1991

3. Assessment of Val id ity In Computer-Based Test Interpretations

Kevin L. Moreland
NCS Professional Assessment Services, Minneapolis

3 Assessment of Validity in Computer-Based Test Interpretations

Kevin L. Moreland
NCS Professional Assessment Services, Minneapolis

The use of computers to interpret psychological tests is a "hot" topic, both within psychology and without. It is hot in the sense of giving rise to an increasing number of books and articles (e.g., Butcher, 1985, 1987; Eyde, 1987; Krug, 1987). It is hot in the sense of giving rise to an ever-increasing number of business enterprises (compare any recent *APA Monitor* with an issue from 1981). It is hot in the sense of capturing the attention of the news media (e.g., Petterson, 1983). And it is hot in the sense of giving rise to increasing controversy within psychology itself. In a *Science* editorial Matarazzo (1983) expressed concern lest computer-based test interpretations (CBTIs) fall into the hands of unqualified users, his bottom line being: "Until more research establishes that the validity of application of these computer products by a health practitioner is not dependent on the practitioner's experience and training in psychometric science, such automated consultations should be restricted to . . . qualified user groups." Matarazzo (1985, 1986) has continued to write in that same vein, causing others to take up the cudgels to defend CBTI (Ben–Porath & Butcher, 1986; Fowler & Butcher, 1986; Murphy, 1987). Lanyon (1984) in his chapter on personality assessment in the *Annual Review of Psychology,* indicated that he was concerned by the proliferation of CBTI systems: "There is a real danger that the few satisfactory services will be squeezed out by the many unsatisfactory ones, since the consumer professionals are generally unable to discriminate among them. . . ." and ". . . lack of demonstrated program validity has now become the norm" (p. 690). Finally the Subcommittee on Tests and Assessment of the American Psychological Association (APA) Committee on Professional Standards and the APA Committee on Psychological Tests and Assessment have developed standards for the area (American Psychological Association, 1986). I published an

article describing attempts to establish the validity of CBTIs and made some suggestions regarding the shape future attempts might take (Moreland, 1985). The heat generated by the debate over CBTI seems not to have dissipated; however, some light seems to have been shed on the field since I was writing in 1984. In view of all this, a revision and expansion of my earlier efforts seems timely.

SOME HISTORY

The use of machines to process psychological test data is not a recent innovation (Fowler, 1985). A progression from hand scoring materials through a variety of mechanical and electronic "scoring machines" to the digital computer, has freed successive generations of beleaguered secretaries and graduate students from laborious hand scoring of objective tests. The first information concerning scoring machines for the Strong Vocational Interest Blank (SVIB) appeared in 1930 (Campbell, 1971). These initial machines were very cumbersome, involving the use of 1,260 Hollerith cards to score each protocol. In 1946, Elmer Hankes, a Minneapolis engineer, built the analogue computer that was the first automatic scoring and profiling machine for the SVIB (Campbell, 1971). A year later, he adapted the same technology to the scoring of the Minnesota Multiphasic Personality Inventory (MMPI) (Dahlstrom, Welsh, & Dahlstrom, 1972). In the mid 1950s, E. F. Lindquist's Measurement Research Center in Iowa City began to use optical (answer sheet) scanning devices instead of card-based scoring equipment. In 1962, National Computer Systems linked an optical scanner with a digital computer and began scoring both the SVIB and the MMPI (Campbell, 1971; Dahlstrom et al., 1972). Most automated test scoring still employs optical scanning/digital computer technology and the number and types of tests scored by this method have grown exponentially during the last three decades. Though automated scoring is most easily accomplished for objective tests with a limited number of response alternatives, sophisticated computer programs have also been developed to score the narrative responses elicited by projective techniques (e.g., Gorham, 1967). Prior to the advent of these programs, extensive training, if not professional expertise, was required to score projective tests. Similar programs have also been developed to evaluate other types of complex verbal productions (e.g., Tucker & Rosenberg, 1980).

In addition to keeping nerves from becoming frayed, automated scoring frees psychologists to spend more time in other functions, such as psychotherapy, where computer technology is not so advanced (see, however, Colby, 1980). It also enables more individuals to undergo psychological assessment. Finally, though not completely immune from the slings and arrows of human imperfections (e.g., Fowler & Coyle, 1968; Grayson & Backer, 1972; Weigel & Phillips, 1967), computer scoring appears to be more reliable than that done solely by

humans (Greene, 1980, pp. 25–26; Klett, Schaefer, & Plemel, 1985). A computer, once correctly programmed, will apply scoring rules with slavish consistency, whereas fatigue and other human frailties may render the psychologist, graduate student, or secretary inconsistent in the application of even the most objective scoring rules (Kleinmuntz, 1969).

In the late 1950s, a group of psychologists and psychiatrists decided that similar advantages might accrue if tests were interpreted by computer. Thus the first CBTI system was developed at the Mayo Clinic in Rochester, Minnesota (Rome, Mataya, Pearson, Swenson, & Brannick, 1965; Pearson, Swenson, Rome, Mataya, & Brannick, 1965). The MMPI was administered on special IBM cards that could be marked by the patient and read into the computer by a scanner. The computer then scored the MMPI and printed a series of descriptive statements from among a library of 62 statements, most of which were associated with elevations on single MMPI scales. Soon after the Mayo system was reported in the literature the first CBTI system to receive widespread professional use was developed by Fowler (1966) at the University of Alabama. In 1965, the Roche Psychiatric Service Institute (RPSI), established by Roche Laboratories to make the Fowler system commercially available to psychologists and psychiatrists, initiated the first national mail-in MMPI CBTI service. During the 17 years RPSI operated, approximately one-fourth of the eligible psychiatrists and psychologists in the United States used the service.

The Behaviordyne system (Finney, 1966) and Caldwell Report (Caldwell, 1970) have received wide use in the United States, and are still available. Later MMPI interpretation systems were developed by Lachar (1974b) and Butcher (University of Minnesota, 1982, 1984). Other prominent CBTI systems which have been marketed commercially in the United States interpret the 16 Personality Factor Questionnaire (Karson & O'Dell, 1975, 1987); the Rorschach (Exner, 1987); the Personality Inventory for Children (Lachar, 1987); and the Millon instruments: the Millon Multiaxial Clinical Inventory, Millon Behavioral Health Inventory, and Millon Adolescent Personality Inventory (National Computer Systems, 1989), among others.

TYPES OF CBTI SYSTEMS

CBTI systems can be usefully characterized along two dimensions, the amount of information they provide and the method used to develop them.

Information Provided by CBTIs

Descriptive reports may be distinguished from other types of CBTIs by two factors: Each scale on the instrument is interpreted without reference to the other, scale by scale, and comments on any one scale are usually quite cryptic. These

interpretations often involve no more than an adverb modifying the adjectival form of the scale name. Such an interpretation of a high score on an anxiety scale might, for example, read: "Mr. Jones reports that he is very anxious." Thus the interpretive comments directly reflect empirical data. The interpretive statements are as valid as the scales themselves. At first blush, this kind of report may seem so simple minded as to be unhelpful. Not so. This type of report can be especially helpful when a test has a large number of scales or when a large number of tests need to be interpreted in a short period of time. They allow the practitioner to identify quickly and easily the most deviant scales. This kind of report is most helpful if an instrument contains scales that are reported in terms of different types of standard scores (e.g., Ripley & Ripley, 1979) or different normative samples (e.g., Hansen, 1987). The MMPI report developed at the Mayo Clinic was the first report of this type.

Screening reports, like descriptive reports, are cryptic. They are distinguished from descriptive reports in that relationships among scales are usually considered in the interpretation and the interpretive comments are not usually couched in terms of a single scale name. The Minnesota Personnel Screening Report for the MMPI (University of Minnesota, 1984) is a screening report in this sense. The main body of that report is very cryptic—five 6-point rating scales. None of the rating scales corresponds directly to an MMPI scale, however. In fact, the rating on each of the five scales is determined by the configuration of a number of MMPI scales. The rules governing the "Content Themes" presented in that report are also complex. The comment that the client "may keep problems to himself too much" results from consideration of the following set of rules:

L and *K* are greater than *F* and
F is less than 55*T* and
D, Pa, Pt, and *Sc* are less than 65*T* and
  Hy is greater than 69*T* or
  *Hy2*is greater than 63*T* or
  Hy is greater than 64*T,* and *Hy1* or *Hy5* is greater than 59*T* or
  R is greater than 59*T* or
  D5 is greater than 59*T*

Screening reports are most helpful in situations where the same decision can be reached by multiple paths. Take the example of screening commercial pilots for emotional fitness. A screening report such as the Minnesota Report may deem a candidate's emotional fitness "suspect" if he or she: (1) seems to be a thrill-seeking individual; (2) is so obsessive that he or she is unlikely to respond promptly to in-flight emergencies; or (3) may have a drinking problem. Because of this multifaceted approach to the assessment problem, such reports are also

likely to be most helpful when they are truly used for screening rather than for making final decisions. They are too deliberately cryptic to be used for the latter purpose. Further investigation, triggered by a screening report, may lead one to discover that a suspect candidate is a recovered alcoholic who has been dry for 10 years.

Like descriptive reports, the output of screening reports is limited. However, the validation of screening reports is not simple and straightforward. As has been illustrated, the simple output may be generated by extensive, complex sets of rules, each of which must be validated.

Dahlstrom et al. (1972) contrasted *consultative reports* for the MMPI to screening reports in the following fashion: "The intent [of consultative reports] is to provide a more detailed analysis of the test data in professional language appropriate to communication between colleagues" (p. 313). In other words, consultative reports are designed to mimic as closely as possible the reports generated every day by human test interpreters. Well-developed reports of this type are characterized by the smoothly flowing prose and detailed exploitation of the data that would be expected from an expert human consultant. Indeed, the chief advantage of these reports is that they can provide busy practitioners with a consultation from someone who has spent years studying and using the instrument in question—an expert to whom the average practitioner would not ordinarily have access. Fowler's system for the MMPI produced the first CBTIs of this type. It is these types of reports that come to most minds upon hearing the phrase "computer-based test interpretations." It is these types of reports that will be the subject of most of this chapter.

How CBTIs Are Developed

In 1956, Paul Meehl called for a good "cookbook" for test interpretation. He was advocating the actuarial approach to prediction and description defined by Sines (1966) as "the empirical determination of the regularities that may exist between specified psychological test data and equally clearly specified socially, clinically, or theoretically significant non-test characteristics of the persons tested" (p. 135). This approach to CBTI development can best be illustrated through the example of one such system.

Unlike the MMPI and most other popular psychological tests, which were developed prior to the computer age, Lachar's CBTI system for the Personality Inventory for Children (PIC) was developed without a considerable "clinical lore" concerning the performance of the PIC scales (Lachar, 1987). (Fowler [1986] considers the concurrent development of test and interpretive system an "ideal" strategy, test development efforts enriching the evolving interpretation system.)

Efforts to compile a data base that would allow the development of empirically supported interpretive guidelines were initiated before the PIC was

published. Criterion data collection forms (see Lachar & Gdowski, 1979, Appendix A) were accepted by the staff of an active teaching service as performing clinically meaningful functions. An application form gathered presenting complaints, developmental history, and facts concerning pregnancy and birth. A form mailed to the child's school recorded teacher observations, estimates of achievement, and judgments as to the etiology of observed problems as well as suggested solutions. A final form was completed by the psychiatry resident or psychology intern who conducted the initial evaluation of the child or adolescent and parents. The latter form allowed the collection of dichotomous ratings (present/absent) of descriptors most of which could be arrayed under the following headings: affect, cognitive functioning, interpersonal relations, physical development and health, family relations, and parent description. Psychiatric diagnoses and ideal treatment recommendations were also recorded. Collection of data using these three forms resulted in an actuarial analysis of the PIC scores of 431 children and adolescents (Lachar & Gdowski, 1979).

Development of Lachar's CBTI system for the PIC first focused on the correlates of each scale on the basic PIC profile (Lachar, 1982; Lachar & Gdowski, 1979). The initial goal was to construct an interpretive system similar to the Mayo Clinic MMPI system (see Marks & Seeman, 1963, Appendixes E & F), in which each scale is individually interpreted. The individual scale approach resulted in an interpretation for every PIC profile, while actuarial interpretive systems based on the total profile configuration have proven, in the case of the MMPI, to be of limited value because a significant number of profiles usually remain unclassified.

The actuarial data base that provided the interpretive paragraphs and paragraph assignment to scores was generated in two phases. In the first phase, the 322 descriptive variables from the parent, teacher, and clinician forms were correlated with each of 20 profile scales to develop scale correlates. In the second phase, each identified correlate was studied to determine the relationship between the correlate and PIC scale *T*-score ranges. That is, correlate frequency was tabulated within a number of contiguous *T*-score ranges, usually 10 points in width. The goal of this process was to identify appropriate *T*-score ranges to which a given correlate could be applied, as well as to obtain an estimate of the frequency of each correlate within the *T*-score ranges. Rules were established to lead to correlate classification rates similar to their base rates within the study sample. A similar analysis determined frequent patterns of elevated *T*-score ranges and allowed the development of narrative paragraphs that reflected the elevation of two or more profile scales. Those efforts produced interpretive correlates like those in Table 3.1. Those correlates form the basis of the CBTI system for the PIC sold by Western Psychological Services (Western Psychological Services, 1984). It is easy to see that this system conforms with Sines's (1966) definition of an actuarial system. It is also easy to understand Meehl's

TABLE 3.1
Actuarial Correlates of the Personality Inventory for Children Delinquency Rate

Descriptor [1]	*Correlations* [2]			*T-Score Ranges*								
			Base Rate	*30-59*	*60-69*	*70-79*	*80-89*	*90-99*	*100-109*	*110-119*	*>120*	*Decision Rule*	*Classification Rate*
Impulsive Behavior	.25,	39	68[3]	40	57	61	72	76	72	84	100	>79T	79%
Temper Tantrums	.27,	.25	43	18	42	40	38	44	63	64	69	> 99T	66%
Involved with Police	.44,	.49	17	0	4	6	10	21	19	58	63	(< 60T) > 109T	(47%) 15%
Dislikes School	.18,	.38	39	28	28	28	30	48	55	63	70	> 89T	57%
Mother Inconsistent in Setting Limits	.26,	.3	59	27	45	61	59	64	82	89	67	(> 99T) < 60T	(79%) 63%

Adapted from Lachar and Gdowski (179).

[1] Clinician ratings.

[2] Ns = 215 and 216, respectively.

[3] Percentage of clients rated as displaying the characteristics.

(1956) enthusiasm for the actuarial approach to test interpretation: the interpretations are, ipso facto, valid within known limits.[1]

Combination of automated scoring and automated, actuarial interpretation would seem to be a marriage made in Assessment Heaven. Unfortunately, this relationship remains in the courtship stage. In spite of the fact that this is the way CBTI systems *should* be developed, only two such CBTI systems are commercially available, that for the PIC and one for the Marital Satisfaction Inventory (Western Psychological Services, 1984). After Meehl published his want ad there were several major attempts to produce actuarial cookbooks for the MMPI (Drake & Oetting, 1959; Gilberstadt & Duker, 1965; Gynther, Altman, & Sletten, 1973; Marks & Seeman, 1963; Marks, Seeman, & Haller, 1974). These herculean efforts have fared poorly outside the settings in which they were developed. Application of the complex profile classification rules necessary for actuarial interpretation causes the bulk of the tests to go unclassified (e.g., Briggs, Taylor, & Tellegen, 1966; Cone, 1966). Even when the cookbooks published by Marks and Seeman, and by Gilberstadt and Duker have been

[1]Generalizability is the most pressing question to be answered about actuarial CBTI systems. That is, are there extraneous factors that were not considered in the development of the actuarial CBTI system (e.g., setting) that affect its validity.

combined, the majority of tests have failed to find an interpretive niche (e.g., Payne & Wiggins, 1968). Although ignoring some of the classification rules allowed a greater number of tests to be classified, Payne and Wiggins still could not classify all of their sample. That is to say nothing of the decrement in validity that has been shown to occur when the actuarial correlates are generalized to populations differing in base rates of psychopathology, demographic characteristics, and other important factors (cf. Fowler & Athey, 1971; Gynther & Brilliant, 1968; Palmer, 1971). This state of affairs led some psychologists who were determined to exploit the advantages of automated test interpretation, such as Fowler, to advocate the "automated clinician . . . until the actuary comes" (1969, pp. 109–110).

The essential difference between the automated actuarial and automated clinical approaches is that the former method assigns interpretive statements on the basis of their statistical association with test data, while statements chosen by the latter approach are a function of human decision making. The psychologist who devises the statements and assignment rules in the automated clinical approach typically makes use of available actuarial data but, as suggested by the fate of the actuarial cookbooks discussed herein, is sometimes forced to rely on his or her practical experience in order to ensure that all tests are interpreted (Fowler, 1969). Fowler assumed that even though practical experience must sometimes be resorted to, the psychologist developing the interpretive statements usually possesses greater experience and, presumably, expertise than the average psychologist. (Unfortunately, the advent of microcomputers has made that assumption less tenable than it was when Fowler was writing; cf. Moreland, 1987.) Although undoubtedly not as good as actuarial interpretation, automated clinical interpretation possesses several advantages over human interpretation. In addition to those advantages that have been noted in the context of automated scoring of test data, automated interpretation has an advantage over human interpretation when large and varied populations are involved. Fowler (1969) noted that computers can store tremendous volumes of material and can retrieve them more rapidly and reliably than humans. Thus, while the average psychologist is typically limited in the research literature and population samples to which he or she is exposed and the information about them he or she can retain, the expert human interpreter can see to it that the computer adjusts for relevant demographic and other nontest variables.

The promise of the "automated clinician" has been realized in a number of studies, some employing the MMPI (e.g., Goldberg, 1965, 1970; Kleinmuntz, 1963) and many involving other types of clinical judgments (e.g., Bleich, 1973; DeDombal, 1979; Greist et al., 1973, 1974; McDonald, 1976; but see Blois, 1980; Kleinmuntz, 1968; Weizenbaum, 1976 for counterexamples). It comes as no surprise then, that automated clinicians to interpret psychological tests have proliferated. Several CBTI systems have been developed that interpret, but do not score, the Rorschach (Century Diagnostics, 1980; Exner, 1987; Harris,

Niedner, Feldman, Fink, & Johnson, 1981; Piotrowski, 1964). There has also been work on an interpretive program for the Holtzman Inkblot Technique (Holtzman, 1975), a projective technique that can also be computer-scored (Gorham, 1967). Automated clinical prediction systems have also been developed for the Halstead–Reitan Neuropsychological Battery (Adams, 1975; Finkelstein, 1977). By far the majority of automated interpretive systems have, however, been developed for objective tests. Fowler (1969) has suggested that this is because the administration, scoring, and interpretation of projective techniques is often highly individualistic and based heavily on intuition and clinical experience (cf. Exner & Exner, 1972). Scoring of ability tests such as the Halstead–Reitan often requires professional judgment. By contrast, objective tests have traditionally emphasized standardized administration and scoring, and have emphasized an objective, empirical approach to interpretation.

Of the objective tests, personality inventories have most often been the subjects of automated interpretation. The reasons for this are unclear, but I would speculate that it is due to the fact that data from many scales and indexes, as well as nontest data (e.g., demographic characteristics), are often combined to arrive at complex and lengthy interpretations (cf. Kleinmuntz, 1975). The complexity of this task allows for the fullest use of the advantages conferred by automation noted previously. Of these tests, computer interpretation of the MMPI has been most frequently attempted (Fowler, 1985).

It should come as no surprise then, that MMPI systems have been the subject of most investigations of the validity of CBTIs. These investigations appear to be representative of the few attempts to study the validity of clinical CBTIs and they will provide the focus for most of the remainder of this chapter (but see Adams, Kvale, & Keegan, 1984; Anthony, Heaton, & Lehman, 1980; Goldstein & Shelly, 1982; Green, 1982; Harris et al., 1981; Heaton, Grant, Anthony, & Lehman, 1981; Katz & Dalby, 1981; Klingler, Johnson, & Williams, 1976; Klingler, Miller, Johnson, & Williams, 1977; Moreland & Onstad, 1987a; Mules, 1972; O'Dell, 1972).

VALIDITY STUDIES TO DATE

To date the accuracy of clinical CBTIs has been evaluated in several ways. Some writers have compared CBTIs with test interpretations generated by human interpreters. Most of these comparisons have been anecdotal, often involving several automated interpretations but usually based on only a single case (Adair, 1978; Butcher, 1978; Dahlstrom et al., 1972; Eichman, 1972; Eyde, 1985; Goldstein & Reznikoff, 1971; Graham, 1977; Greene, 1980; Kleinmuntz, 1972; Labeck, Johnson, & Harris, 1983; Manning, 1971; Nichols, 1985; Sundberg, 1985a, 1985b). These comparisons are informative because of the extensive analysis they permit and the fact that the analysis is usually provided by a recognized

expert in MMPI interpretation. Obviously, however, this work lacks scientific rigor and, therefore, will not be considered further in this chapter. A few studies have compared CBTIs with human interpretations using more rigorous standards (Bringmann, Balance, & Giesbrecht, 1972; Glueck & Reznikoff, 1965; Johnson, Giannetti, & Williams, 1978). Reports prepared by human interpreters provide a poor criterion against which to judge the validity of CBTIs. The validity of clinicians' interpretations is low enough that a CBTI could be at serious variance with a clinician's interpretation and still be quite valid (cf. Golden, 1964; Graham, 1967; Kostlan, 1954; Little & Shneidman, 1959; Sines, 1959). There is also abundant evidence that clinicians may agree on the meaning of test scores although the presumed relationship between the test sores and the criterion does not, in fact, exist (e.g., Chapman & Chapman, 1967, 1969; Dowling & Graham, 1976; Golding & Rorer, 1972; Kurtz & Garfield, 1978). Hence, this approach will also not be discussed further here. A handful of writers have asked report consumers to fill out symptom checklists or complete Q-sorts based on CBTIs, subsequently comparing those ratings with analogous ratings made by clinicians familiar with each patient. Those studies will be considered subsequently. Most of the more rigorous studies that have employed nontest criteria have involved asking the recipients of CBTIs to rate the accuracy of various elements of the reports. Though disparaged by some writers (Lanyon, 1984; Matarazzo, 1983), these studies are considered promising by other experts (cf. Adair, 1978), especially if slightly modified (cf. Butcher, 1978; Moreland, 1985; O'Dell, 1972; Webb, Miller, & Fowler, 1970), and so merit further consideration.

External Criterion Studies

Several studies have compared rating scale or Q-sort data based on patient contact with the same data generated from computer-based MMPI interpretations. The first such study employed the Roche system (Anderson, 1969). In this study, 24 MMPI experts were asked to rate 12 psychological variables such as ego strength, impulsivity, and motivation for psychotherapy. The 12 variables were culled from a previously studied 27-item list on the basis of criterion rater's perceptions of their importance for treatment. The MMPI experts independently rated the patients' basic MMPI profiles and CBTIs. The patients' psychotherapists provided criterion ratings after 10 hours of individual psychotherapy or 30 days of inpatient treatment or both.

In several respects, this study was one of the best of its kind. A large sample of raters was employed (11 criterion raters, in addition to the 24 MMPI raters), and a comparatively large sample of MMPI respondents ($N = 28$) was studied. Moreover, each patient's basic MMPI profile and CBTI were rated by 6 individuals. Thus, although Anderson chose not to, assessment of interrater reliability of the report- and profile-based ratings was possible. In addition, the assess-

ment of individual differences in rater accuracy was possible. Anderson also took the unusual step of assessing the reliability, over 30 days, of the criterion ratings. The data were analyzed both within individuals, across variables and across individuals, within variables. The former analysis facilitated the detection of inaccurate reports, whereas the latter allowed Anderson to pinpoint variables that could not be accurately rated from the MMPI. If such had been the case, he also could have detected individuals or variables more accurately characterized by the human interpreters than by the CBTIs and vice versa. Anderson also collected average patient ratings from the MMPI raters in an attempt to deal with the problem of discriminant validity. He chose not to analyze those ratings, however, because the genuine ones were so poorly correlated with the criteria (mean r = .22).

Anderson did not fully use the multitude of MMPI-based ratings available to him. Knowing how well the average of the MMPI-based ratings or, alternatively, the most reliable ones, correlated with the therapists' ratings would have been useful, particularly because inspection of both the variables and some of Anderson's analyses suggest that some of the variables (e.g., ability to "stay with" feelings) were difficult to rate from the MMPI. The generalizability of the study was limited by the use of MMPI experts to render judgments, rather than using typical MMPI interpreters and CBTI consumers.

Hedlund, Morgan, and Master (1972) attempted to cross-validate the MMPI interpretive system developed at the Mayo Clinic and subsequently modified at the Institute of Living (Glueck, 1966). Two criterion raters completed a 33-item symptom checklist for each case by consulting the final summaries of 100 psychiatric inpatients at a military hospital. Disagreements were resolved by obtaining a consensus among the two raters and a third clinician. Checklist ratings were then compared with the 38 different statements (out of a possible 59) available from the patients' MMPI reports. Three interpretations were prepared for each patient, each based on a different set of MMPI norms.

A number of factors make this study a well-crafted attempt to validate a CBTI system. The sample of patients (N = 100) was the largest yet studied in this kind of research. Items were selected that could be rated with high reliability and that appeared especially relevant to the MMPI interpretations under evaluation. Expected relations of criteria to MMPI-based statements were established by consensus of the authors. A number of cases were rated prior to beginning the study to ensure adequate interrater reliability. Some of the cases chosen for the study were discarded before the data were analyzed because the raters believed that they had insufficient information on which to base their judgments or because the cases yielded low interrater agreement. The development of three different reports for each patient also allowed some estimation of the discriminant validity of the system.

The study of Hedlund et al. was not without some shortcomings, most notably the "file drawer" nature of the criterion data. Gdowski, Lachar, and Butkus

(1980) noted that data collected systematically at the time of evaluation often dramatically differs from the same ratings made from records. Moreover, when these differences occur, symptoms and behaviors usually are noted less frequently in records. Thus, the 62% false positive rate of Hedlund et al. might have been due to underrecording of the relevant data in the patients' records. Also important to keep in mind is that the MMPI data were obtained on admission, whereas the final summaries covered the patients' entire hospitalization. As a result, some report-based ratings (e.g., ratings of acute symptoms) might have been deemed inaccurate because they were compared with criterion ratings based on data collected long after the MMPI data. Although this criticism is highlighted in regard to the study by Hedlund et al., it also is applicable to some extent to many of the studies reviewed in this chapter.

The authors of the CBTI system examined by Hedlund and his colleagues could justifiably complain that a significant part of their system (36%) was ignored in the study. Although this shortcoming is common to all of the studies reviewed in this chapter, it is especially serious in regard to this study because of the small size of the interpretive statement library under consideration. Caldwell Report, by way of contrast, contains more than 30,000 sentences (A. B. Caldwell, personal communication, March 8, 1984), and other commercial services also claim large statement libraries.

Chase (1974) compared MMPI data with clinicians' ratings using a 59-item subset of the Minnesota–Ford phenotypic item pool (Meehl et al., 1962). Each patient's MMPI was interpreted in six different ways. MMPI experts wrote interpretive reports and, several weeks later, characterized the patients' MMPIs using the Minnesota–Ford items. Reports were prepared, using the actuarial atlas developed by Marks and Seeman (1963) and CBTIs were supplied by three commercial services: Roche, Behaviordyne (formerly called OPTIMUM), and Caldwell Report. All the reports were then characterized via ratings on the Minnesota–Ford items by 3 of 21 raters from four professions: clinical psychology, psychiatry, social work, and psychiatric nursing. Criterion ratings were supplied by two psychologists who either had worked with the patients or had studied their histories.

Chase's study is notable in that it involved more methods of interpreting the MMPI than any other study to date. Although Chase's method might be faulted because it was MMPI-based, her pool of rating items was selected carefully. She asked three MMPI experts to use the Minnesota–Ford items to rate the modal MMPI profiles for the three Marks and Seeman profile types under study. The items judged most and least characteristic of individuals producing the modal profiles were retained for the study. Consequently, unlike the other investigations discussed in this section, Chase can plausibly argue that her criterion items adequately covered at least the phenotypical behavioral domain germane to the reports studied. Her use of three raters for each report and two criterion raters also is noteworthy. The fact that she averaged the ratings across all raters before intercorrelating them considerably enhances confidence in the reliability of her

findings. Her data also allowed for assessment of interrater agreement, individual differences in rater accuracy, and differential accuracy among professions, although she chose not to explore those areas. Chase did present her data in the form of a multiinterpretation–multirating intercorrelation matrix, thus allowing an evaluation of both convergent and discriminant validity.

Another interesting facet of Chase's study is that she found a comparatively large difference between the accuracy of the ratings made from the psychologists' narrative interpretations (.32) and those same psychologists' rating-scale characterizations (.45). This shrinkage suggests that CBTIs are most fairly compared with interpretations generated in the traditional manner, not ratings made directly from test results.

A study similar to Chase's was performed by Crumptom (1974). She submitted the MMPIs of nine randomly selected patients being seen privately for psychotherapy to Caldwell Report, Roche, and the Institute for Clinical Analysis (Butcher, 1978). After four therapy sessions, each therapist characterized his or her patients via the Marks Q-Sort (cf. Marks & Seeman, 1963, Appendix C). Two recently graduated clinical psychologists and a clinical psychology graduate student who had completed all course work used the Q-sort to summarize each of the computer-based MMPI interpretations.

Crumpton's study is most noteworthy for her assessment of interrater reliability of the report ratings (as opposed to the criterion ratings). Her mean reliability coefficient of .62 suggests that this kind of reliability is indeed a factor to be considered in these studies. Validity coefficients in the .50s can hardly be faulted in the face of that kind of reliability! Like Chase, Crumpton averaged the report ratings across all raters before intercorrelating them; however, the criterion ratings were made by only one individual. Crumpton addressed the issue of discriminant validity by assessing the effects of shared patient stereotypes on the report raters' Q-sorts. The low mean interrater correlation of .22 suggests that commonly held stereotypes did not greatly influence Crumpton's results. Two further analyses also would have been of interest: (a) Would there have been as much disagreement about the typical patient among the therapists and between the therapists and the report raters? (b) How did the Q-correlations between the report-rater and therapist sorts compare with the correlations between the stereotype and therapist sorts? Crumpton's design also permitted her to assess therapist and patient-within-therapist effects in addition to the accuracy of the reports.

Crumpton's study, like Chase's, is subject to criticism on the ground of small sample size. This problem is compounded by the fact that the profiles of five of the nine MMPIs were very similar, and Crumpton's data indicate that they yielded very similar interpretations. Her study also can be faulted for using report raters familiar with the MMPI but with little clinical experience. Crumpton analyzed her data across subjects, within the nine conceptual categories of Q-sort items (cf. Marks & Seeman, 1963, Appendix C), but she used the categories as independent variables in an analysis of variance rather than as dependent variables in a multivariate analysis of variance.

Detailed next is a study conducted with the intention of capitalizing on the positive aspects of all foregoing work and improving upon it in several ways (Moreland, 1983). A large (N = 1186) initial sample was culled in an effort to gather a representative sample of interpretations. The final sample comprised 70 profiles, divided evenly among the five categories in Lachar's (1974b) MMPI profile typology: within normal limits, psychotic, neurotic, characterological, and indeterminate. Seasoned clinicians who were not familiar with the MMPI were solicited as report raters. Assurance was obtained that none of the raters had used either of the computer services under investigation—Roche and Lachar's CBTI system, which was first sold by Automated Psychological Assessment and is now sold by Western Psychological Services (Lachar, 1974b)—because such prior experience could bias the ratings made in the study. Moreover, no report rater received two reports on the same patient to avoid a recognition problem that could contaminate the report ratings. Criterion ratings were made at the time the patient took the MMPI. Discriminant validity was assessed by comparing report-based ratings with "stereotypical patient" ratings. Both interrater and intrarater (report) reliability data were collected. Profile type was employed as an independent variable in the data analyses.

Needless to say, this study did not avoid all of the shortcomings of its predecessors. To obtain a relatively large and diverse sample, data previously collected for other purposes had to be used. As a result, reliability data were not available for the criterion ratings. The criterion instruments themselves also were less than optimal for a study of computer-based MMPI interpretations. As in Anderson's study, both inspection of the variables and some of the analyses suggest that some of the variables were very difficult to rate from the MMPI.

Another factor noted in this study that may contribute to the low validity coefficients commonly found in studies of this type was the poor metric qualities of the criterion instruments. None of the distributions of criterion ratings approximated normality—a finding typical of psychiatric rating scales (Maxwell, 1971)—whereas the CBTI-based ratings did. If the report raters had received information about the score distributions characteristic of the criterion instruments or, better yet, if they had received actual base rate data, the validity coefficients might have averaged higher than .36. The report raters complained of another metric problem. They pointed out the difficulty of converting terms such as "mild" and "often" into metric ratings. The low interrater reliabilities obtained in this study (generally in the .50s) also attest to this problem. The problem could have been alleviated in two ways. First, pilot cases could have been employed in the manner of Hedlund et al. to ensure that all raters meant the same thing when they checked a statement (e.g., "mild X"). Second, contrary to the assumption made when this study was designed, report raters should have received as much experience as possible with the two CBTI systems prior to beginning the study. In that way, some assurance would have been gained that the raters knew what "severe Y" in a test interpretation looked like in a patient.

A manipulation check suggested that some raters may not have been weighing the various parts of the reports in the same manner as typical consumers. For example, one rater reported that she ignored the entire narrative, considering only the listing of critical items endorsed by the MMPI respondent. This finding calls into question the external validity of the study.

In closing, the most serious shortcoming of the foregoing study and, in fact, of all of the external criterion studies is that none actually evaluated an entire interpretive system, although the investigator's conclusions often suggest they did so. Not only did these studies evaluate only small proportions of the statements available in the interpretive systems but they usually did so using criterion instruments that did not adequately map the behavioral domain covered by the systems.

Having personally attempted an external criterion study, I now believe many of the problems that have been noted are, as a practical matter, insurmountable. Future attempts to validate clinical (as opposed to actuarial) CBTIs are likely to produce more useful data if the external criterion method is abandoned in favor of the "customer satisfaction" method described below.

Customer Satisfaction Studies

The early work in this area was conducted to assess the CBTI system for the MMPI that was developed by Fowler and later sold by the Roche Psychiatric Service Institute and, in a slightly embellished version, by Psychological Assessment Services (Adair, 1978; Butcher, 1978).[2] Webb and his colleagues (Webb, 1970; Webb, Miller, & Fowler, 1969, 1970) asked consumers to use a 5-point rating scale to indicate each report's clarity, accuracy, and usefulness and to note how the CBTIs compared with reports prepared in the usual manner. The specific areas explored in one of these studies can be found in Table 3.2. Bachrach (cited in Fowler, 1966) also studied Fowler's MMPI reports; however, Bachrach asked raters only for a single set of ratings for a group of reports. The foregoing studies, as an aggregate, involved a large, diverse array of clinical raters and patients. Webb and colleagues' use of numerous queries about each CBTI improved upon Bachrach's request for a single set of ratings for a group of reports.

Although useful, these studies were not without major faults. Lachar (1974a) noted that because the reports were rated according to content areas (e.g., psychosomatic symptoms) rather than statement by statement or paragraph-by-paragraph, systematic isolation of weaknesses in the CBTI system was difficult. Some of the studies were large enough to permit breakdown of the ratings according to test or patient characteristics (e.g., MMPI profile type or clinical

[2]Similar studies have been conducted to evaluate European adaptations of both Fowler's system (Fowler & Blaser, 1972) and Lachar's system (Engel, 1977).

TABLE 3.2
Questions Used in Some of the Validation Studies of Fowler's MMPI Interpretation System

The report is well organized and its descriptions are clear.
The report gives a valid overall description of this patient.
The behaviors described are characteristic of this person.
The report is helpful in planning this patient's treatment.
The symptoms reported are accurate.
I could find little good in this report.
Major symptoms of this person are omitted.
The report is in error regarding this person's physical complaints (if described in the report).
This person's mood and feelings are accurrately described.
The report misrepresents this person's relations with family members (if described in the report).
Useless information was included.
The severity of personality desorder was accurately described.
Parts of the report contradicted each other.
The report's prediction of response to therapy was accurate (if described in the report).
The report pointed out things about the patient I had not noticed previously.
I know this patient: very well, well, moderately, somewhat, scarcely at all.
This report, compared with most noncomputerized psychological reports I have seen is: much worse, worse, equal, better, much better.

Note. Adapted from Table 1, Webb, Miller, and Fowler (1970). Unless otherwise noted, raters indicated: strongly disagree, mildly disagree, neutral, mildly agree, strongly agree.

diagnosis). If this had been done, the detection of inaccurate report types or types of patients for whom the reports were inaccurate would have been possible.

Three studies have been conducted to assess the adequacy of the CBTI system for the MMPI developed by Gilberstadt (1970) for the Veterans Administration. Klett (1971) conducted a study virtually identical in approach to that of Bachrach. Thus, the same comments apply to both. The other two studies were conducted by Lushene and Gilberstadt (1972). In their initial study, they collected accuracy ratings on each interpretive statement. They also collected overall report-accuracy ratings on a 6-point scale. They then revised those statements that were rated as accurate less than 60% of the time. The revised system was then studied in the same manner.

The outstanding feature of the work by Lushene and Gilberstadt is that they conducted a second study to assess the adequacy of the revisions prompted by the first. Lushene and Gilberstadt's studies were similar in method to those conducted by Webb et al. Therefore, the same criticisms apply with the exception of one. Because Lushene and Gilberstadt asked raters to judge each statement in each report, they were able to pinpoint weaknesses in Gilberstadt's system. A criticism unique to Lushene and Gilberstadt's studies involves their rating procedure. They asked raters to check one of eight adjectival phrases to describe each interpretive statement: correct, incorrect, irrelevant, redundant, contradictory, base rate, unclear, and don't know. The raters, perhaps believing the accuracy or inaccuracy of the statements to be the crucial consideration, selected the correct and incorrect categories an average of 91% of the time. Unfortunately, the eight categories were not mutually exclusive. For example, correct and incorrect overlapped with redundant and contradictory. The studies would have

been more informative had raters been requested to make all applicable ratings (e.g., indicate that statements were both correct and redundant).

Lachar (1974a, 1974b) was able to overcome some of the shortcomings of the studies of the systems developed by Fowler and Gilberstadt in his initial attempt to demonstrate the validity of his CBTI system. Drawing a lesson from the work of Lushene and Gilberstadt, Lachar asked his raters to indicate whether each paragraph of each report was accurate or inaccurate. He also asked that the overall accuracy of each report be rated on a 6-point scale. Moreover, Lachar used a factorial design that included both MMPI and patient characteristics as independent variables. This approach allowed him to determine that some paragraphs in his system were relatively inaccurate, compared with other elements of the system, particularly for certain types of MMPI profiles and certain types of patients.

An outstanding feature of Lachar's study (1974a) is that the accuracy of each interpretive paragraph (the unit of selection is his system) was independently assessed. His conclusions receive added force by the large sample ($N = 1410$) used in the study, which included subsamples from several populations. These positive aspects of the study are tempered somewhat by the fact that 75% of the patients were men and 85%, patients in military medical facilities. Hence, Lachar's sample was not representative of medical and psychiatric patients in the United States, the population with which his reports currently are used.

Two studies of Lachar's system used slight twists on his original methodology. Adams and Shore (1976) completed a partial replication of Lachar's initial study. Their small sample ($N = 100$) did not permit a factorial design, but they asked more of their raters than did Lachar. Each paragraph was rated on a 6-point scale. This innovation allowed Adams and Shore to note that paragraphs containing specific predictions or treatment recommendations usually were given extreme ratings, whereas the ratings of general statements were more evenly distributed. Lachar, Klinge, and Grisell (1976) had clinicians rate the overall accuracy of two types of reports for each of their adolescent patients. One report was based on standard MMPI norms and the other on adolescent norms. This approach permitted the researchers to conclude that Lachar's system was most useful with adolescents when age-appropriate norms were employed.

Although the studies of Lachar's system improved on the investigations of Fowler and Gilberstadt systems, they also contained some weaknesses not apparent in the latter studies. Most important, Lachar (1974b, p. 159) instructed his raters to consider his paragraphs accurate when no criterion information was available. This raises the possibility that some elements of Lachar's system received spuriously high ratings due to a frequent absence of relevant criterion information. Two factors heighten this concern. First, Lachar's article indicates that some ratings were made after as little as 1 hour of contact with the patient. (Limited patient contact is a problem in most of the studies reviewed in this chapter.) Second, some of Lachar's interpretations appear to be impossible to

judge without a great deal of information, sometimes longitudinal in nature (see Table 3.3). By contrast, Webb et. al offered, and their raters frequently used, a neutral category that "may [have] represent[ed] a rater's unfamiliarity with the patient" (1970, p. 212). Though seldom used, a don't-know category also was available to those rating Gilberstadt's interpretations.

Less important criticisms of the work of Lachar and his colleagues involved the assessment of report usefulness instead of report validity per se. Asking raters to indicate simply the accuracy or inaccuracy of each paragraph and each report, rather than using multifaceted ratings such as those employed by Webb and his colleagues, involved a tradeoff. It allowed inaccurate paragraphs to be pinpointed but did not permit the identification of those reports that omitted important information. (This same criticism also may apply to the work of Lushene and Gilberstadt, although it cannot be established on the basis of their report.) Lachar's raters also could not point out useless information.

In her doctoral dissertation, Chase (1974) employed the customer satisfaction approach to CBTI validity as an adjunct to the external criterion work described earlier. Clinicians familiar with the patients rated the accuracy of the interpretations globally on a 5-point scale. The Roche and Caldwell reports were judged superior, whereas those from Behaviordyne, poor. The evaluation of the same reports using external criterion ratings reversed this trend, however (see External Criterion Studies section). Although the scope of Chase's study was limited, her findings indicate that the results of most customer satisfaction studies are best viewed skeptically.

Chase's study is unique in gathering both global report ratings and using external criterion ratings. The fact that Chase studied three different CBTIs is also a plus. The selection of cases that cover a broad range of psychopathology is another positive feature of her study, although the examination of only three MMPIs severely restricts the generality of any conclusions drawn from the study.

TABLE 3.3
Excerpts from Lachar's CBTI System for the MMPI in Which Accuracy Appears Difficult to Rate

Response to chemotherapy, psychotherapy, and environmental manipulation is often good.
Rationalization and Intellecutalization are common defense mechanisms.
Chronic adjustment utilizing repression, denial, somatization, and a passive-dependent orientation make any psychological intervention, except temporary supportive measures extremely difficult.
Inconsistency and unpredictability are characteristic. These individuals appear demanding and resistant in therapy.
While the insight these persons show may be good and their protestation of resolve to do better seem genuine, long-range prognosis for behavior change is poor.
These individuals are attempting to deny lowered abilities through overactivity and over-production.
Hostility is likely to be expressed in an indirect manner.
Excessive fantasy is often used as an escape from the direct expression of unacceptable impulses.

Adapted from Lachar (1974b).

TABLE 3.4
Areas Explored in Green's Study of CBTI Validity

Information Adequacy[1]
1. Confirmation of knowledge
2. Addition of relevant information
3. Clarification of case
4. Exclusion of important information
5. Inclusion of trivial information
6. Inclusion of misleading information

Descriptive Accuracy[2]
1. Interpersonal attitudes and relationships
2. Affective tone and moods
3. Personality traits and behaviors
4. Self-image
5. Primary symptoms and complaints
6. Styles of coping
7. Stress or areas of conflict
8. Throght processes
9. Severity of disturbances

Report Format and Utility[2]
1. Internal Consistency
2. Organization
3. Intelligibility and clarity
4. Helpful in treatment

Adapted from Green (1982).
[1] Raters indicated: substantial, moderate, minimal, none.
[2] Raters indicated: excellent, good, adequate, poor.

Green (1982) compared the accuracy and usefulness of MMPI CBTIs with reports from Millon's CBTI system for the Millon Clinical Multiaxial Inventory (MCMI; Millon, 1982). Her 23 raters rated 100 Roche reports, 100 MCMI reports, and 50 Mayo Clinic reports, using a set of 19 thoughtful queries about information adequacy, descriptive accuracy, and report format and utility (see Table 3.4).

Green's study was unique and pioneering in two respects. First, she compared CBTIs based on two different tests. Her study is useful in pointing up the dangers of doing so. The MCMI was designed to assess the personality styles hypothesized by Millon (cf. Millon, 1981). Thus, it should come as no surprise that the MCMI CBTIs were superior when it came to describing personality traits and coping styles. On the other hand the CBTIs based on the MMPI, which was built primarily to assess major mental illness, provided the most accurate descriptions of primary symptoms and though processes. It should also come as no surprise that the two consultative CBTIs (Roche and MCMI) outstripped the screening CBTI (Mayo) virtually across the board. When setting up a horse race of this sort it is important to make sure that none of the horses are hobbled. Another pioneering feature of Green's study was her effort to rule out base-rate accuracy as an important influence on her results. Of that, more to come. A further positive aspect of Green's study was her effort to make sure her raters were knowledgeable about the clients whose reports they rated. She required that the raters have

at least 4 hours of client contact prior to rating the reports. Meehl (1960) has demonstrated that clinicians' views of clients stabilize after 4 hours.

Vincent and Castillo (1984) asked 13 nurses and 1 social worker to indicate their preference for Lachar's CBTI or one developed by the first author (Vincent, Wilson, & Wilson, 1983). Specifically, the raters were asked to "rate [the CBTIs] as to whether you prefer A or B, or A and B are equal, taking into account the report's overall consistency, organization, clarity, readability, and . . . overall usefulness" (p. 30). They were asked to rate reports only for those patients with whom they had "significant personal contact." These instructions led to ratings of pairs of reports on 32 patients out of 50 that were originally eligible for the study.

This study is most noteworthy for its explicitly ipsative, "horse race" character. The results indicated that the raters felt Vincent's CBTI to be superior to Lachar's in most instances but we have no way of knowing, in any absolute sense, how satisfactory they felt either report to be. On the other hand, confidence in the ratings that were made is increased by the fact that the raters were asked to, and did, decline to rate reports on patients with whom they were unfamiliar.

Widespread Problems

Reviewers appear to agree that the major shortcoming of the customer satisfaction studies is what Webb et al. (1970) characterized as the lack of information on base-rate accuracy of the reports (cf. Butcher 1978; Eichman, 1972). Webb and his colleagues were concerned that raters would characterize reports as accurate not because the reports were pointed descriptions of the individuals at issue, but rather because they contained glittering generalities (cf. Baillargeon & Danis, 1984). Butcher (1978) offered the following colorful description of this problem:

> The problem here is very similar to the situation presented by the overzealous, rookie policeman who blows a case by prejudicing the witness as follows: The policeman takes a photograph (and only one photograph) of the suspect to the prime witness and asks if this is the person who committed the crime. Even the police, whose methods and intent are frequently questioned, do not try to get away with this type of validation. Most often they are required to utilize more rigorous methods of gathering evidence that will hold up in court, such as "having the witness pick the guilty person from a lineup." (pp. 617–618)

I referred to this same issue, in the context of the external criterion studies, as the problem of discriminant validity.

This concern is lent credibility by Chase's finding that global accuracy ratings sometimes disagree sharply with the results of external criterion ratings. Thus, the customer satisfaction studies reviewed so far provide only half of the picture. They may correctly indicate that CBTIs have high convergent validity, but they

afford little or no information concerning the reports' discriminant validity. The focused questions employed in the evaluations of Fowler's system—especially the one dealing with the inclusion of useless information—might have reduced this problem (see also Green, 1982); it is doubtful that they completely eliminated the problem. Lushene and Gilberstadt's provision of irrelevant and base-rate categories might have ameliorated this problem had raters used these categories more often. Lachar and his colleagues (1976) were afforded some protection from this problem by their request that clinicians rate two reports on each patient. Although the clinicians's ratings of the two types of reports differed only slightly, the reports themselves frequently differed radically (Lachar et al., 1976, Table 2, p. 22). It may be argued that Chase's use of three different CBTI systems allowed some appraisal of base-rate accuracy; however, this argument ignores the fact that differential ratings may result from differences among the reports irrelevant to the question of their validity. Indeed, the comments of Chase's raters provide support of this hypothesis. They complained about the infelicitous use of the English language in the low-rated Behaviordyne reports and praised Caldwell's use of the same. When, on the other hand, the reports were subjected to scrutiny via external criteria, Behaviordyne was found superior.

Green (1982) made the first self-conscious effort to deal with the problem of base-rate accuracy. She had 32 clinical psychology graduate students simulate the responses of two different types of patients on the MCMI. The students then rated the accuracy of two CBTIs, one generated from one of the tests completed by the student and one, with the student unaware, selected at random. Green reported that the students rated the genuine reports excellent or good more than three times more often than the random reports. Notwithstanding the work involved, this approach to the problem of base-rate accuracy is flawed in several ways.

First, the subjects were not clinical clients. They were graduate students with considerable exposure to Millon's personality theory. Their MCMI responses could be expected to reflect those of prototypical patients of the sort they were simulating. Such prototypical cases seem to be the exception rather than the rule in clinical practice, as demonstrated by the poor classification rates usually obtained using the MMPI cookbook prototypes discussed earlier. It is also important to note that the students were rating reports ostensibly based on their own responses to the test. Thus, they were a giant step closer to the raw test responses that led to the CBTIs than are clinicians evaluating clients' test responses. This problem seems especially salient when one recalls that Chase (1974) experienced a 50% decrease in percentage of variance accounted for when taking the step from Q-sorts based on MMPI profiles to Q-sorts by other raters based on narrative reports. Finally, the graduate students were not clinical clients, nor were they the full-fledged practitioners who served as raters in the main part of the study. Because of these problems, Green's efforts can probably best be thought of as yielding a lower-bound estimate of the influence of base rate accuracy in

studies of this type. She found that the genuine CBTIs were rated good or excellent more than three times as frequently as the randomly chosen reports.

A recent study of mine provides direct evidence of the importance of assessing the degree to which high base rates contribute to high accuracy ratings (Moreland & Onstad, 1985, 1987b). Seven psychologists and one psychiatrist rated the accuracy of each section of 86 pairs of reports generated by the Minnesota Report: Adult Clinical System developed by Butcher (University of Minnesota, 1982). One report was based on the patient's MMPI profile while the other was based on a test profile similar to, but not the same as, the patient's. Raters believed they were rating one CBTI prepared in the usual manner and one "experimental" CBTI. They did not know which was which. The results of that study are presented in Table 3.5. Those results clearly indicate the importance of having a means of assessing CBTIs' discriminant validity. A recent study by Wimbish (1985) supports this point.

A second serious question about the studies under discussion involves reliability: None of the foregoing customer satisfaction studies assessed the reliability of the ratings across either raters or time. The work of Eyde, Kowal, and Fishburne, detailed elsewhere in this volume, makes it clear that this is an important consideration. The average reliability of pairs of raters for their four cases ranged from .16 to .49. On the other hand, their ratings reached acceptable levels of reliability

TABLE 3.5
Comparative Validity of Genuine and "Experimental" Minnesota Report CBTIs by Section

	Percentage "Accurate"[1]				
Report Section	*Genuine Report*	*"Experimental" Report*	*G-"E"*	z^2	p^3
Profile validity	90% (70/78)	79% (60/76)	11%	1.90	.0300
Symptomatic pattern	74% (62/84)	43% (35/81)	31%	4.08	.0001
Interpersonal relations	80% (61/76)	61% (50/82)	19%	2.60	.0050
Behavioral Stability	90% (65/72)	75% (59/79)	15%	2.38	.0090
Diagnostic Considerations	82% (56/68)	48% (33/69)	34%	4.15	.0001
Treatment Considerations	76% (56/74)	53% (40/75)	23%	2.91	.0020

Adapted from Moreland and Onstad (1985, 1987b).

[1] Accurate/Accurate + Inaccurate.

[2] Test of the difference between correlated proportions.

3 One-tailed.

TABLE 3.6
High- and Two-Point Code Paragraphs Rated Fewer than Ten Times in Lachar's CBTI System for the MMPI

Scale	*Rule*	*Number of Ratings*
1	>69T	5
7	>69T	6
8	>69T	5
1/6	both >69T	0
1/7	both >69T	6
1/9	both >69T	7
4/3	both >69T, and 4 > 3 by at least 6T	6
3/6	both >69T	4
3/7	both >69T	6
3/8	both >69T	8
3/9	both >69T	7
6/7	both >69T	2
6/9	both >69T	9
7/8	both >69T and 7 > 8 by at least 6T	6
7/9	both >69T	4

Adapted from Lachar (1974b). Patient sample size = 1410; High- and 2-point code paragraph sample size = 51. For high-point codes other clinical scales must be < 70 T; for 2-point codes other clinical scales must be less than or equal to those in the code, ties broken as in the Welsh Code.

when aggregated across raters (range = .70–.92). Taken as aggregates, the studies of the systems developed by Fowler, Gilberstadt, and Lachar involved relatively large, diverse groups of raters. A fair speculation is that such groups might have reduced the problem presented by the lack of data on reliability across raters; however, a large number of raters does not render interrater reliability data completely unnecessary. Consider that even in Lachar's (1974a, 1974b) large study, 15 of the 40 paragraphs composing the heart of his system were rated less than 10 times (see Table 3.6). To be sure, most of these paragraphs pertain to rare configurations of scores, but several pertain to configurations that are quite common in some settings. This problem can only have been much worse in the other, smaller studies reviewed in this section.

The reliability of the reports themselves, both across time and internally, also merits consideration (cf. Hofer & Bersoff, 1983). Because test scores and configurations of test scores are unreliable over time (e.g., Graham, Smith, & Schwartz, 1986), CBTIs are likely to be unreliable, too. The unfailing accuracy with which computers apply rules makes reliability of reports across time a significant consideration because even a 1-point difference on a single scale can cause a radical change in a CBTI (see Table 3.7). Through provision of a contradictory category, Lushene and Gilberstadt did attempt to investigate the internal consistency of Gilberstadt's interpretations. Given the apparent frequency with which CBTI consumers comment on internal inconsistencies, that other researchers have not investigated this problem is surprising.

Problems with the report raters also made the studies reviewed in this section less useful than they might have been. A number of the raters were not usually

TABLE 3.7
Interpretations of Two Very Similar Profiles by Lachar's CBTI System for the MMPI

Clinical Profile 1: 2, 3 > 69*T*; 1, 4, 8, 9 < 65*T*; 5, 6, 7 0 < 60*T*:

Individuals who obtain similar profiles are characterized by the ineffective use of repressive defenses and hysteroid mechanisms. Such individuals may show symptoms of apathy, dizziness, and lowered efficiency as well as symptomatic depression. Chronic tension, feelings of inadequacy and self-doubt, bottled-up emotion and general over control are frequently characteristic. He or she may have a hysterical quality. Sexual maladjustment, immaturity and dependency are often characteristic. In general these individuals have little insight, are resistant to psychodynamic formulations of their problems and have little genuine motivation to seek help.

Neuroses are common and characterological impressions are rare. Prognosis is poor.

Clinical Profile 2: 2 > 69*T*; 3 = 69*T*; 1, 4, 8, 9 < 65*T*; 5, 6, 7, 0 < 60*T*

Individuals who obtain similar profiles are often significantly depressed, worried and pessimistic. Feelings of inadequacy and self-depreciation are likely present. These people internalize stress and usually withdraw when put under pressure. An acute reactive depression is suggested. If depression is denied by this patient, its effects should still be carefully evaluated. Response to chemotherapy, psychotherapy and environmental manipulation is often good.

Reactive depression is suggested.

Note. Adapted from Lachar (1974b).

direct consumers of computer-based MMPI interpretations (e.g., nurses). In addition, a number of raters were students (e.g., psychiatry residents) who probably did not possess an expertise in evaluating the reports that would be commensurate with that of fully qualified clinicians. Finally, none of the studies examined other potential rater effects. For example, biologically oriented psychiatrists could be envisioned as giving high marks to those statements suggesting chemotherapy and low ratings to those with psychodynamic inferences. The converse may hold true for psychoanalytically oriented clinicians, regardless of the accuracy of the statements.

HOW TO VALIDATE "AUTOMATED CLINICAL" CBTIS

Consideration of the pros and cons of the customer satisfaction validation studies completed to date precipitated the formulation of this list of desirable characteristics of future such studies, some of which also can be found elsewhere (e.g., Harris, 1984; Hofer & Bersoff, 1983; Moreland, 1985, 1987):

1. Raters should have prior experience with the interpretive systems under study.
2. Raters should have prior experience with the ratings they are to make.
3. The sample of raters should be representative of those using the report in applied contexts. The sample can be random or stratified, depending on the inferences one wishes to draw.

The relaxation of Guidelines 1–3 may be useful in some cases. For example, attempts to validate jargon-free CBTIs based on normal personality tests may make advantageous the use of ratings completed by individuals who know the test respondent well or the test respondent.

4. The sample of test respondents (or interpretations) should be representative of those found in applied settings. The sample can be random or stratified, depending on the inferences one wishes to draw.

5. Ratings should be completed keeping the appropriate time frame in mind. For example, care should be taken to ensure that ratings of acute symptoms are made, considering only that phase of a patient's illness.

6. Discriminant validity of the interpretations should be assessed. This guideline can be fulfilled by having each rater judge two reports (per test respondent) from the same interpretive system, one of the reports being genuine and the other, bogus. Of course, raters should not know which report is which until after competing the ratings.

7. Interrater reliability should be assessed. Raters should be given access to the same criterion information (e.g., case records).

8. Intrarater reliability should be assessed. Some of the inferences made in CBTIs may remain valid for only a short period of time due to actual changes in the test respondent. Hence, intrarater reliability should be assessed over a short period of time. Raters also should be asked to keep in mind when the test was administered when they are making reliability ratings.

9. Reliability, across time, of the CBTIs themselves should be assessed.

10. Studies should make provisions for indicating contradictory elements of interpretations.

11. Studies should make provisions for indicating useless elements of interpretations.

12. Studies should make provisions for indicating when interpretations omit significant information as well as the nature of that information. Studies with this aim should employ expert test interpreters either to rate the CBTIs or to decide, post hoc, whether the interpretations could have been expected to include such information.

13. Each element of an interpretive statement that is produced by different decision rules should be assessed independently.

EPILOGUE

The attentive reader will have noticed that I have not critiqued the three most recent customer satisfaction studies (Eyde, Kowal, & Fishburne, this volume; Moreland & Onstad, 1985, 1987b; Wimbish, 1985) in detail, as I did the earlier studies. The three most recent studies were designed with the advice offered in

my 1985 article in mind. I invite the reader to evaluate for oneself the degree to which those three studies improved upon their predecessors.

ACKNOWLEDGMENT

This chapter represents the views of the author and not necessarily those of National Computer Systems.

REFERENCES

Adair, F. L. (1978). Computerized scoring and interpreting services [Re: Minnesota Multiphasic Personality Inventory]. In O. K. Buros (Ed.), *Eighth mental measurements yearbook* (Vol. 1, pp. 940–942, 945–949, 952–953, 957–960). Highland Park, NJ: Gryphon Press.

Adams, K. M. (1975). Automated clinical interpretation of the neuropsychological test battery: An ability based approach. *Dissertation Abstracts International, 35,* 6085B. (University Microfilms No. 75–13, 289).

Adams, K. M., Kvale, V. I., & Keegan, J. R. (1984). Relative accuracy of three automated systems for neuropsychological interpretation based on two representative tasks. *Journal of Clinical Neuropsychology, 6,* 413–431.

Adams, K. M., & Shore, D. L. (1976). The accuracy of an automated MMPI interpretation system in a psychiatric setting. *Journal of Clinical Psychology, 32,* 80–82.

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.

Anderson, B. N. (1969). *The utility of the Minnesota Multiphasic Personality Inventory in a private psychiatric hospital setting.* Unpublished master's thesis, Ohio State University.

Anthony, W. Z., Heaton, R. K., & Lehman, R. A. W. (1980). An attempt to cross-validate two actuarial systems for neuropsychological test interpretation, *Journal of Consulting and Clinical Psychology, 48,* 317–326.

Baillargeon, J., & Danis, C. (1984). Barnum meets the computer: A critical test. *Journal of Personality Assessment, 48,* 415–419.

Ben–Porath, Y. S., & Butcher, J. N. (1986). Computers in personality assessment: A brief past, an ebullient present, and an expanding future. *Computers in Human Behavior, 2,* 167–182.

Bleich, H. L. (1973). The computer as consultant. *New England Journal of Medicine, 223,* 308–312.

Blois, M. S. (1980). Clinical judgment and computers. *New England Journal of Medicine, 303,* 192–197.

Briggs, P. F., Taylor, M., & Tellegen, A. (1966). *A study of the Marks and Seeman MMPI profile types as applied to a sample of 2,875 psychiatric patients* (Research Laboratories Report No. PR–66–5). University of Minnesota, Department of Psychiatry.

Bringmann, W. G., Balance, W. D. G., & Giesbrecht, C. A. (1972). The computer vs. the technologist: Comparison of psychological reports on normal and elevated MMPI profiles. *Psychological Reports, 31,* 211–217.

Butcher, J. N. (1978). Computerized scoring and interpreting services [Re: Minnesota Multiphasic Personality Inventory]. In O. K. Buros (Ed.), *Eighth mental measurements yearbook* (Vol. 1, pp. 942–945, 947–956, 958, 960–962). Highland Park, NJ: Gryphon Press.

Butcher, J. N. (Ed.). (1985). Perspectives on computerized psychological assessment (special series). *Journal of Consulting and Clinical Psychology, 53,* 746–848.

Butcher, J. N. (Ed.). (1987). *Computerized psychological assessment: A practitioner's guide*. New York: Basic Books.

Caldwell, A. B. (1970). *Recent advances in automated interpretation of the MMPI*. Paper presented at the fifth annual Symposium on Recent Developments in the Use of the MMPI, Mexico City.

Campbell, D. P. (1971). *Handbook for the Strong Vocational Interest Blank*. Stanford, CA: Stanford University Press.

Century Diagnostics. (1980). *Computer interpreted Rorschach*. Tempe, AZ: Author.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72,* 193–204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271–280.

Chase, L. L. S. (1974). An evaluation of MMPI interpretation systems. *Dissertation Abstracts International, 35,* 3009B. (University Microfilms No. 74–26, 172).

Colby, K. M. (1980). Computer psychotherapists. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), *Technology in mental health care delivery systems*. Norwood, NJ: Ablex.

Cone, J. D. (1966). A note on Marks' and Seeman's rules for actuarially classifying psychiatric patients. *Journal of Clinical Psychology, 22,* 270.

Crumpton, C. A. (1974). An evaluation and comparison of three automated MMPI interpretive reports. *Dissertation Abstracts International, 35,* 6090B. (University Microfilms No. 75–11, 982).

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook. Vol. 1: Clinical applications*. Minneapolis: University of Minnesota Press.

DeDombal, F. T. (1979). Computers and the surgeon: A matter of decision. *The Surgeon, 33,* 57.

Dowling, J. F., & Graham, J. R. (1976). Illusory correlation and the MMPI. *Journal of Personality Assessment, 40,* 531–538.

Drake, L. E., & Oetting, E. R. (1959). *An MMPI codebook for counselors*. Minneapolis: University of Minnesota Press.

Eichman, W. J. (1972). Computerized scoring and interpreting services [Re: Minnesota Multiphasic Personality Inventory]. In O. K. Buros (Ed.), *Seventh mental measurements yearbook* (Vol. 1, pp. 105–110). Highland Park, NJ: Gryphon Press.

Engel, R. R. (1977, August). *Cross-national accuracy of automated MMPI reports*. Paper presented at the sixth World Congress of Psychiatry, Honolulu.

Exner, J. E., Jr. (1987). Computer assistance in Rorschach interpretation. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 218–235). New York: Basic Books.

Exner, J. E., & Exner, D. E. (1972). How clinicians use the Rorschach. *Journal of Personality Assessment, 36,* 403–408.

Eyde, L. D. (1985). Review of the Minnesota Report: Personnel Selection systems for the MMPI. In J. V. Mitchell, Jr. (Ed.), *Ninth mental measurements yearbook* (Vol. 2, pp. 1003–1005). Lincoln, NE: Buros Institute of Mental Measurements.

Eyde, L. D. (Ed.). (1987). *Computerised psychological testing*. London: Lawrence Erlbaum Associates.

Finkelstein, J. N. (1977). BRAIN: A computer program for interpretation of the Halstead–Reitan Neuropsychological Test Battery. *Dissertation Abstracts International, 37,* 5349B. (University Microfilms No. 77–88, 8864).

Finney, J. C. (1966). Programmed interpretation of MMPI and CPI. *Archives of General Psychiatry, 15,* 75–81.

Fowler, R. D. (1966). *The MMPI notebook: A guide to the clinical use of the automated MMPI*. Nutley, NJ: Roche Psychiatric Service Institute.

Fowler, R. D. (1969). Automated interpretation of personality test data. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications* (pp. 105–126). New York: McGraw–Hill.

Fowler, R. D. (1985). Landmarks in computer-assisted psychological assessment. *Journal of Consulting and Clinical Psychology, 53,* 748–759.

Fowler, R. D. (1987). Developing a computer-based test interpretation system. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 50–63). New York: Basic Books.

Fowler, R. D., & Athey, E. B. (1971). A cross-validation of Gilberstadt and Duker's 1–2–3–4 profile type. *Journal of Clinical Psychology, 27,* 238–240.

Fowler, R. D., & Blaser, P. (1972). *Around the world in 566 items.* Paper presented at the seventh annual Symposium on Recent Developments in the Use of the MMPI, Mexico City. Cited in J. N. Butcher & P. Pancheri (1976), *A handbook of cross-national MMPI research* (pp. 194–196). Minneapolis: University of Minnesota Press.

Fowler, R. D., & Butcher, J. N. (1986). Critique of Matarazzo's views on computerized testing: All sigma and no meaning. *American Psychologist, 41,* 94–96.

Fowler, R. D., & Coyle, F. A. (1968). Scoring error on the MMPI. *Journal of Clinical Psychology, 24,* 68–69.

Gdowski, C. L., Lachar, D., & Butkus, M. (1980). A methodological consideration in the construction of actuarial interpretation systems. *Journal of Personality Assessment, 44,* 427–432.

Gilberstadt, H. (1970). *Comprehensive MMPI code book for males* (MMPI Research Laboratory Rep. No. 1B 11–5). Minneapolis: Veterans Administration Hospital.

Gilberstadt, H., & Duker, J. (1965). *A handbook for clinical and actuarial MMPI interpretation.* Philadelphia: W. B. Saunders.

Glueck, B. C., Jr. (1966). Current personality assessment research. *International Psychiatric Clinic, 3,* 205–222.

Glueck, B. C., Jr., & Reznikoff, M. (1965). Comparison of computer-derived personality profile and projective psychological test findings. *American Journal of Psychiatry, 121,* 1156–1161.

Goldberg, L. R. (1965). Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs, 79*(9, Whole No. 602).

Goldberg, L. R. (1970). Man vs. model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422–432.

Golden, M. (1964). Some effects of combining psychological tests on clinical inferences. *Journal of Consulting Psychology, 28,* 440–446.

Golding, S. G., & Rorer, L. (1972). Illusory correlation and subjective judgment. *Journal of Abnormal Psychology, 80,* 249–260.

Goldstein, A. M., & Reznikoff, M. (1972). MMPI performance in chronic medical illness: the use of computer-derived interpretations. *British Journal of Psychiatry 120,* 157–158.

Goldstein, G., & Shelly, C. (1982). A further attempt to cross-validate the Russell, Neuringer, and Goldstein neuropsychological keys. *Journal of Consulting and Clinical Psychology, 50,* 721–726.

Gorham, D. R. (1967). Validity and reliability studies of a computer-based scoring system for inkblot responses. *Journal of Consulting Psychology, 31,* 65–70.

Graham, J. R. (1967). A Q-sort study of the accuracy of clinical descriptions based on the MMPI. *Journal of Psychiatric Research, 5,* 297–305.

Graham, J. R. (1977). *The MMPI: A practical guide.* New York: Oxford University Press.

Graham, J. R., Smith, R. L., & Schwartz, G. F. (1986). Stability of MMPI configurations for psychiatric inpatients, *Journal of Consulting and Clinical Psychology, 54,* 375–380.

Grayson, H. M., & Backer, T. E. (1972). Scoring accuracy of four automated MMPI interpretation report agencies. *Journal of Clinical Psychology, 28,* 366–370.

Green, C. J. (1982). The diagnostic accuracy and utility of MMPI and MCMI computer interpretive reports. *Journal of Personality Assessment, 46,* 359–365.

Greene, R. L. (1980). *The MMPI: An interpretive manual.* New York: Grune & Stratton.

Greist, J. H., Gustafson, D. H., Stauss, F. F., Rowse, G. L., Laughren, T. P., & Chiles, J.

A. (1973). A computer interview for suicide risk prediction. *American Journal of Psychiatry, 130,* 1327–1332.

Greist, J. H., Gustafson, D. H., Stauss, F. F., Rowse, G. L., Laughren, T. P., & Chiles, J. A. (1974). Suicide risk prediction: A new approach. *Life Threatening Behavior, 4,* 212–223.

Gynther, M. D., Altman, H., & Sletten, I. W. (1973). Replicated correlates of MMPI two-point code types: The Missouri actuarial system. *Journal of Clinical Psychology, 28,* 263–286.

Gynther, M. D., & Brilliant, P. J. (1968). The MMPI K+ profile: A reexamination. *Journal of Consulting and Clinical Psychology, 32,* 616–617.

Hansen, J. C. (1987). Computer-assisted interpretation of the Strong Interest Inventory. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 292–321). New York: Basic Books.

Harris, W. G. (1984, August). Use of computer-based test interpretation: Some possible guidelines. In J. D. Matarazzo (chair), *Computer-based test interpretation: Prospects and problems.* Symposium conducted at the annual convention of the American Psychological Association, Toronto.

Harris, W. G. (1987). Computer-based test interpretations: Some development and application issues. In L. D. Eyde (Ed.), *Computerised psychological testing.* London: Lawrence Erlbaum Associates.

Harris, W. G., Niedner, D., Feldman, C., Fink, A., & Johnson, J. H. (1981). An on-line interpretive Rorschach approach: Using Exner's Comprehensive System. *Behavior Research Methods and Instrumentation, 13,* 588–591.

Heaton, R. K., Grant, I., Anthony, W. Z., and Lehman, R. A. W. (1981). A comparison of clinical and automated interpretation of the Halstead–Reitan Battery. *Journal of Clinical Neuropsychology, 3,* 121–141.

Hedlund, J. L., Morgan, D. W., & Master, F. D. (1972). The Mayo Clinic automated MMPI program: Cross-validation with psychiatric patients in an army hospital. *Journal of Clinical Psychology, 28,* 505–510.

Hofer, P. J., & Bersoff, D. N. (1983). *Standards for the administration and interpretation of automated psychological testing.* Unpublished manuscript. (Available from P. J. Hofer or D. N. Bersoff, Suite 511, 17th St. N.W., Washington, DC 20036)

Holtzman, W. H. (1975). New developments in the Holtzman Inkblot Technique. In P. McReynolds (Ed.), *Advances in psychological assessment,* (Vol. 3, pp. 243–274). San Francisco: Jossey–Bass.

Johnson, J. H., Giannetti, R. A., & Williams, T. A. (1978). A self-contained microcomputer system for psychological testing. *Behavior Research Methods and Instrumentation, 10,* 579–581.

Karson, S., & O'Dell, J. W. (1975). A new automated interpretation system for the 16PF. *Journal of Personality Assessment, 39,* 256–260.

Karson, S., & O'Dell, J. W. (1987). Computer-based interpretation of the 16PF: The Karson Clinical Report in contemporary practice. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 198–217). New York: Basic Books.

Katz, L., & Dalby, J. T. (1981). Computer assisted and traditional assessment of elementary-school-aged children. *Contemporary Educational Psychology, 6,* 314–322.

Kleinmuntz, B. (1963). MMPI decision rules for the identification of college maladjustment. *Psychological Monographs, 77*(14 Whole No. 577).

Kleinmuntz, B. (Ed.). (1968). *Formal representation of human judgment.* New York: Wiley.

Kleinmuntz, B. (1969). Personality test interpretation by computer and clinician. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications* (pp. 97–104). New York: McGraw–Hill.

Kleinmuntz, B. (1972). *Computers in personality assessment.* New York: General Learning Press.

Kleinmuntz, B. (1975). The computer as clinician. *American Psychologist, 30,* 379–387.

Klett, W. (1971, May). The utility of computer interpreted MMPIs at St. Cloud VA Hospital. *Newsletter of Research in Psychology, 13,* pp. 45–47.

Klett, B., Schaefer, A., & Plemel, D. (1985, May). Just how accurate are computer-scored tests? *VA Chief Psychologist, 8,* p. 7.

Klingler, D. E., Johnson, J. H., & Williams, T. A. (1976). Strategies in the evaluation of an on-line computer-assisted unit for intake assessment of mental health patients. *Behavior Research Methods and Instrumentation, 8,* 95–100.

Klingler, D. E., Miller, D. A., Johnson, J. H., & Williams, T. A. (1977). Process evaluation of an on-line computer-assisted unit for intake assessment of mental health patients. *Behavior Research Methods and Instrumentation, 9,* 110–116.

Kostlan, A. (1954). A method for the empirical study of psychodiagnosis. *Journal of Consulting Psychology, 18,* 83–88.

Krug, S. E. (Ed.). (1987). *Psychware Sourcebook* (2nd ed). Kansas City, MO: Test Corporation of America.

Kurtz, R. M., & Garfield, S. L. (1978). Illusory correlation: A further exploration of Chapman's paradigm. *Journal of Consulting and Clinical Psychology, 46,* 1009–1015.

Labeck, L. J., Johnson, J. H., & Harris, W. G. (1983). Validity of an automated on-line MMPI interpretive system. *Journal of Clinical Psychology, 39,* 412–416.

Lachar, D. (1974a). Accuracy and generalization of an automated MMPI interpretation system. *Journal of Consulting and Clinical Psychology, 42,* 267–273.

Lachar, D. (1974b). *The MMPI: Clinical assessment and automated interpretation.*Los Angeles: Western Psychological Services.

Lachar, D. (1982). *Personality Inventory for Children (PIC) revised format manual supplement.* Los Angeles: Western Psychological Services.

Lachar, D. (1987). Automated assessment of child and adolescent personality: The Personality Inventory for Children (PIC). In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 261–291). New York: Basic Books.

Lachar, D., & Gdowski, C. G. (1979). *Actuarial assessment of child and adolescent personality: An interpretive guide for the Personality Inventory for Children profile.* Los Angeles: Western Psychological Services.

Lachar D., Klinge, V., & Grisell, J. L. (1976). Relative accuracy of automated MMPI narratives generated from adult norm and adolescent norm profiles. *Journal of Consulting and Clinical Psychology, 44,* 20–24.

Lanyon, R. I. (1984). Personality assessment. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 35, pp. 667–701). Palo Alto, CA: Annual Reviews.

Little, K. B., & Shneidman, E. S. (1959). Congruencies among interpretation of psychological test and anamnestic data. *Psychological Monographs, 73*(6, Whole No. 476).

Lushene, R. E., & Gilberstadt, H. (1972, March). *Validation of VA MMPI computer-generated reports.* Paper presented at the Veterans Administration Cooperative Studies Conference, St. Louis.

Manning, H. M. (1971). Programmed interpretation of the MMPI. *Journal of Personality Assessment, 35,* 162–176.

Marks, P. A., & Seeman, W. (1963). *The actuarial description of personality: An atlas for use with the MMPI.* Baltimore: Williams & Wilkins.

Marks, P. A., Seeman, W., & Haller, D. L. (1974). *The actuarial use of the MMPI with adolescents and adults.* Baltimore: Williams & Wilkins.

Matarazzo, J. M. (1983, July 22). Computerized psychological testing. *Science, 221,* 323.

Matarazzo, J. M. (1985). Clinical psychological test interpretations by computer: Hardware outpaces software. *Computers in Human Behavior, 1,* 235–253.

Matarazzo, J. M. (1986). Computerized clinical psychological test interpretation: Unvalidated plus all mean and no sigma. *American Psychologist, 44,* 14–24.

McDonald, C. J. (1976). Protocol-based computer reminders, the quality of care and the non-perfectibility of man. *New England Journal of Medicine, 295,* 1351–1355.

Maxwell, A. E. (1971). Multivariate statistical methods and classification problems. *British Journal of Psychiatry. 119,* 121–127.

Meehl, P. E. (1956). Wanted – a good cookbook. *American Psychologist, 11,* 263–272.

Meehl, P. E. (1960). The cognitive activity of the clinician. *American Psychologist, 15,* 19–27.

Meehl, P. E., Schofield, W., Glueck, B. C., Studdiford, W. B., Hastings, D. W., Hathaway, S. R., & Clyde, D. J. (1962). *Minnesota-Ford pool of phenotypic personality items* (August 1962 ed.). Unpublished materials. (Available from P. E. Meehl or W. Schofield, Department of Psychiatry, 393 Mayo Memorial Building, University of Minnesota, Minneapolis, MN 55455.)

Millon, T. (1981). *Disorders of Personality: DSM–III, Axis II.* New York: Wiley.

Millon, T. (1982). *Millon clinical multiaxial inventory manual* (3rd ed.). Minneapolis: National Computer Systems.

Moreland, K. L. (1983, April). *A comparison of the validity of two MMPI interpretation systems: A preliminary report.* Paper presented at the 18th annual Symposium on Recent Developments in the Use of the MMPI, Minneapolis.

Moreland, K. L. (1985). Validation of computer-based test interpretations: Problems and prospects. *Journal of Consulting and Clinical Psychology, 53,* 816–825.

Moreland, K. L. (1987). Computer-based test interpretations: Advice to the consumer. In L. D. Eyde (Ed.), *Computerised testing.* London: Lawrence Erlbaum Associates.

Moreland, K. L., & Onstad, J. A. (1985, March). *Validity of the Minnesota Report, 1: Mental health outpatients.* Paper presented at the 20th annual Symposium on Recent Developments in the Use of the MMPI, Honolulu.

Moreland, K. L., & Onstad, J. A. (1987a). Validity of Millon's computerized interpretation system for the MCMI: A controlled study. *Journal of Consulting and Clinical Psychology, 55,* 113–114.

Moreland, K. L. & Onstad, J. A. (1987b, summer). A controlled study of the Minnesota Report: Adult Clinical System. *Network News, 1,* 1, 6, 11. (Available from National Computer Systems, P. O. Box 1416, Minneapolis, MN 55440.)

Mules, W. C. (1972). A comparison of conventional modes of interpreting Strong Vocational Interest Blank results to modes which employ a computer generated, prose interpretation. *Dissertation Abstracts International, 33,* 1445a.

Murphy, K. R. (1987). The accuracy of clinical versus computerized test interpretations. *American Psychologist, 42,* 192–193.

National Computer Systems (1989). *Professional Assessment Services 1989 Catalog.* Minneapolis: Author.

Nichols, D. (1985). Review of the Minnesota Report: Personnel Selection System. In J. V. Mitchell, Jr. (Ed.), *Ninth mental measurements yearbook* (Vol. 2, pp. 1008–1009). Lincoln, NE: Buros Institute of Mental measurements.

O'Dell, J. W. (1972). P. T. Barnum explores the computer. *Journal of Consulting and Clinical Psychology, 38,* 270–273.

Palmer, W. H. (1971). Actuarial MMPI interpretation: A replication and extension. *Dissertation Abstracts International, 31,* 3265B.

Payne, F. D., & Wiggins, J. S. (1968). Effects of rule relaxation and system combination on classification rates in two MMPI "cookbook" systems. *Journal of Consulting and Clinical Psychology, 32,* 734–736.

Pearson, J. S., Rome, H. P., Swenson, W. M., Mataya, P., & Brannick, T. L. (1965). Development of a computer system for scoring and interpretation of MMPI in a medical clinic. *Annals of the New York Academy of Sciences, 126,* 684–692.

Petterson, J. (1983, November 9). Computer testing spurs writing of ethics codes. *Kansas City Times,* pp. A1, A11.

Piotrowski, Z. A. (1964). A digital computer administration of inkblot test data. *Psychiatric Quarterly, 38,* 1–26.

Ripley, R. E., & Ripley, M. J. (1979). *Career families: Interpretation manual for the World of Work Inventory* (rev. ed.). Scottsdale, AZ: World of Work.

Rome, H. P., Mataya, P., Pearson, J. S., Swenson, W., & Brannick, T. L. (1965). Automatic personality assessment. In R. W. Stacy & B. Waxman (Eds.), *Computers in biomedical research* (Vol 1., pp. 505–524). New York: Academic Press.

Sines, L. K. (1959). The relative contribution of four kinds of data to accuracy in personality assessment. *Journal of Consulting Psychology,* 1959, *23,* 483–492.

Sines, J. O. (1966). Actuarial methods in personality assessment. In B. Maher (Ed.), *Progress in experimental personality research* (Vol. 3, pp. 133–193). New York: Academic Press.

Sundberg, N. D. (1985a). Review of Behaviordyne Psychodiagnostic Laboratory Service. In J. V. Mitchell, Jr. (Ed.), *Ninth Mental Measurements Yearbook* (pp. 1003–1005). Lincoln, NE: Buros Institute of Mental Measurements.

Sundberg, N. D. (1985b). Review of WPS Test Report. In J. V. Mitchell, Jr. (Ed.) *Ninth Mental Measurements Yearbook* (pp. 1009–1011). Lincoln, NE: Buros Institute of Mental Measurements.

Tucker, G. J., & Rosenberg, S. D. (1980). Computer analysis of schizophrenic speech: An example of computer usage in the study of psychopathologic processes. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.). *Technology in mental health care delivery systems,* Norwood, NJ: Ablex.

University of Minnesota. (1982). *User's guide for the Minnesota Report.* Minneapolis: National Computer Systems.

University of Minnesota. (1984). *User's guide for the Minnesota Report: Personnel Selection System.* Minneapolis: National Computer Systems.

Vincent, K. R., & Castillo, I. M. (1984). A comparison of two MMPI narratives. *Computers in Psychiatry/Psychology, 6*(4), 30–32.

Vincent, K. R., Wilson, A. L., & Wilson, J. L. (1983). *Automated interpretation program for the MMPI.* Houston: Psychometric Services.

Webb, J. T. (1970). Validity and utility of computer-produced MMPI reports with Veterans Administration psychiatric populations (Summary). *Proceedings of the 78th annual convention of the American Psychological Association, 5,* 541–542.

Webb, J. T., Miller, M. L., & Fowler, R. D. (1969). Validation of a computerized MMPI interpretation system (Summary). *Proceedings of the 77th annual convention of the American Psychological Association, 4,* 523–524.

Webb, J. T., Miller, M. L., & Fowler, R. D. (1970). Extending professional time: A computerized MMPI interpretation service. *Journal of Clinical Psychology, 26,* 210–214.

Weigel, R. G., & Phillips, M. (1967). An evaluation of MMPI scoring accuracy by two national scoring agencies. *Journal of Clinical Psychology, 23,* 101–103.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* San Francisco: Freeman.

Western Psychological Services. (1984). *1985–1986 catalog.* Los Angeles: Author.

Wimbish, L. G. (1984). The importance of appropriate norms for the computerized interpretation of adolescent MMPI profiles. *Dissertation Abstracts International, 46,* 3234B. (University Microfilms No. 85-26, 277).