The Use of Research Instruments in Psychopathological Assessment: Some Historical Perspectives

Joseph Zubin\(^1\) and Raymond J. McCall\(^2\)
Veterans Administration Medical Center
and
\(^1\)Department of Psychiatry, University of Pittsburgh
\(^2\)Department of Psychiatry and Mental Health Sciences
Medical College of Wisconsin

Delivered at the Symposium on "Contemporary Assessment of Psychopathology"
Thursday, May 14, 1987, 2 to 5 PM
At the Annual Meeting of the
American Psychiatric Association, Chicago Illinois

The beginnings of classification, assessment, and measurement of human behavior including psychopathological behavior is hidden in the dawn of history. There are, however, bits of cultural evidence in customs and laws indicating that at least deviant behavior was recognized and sanctioned. But even more than recognizing the presence of the deviant behavior, tests to establish the deviation were also applied quite early, e.g. the Shibboleth test at the forks of the Jordan and the Talmudic test for mental competence consisting of the placement of a nut and a shaving of wood in front of the suspect.

Perhaps the earliest assessment of personality, including psychopathology, was through astrology dating back 25 centuries in Mesopotamia. Palmistry began 3,000 years ago in China. Handwriting analysis came much later and phrenology came in the middle of the 18th century in France—pioneered by Gall.
These pseudo sciences are still part of folk lore and have left scientific impacts. Thus, astrology is probably responsible for the studies of Locus of Control, palmistry and handwriting analysis are probably responsible for identification of individuals through palm prints, finger prints and signatures, and phrenology was the forerunner of brain localization. Divining personality through animal entrail examination, cloud reading, and flame inspection are the forerunners of projective techniques.

Palmistry has provided an apocryphal story about Yasser Arafat. He allegedly inquired about the day of his death from a palmist only to be told that he would die on a Jewish holiday. What holiday, he inquired, only to be told any day he dies will become a Jewish holiday.

Assessment based on handwriting analysis is another art which goes back to earliest times. Quite a literature has developed describing the meanings or be attached to various elements in a handwriting specimen including such characteristics as length of the middle zone of the letters, width, slant, curvature, pressure on the writing instinctual which was usually interpreted as reflecting instinctual drive.

The evaluation of handwriting and similar projective, as well as objective techniques, became a part of the senior author's major preoccupation. There wasn't a single technique that crossed over into the psychopathological assessment domain that was not investigated and the same
held true of therapeutic techniques. Unfortunately, very few of these techniques could meet the standards of reliability, validity, and objectivity that one looks for in a measuring device. As for handwriting, I undertook to evaluate it at the behest of Dr. Nolan D.C. Lewis, then the Director of the NYSPI because a Viennese graphologist named Thea Stein Lewinson had analyzed the handwriting of some of his patients and reported apparently to his satisfaction their personality characteristics based only on the inspection of the handwriting samples. As a result, we developed some objective measures of handwriting characteristics of high reliability, but could not establish their validity. (Lewison, T.S. & Zubin, J. Handwriting analysis: A series of scales for evaluating the dynamic aspects of handwriting. New York: Kings Crown Press, 1942)

How did classification, assessment and measurement of behavior get their start? Classification is the way infants get a handle on the blooming confusion which the world presents. To live in this world they must classify, identify and evaluate their surround even as the anthropologist entering a new culture must learn how the natives perceive and organize their world. The classification of the tribal connections in the form of kinships is perhaps the first classification of human beings. As for objects, Aristotle was perhaps the first to establish categories for the classification of objects and insisted that each category has an essence which must be
shared by the members in the category. This establishment of qualitative categories permitted comparison between objects in the form of the positive, comparative and superlative levels of adjectives e.g. red, redder, reddest.

But how did the categories of objects arise and how was class-membership determined? Perhaps the judgments of similarity and difference were the underpinnings of the classification system. The judgment of similarity between objects may have rested on the criterion of whether the two objects served the infant equally well for the task at hand. If the string and the long handled fork were equally good for fetching a toy, the judgment of "similarity" applied. If only one of two objects suited the purpose of the individual, the objects were regarded as "different". However, as development continued, the self-referred criterion of similarity gradually was replaced by more objective external criteria independent of whether the objects under scrutiny depend upon self-reference as a basis. Thus an ounce of gold and an ounce of coal are regarded as equal in weight even though self reference would make one prefer the gold.

As a matter of interest, the psychological rationale for sorting tests such as the Goldstein and Scherer, used in differentiating schizophrenics from normals, may have arisen on the basis of infantile self-reference which the adult schizophrenic may regress to when the disorder develops.

(Zubin & Thompson, 1941)
But how did the categorical qualitative judgment of similarity and difference lead to measuring scales? Here again, history has no record of the development. However, we can safely assume that even before the scales for measuring such physical characteristics as warmth, weight, and length developed, they were measured subjectively. There are no data to substantiate this suggestion, but we can use our imagination in describing how it may have happened.

Let us look at the measurement of subjective warmth. Since there is no documentary evidence for man's initial attempt at gauging the cause and degree of subjective warmth, we shall resort to fantasy. Come with me to a cave in some prehistoric ice age, before Prometheus, and listen to a symposium on the origin of the experience of subjective warmth. One savant declares that warmth depends on the number of skins covering the body. Another claims that it depends on duration of exposure to the sun. A third postulates swift running as the source of warmth and the distance covered as a measure. The medicine man in their midst raises a controversy because of his claim that his patients often report experiencing warmth without the benefit of skins, sunshine and running. The symposium ends without a resolution. But Prometheus later discovers fire and it becomes possible to demonstrate that by adding faggots to the fire, the experience of warmth is raised in all the inhabitants of the cave and by the same process
removal of faggots reduces it. The first breakthrough has occurred. Man can manipulate an external agent to raise and lower subjective warmth at will. But this is still far from measurement. The first known historical breakthrough occurs in ancient Egypt, where after the linguistic development of ratings in the grammatical form of positive comparative and superlative adjectives (warm, warmer, warmest) took place and rating scales were developed for measuring warmth in four steps - with warmth anchored at one extreme to the hottest day of summer, and at the other extreme to the coldest day of winter. Eventually, the expansion of mercury with increase of experienced warmth is noted, and the thermometer is born, and finally, humidity and air pressure are recognized as important factors, and the present-day discomfort index emerges. What were the essential steps in the process? First, the discovery of means of inducing changes in a sensory experience by external control (fire). Second, the development of an external criterion for measurement independent of the self-referred subjective experience (mercury). The same process, no doubt, held true of the other measures such as weight and length which have attained great objectivity. Pain is still a subjective phenomenon without an external criterion, and without measurable ways of inducing it, though recent efforts along these lines show considerable promise. Intelligence tests became objective when mental age scales were substituted for teachers' subjective impressions. Anxiety, depression,
elation are still in the rating stage. In my fantasy, I sometimes imagine that we may find a life-bearing planet, somewhere in space, where anxiety has already been measured, but where warmth or length are still on the intuitive level (Zubin, 1965).

It is interesting to note that psychology as well as psychopathology cede their hard-won gains to other fields, once the subjective experience has been transmuted to objective data. Thus, in psychology, a large part of psychophysics has become the domain of illuminating and acoustical engineers; educational psychology has been taken over by the educators; while in psychopathology, general paresis, as soon as its etiology became known, was lost to the field, as were pellagra with psychosis, epilepsy, phenylketonuria, etc. Psychopathology retains only the diseases of unknown etiology, and psychology deals only with the still generally intractable aspects of experience. The applied fields take over everything susceptible of measurement.

II. Early Development of Modern Rating Scales, tests and interviews.

Two traditions have developed in assessment: The psychometric and the clinical traditions. These two trends differ essentially in the contrast between the dimensional and the categorical approaches. The dimensional approach arose from the psychometric tradition and the categorical from the clinical tradition.
It might be well to pause here for a moment and inquire as to which approach — the categorical (typological) or the dimensional—is the truest. By the dimensional approach I mean the assumption that types do not exist and instead the hyperspace of personality variables has continuous hypersurfaces without any hills nor valleys, no multimodal distributions, nor discontinuities.

I had raised this question in 1965 at the Conference on the Role and Methodology of Classification in Psychiatry and Psychopathology (Katz, M.M., Cole, J.O. and Barton, W.E., Eds.) and as a result of the heated discussion which ensued, could not fall asleep that night. I turned on the TV and caught the Farmer's Hour with a program on pomology. The reporter indicated thaters that the genes of the apple may determine its acceptability we again return to a typology of alleles. But since the alleles secrete some amino acids, we can measure their quantity and return to dimensional tests again. However, the amino acids consist of patterns of molecules and these again yield a typological basis for discrimination and so on ad infinitum. Apparently, the struggle between typology and dimensionality is based on a pseudo-problem dependent upon the state of the art at any given timeers that the genes of the apple may determine its acceptability we again return to a typology of alleles. But since the alleles secrete some amino acids, we can measure their quantity and return to dimensional tests again. However, the amino acids consist of patterns of molecules
and these again yield a typological basis for discrimination and so on ad infinitum. Apparently, the struggle between typology and dimensionality is based on a pseudo-problem dependent upon the state of the art at any given time.

A. Rating Scales and Personality Inventories

Following the pattern of the ancient Egyptians in providing a rating scale for the measurement of warmth, similar attempts form the basis for the ratings of human behavior. Thus, ordinal rating scales were the initial steps. The first recorded attempt was due to Heymans and Wiersma in Holland, who drew up an inventory of symptoms through a self reporting instrument which was revised by Hoch and Amsden and subsequently by Wells. This eventually led to the Woodworth Personal Data Sheet, the father of modern personality inventories leading up to the MMPI. Kraepelin introduced measurement of emotional complexes through the Word Association Test and also introduced psychological testing to help in diagnosis in the following words:

"As soon as our methodology has sufficiently proved itself through experience with healthy individuals, it would be possible to approach the actual ultimate goal of these efforts, the investigation of the sick personality, especially of the inborn pathological disposition . . . We, therefore, have first of all to investigate whether it is possible by means of psychological tests to determine individual deviations, which cannot be recognized by ordinary observation. If that succeeds, we would be in the position through the quantitative determinations at our disposal, to establish the borderline between health and disease much more precisely and more validly than has been possible so far."
In psychopathology, measurement lagged behind the measurement of psychometric traits primarily because the changes in patient behavior were not sufficiently large to observe. Patients stagnated for the most part, and static description of their behavior did not require rating or measurement. However, with the arrival of somatic therapies changes in behavior were so striking and rapid that a demand arose for providing such measures.

The speed of the behavioral changes under the influence of drugs was so dramatic that it challenged the therapist to provide measures of the degree of change and the corresponding relation to dosage. For the first time in psychiatric history, psychotic behavior could be altered before one's very eyes and the need for measurement and evaluation became apparent. This caught the measurement experts--mostly psychologists--unprepared, since all the available clinical tests and tools provided only static descriptive measurements like intelligence tests or projective techniques, geared to constant trait rather than varying state measurement. Rating scales had to be produced practically overnight by such pioneers as Lorr, Wittenborn and Malamud. There is probably an apocryphal story told about the physiologist Hudson Hoagland, who found some biochemical change in Malamud's patients at the Worcester State Hospital and asked to see Malamud's data on the same patients. Malamud handed him his voluminous case histories. "Are there no numbers here on their behavior," Hoagland
inquired, to which Malamud cried, "No, but if you want numbers I'll make them up for you." That night with the help of Sands he converted his descriptions of behavior into rating scales.

The success of the drug therapies also required better classification of patients in order to obtain more homogeneous groups. To fill this need, the free floating clinical interview had to be converted from a blunderbuss into a sharpshooting rifle. Systematic semi-structured and structured interviews resulted. Their use in the US-UK Diagnostic Project and in the WHO International Pilot Study of Schizophrenia were the proving grounds in which these instruments were fire-tested. These projects aroused diagnosis from its academic lethargy and finally led to the developments of such instruments as the SADS, DIS, Research Diagnostic Criteria, and the DSM-III classification system.

It must be pointed out, however, that the revolution in psychopathology was not due to pharmacological intervention alone. Even before the drug era, the opening of the doors in European hospitals, especially in England, had brought about better outcome results. But, the dramatic effect of the drugs electrified the field.

It is clear that the advances in descriptive psychiatry through the improvement in diagnosis and the introduction of neuroleptics stands at the top of the list of recent achievements.

B. Tests
Perhaps the best examples of measurement in human behavior is the intelligence test. Before Binet, attempts were made to rate intelligence through adjectival ratings. Eugen Bleuler, at the turn of the century, would ask his residents to assess the intelligence of a new patient. One resident began his questioning by asking a female patient where she came from and how she got to the clinic. Bleuler interrupted scoldingly with "never ask a lady about geography, she has no sense of it". (Remember that this was 1900 in a country that has even now denied women votes.) Apparently Bleuler laid down the first requisite of a test, it should deal with knowledge that the patient possesses or should possess. Following our discussion of the development of temperature scales, what are the criteria that the test would satisfy? There has to be some way in which we could manipulate the trait under investigation. We can not, of course vary intelligence through manipulation but can instead use variation in chronological age as a quasi-manipulation, since it is clear that intelligence, at least during the developmental period, varies with age. That is how the concept of mental age was born. Thus, instead of varying the subjective experience, as was the case with subjective warmth, the natural variation in intelligence with age was used to manipulate the degree of intelligence. The external criterion for intelligence corresponding to the expansion of the mercury column in the case of subjective
The external criterion for intelligence corresponding to the expansion of the mercury column in the case of subjective warmth, was school progress which presumably varied with intelligence and thus served the ends of construct validity.
Three Descriptive Psychiatry

6. The interview

The Oxford English Dictionary states that the
term interview first appeared in 1514 to designate a
meeting of persons face to face, especially for the
purpose of formal conference on some point. In 1548 it
meant ‘to have a personal meeting with’ and ‘to meet
together in person’. In 1869 it was used to designate
a meeting between a representative of the press and
someone from whom statements were sought for
publication.

The clinical interview is the psychiatric tool par
excellence for arriving at a diagnosis and the conduct
of psychotherapy. None the less, until recently little
research effort has been directed at investigating the
reliability, validity, and usefulness of this basic tool.
The interview has a short history but a long past
dating back to antiquity. Interview-like conversations
consisting of questions and answers are found in the
Bible. The Socratic method of interviewing in the
search for truth is well known. Celsus, a layman in the
first century A.D., described visual hallucinations as
reported by patients, and Soranus formulated rules
for conversing with mental patients, recommending
that labourers should be engaged in conversation about
cultivation, sailors about navigation, and so on (Zilboorg & Henry, 1941). Interviewing techniques have
varied with the social-cultural milieu and Zeitgeist. The
ancient Greeks regarded mental disorder as due to
either perversity, possession, or drug-induced states,
and verbal reports were regarded as random effects
which had little or no interest for the physician. Inter-
est in interviewing languished during this period.
During the Middle period in Spain diagnosis was
based on mystical and intuitive approaches and the
patient’s utterances were interpreted by seers on the
basis of these criteria, rather than on the basis of the
patient’s needs and difficulties. Interest in the patient’s
attitudes and beliefs rose in the Western Caliphate
(Seville and Cordova) when the problem of personal
responsibility came to the fore. During the Spanish
Inquisition the accused were interrogated to establish
their heresy and the guide for enticing heretics to admit
their guilt was formulated as a systematic interview
(Lex, 1888). With the rise of the idea of personal
responsibility, implying that only those who are
mentally competent can be punished, there arose a
need to interview the patient in order to assess
thoughts and feelings. In the early seventeenth cen-
tury Bacon used certain test questions. He enquired,
for example, whether the patient knew his own age
and name, thereby initiating psychological interview-
ing with regard to time and personal orientation. These
questions were eventually incorporated into the men-
tal status examination for determining orientation for
person, time, and place. Freud first introduced sugges-
tion and hypnotism to induce patients to speak more
freely before developing the psychoanalytical inter-
view, in which all constraints on the conversation were
removed and the free associations encouraged, thus
permitting unconscious as well as conscious commu-
nication.

The interview is a form of natural conversation,
and its status as a special form of inter-personal com-
munication did not become apparent until recently.
Certain criteria separate the interview from ordinary
conversation or communication. For the purposes of
this chapter, we can differentiate it from other types
of human encounter as follows:

1) it consists of a meeting of at least two individ-
uals, the interlocutors, face to face;

2) it is directed at a particular purpose by either
one or all interlocutors;

3) it employs conversation and observation;

4) it involves a non-reciprocal relation between
the interlocutors, in so far as one is the interviewer
and the other the interviewee.

The definition of the interview in psychopathology is
to be found in a variety of sources (Zubin, 1965; Matarazzo, 1978). For a long time the psychia-
tric interview remained a free floating conversation
between clinician and patient, and though this is still
the vogue today, the increasing use of the interview
for research purposes has led to a considerable metamorphosis of this communication technique. The impetus for improving the interview was provided by several factors, including the need for yardsticks to assess changes due to pharmacological treatment, and by individual studies like the bi-national US UK diagnostic project (Cooper et al., 1972) and the International Pilot Study of Schizophrenia (WHO, 1973). Previously interviewing methods and diagnoses were of more academic interest since they led to no testable consequences.

The aims of the psychiatric interview are fourfold: first, to elicit as much as possible information about presenting symptoms and the antecedents of current episodes in order to cast light on possible causes, secondly, to determine the patient's feelings and attitudes about his current status and symptoms, and, thirdly, to note the pertinent non-verbal behaviour which can help establish the nature of the problem. In addition, fourthly, since an episode of mental illness reflects a disturbance in interpersonal relationships, the behaviour sampled during the interview offers information on the disturbance in interpersonal relationships. The degree to which the interview taps these main sources of information is a measure of its success, which is why psychiatric interviewing cannot be left to the novice or be based on computers alone. While sophisticated systematic interviewing can elicit factual information, the more subtle attitudinal factors and the assessment of interpersonal relationships and the detection of abnormal non-verbal signs require a flexible give-and-take possessed only by experienced clinicians.

The success which psychiatry attained in providing psychometric methods for first assessing intelligence and then personality by the use of systematic tests and inventories finally challenged the unsystematic way in which interviews were conducted by psychiatrists. Several additional forces were at work to improve the quality of the interview. First came research soon after World War I demonstrating that the clinical interviews used for diagnosis or classification were neither reliable nor valid.

The proven unreliability of the early interviews prompted attempts to develop self-rating inventories and scales designed to be filled out by the patient, and checklists and rating scales to be completed by the interviewer. Self-reporting instruments have the advantage of eliminating interviewer bias and interpretation by putting the onus on the individual's response rather than on the response of the inter-

viewer to the patient's behaviour. This strategy does not, of course, guarantee a better result, and obvious self-reporting inventories have generally not been found to be useful for the classification of patients although they have been found quite useful in dealing with minor disorders encountered primarily among student populations and in surveys of the general population. In a recent study, Dohrenwend and his colleagues (Dohrenwend et al., 1979) found that the common denominator in such self-reports seems to be related more to Frank's (1973) demoralization factor than to clinical psychopathology.

The checklists and rating scales used by interviewers did not utilize standardized methods for obtaining the information needed for the ratings and consequently results were not always comparable from patient to patient. To obtain a more systematic method for the collection of the information, a new clinical interview with a systematic structured format was developed.
This began with an attempt to utilize a check-list for the Mental Status Examination utilized by the residents and staff of the New York State Mental Hygiene Department. This "redbook" as it came to be called underwent a series of revision by each succeeding commissioner of the Department of Mental Hygiene and when Dr. Nolan D.C. Lewis became commissioner he invited me to help the revision. I introduced a systematic checking schedule for recording the presence of the symptoms observed during the course of the examination. World War II interfered with its application but soon after the end of the war, Drs. Burdock and Hardesty took over this task and finally with the help of two of their trainees - Robert Spitzer and Joseph Fleiss prepared in 1964 the first in the series of systematic semi-structured interviews from which the SADS finally emerged and the revision of DSM and provision of RDC followed.
A first step was to standardize the usual mental status examination (Spitzer, Fleiss, Burdock & Hardesty, 1964). The questions to be asked of the patient were then specified in a proper sequence and directions for coding the responses were laid down. Thus the loose mental status examination, formerly the psychiatrist's mainstay, was converted into a systematic, structured interview which yielded high reliability in the scoring of items and carried some degree of validity.

This technique has undergone a series of modifications (Zubin, 1972). Three types of interviews were developed: (1) the non-probing Structured Clinical Interview (SCI) (Burdock & Hardesty, 1969), (2) the two medium-probing schedules, the Mental Status Schedule (MSS) (Spitzer et al., 1964) and the Psychiatric Status Schedule (PSS) (Spitzer, Endicott, Fleiss & Cohen, 1970), and (3) a deep-probing Present State Examination (PSE) (Wing, Birley, Cooper, Graham & Isaacs, 1967). These interviewing methods have demonstrated their value in the two major international studies mentioned previously.

The US–UK project was initiated to determine why the national statistics show such disproportionate frequencies of affective disorders in the United Kingdom and of schizophrenia in the United States. When the newly developed systematic interviews (Gurland, cited in Zubin, Salzinger, Fleiss et al., 1975) were applied to samples of patients admitted to hospitals in the two countries, the cross-national differences turned out to reflect different diagnostic practices of psychiatrists rather than differing characteristics
not ensure advances in diagnosis because for diagnostic purposes the data provided by the interviewers had to be integrated systematically. The first step was to provide for each diagnosis specific criteria which had to be met by the interview data before being applied. The St. Louis group (Feighner et al., 1972) pioneered in this effort, the US-UK Project, as well as the WHO Project, also provided tentative bases for arriving at a diagnosis by analyzing the interview results into underlying dimensions arrived at either by clinical judgement (FSE) or by both factor analysis and clinical judgement (US-UK).

The most recent developments in diagnosis include the Research Diagnostic Criteria (Spitzer et al., 1978), a distillation of the earlier results obtained from the above-mentioned sources which provide the criteria which have to be satisfied in a specific manner for a given diagnosis.

Several other developments have further advanced the problem of diagnosis. First, a measure of agreement between diagnosticians was provided by a new statistical index, Kappa, especially designed for this purpose by Cohen (1960). Formerly the degree of agreement between two clinicians was based on the percentage of cases in which both of them agreed on the presence or absence of the disorder. However, this percentage agreement did not take into consideration chance agreement between the two judgements. Kappa eliminates chance. Kappa has also been extended to agreement between more than two clinicians, and standard errors have been provided for estimating confidence levels. A second technique was to present videotaped interviews to an audience of clinicians, in order to obtain simultaneous measures of agreement (Sharpe et al., 1974). Finally, a technique was developed that analysed the process of arriving at a diagnosis by providing a step-wise decision tree in which the successive judgements of the clinician could be simulated on a computer. Thus, after considering specified criteria as to whether the patient is or is not suffering from an organic disorder and excluding such pathology, the computer next examines the evidence for the criteria of schizophrenia. If the data satisfy these criteria, the process of diagnosis is complete. If the criteria for schizophrenia are not met, the search continues until either an alternative diagnosis is reached or the subject is found to be non-classifiable or normal (Spitzer & Endicott, 1968; Wing, Cooper & Sarantou, 1974).

The structured interview technique and research diagnostic criteria have become so widespread that
there is now a trend towards the premature standardization of diagnosis, thereby preventing further growth and development in this area. Though the standardization of the interview introduces uniformity and comparability, innovative findings may be ruled out because of the rigidity of the process, and further developments may be endangered. A biological analogy is afforded by comparing the genetic apparatus of bacteria with that of animal cells (Sbar & Price, 1972). The genetic apparatus of bacteria organizes each of its genes neatly in unbroken genetic codes with no intervening unnecessary parts. Animal genes are more loosely organized and have excess unused DNA, but have nevertheless proved more flexible in the evolutionary sense. Thus, while bacteria were frozen into an inflexible mould for millions of years, animals have scaled the evolutionary ladder. A similar fate could overtake the over-systematic approach to diagnosis offered by a scheme such as DSM-III (American Psychiatric Association, 1980), despite the immediate advantages it offers. For this reason, we must continually return to careful clinical observations if diagnostic advances are to be made.

In an examination of descriptive psychiatry from the point of view of research applications, it must be recalled that the purpose of description is classification or diagnosis. The criteria for good classification are twofold: to serve as a basis for generalization in various types of comparative studies, and to serve as a key to the information storage system (Meyr, 1981). In order to achieve this basis for generalization about a given classification grouping, such groupings should cluster together individuals who are as homogeneous as possible and should summarize the pertinent data for such factors as aetiology, treatment and outcome. One of the reasons why the task of the diagnostician in psychopathology is more onerous than that of the diagnostician in better understood physical disorders is that the latter deals with more objective indicators. Behavioural diagnosis, even when it is based on specific verifiable indicators, suffers from the fact that some observed behaviour stems not from the deviations due to psychopathology, but from the individual differences reflecting pre-morbid personality. Our inability, at the present time, to separate these two sources of observed behaviour underlies much of the heterogeneity observed in patient groups even when they suffer from the same disorder. This source of heterogeneity also interferes with the second goal of diagnosis, that of summarizing the variables pertinent to the diagnostic category concerned.

In choosing strategies for improving classification, we might reasonably consider the paths followed by taxonomists in biology where, as Mayr (1981) has noted, there are three primary approaches that are now in conflict: (1) phenetic, (2) cladistic, and (3) evolutionary. The phenetic approach refers to the overall similarity between the members of a class, regardless of how this similarity developed. The cladistic approach bases classification exclusively on genealogy, that is, on the branching pattern of phylogeny. This phylogeny consists of a sequence of dichotomies, each representing the splitting of a parental species into two daughter species. The ancestral species ceases to exist at the time of the dichotomy; sister groups must be given the same categorical rank, and the ancestral species together with all of its descendants must be included in a single 'holophyletic taxon'. The evolutionary classification is the more traditional, following Darwin's development of the theory of evolution which bases its classification on observed similarities and differences among groups of organisms, evaluated in the light of their inferred evolutionary history. At the present time, Mayr concludes, each of these approaches serves a valuable purpose in classification and the best system might be a synthesis of the more promising components of each.

Certain parallels can be drawn between these three approaches in biology to the current state of classification in psychopathology. The phenetic approach is the one that is most in vogue; much of DSM-III, for example, is based primarily on observed similarities in behaviour. The cladistic approach has characterized so much of the description of the behaviour as the decisions applied to the observations by the observer. This approach may be likened to the making of decision trees which serve to simulate the process of arriving at clinical diagnosis. Several attempts at providing computerized diagnoses have followed this path, for example, DIAGNO (Spitzer & Endicott, 1968, 1969) and CATEGO (Wing et al., 1974).

The evolutionary approach has its counterpart in the search for aetiology. Thus, even if the phenetic approach gives a careful description of the terrain which psychopathology covers, it would only be a surface picture of the topography of hills and valleys, streams and islands. The phenetic approach gives no indication of the geological forces that have produced the features of the terrain. Similarly, in psychopathology complete understanding cannot be attained until
the aetiology of the disorder is unravelled. This problem will be discussed further in the section on aetiology.

Other advances in descriptive psychopathology have included (1) the use of computers that simulate clinical decision processes and arrive at a diagnosis (Spitzer & Endicott, cited in Zahn et al., 1975; Wing et al., 1974), (2) the provision of mathematical methods for clustering individuals with shared psychopathological characteristics into more homogeneous subgroups (Fleiss & Zahn, 1969; Wing & Nixon, 1975), and (3) the development of analytical descriptions of patients' behavioural characteristics (Kanfer & Savlow, 1969; Salzinger, cited in Zahn et al., 1975). Rather than emphasizing the presence of symptoms per se, the analytical description of behaviour relates deviant behaviour to environmental contingencies that initiate and sustain it. Description is therefore focused on aspects of the individual and his environment that are of immediate relevance to behaviour modification therapy.

According to the philosophy of scientific taxonomy (Alexander Wolf (1929) and Carl G. Hempel (1961)), mutually exclusive and exhaustive categories should be based on the most important characteristics of the thing concerned (in this case, the disorder) and on the actual relations between them. Natural classification systems should be preferred to artificial constructs. Judged by such criteria, achievement in descriptive psychopathology has been quite limited. A dispassionate view of descriptive psychopathology today leads to the conclusion that some progress has been made, perhaps as much as can be expected without more aetiological knowledge. Nonetheless, through the use of systematic structured interviews and operational criteria for selecting a diagnosis, the characteristics of patients can be described objectively so that the replication of basic research and treatment findings on similar patients becomes possible. Moreover, it can be demonstrated that individuals suffering from syndromes like schizophrenia can be found in parts of the world that differ widely in respect of cultural and ecological conditions.

However, it cannot be assumed that the reliable description of a syndrome implies a valid understanding of its underlying cause. Description alone is inadequate for a durable classification of mental/behavioural disorders, which must be based on aetiology. But how can progress be made when, with a few exceptions, so little is known about the causes of mental disorder and the efficacy of treatment? When faced with such ignorance one must contemplate possible or as it causes, formulate them into parsimonious scientific models and proceed to test the hypotheses they generate. (Zahn, 1972)
IV Personality

In similar fashion, attempts were made to devise techniques for measuring personality and psychopathology. Unfortunately, means for utilizing the first procedure in measuring subjective experience -- manipulation of the subjective or internal phenomena -- was not always available as it was in the case of temperature and intelligence. The second procedure, that of establishing an independent criterion (the expansion of mercury in the case of subjective warmth) was also rarely available. For these reasons, progress in tests for personality and in psychopathology has been slow. Perhaps the most successful test for psychopathology is the MMPI. The varying of the degree of the subjective phenomena under measurement was accomplished by studying contrasted groups of diagnostic categories including normalcy. The difficulties presented by this approach inhere in the fact that the clinical categories were based on interviewing, which at the time of the preparation of the MMPI were in a low state of reliability and validity. Perhaps a restandardization based on Research Diagnostic criteria might help. Construct validity (corresponding to the expansion of mercury) is also problematical, and very little research has been done along these lines because we do not have any consensus on what the constructs might be. The problems of developing tests of personality suffer from the same difficulties facing tests
of psychopathology and we are reduced to returning to interviewing as our mainstay.

One of the reasons for the failure of the Rorschach to qualify as a diagnostic instrument is the fact that in the last analysis the scoring of the responses and their clinical significance was originally based on the interview by means of which a clinical diagnosis was reached. We have already pointed out that the clinical interview in Rorschach's day was a weak rod to lean on. However, if the Rorschach protocols are regarded as interviews and analyzed for their content, the resulting dimension have been shown to correlate with diagnoses because both the Rorschach protocols and the diagnoses are based on interview content.

V RELIABILITY

It is well known, that the greatest advances in science come not from new theories, but from new technological inventions. The development of audiotaping and videotaping have had the most important impact on the study of reliability. I recall that I was once asked by Clarence P. Oberndorf to help him evaluate psychoanalysis. He invited me to his penthouse for dinner to discuss the problem. I suggested that he pick out one of his analysands and record several interviews with the patient so that we could together go over the recording and develop a content analysis of the interview to see what dimensions we could discover for evaluation. He went out and bought an expensive Stromberg Carlson recorder and started recording
his interviews. After several months he again invited me for dinner and at the end, pulled out a cigar, lit it and puffed on it and then gingerly pushed the button to start the recording. No sooner had the record started when I noticed that Oberndorf grew uneasy and as time went on his face reddened and finally he rose, turned off the sound saying: "This never happened -- it's impossible." That was the end of our attempt at evaluation.

I had better luck in studying the reliability of the ratings on interviews with the aid of video tapes in our US-UK project. Instead of having a second interviewer sit with the main interviewer to get reliability ratings, we offered the video tapes as a basis for reliability studies. Furthermore, groups of raters could be obtained either individually by mailing them the videotapes or we could gather raters at a convention and obtain their ratings simultaneously. In one such experiment which Dr. Katz and I conducted we found large discrepancies in the judgement of whether the patient was neurotic or psychotic. An analysis of the ratings indicated that the differential was based on the rating of apathy. Those who gave a high rating on apathy judged the patient to be psychotic while those who rated the patient low in apathy judged him to be neurotic.
VI SUMMARY

A. Classification

B. Development of modern rating scales test and interviews

C. Diagnoses

D. Personality

E. Reliability

F. Validity?