmonary shadows, which these authors subtitle "A Revelation of Unreliability in the Roentgenographic Diagnosis of Tuberculosis" (**40**). I must say that my morale was improved after reading this article.

In an effort to conclude these ruminations on a more encouraging note, let me try to pull together some positive suggestions. Briefly and dogmatically stated, my constructive proposals would include the following:

1. Rather than decrying nosology, we should become clinical masters of it, recognizing that some of our psychiatric colleagues have in recent times become careless and even unskilled in the art of formal diagnosis.

2. The quantitative methods of the psychologist should be applied to the refinement of taxonomy and not confined to data arising from psychological tests. (I would see the work of Wittenborn (**59**) and of Lorr and his associates (**30**) as notable beginnings in this direction.)

3. While its historical development typically begins with syndrome description, the reality of a diagnostic concept lies in its correspondence to an inner state, of which the symptoms or test scores are fallible indicators. Therefore, the validation of tests as diagnostic tools involves the psychiatrist's diagnosis merely as one of an indicator family, not as a "criterion" in the concurrent validity sense. Accumulation of numerous concurrent validity studies with inexplicably variable hit-rates is a waste of research time.

4. Multiple indicators, gathered under optimal conditions and treated by configural methods, must be utilized before one can decide whether to treat inter-observer disagreement as showing the unreality of a taxonomy or merely as diagnostic error.

5. We must free ourselves from the almost universal assumption that when we elucidate the motives and defences of a psychiatric patient, we have thereby explained why he has fallen ill. As training analysts have observed for years, patients and "normals" tend to have pretty much the same things on their minds, conscious and unconscious.

6. The relative power, for prognosis and treatment selection, of formal diagnosis, non-nosological taxonomies based upon trait clusters, objective life-history factors, and dynamic understanding via tests, is an empirical question in need of study, rather than a closed issue. We must face honestly the disparity between current clinical practice and what the research evidence shows about the relatively feeble predictive power of present testing methods.

7. There is some reason to believe that quantitative treatment of life-history data may be as predictive as psychometrics in their present state of development. Research along these lines should be vigorously prosecuted.

8. It is also possible that interview-based judgments at a minimally inferential level, if recorded in standard form (for example, Q-sort) and treated statistically, can be made more powerful than such data treated impressionistically as is currently the practice.

9. While maximum generalizability over populations hinges upon high construct validity in which the test's functioning is imbedded in the network of personality theory, there is a pressing interim need for empirically derived rules for making clinical decisions (that is, "clinical cookbooks"). Research is needed to determine the extent to which such cookbooks are tied to specific clinic populations and how the recipes can be adjusted in moving from one population to another.

10. Perhaps there are mathematical models, more suitable than the factor-analytic one and its derivatives, for making genotypic inferences, and especially inferences to nosology. Investigation of such possibilities must be pursued by psychologists who possess a thorough familiarity with the intellectual traditions of medical thinking, a solid grasp of psychodynamics, and enough mathematical skill to take creative steps along these lines.

11. From the viewpoint of both patients' welfare and taxpayers' economics, the most pressing *immediate* clinical research problem is that of determining the incremental information provided by currently used tests, especially those which consume the time of highly skilled personnel. We need not merely validity, but incremental validity; further, the temporal factor, "Does the test tell us something we are not likely to learn fairly early in the course of treatment?" should be investigated; finally, it is well within the capacity of available research methods and clinical facilities to determine what, if any, is the pragmatic advantage of a personality assessment being known in advance by the therapist.

12. In pursuing these investigations we might better avoid too much advertising of the results since neither psychiatrists nor government officials are in the habit of evaluating the efficiency of their own procedures, a fact which puts psychologists at a great propaganda disadvantage while the science is still in a primitive stage of development.

## REFERENCES


**1.** APA COMMITTEE ON Test STANDARDS. Technical recommendations for psychological tests and diagnostic techniques. *Psychol. Bull. Suppl.,* 1954, *51*, 2, Part 2, 1-38.

**2.** ASH, P. The reliability of psychiatric diagnosis. *J. abnorm. soc. Psychol.,* 1949, *44*, 272-278.

**3.** BLEULER, E. *Dementia praecox.* New York: International Univer. Press, 1950.

**4.** BLOCK, J., & BAILEY, D. *Q-sort item analyses of a number of MMP1 scales.* Technical Memorandum OERL-TM-55-7 Officer Education Research Laboratory. Air Force Personnel and Training Research Center, Air Research and Development Command, Maxwell Air Farce Base, Alabama, 1955.

**5.** CATTELL, R. B. *Description and measurement of personality.* New York: World Book Company, 1946.

**6.** CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.,* 1955, *52*, 281-302.

**7.** DAILEY, C. A. The effect of premature conclusion upon the acquisition of understanding a person. *J. Psychol,* 1952, *33*, 133-152.

**8.** DAILEY, C. A. The practical utility of the clinical report. *J. consult. Psychol,* 1953, *17*, 297-302.

**9.** DAVENPORT, BEVERLY F. The semantic validity of TAT interpretations. *J. consult. Psychol,* 1952, *16*, 171-175.

**10.** ERON, L. D. Frequencies of themes and identifications in the stories of schizophrenic patients and non-hospitalized college students. *J. consult. Psychol.,* 1948, *12*, 387-395.

**11.** FORER, B. R. The fallacy of personal validation: A classroom demonstration of gullibility. *J. abnorm. soc. Psychol.,* 1949, *44*, 118-123.

**12.** FREUD, S. Further remarks on the defense neuro-psychoses. *Collected papers*, I, 155-182. London: Hogarth Press, 1948.

**13.** GARFIELD, S. *Introductory clinical psychology.* New York: Macmillan, 1957.

**14.** GOUGH, H. G., MCKEE, M. G., & YANDELL, R. J. *Adjective check list analyses of a number of selected psychometric and assessment variables.* Institute of Personality Assessment and Research. Berkeley: Univer. California, 1953.

**15.** HALBOWER, C. C. A comparison of actuarial versus clinical prediction to classes discriminated by MMPI. Unpublished Ph.D. thesis, Univer. Minnesota, 1955.

**16.** HASTINGS, D. W. Follow-up results in psychiatric illness. *Amer. J. Psychiat.,* 1958, *114*, 1057-1066.

**17.** HATHAWAY, S. R. A study of human behavior: the clinical psychologist. *Amer. Psychologist,* 1958, *13*, 257-265.

**18.** HEMPEL, C. G. Problems and changes in the empiricist criterion of meaning. *Revue internat, philosophie,* 1950, *4*, 41-63.

**19.** HEMPEL, C. G. Fundamentals of concept formation in empirical science. *International encyclopedia of unified science,* II, no. 7. Chicago: Univer. Chicago Press, 1952.

**20.** HEMPEL, C. G. A logical appraisal of operationism. *Scientific Mon.,* 1954, *79*, 215-220.

**21.** HOCH, P. & POLATIN. Pseudoneurotic forms of schizophrenia. *Psychiat. Quart.,* 1949, *23*, 248-276.

**22.** HORST, P. Pattern analysis and configural scoring. *J. clin. Psychol.,* 1954, *10*, 3-11**.**

**23.** KAPLAN, A. Definition and specification of meaning. *J. Philosoph.,* 1946, *43*, 281-288.

**24.** KELLY, E. L. A FISKE, D. W. *The prediction of performance* in *clinical psychology.* Ann Arbor, Mich.: Univer. Michigan Press, 1951.

**25.** KOSTLAN, A. A method for the empirical study of psychodiagnosis. *J. consult. Psychol.,* 1954, *18*, 83-88.

**26.** KOSTLAN, A. A reply to Patterson. *J. consult. Psychol.,* 1955, *19*, 486.

**27.** KRAEPELIN, E. *General paresis* (Trans. J. W. MOORE). New York: Nervous and

Mental Disease Publishing Co., 1913.

**28.** LOEVINGER, JANE. Objective tests as instruments of psychological theory. *Psychol. Reports, Monogr. Suppl. 9,* 1957, *3*, 635-694.

**29.** LONDON, I. D. Some consequences for history and psychology of Langmuir's concept of convergence and divergence of phenomena. *Psychol. Rev.,* 1946, *53*, 170-188.

**30.** LORR, M. & RUBINSTEIN, E. A. Factors descriptive of psychiatric outpatients. *J. abnorm. soc. Psychol*., 1955, *51*, 514-522.

**31.** LYKKEN, D. T. A study of anxiety in the sociopathic personality. *J. abnorm. soc. Psychol.,* 1957, *55*, 6-10.

**32.** MACCORQUODALE, K. & MEEHL, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.,* 1948, *55*, 95-107.

**33.** MAJOR, R. H. *Classic descriptions of disease.* Springfield, Ill.: Charles C. Thomas, 1932.

**34.** MASSERMAN, J. H. & CARMICHAEL, H. T. Diagnosis and prognosis in psychiatry with a follow-up study of the results of short-term general hospital therapy of psychiatric cases. *J. ment. Sci*., 1939, *84*, 893-946.

**35.** MEEHL, P. E. Wanted—a good cookbook. *Amer. Psychologist,* 1956, *11*, 263-272.

**36.** MEEHL, P. E. When should we use our heads instead of the formula? *J. consult. Psychol.,* 1957, *4*, 268-273.

**37.** MEEHL, P. E. & ROSEN, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.,* 1955, *52*, 194-216.

**38.** MEHLMAN, B. The reliability of psychiatric diagnosis. *J. abnorm. soc. Psychol.,* 1952, *47*, 577-578.

**39.** NASH, L. K. *The atomic-molecular theory.* Cambridge: Harvard Univer. Press, 1950.

**40.** NEWELL, R. R., CHAMBERLAIN, W. E**.,** & RIGLER, C. Descriptive classification of pulmonary shadows: a revelation of unreliability. *Amer. Rev. Tuberculosis,* 1954, *69*, 566-584.

**41.** PAP, A. Reduction sentences and open concepts. *Methodas,* 1953, *5*, 3-30.

**42.** PATTERSON, C. H. Diagnostic accuracy or diagnostic stereotype? *J. consult. Psychol.,* 1955, *19*, 483-485.

**43.** PEPPARD, T. A. Mistakes in diagnosis. *Minnesota Med.,* 1949, *32*, 510-11.

**44.** ROGERS, C. R. & DYMOND, R. F. *Psychotherapy and personality change.* Chicago: Univer. Chicago Press, 1954.

**45.** ROLLER, D. E. *The development of the concept of electric charge.* Cambridge: Harvard Univer. Press, 1954.

**46.** ROLLER, D. E. *The early development of the concepts of temperature and heat.*

Cambridge: Harvard Univer. Press, 1950.

**47.** ROZEBOOM, W. Mediation variables in scientific theory. *Psychol. Rev.,* 1956, *63*, 249-264.

**48.** SCHMIDT, H. O. & FONDA, C. P. Reliability of psychiatric diagnosis: A new look. *J. abnorm. soc. Psychol.,* 1956, *52*, 262-267.

**49.** SCRIVEN, M. Definitions, explanations, and theories. In H. FEIGL, M. SCRIVEN, & G. MAXWELL, *Concepts, theories and the mind-body problem.* Minnesota Studies in the Philosophy of Science, II. Minneapolis: Univer. Minnesota Press, 1958, pp. 99-195.

**50.** SIMONS, D. J. & DIETHELM, O. Electroencephalographic studies of psychopathic personalities. *Arch. Neurol. & Psychiat.,* 1946, *55*, 619-627.

**51.** SINES, L. K. An experimental investigation of the relative contribution to clinical diagnosis and personality description of various kinds of pertinent data. Unpublished Ph.D. thesis, Univer. Minnesota, 1957.

**52.** SOSKIN, W. F. Bias in past-diction from projective tests. *J. abnorm. soc. Psychol.,* 1954, *49*, 69-74.

**53.** STEPHENSON, W. The significance of Q-technique for the study of personality. In M. L. REYMERT (ed.), *Feelings and emotions.* New York: McGraw-Hill, 1950.

**54.** SUNDBERG, N. The acceptability of fake *versus* bona fide personality test interpretations. *J. abnorm. soc. Psychol.,* 1955, *50*, 145-147.

**55.** TALLENT, N. On individualizing the psychologist's clinical evaluation. *J. clin. Psychol.,* 1958, *14*, 243-44.

**56.** TAYLOR, L. W. *Physics, the pioneer science.* New York: Houghton Mifflin Co., 1941.

**57.** WAISMANN, F. Verifiability. *Proc. Aristotelian Soc. Suppl.,* 1945, *19*, 119-150.

**58.** WINCH, R. F. & MORE, D. M. Does TAT add information to interviews? Statistical analysis of the increment. *J. clin. Psychol.,* 1950, *12*, 316-321.

**59.** WITTENBORN, J. R. *Wittenborn Psychiatric Rating Scales. New* York: Psychological Corp., 1955.

**60.** WITTGENSTEIN, L. *Philosophical investigations*. Oxford: Blackwell, 1953.

**61.** ZUBIN, J. & WINDLE, C. Psychological prognosis of outcome in the mental disorders. *J. abnorm. soc. Psychol.,* 1954, *49*, 272-281.