Psychological Assessment
March 2000 Vol. 12, No. 1, 6-18

Validity and Utility of Computer-Based Test Interpretation

James N. Butcher
Department of Psychology University of Minnesota,Twin Cities Campus
Julia N. Perry
Department of Psychology University of Minnesota,Twin Cities Campus
Mera M. Atlis
Department of Psychology University of Minnesota,Twin Cities Campus

ABSTRACT

Computers have been important to applied psychology since their introduction, and the application of computerized methods has expanded in recent decades. The application of computerized methods has broadened in both scope and depth. This article explores the most recent uses of computer-based assessment methods and examines their validity. The comparability between computer-administered tests and their pencil-and-paper counterparts is discussed. Basic decision making in psychiatric screening, personality assessment, neuropsychology, and personnel psychology is also investigated. Studies on the accuracy of computerized narrative reports in personality assessment and psychiatric screening are then summarized. Research thus far appears to indicate that computer-generated reports should be viewed as valuable adjuncts to, rather than substitutes for, clinical judgment. Additional studies are needed to support broadened computer-based test usage.

Computers have played an integral role in scoring psychological tests virtually since their introduction, almost a half-century ago. Initially, computer-based applications involved scoring and data processing; however, as their general use became more widespread, their potential advantage to the process of interpretation came to be recognized. Over the past several decades, their presence in mental health care settings has broadened, and computers have served as useful and necessary aids to the evaluation process. The benefits of computers to the field continue to expand as technology becomes more advanced, allowing for more sophisticated operations.

How can such a complex human cognitive process as test interpretation be performed by an electronic device? A theoretical rationale for computerized evaluation was provided by the introduction of the concept of actuarial interpretation. In 1954 , Meehl published a monograph in which he debated the merits of actuarial/statistical (objective) decision-making methods versus clinical (subjective) ones. His analysis led to the conclusion that decisions based on objectively derived information ultimately were more valid than judgments rooted in decision-making rules that were derived and applied subjectively. As a result of this and his subsequent collaborative works (Dawes, Faust, & Meehl, 1989 ; Grove & Meehl, 1996), objective assessment procedures generally are considered to be equal or superior to subjective methods. After reviewing 136 studies, a recent meta-analysis by Grove, Zald, Lebow, Snitz, and Nelson (2000 ; this issue) concluded that the advantage in accuracy for statistical prediction over clinical prediction was about 10%.

Despite their common roots and comparable rationales, however, *actuarial assessment* and *computerized assessment* are not strictly equivalent concepts. Therefore, establishing the validity of the former does not mean that the veracity of the latter can be assumed in all cases. A computer-based test interpretation (CBTI) system is actuarial only if its interpretive output is wholly determined by statistical rules that have been demonstrated empirically to exist between the output and the input data (Sines, 1966). Other types of CBTI systems, which are not actuarial in nature, draw conclusions on the basis of work of a clinician who creates interpretations using published research and clinical experience (Wiggins, 1973). Because of such distinctions, the validity of each type of computerized assessment must be examined in its own right, particularly in light of computerized procedures' standard use by modern-day clinicians and the degree to which the clinicians' results are relied on to make crucial health care decisions.

In 1984, the American Psychological Association Committee on Professional Standards cautioned psychologists who used interpretive reports in business and school settings against using narrative summaries in the absence of adequate data to validate their accuracy; however, the American Psychological Association's caveat is important for practitioners in all areas of psychology, necessitating exploration of the validity of computerized-assessment procedures. In addition to examining computer-based reports, this review also investigates computer interviews that have been used to make diagnoses.

Previous reviews of research in this area have pointed to the limitations of computer-based interpretations (Maddux & Johnson, 1998). For example, Moreland (1985) reviewed validity studies of computer-based interpretations and pointed to the fact that conclusions drawn from the research must be evaluated in light of several problems. Among those he discussed were the following: small sample sizes; inadequate external criterion measures with which to compare the CBTI statements; a lack of information regarding the reports' base-rate accuracy; failure to assess the ratings' reliability across time or across raters; failure to investigate the internal consistency of the reports' interpretations; and issues pertaining to the report raters, such as lack of familiarity with the interpretive system used, lack of expertise in the area of interest, and possible bias secondary to the theoretical orientation held by the rater. For precisely these types of reasons, Matarazzo cautioned that computerized testing must be subjected to the "checks and balances of democratic society" (1986 , p. 14) in order to preserve the integrity of psychological assessment. He also provided several key criticisms of his own with respect to computerized test interpretations, including the interpretations' failure to recognize test takers' uniqueness, the tendency for the interpretations to be unsigned (ostensibly leaving no one directly accountable for the interpretations' contents), and the tendency to be viewed as an end rather than as means to an end. Even more recently, Snyder, Widiger, and Hoover (1990) voiced their own concerns with CBTIs, concluding that "[w]hat the [CBTI] literature lacks (and desperately needs) are rigorously controlled experimental studies examining methodological issues in CBTI validation per se" and recommending specifically that future studies "include representative samples of both [CBTI] consumers and test respondents" and "use characteristics of each as moderator variables in analyzing reports' generalizability" (p. 476).

Equivalence of Computer-Administered Assessment Methods

Throughout the past decade, the areas that computer programs have addressed have been as varied as the individuals using them. At the most fundamental level, research can examine the validity of computerized tests in general and the more specific issue of the tests' equivalence to paper-and-pencil measures. Research on this topic is particularly important because, as noted by Hofer (1985) , there are several factors specific to computerized test administration that could be instrumental in yielding noncomparable results. Foremost among these may be some individuals' discomfort with computers and consequent awkwardness when dealing with them. In addition, there are such factors as the type of equipment used and the nature of the test material. Regarding the latter, Hofer noted that when item content deals with sensitive and personal information, respondents may be more willing to reveal their true feelings to a computer than to a human being, which may lead to atypical results when computerized assessment is used. Computer administration has other advantages over traditional methods as well. These include potential time savings, elimination of the possibility of respondents making errors while filling out handwritten answer sheets, and elimination of the possibility of clinicians and technicians making errors while hand-scoring items (Allard, Butler, Faust, & Shea, 1995). However, there are disadvantages as well. Chief among the disadvantages is the possibility of excessive generality of results and the high potential for the results' misuse because of their increased availability (Butcher, 1987). Whether the distinctions between computerized and traditional assessment methods seem positive or negative, however, they leave room for questions regarding comparability and, ultimately, validity. Below, we review studies comparing computer-administered tests with paper-and-pencil tests or other traditional methods of data collection.

Psychiatric Screening

In recent years, computerized assessment frequently has been used for psychiatric screening. This is in part due to research that demonstrates mental health clients' comfort with (and even preference for) automated assessments (e.g., Hile & Adkins, 1997). Because automated procedures have become common components of psychiatric assessment batteries, their validity requires careful evaluation.

Several recent studies have investigated the accuracy with which computerized assessment programs were able to diagnose the presence of behavioral problems. In one, Ross, Swinson, Larkin, and Doumani (1994) examined the Computerized Diagnostic Interview Schedule (C-DIS). In their study, 173 substance abusers were assessed using both the C-DIS and a clinician-administered Structural Clinical Interview for the *Diagnostic and Statistical Manual for Mental Disorders* (3rd ed., rev.; *DSM—III—R;* American Psychiatric Association, 1987). The diagnostic agreement between the two instruments was then examined. Results showed that, with the exceptions of substance abuse disorders and antisocial personality disorder, levels of agreement (assessed using kappa coefficients) between the C-DIS and the Structural Clinical Interview for *DSM—III—R* were poor. For alcohol, cocaine, and opiate diagnoses, however, the kappa values were .71, .75, and .81, respectively. In addition, the C-DIS was able the rule out the presence of comorbid disorders in the sample with approximately 90% accuracy.

Using a somewhat more complex design, Jemelka, Wiegand, Walker, and Trupin (1992) administered a brief mental health and personal history interview, the group form of the Minnesota Multiphasic Personality Inventory (MMPI), the Revised Beta IQ Examination, the Suicide Probability Scale, the Buss—Durkee Hostility Inventory, the Monroe Dyscontrol Scale, and the Veteran's Alcohol Screening Test to 100 incarcerated male felony offenders. Algorithms from a CBTI system were then used to assign rankings to each participant, on the basis of potential for violence, substance abuse, suicide, and victimization. The algorithms were also used to identify the presence of clinical diagnoses based on the basis of the *DSM—III—R.* Clinical interviewers then rated the members of the sample on the same five dimensions. An additional 109 participants were administered eight sections of the Diagnostic Interview Schedule (DIS). Agreement between the CBTI ratings and the clinician ratings was fair, and the CBTI's algorithms for violence and suicide potential were judged to be too sensitive. There was high agreement between CBTI- and clinician-diagnosed *DSM—III—R* disorders, with an overall concordance rate of 82%. The researchers suggested that a more accurate prediction could be achieved using algorithms that produced continuous scale scores rather than Likert-type ratings.

Farrell, Camplair, and McCullough (1987) evaluated the ability of a computerized interview to identify the presence of a group of target complaints. A face-to-face, unstructured intake interview and the interview component of a computerized mental health information system, the Computerized Assessment System for Psychotherapy Evaluation and Research, were both administered to 103 adult clients requesting outpatient psychological treatment. Response formats included Likert-type scales, behavioral frequencies, and multiple-choice options. Paper-and-pencil assessment measures were also completed by the participants. Results showed that there was fairly poor agreement (mean $r = .33$) between target complaints as reported by clients on computer and as identified by therapists in interviews. However, the complaints identified in the computerized interview correlated adequately with measures of global functioning, with 9 of the 15 correlations significant at the .0009 level. They ranged from .04 to .54.

Personality Assessment

A number of studies related to the assessment of personality focused on evaluating the comparability of computer and booklet administrations of various instruments. For example, a study by F. R. Wilson, Genco, and Yager (1985) used a test-attitudes screening instrument as a representative of all comparable assessment instruments administered through computer. Ninety-eight female college freshman were administered the Test Attitude Battery (TAB), which came in both paper-and-pencil and computer-administered formats. They were also administered the Test Anxiety Questionnaire, whose sole format was paper and pencil. Each participant was assigned to one of three conditions, which corresponded with the three possible pairings of the three instrument formats, and the subsequent order of test administration was counterbalanced. Means, variances, and correlations with construct validating variables were comparable for paper-and-pencil and computerized test versions, demonstrating that the two versions satisfied criteria for serving as alternative forms and could also be considered parallel forms. The paper-and-pencil—computer-administered TAB correlations were all significant at the .001 level, and they ranged from .71 to .95. Despite these findings, however, the authors advised that caution be exercised in generalizing the results too broadly beyond the scope of their study, particularly with regard to clinical samples and less straightforward measures.

Much of the research investigating the comparability of computer-based personality assessment measures has used the MMPI or the restandardized MMPI-2. For example, several studies examined possible differences between paper-and-pencil and computerized testing formats. The findings of Lambert, Andrews, Rylee, and Skinner (1987) , Schuldberg (1988) , and Watson, Juba, Anderson, and Manifold (1990) all indicate that the differences are generally few and small in magnitude, leading to between-forms correlations of .68 to .94 (Watson et al., 1990). Other researchers have obtained close (Sukigara, 1996) or "near perfect" (i.e., 92% to 97%) agreement between computer and booklet administrations (Pinsoneault, 1996). Honaker, Harrell, and Buffaloe (1988) investigated the equivalency of Microtest computer MMPI administration among 80 community volunteers. They found no significant differences between various computer formats for means and standard deviations for validity, clinical, and 27 additional scales. However, analogous to a number of studies investigating the equivalency of computer and booklet forms of the MMPI, the power of their statistical analyses did not provide conclusive evidence regarding the equivalency of the paper-and-pencil and computerized administration format (Honaker et al., 1988).

Most recently, a comprehensive meta-analysis on the topic of psychometric comparability between computerized and standard MMPI testing formats was conducted by Finger and Ones (1999) , who included all research available to them. They meta-analyzed the results of 14 studies, all of which reported on the use of computerized formats of the MMPI or MMPI-2, that were conducted between 1974 and 1996. Across studies, the differences in *T* -score means and standard deviations between test formats were negligible. Corrected mean crossform correlations were consistently near 1.00. On the basis of these findings, Finger and Ones (1999) concluded that " ... the evidence is strong that there is little effect on MMPI scale scores when the test is administered via computer as opposed to when the test is administered via booklet" (p. 19).

In addition to providing a viable alternative format to traditional paper-and-pencil testing, computerized methods offer the option of adapting the test to suit the needs and preferences of the test taker. The comparability and validity of this method (known as computerized adaptive testing) were examined by Roper, Ben-Porath, and Butcher (1995) in their study of the MMPI-2. Five hundred seventy-one undergraduate psychology students were administered three versions of the MMPI-2: a booklet version, an adaptive computerized version, and a conventional computerized version. Each participant took the same format twice, the booklet and adaptive computerized versions (in counterbalanced order), or the convention and adaptive computerized versions (again, in counterbalanced order). There were few statistically significant differences in the resulting mean scale scores between the booklet and adaptive computerized formats. Men scored significantly lower on the adaptive computerized version on clinical Scale 0 (Social Introversion) and on the content scales Bizarre Mentation (*BIZ*) and Depression (*DEP*). Women scored significantly lower on the adaptive computerized version on clinical Scale 0 (Social Introversion) and the Content scales Fears (*FRS*) and Family Problems (*FAM*). They scored significantly higher on the adaptive computerized version on clinical Scale 5 (Masculinity—Femininity) and the Content scale Type A (*TPA*). Regarding the conventional computerized versus adaptive computerized comparisons, there were no significant differences for either men or women. In terms of criterion-related validity, there were no significant differences between the correlations of criterion measures (viz., the Beck Depression Inventory, Trait Anger and Trait Anxiety from the State—Trait Personality Inventory, and nine scales from the Symptoms Checklist–Revised) with MMPI-2 scales from the adaptive computerized administration and as compared with the booklet version.

Holden and Hickman (1987) investigated computerized and paper-and-pencil versions of the Jenkins Activity Scale, a measure that assesses behaviors related to the Type-A personality. Sixty male undergraduate students were assigned to one of the two administration formats. The stability of scale scores was comparable for both formats, as were mean scores, variances, reliabilities, and construct validities, and none of them demonstrated statistically significant differences.

The 72 German-speaking participants in Merten and Ruch's (1996) study of German versions of the Eysenck Personality Questionnaire–Revised and the Carroll Rating Scale for Depression completed half of each instrument using paper and pencil and the other half on computer (with orders counterbalanced). They compared the results from the two formats to one another as well as to data from another German sample, consisting of individuals who were administered only the paper-and-pencil version of the Eysenck Personality Questionnaire–Revised. Once again, means and standard deviations were comparable across computerized and more traditional formats.

Neuropsychology

Pelligrino, Hunt, Abate, and Farr's (1987) study comparing a battery of 10 computerized tests of spatial abilities with traditional paper-and-pencil counterparts suggested that computer-administered static spatial reasoning tasks can augment current paper-and-pencil procedures. The study indicated that "computer-based static tasks and measures encompass the variance contained in traditional paper-and-pencil tests and also provide measures of unique variance associated with speed versus accuracy of processing" (p. 235). Choca and Morris (1992) used a sample of 46 neurologically impaired men to compare a computerized version of the Halstead Category Test to its standard counterpart. All participants were administered both versions of the test. Results generally supported the measures' comparability, as the difference between the mean number of errors obtained on each version (85.22 for the standard vs. 90.22 for the computerized) was nonsignificant.

The compatibility of a computerized neuropsychological test with the original measure can be influenced by the differences between standard and computerized administration procedures. For example, research conducted by French and Beaumont (1990) investigated the validity of computerized versions of the Mill Hill Vocabulary Test and the Standard Progressive Matrices Test by administering computerized and standard versions of the former to 274 patients and computerized and standard versions of the latter to an additional 184 patients. On the Mill Hill Vocabulary Test, there were no significant score differences between the computerized and standard versions. However, participants obtained significantly lower scores on the computerized version than on the standard version of the Standard Progressive Matrices Test, indicating that these two measures cannot be used interchangeably. The researchers noted that the poor resolution of the computer graphics system was a likely explanation of the difference. French and Beaumont (1990) also found that for the Mill Hill test, participants expressed a clear preference for keyboard to the touch-screen responding. Additionally, those participants who took the computer version first reported that they would be more likely to take another similar test.

Personnel Psychology

In the field of personnel psychology, automated procedures typically are used for such purposes as personnel selection and job performance prediction. McKee and Levinson's (1990) review addressed the adaptation of paper-and-pencil measures to computerized formats in general and included a specific discussion of the computerized form of the Self-Directed Search, an instrument commonly used in career counseling. They cited the sole study investigating the computerized measure's equivalence to the standard version (Reardon & Loughead, 1988) and advised that additional research (particularly with respect to reliability and validity) be conducted.

Some of the recent research in this domain addressed the role of computerized assessment in the selection and classification of military personnel. Carretta (1989) examined the usefulness of the computerized test battery, Basic Attributes Battery, for selecting and classifying United States Air Force Pilots. A total of 478 Air Force officer candidates completed a paper-and-pencil qualifying test and the Basic Attributes Battery, and they were also judged on the basis of undergraduate pilot training performance. The validation of the Basic Attributes Battery was ongoing at the time of the report, but the data then available demonstrated that it was adequately assessing abilities and skills related to flight training performance, as measured by its correlations with the other military assessment procedures, such as dot estimation, memory for digits, encoding speed, and self-crediting word knowledge. However, despite the significance of many of the multiple correlations at the .001 level, they ranged in magnitude from .062 to only .498.

Conclusions

The comparability of computerized and standard administration of various measures appears to vary, with the most promising results in the area of personality assessment. Studies examining a variety of instruments and models have demonstrated that scores from these computerized measures, and decisions made using them, are typically equivalent to more traditional methods whose validity has already been established (although there are some notable exceptions). However, there remains a need to investigate each measure on a case-by-case basis (Hofer & Green, 1985) and to establish the repeatability of the resulting data, particularly within the areas of psychiatric screening and neuropsychology. Additional pertinent issues, such as the impact of ergometric factors on the computer administration process, also warrant further consideration.

Validity of Computerized Narrative Reports

Studies of validity of CBTI systems frequently highlight the ease with which computerized assessment procedures can provide extensive interpretations of findings. Numerous computer programs that are now in use not only provide assessment and rudimentary interpretation but also incorporate test findings into extensive narrative reports. It is the validity of these applications to which we turn.

Narrative reports generated by automated assessment appear to fall into two broad categories: descriptive and consultative. Descriptive reports (e.g., for the 16 Personality factors [16PF]) are generated for each individual scale without being moderated by scores on other scales. Consultative reports (e.g., those for the MMPI and Decision Tree [DTREE]) provide detailed analyses of all test data; they are designed to mimic as closely as possible an expert human consultant.

The validity of narrative CBTIs has been investigated extensively in the fields of personality assessment and psychiatric screening. In personality assessment, the validity of MMPI narratives has been described most extensively (for a detailed review of this topic, see Moreland, 1987). However there also has been research aimed at exploring the accuracy of narrative reports for other computerized personality tests, such as the 16PF (e.g., Guastello & Rieke, 1990 ; O'Dell, 1972), the Millon Clinical Multiaxial Inventory (MCMI; e.g., Moreland & Onstad, 1987), and the Rorschach Inkblot Test (e.g., Harris et al., 1981 ; Prince & Guastello, 1990). In psychiatric screening, studies of the computer outputs of the DIS (e.g., First, 1994) are most common.

Narrative Reports in Personality Assessment

Moreland (1987) summarized a number of studies investigating the accuracy ratings for computer-generated MMPI narrative reports. A few studies compared computer-generated narrative interpretations with *criterion* reports generated by human interpreters of the MMPI. One methodological limitation of such studies arises from the possibility that the validity of a clinician's interpretation is often low enough so that a CBTI can have low validity yet can still be judged as accurate (Moreland, 1987). For example, Labeck, Johnson, and Harris (1983) asked three clinicians (each with at least 12 years of clinical experience) to rate independently the quality and the accuracy of code-type interpretations and associated rules using an automated MMPI program (the clinicians did not rate the fit of a narrative to a particular patient, however). Results indicated that the MMPI code-type, diagnostic, and overall profile interpretive statements were consistently rated highly by the expert judges. The narratives generated by automated MMPI programs were judged to be substantially better than average when compared with the "blind" interpretations of similar profiles that were produced by the expert clinicians. Unfortunately, the report did not state how the researchers judged the quality of the blind interpretation. Further, researchers did not investigate the possibility that statements in the blind interpretation could have been so brief and general (especially when compared to a two-page narrative CBTI) that they could have artificially inflated the ratings of the CBTI reports. Despite these limitations, however, this research design was useful in evaluating the congruence of computer-generated decision and interpretation rules.

In most published studies of CBTI validity, the CBTIs' recipients (usually clinicians) rate the accuracy of computer interpretations on the basis of the clinicians' knowledge of test respondents (refer to Moreland, 1987 , p. 42, for a summary of these studies). A study by Butcher et al. (1998) explored the utility of computer-based MMPI-2 reports in Australia, France, Norway, and the United States. In all four countries, clinicians administered MMPI-2 to their patients (whom they were seeing for psychological evaluation or therapy) using a booklet format, in the language of each country. Item responses for each patient were analyzed by the Minnesota Report using the United States norms. Each training clinician then rated the amount of information in each narrative section as *insufficient, some, adequate, more than adequate,* or *extensive.* For each report, clinicians also indicated the percentage of accurate descriptions of the patient and were asked to respond to open-ended questions regarding ways to improve the report. Relatively few raters found the reports inappropriate or inaccurate. In all four countries, Validity Considerations, Symptomatic Patterns, and Interpersonal Relations sections of the Minnesota Report were found to be the most useful sections in providing detailed information about the patients. Over two thirds of the records were found to be highly accurate, which indicated that clinicians judged 80% to 100% of the computer-generated narrative statements in them to be appropriate. In 87% of the reports, at least 60% of the computer-generated narrative statements were believed to be appropriate.

In one of the few studies comparing clinical and computer-based assessment decisions, Butcher (1988) investigated the usefulness of computerized MMPI assessment for screening in personnel settings. In his study, 262 airline-pilot applicants were evaluated by both expert clinicians and computer-based decision rules. Each applicant's overall level of adjustment was rated by experts (using only an MMPI profile) on a Likert-type scale with three categories: *adequate, problems possible,* and *problems likely.* The computer-based decision rules were also used to make determinations about the applicants. Here, the categories of *excellent, good, adequate, problems possible,* and *poor* were used. The study showed high agreement between the computer-based decisions and those made by clinicians in rating overall adjustment. Over 50% of individuals falling into the *adequate* category on the basis of the computer-based rules were given ratings of *adequate* by the clinicians. There was agreement between the computer rules and clinician judgment on the possibility of problems being present in 27% of cases. Over 60% of individuals rated as *poor* by the computer rules were given *problems likely* ratings by the clinicians. No kappa values were calculated. This study showed strong agreement between clinicians and the computer; however, no external criteria were available to allow for an assessment of the relative accuracy of each method, thereby leaving in question the validity of both the computer-generated and the clinician-generated ratings.

One methodological problem involving accuracy ratings that is commonly found in field studies is that estimates of interrater reliability are difficult, if not impossible, to obtain. In addition, clinicians are frequently asked to judge the accuracy of computer interpretations for their patients without providing descriptions of how these judgments were made and without verifying the appropriateness of such judgments with information from the patients themselves and from other sources (such as physician reports or patients' family members). Furthermore, to study the validity of computer-generated narrative reports, raters should evaluate individual interpretive statements, because global accuracy ratings may limit the usefulness of ratings in developing the CBTI system (Moreland, 1985).

Moreland's (1987) suggestions for evaluating the validity of narrative CBTIs were adopted by Eyde, Kowal, and Fishburne in their 1991 investigation of the comparative validity of the narrative outputs for CBTI systems. These researchers used case histories and self-report questionnaires as a criterion against which narrative reports generated by seven MMPI computer interpretation systems were evaluated. Each of the clinicians rated six protocols on the basis of profile type. Some of the cases were assigned to all raters; they consisted of a pair of Black and White patients who were matched for 7-2 (Psychasthenia—Depression) code-type and a pair of Black and White soldiers who had all clinical scales in the subclinical range ($T < 70$). Clinicians rated the relevance of each sentence in the narrative CBTI as well as the global accuracy of each report. Some CBTI systems studied showed a high degree of accuracy. However, the overall results indicated that the validity of the narrative outputs varied, with the highest accuracy ratings being associated with narrative lengths in the short-to-medium range. For different CBTI systems, results for both sentence-by-sentence and global ratings were consistent, but they differed for the clinical and subclinical normal profiles. For instance, the subclinical normal cases presented to all clinicians had a high percentage (*Mdn* = 50%) of unratable sentences, whereas the 7-2 profiles had a low percentage (*Mdn* = 14%) of sentences that could not be rated. A potential explanation for such differences may come from the fact that the clinical cases were inpatients for whom more detailed case histories were available. In addition, since the length of time between the preparation of the case histories and the administrations of the MMPI varied, it was impossible to control for changes (i.e., some patient might have gotten better or worse) in acute symptoms over time (Eyde, Kowal, & Fishburne, 1991).

The research by Eyde et al. (1991) represents a group of studies in which participants rated the accuracy of their narrative feedback after they had taken a test. A possible shortcoming of this design is that there was a failure to control for a phenomenon alternately known as the P. T. Barnum effect (e.g., Meehl, 1956) and Aunt Fanny effect (e.g., Tallent, 1958), which holds that a narrative report may contain high base-rate descriptions that will apply to virtually anybody. Research on the Barnum effect has shown that participants can detect the nature of the overly general feedback if asked the appropriate questions about it (Furnham & Schofield, 1987 ; Layne, 1979). However, research also has demonstrated that people typically are more accepting of favorable Barnum feedback than of unfavorable feedback (Dickson & Kelly, 1985 ; Furnham & Schofield, 1987 ; Snyder & Newburg, 1981). Further, people have been shown to perceive favorable descriptions as more appropriate for themselves than for people in general (Baillargeon & Danis, 1984). Dickson and Kelly's (1985) research demonstrated that test situations, such as the type of assessment used, can be significant in eliciting acceptance of Barnum statements, although Baillargeon and Danis (1984) found no interaction between the type of assessment device and the favorability of statements. Research also demonstrates that people are more likely to personalize Barnum descriptions delivered by persons of authority or expertise (Lees-Haley, Williams, & Brown, 1993). This may be because in group settings, feedback from people of high status might be better accepted and perceived as more accurate (Snyder & Newburg, 1981). Personality-related variables such as extraversion, introversion, and neuroticism (Furnham, 1989), as well as the extent of private self-consciousness (Davies, 1997), also have been found to be connected to individuals' acceptance of Barnum feedback.

In an attempt to deal with Barnum effects on narrative CBTIs, some studies have compared the accuracy of CBTI ratings to a stereotypical patient or an average subject. Researchers have also tried dealing with Barnum-effect problems by using multireport—multirating intercorrelation matrices (Moreland, 1987) and by looking at differences in perceived accuracy between bogus and real reports (Moreland & Onstad, 1987 ; O'Dell, 1972).

Some studies that compared bogus and real reports found the reports to be statistically significantly different. For example, Guastello, Guastello, and Craft (1989) asked 64 undergraduate students to complete Comprehensive Personality Profile Compatibility Questionnaire. One group of students rated the real computerized test interpretation of the Comprehensive Personality Profile Compatibility Questionnaire, and another group rated a bogus CBTI. The difference between the accuracy ratings for the bogus and real profiles (58% and 75%, respectively) was statistically significant. In a study by Guastello and Rieke (1990) , 54 undergraduate students enrolled in an industrial psychology class evaluated a real computer-generated Human Resources Development Report of the 16PF and a bogus report generated from the average 16PF profile of the entire class. Results indicated no statistically significant difference between the ratings for the real reports and the bogus reports (which had mean accuracy ratings of 71% and 71%, respectively). However, when analyzed separately, four out of five sections of the real 16PF output had significantly higher accuracy ratings than did the bogus report. Unlike the aforementioned groups of researchers, Prince and Guastello (1990) found no statistically significant differences between bogus and real CBTI interpretations when they investigated a computerized Exner Rorschach interpretation system. In this study, four psychiatrists evaluated real and bogus computer-generated outputs for 12 psychiatric outpatients. Analyses indicated that the discriminant power of the CBTI for any single patient was 5%, with 60% of interpretive statements only describing characteristics of the "typical" outpatient.

In a study of the validity of Millon's Computerized Interpretation System for the MCMI, Moreland and Onstad (1987) asked eight clinical psychologists to rate real MCMI computer-generated reports and randomly generated reports. The judges rated the accuracy of the reports as a whole as well as the global accuracy of each section (viz., Axis I narrative, Axis II narrative, Axis I diagnoses, Axis II diagnoses, Axis IV stressors, Severity, and Therapeutic Implications). When considered one at a time, five out of seven sections of the report exceeded chance accuracy. Axis I and Axis II sections demonstrated the highest incremental validity. There was no difference in accuracy between the real reports and the randomly selected reports for the Axis IV psychosocial stressors section. The overall pattern of findings suggests that MCMI can exceed chance accuracy. However, further research is needed to determine the conditions that limit its accuracy (Cash, Mikulka, & Brown, 1989 ; Moreland & Onstad, 1987 , 1989).

The recent studies largely uphold the findings of their predecessors. Research into computer-generated narrative reports for personality assessment generally has revealed that the interpretive statements contained in the reports are comparable to clinician-generated statements. Research also points to the importance of controlling for the degree of generality of the reports' descriptions, because of the confounding influence of the Barnum effect.

Computerized Structured Interviews

The work in computer-assisted psychiatric screening has focused mostly on the development of statistical and logic-tree decision models (Erdman, Klein, & Greist, 1985). Computer outputs based on statistical models such as Bayes's probabilities and discriminant function analysis (Hirshfeld, Spitzer, & Miller, 1974) usually provide a list of probable diagnoses and do not include narrative descriptions of the decision-making process. Logic-tree systems, on the other hand, are designed to establish the presence of symptoms specified in diagnostic criteria and arrive at a particular diagnosis (First, 1994). These systems adopt an expert system approach, in which reasons for asking questions and the diagnostic significance of the criterion in question are available after the assessment process. For instance, DTREE is a recent program designed to guide the clinician through the diagnostic process (First, Williams, & Spitzer, 1989). It provides diagnostic consultation both during and after the assessment process. The narrative report from DTREE includes *DSM—III—R* diagnoses and the ruled-out diagnoses, as well as the extensive narrative explaining the reasoning behind diagnostic decisions.

Studies assessing the validity of logic-tree programs frequently compare diagnostic decisions made by a computer and diagnostic decisions made by clinicians. For example, a pilot study by First et al. (1993) evaluated DTREE in an inpatient setting by comparing results of expert clinicians' case conferences with DTREE output. Twenty inpatients were evaluated by a consensus case conference and by five treating psychiatrists who used DTREE software. On the primary diagnosis, perfect agreement was reached between the DTREE and the consensus case conference in 75% of cases (*N* = 15). However, the number of cases within each of the diagnostic categories was small. Also, the agreement was likely to be inflated because some of the treating psychiatrists participated in both the DTREE evaluation and the consensus case conference. Despite these limitations, this preliminary analysis suggests that DTREE might be useful in education and in evaluation of diagnostically challenging clients (First et al., 1993).

Another logic-tree program is a computerized version of the World Health Organization Composite International Diagnostic Interview (CIDI-Auto).

Peters and Andrews (1995) have investigated the procedural validity of the CIDI-Auto in the *DSM—III—R* diagnoses of anxiety disorders. Prior to entering an anxiety disorders clinic, 98 patients were interviewed by the first clinician in a brief clinical intake interview. The patients returned for the second session to self-administer a CIDI-Auto and to be interviewed by the second clinician. The order in which CIDI-Auto was completed varied, depending on the availability of the computer and the second clinician. At posttreatment, clinicians reached consensus about the diagnosis in each individual case (*k* = .93). When such agreement could not be reached, diagnoses were not recorded as the longitudinal, expert, and all data (LEAD) standard against which CIDI-Auto results were evaluated. The agreement between clinicians and CIDI-Auto ranged from *poor* (*k* = .002) for Generalized Anxiety Disorder to *excellent* (*k* = .88) for obsessive—compulsive disorder. Modest overall agreement (*k* = .40) is consistent with paper-and-pencil versions of CIDI as well as studies assessing various computerized diagnostic interviews. LEAD diagnostic procedure yielded 1.56 diagnoses per patient, while CIDI-Auto resulted in 3.05 diagnoses per patient. Peters and Andrews (1995) suggested that the CIDI overdiagnosis might have been caused by clinicians who used more strict diagnostic rules in the application of duration criteria for symptoms.

In another international study, 37 inpatients from the general wards of one of the hospitals in London completed a structured computerized interview assessing their psychiatric history (Carr, Ghosh, & Ancill, 1983). Case records and clinician interview confirmed 90% of the information elicited by the computer. Most patients (88%) found computer interview no more demanding than a traditional interview, and 33% found the computer interview to be easier. Problems with reading English occurred in 8% of the sample. Some patients felt that their responses to the computer were more accurate because they did not have to make a clinician wait while they deliberated their answers. The computer program in this study elicited an average of 5.5 items per patient that were unrecorded by other assessment methods; in other words, 10% of computer-gathered items were unknown to a clinician. Carr et al. (1983) indicated that this finding should not be viewed as a criticism of clinical care because patients vary in how much and what kind of information they are willing to reveal to a clinician as compared to a computer. The researchers also noted that, during the 4 weeks, the patients in this study were assessed by a registrar for several hours and by a consultant, senior registrar, and a social worker on at least one occasion. Because the computerized interview was conducted after the patients were admitted to the hospital, it is possible that it provided an opportunity to reveal forgotten material.

One of the advantages of such logic-tree programs as DTREE and CIDI-Auto is that they potentially can make use of the unlimited number of diagnostic categories. However, unlike the statistical models that are frequently limited to a sample from which they are derived, logic-tree models cannot assign a numeric likelihood of a given diagnosis for an individual patient (Altman, Evenson, & Cho, 1976). Furthermore, a number of researchers have cautioned against using clinician diagnoses as a criterion against which the validity of another diagnostic instrument is validated (e.g., Altman et al., 1976 ; Peters & Andrews, 1995), due to the fact that the imperfect reliability of clinician diagnoses is likely to reduce the upper level of agreement between the new diagnostic instrument and the clinician diagnoses.

In psychiatric screening research, evaluating computer-administered versions of the DIS appears to be the most common. Research has shown that patients tend to hold favorable attitudes toward DIS systems, though diagnostic validity and reliability are questioned when such programs are used alone (First, 1994). Table 1 lists studies that investigate the various administration formats for this instrument. It includes the kappa coefficients associated with a variety of *DSM—III* diagnoses and demonstrates that overall reliabilities for computerized administrations are variable, ranging from .49 to .68.

Differences in diagnostic style, method of data analysis, and diagnostic system used to arrive at a decision all can potentially influence agreement between computers and clinicians. For instance, as shown in Table 1 , Wyndowe (1987) found little diagnostic agreement between the computerized DIS and the nonstructured clinical interview. Another study of the computerized DIS, by Mathisen, Evans, and Meyers (1987) , revealed that the overall agreement between the psychiatrist and the computerized DIS diagnosis for all disorders was low, as indicated by a kappa coefficient of .34. The overall kappa coefficient increased to .48 when the researchers limited their analysis to primary diagnoses. When the analysis was limited to primary diagnosis among patients who received both a clinical diagnosis and a positive computer diagnosis, the kappa coefficient was .67, a figure comparable to the clinician agreement reported in *DSM—III* trials.

Table 2 provides a summary of several studies investigating the C-DIS. Again, the results of this research vary. Diagnostically, the C-DIS reports generally were regarded as accurate. However, C-DIS yielded more diagnoses per patient than a clinician-administered interview, and, as compared to clinicians, it appears to be more likely to fail to assign any diagnoses to the interviewee (e.g., Mathisen et al., 1987). Also, the C-DIS administration is longer (on average) than a more traditional face-to-face interview, although patients' perceptions of their comparative lengths vary (e.g., Erdman et al., 1992 ; Greist et al., 1987).

Neuropsychological Assessment

Another area in which computerized testing batteries have been used to assist in assessment decisions is that of neuropsychology. The 1960s marked the beginning of the investigations into the applicability of computerized testing to this field (e.g., Knights & Watson, 1968). Traditionally, the belief has been that neuropsychology programs have not "produced results that are either satisfactory or equal in accuracy to those achieved by human clinicians" (Adams & Heaton, 1985 , p. 790; see also Golden, 1987). Garb and Schramke's (1996) narrative review and meta-analysis included a section that discussed the utility of computer programs for improving neuropsychological assessment. Although they characterized automated assessment as "promising," they also concluded that there is room for improvement. Specifically, they suggested that new programs be created and prediction rules be modified to include such data as patient history and clinician observation, in addition to the psychometric and demographic data that are more customarily incorporated into the prediction process.

A recent review by Russell (1995) indicated that the ability of computerized testing procedures to detect and locate brain damage is quite accurate, though not as accurate as clinical judgment. A number of recent studies conducted in the area of neuropsychology investigated specific assessment measures. For example, the Right Hemisphere Dysfunction Test and Visual Perception Test were the focus of one study (Sips, Catsman-Berrevoets, van Dongen, van der Werff, & Brook, 1994). These computerized measures were created for the purpose of assessing right-hemisphere dysfunction in children and were intended to have the same validity as the Line Orientation Test and Facial Recognition Test had for adults. Fourteen children with acquired cerebral lesions were administered all four tests. Findings indicated that the Right Hemisphere Dysfunction Test and Visual Perception Test together were sensitive (at a level of 89%) to right-hemisphere lesions, had relatively low specificity (40%), had high predictive value (72%), and accurately located the lesion in 71% of cases. Fray, Robbins, and Sahakian (1996) reviewed findings regarding a computerized assessment program, the Cambridge Neuropsychological Test Automated Batteries (CANTAB). Although specificity and sensitivity were not reported, the reviewers concluded that CANTAB can detect the effects of progressive, neurogenerative disorders sometimes before other signs manifest themselves. In particular, CANTAB has been found successful in detecting early signs of Alzheimer's, Parkinson's, and Huntington's diseases (Fray et al., 1996).

Concluding Comments

On the basis of the findings of the various studies discussed, several conclusions can be drawn. For one, research has supported the position that, by and large, computer-administered tests are essentially equivalent to booklet-administered instruments. There are some reported inconsistencies between computerized and other modes of test administration, but the nonequivalence issues are typically small or nonexistent for most computerized adaptations of paper-and-pencil tests (Finger & Ones, 1998; Moreland, 1987 ; see also Hofer & Green, 1985 ; and Honaker, 1988). However, the marriage between computers and psychological test interpretation has not been a perfect union, which is indicative of the fact that past efforts at computerized assessment have not made optimal use of the flexibility and power of computers in making complex decisions. To date, CBTIs largely perform a "look up and list out" function. That is, a broad range of interpretations are stored in the computer for various test indexes, and the computer simply lists out the stored information for appropriate scale score levels: Computers are not involved as much in decision making.

Computerized applications are limited, to some extent, by the available technology. To date, computer—human interactions are confined to written material. Consequently, potentially critical nonverbal cues, such as speech patterns, vocal tone, and facial expressions, cannot be accounted for in CBTIs as they currently exist. Furthermore, the verbal choices are usually provided to the test-taker in a fixed format (e.g., true—false on an MMPI-2). Temporal relationships relating to the course of a disorder, as well as quantitative relationships between various clusters of symptoms, are crucial in defining disorders and are quite difficult for computers to evaluate (First, 1994). As suggested by Sawyer (1966) , the best predictions are made when clinicians are allowed to gather information and decide how to use it, and computers are programmed to provide statistical summaries and compare results with a data bank.

Extensive research on computer-assisted decisions in several applied contexts (such as neuropsychology, personnel screening, and clinical diagnosis) has shown that automated procedures have the capacity to provide more or less valid and accurate descriptions and predictions. Research on the accuracy of some computer-based systems (particularly those based on the MMPI—MMPI-2, which have been subjected to more studies) has shown promising results with respect to accuracy. However, instruments and situations vary in their reliability and utility, as illustrated by Eyde et al. (1991) in their extensive study of the accuracy of CBTIs. Consequently, each computer-based application needs to be evaluated carefully. Simply put, just because a report comes from a computer, does not necessarily mean that it is valid. The caution required in assessing the utility of computer-based applications brings about a distinct need for specialized training in their evaluation. It is apparent that some sort of instruction in the use (and avoidance of misuse) of CBTIs is essential for all professionals who use them (Hofer & Green, 1985). There is also a need for further research focusing on the accuracy of the information contained in the reports.

Moreover, even though computer-based reports have been validated in some settings, this does not guarantee their validity and appropriateness for all applications. In their discussion of the misuse of psychological tests, Wakefield and Underwager (1993) cautioned against using computerized test interpretations of the MCMI and MCMI-II, designed for clinical populations, in other settings, such as for forensic evaluations. If a test has not been developed for or validated in a particular setting, then computer-based applications of it in that context are not warranted. The danger of misusing data applies to all psychological test formats, but the risk seems particularly high when one considers the convenience of computerized outputs and (as noted by Garb, 1998) the fact that some of the consumers of CBTI services are nonpsychologists who are unlikely to be familiar with the validation research. It is important for scoring and interpretation services to provide computer-based test results only to qualified users (however they are defined).

Even though decision-making and interpretation procedures may be automated with computerized testing, personal factors must still be considered in some way. Styles's (1991) research investigated the importance of a trained psychologist during computerized testing with children. Her study of Raven's Progressive Matrices demonstrated the need for the psychologist to maintain rapport and interest prior to, during, and after testing. These factors were found to have important effects on the reliability and validity of the test data, insofar as they affected test-taking attitudes, comprehension of test instructions, on-task behavior, and demeanor. Carson (1990) has also argued for the importance of a "sound clinicianship," both in the development of psychological test systems and in their use.

All of this points to the conclusion that computer-generated reports should be viewed as valuable adjuncts to clinical judgment rather than "stand-alone" substitutes for a skilled clinician (Fowler, 1969). It should be noted that attempting to tailor test results to unique individual characteristics is a complex process and may not always increase their validity. In addition, it bears noting that CBTIs do not necessarily make clinicians' lives easier (as they ostensibly intend to), because in real-life clinical practice, the evaluation of computerized outputs, especially in narrative formats, requires special expertise and extra time. This consideration may help to explain why a number of studies indicate that clients' attitudes toward the use of computers in clinical activities are not only favorable, but are more favorable than attitudes among the professionals who serve them (R. Wilson, Omeltchenko, & Yager, 1991). However, an alternative explanation may be that professionals are likely to have more knowledge regarding the limitations of CBTIs and, thus, are more skeptical about their use than are nonprofessionals.

It is apparent that whatever else computerized assessment has done for the field of psychology, it clearly has focused attention on accurate assessment in the fields of clinical evaluation and diagnosis. Whether decisions are based on computer interpretation strategies or clinical judgment, psychological assessment, as a discipline, requires further evaluation and continual refinement.

References

Adams, K. M. & Heaton, R. K. (1985). Automated interpretation of the neuropsychological test data. *Journal of Consulting and Clinical Psychology, 53*, 790-802. PsycINFO

Allard, G., Butler, J., Faust, D. & Shea, M. T. (1995). Errors in hand scoring objective personality tests: The case of the Personality Diagnostic Questionnaire–Revised (PDQ—R). *Professional Psychology: Research and Practice, 26*, 304-308. PsycINFO

Altman, H., Evenson, R. C. & Cho, D. W. (1976). New discriminant functions for computer diagnosis. *Multivariate Behavioral Research, 11*, 367-376. PsycINFO

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* ((3rd ed., rev.). Washington, DC: Author)

American Psychological Association, Committee on Professional Standards. (1984). Casebook for providers of psychological services. *American Psychologist, 39*, 663-668.

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

Blouin, A. G., Perez, E. L. & Blouin, J. H. (1988). Computerized administration of the Diagnostic Interview Schedule. *Psychiatry Research, 23*, 335-344. PsycINFO

Butcher, J. N. (1987). The use of computers in psychological assessment: An overview of practices and issues.(In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide.* New York: Basic Books.)

Butcher, J. N. (1988). *Use of the MMPI in personnel screening.* (Paper presented at the 22nd Annual Symposium on Recent Developments in the Use of the MMPI, St. Petersburg, FL)

Butcher, J. N., Berah, E., Ellersten, B., Miach, P., Lim, J., Nezami, E., Pancheri, P., Derksen, J. & Almagor, M. (1998). Objective personality assessment: Computer-based MMPI-2 interpretation in international clinical settings.(In C. Belar (Ed.), *Comprehensive clinical psychology: Sociocultural and individual differences.* New York: Elsevier Science.)

Carr, A. C., Ghosh, A. & Ancill, R. J. (1983). Can a computer take a psychiatric history? *Psychological Medicine, 13*, 151-158. PsycINFO

Carretta, T. R. (1989). USAF pilot selection and classification systems. *Aviation, Space, and Environmental Medicine, 60*, 46-49. PsycINFO

Carson, R. C. (1990). Assessment: What role the assessor? *Journal of Personality Assessment, 54*, 435-445. PsycINFO

Cash, T. F., Mikulka, P. J. & Brown, T. A. (1989). Validity of Millon's computerized interpretation system for the MCMI: Comment on Moreland and Onstad. *Journal of Consulting and Clinical Psychology, 57*, 311-312. PsycINFO

Choca, J. & Morris, J. (1992). Administering the Category Test by computer: Equivalence of results. *The Clinical Neurologist, 6*, 9-15.

Davies, M. F. (1997). Private self-consciousness and the acceptance of personality feedback: Confirmatory processing in the evaluation of general vs. specific self-information. *Journal of Research in Personality, 31*, 78-92. PsycINFO

Dawes, R. M., Faust, D. & Meehl, P. E. (1989, March 31). Clinical versus actuarial judgment. *Science, 243*, 1668-1674. PsycINFO

Dickson, D. H. & Kelly, I. W. (1985). The "Barnum effect" in personality assessment: A review of the literature. *Psychological Reports, 57*, 367-382. PsycINFO

Erdman, H. P., Klein, M. H. & Greist, J. H. (1985). Direct patient computer interviewing. *Journal of Consulting and Clinical Psychology, 53*, 760-773. PsycINFO

Erdman, H. P., Klein, M. H., Greist, J. H., Skare, S. S., Husted, J. J., Robins, L. N., Helzer, J. E., Goldring, E., Hamburger, M. & Miller, J. P. (1992). A comparison of two computer-administered versions of the NMIH Diagnostic Interview schedule. *Journal of Psychiatric Research, 26*, 85-95. PsycINFO

Eyde, L., Kowal, D. M. & Fishburne, F. J. (1991). The validity of computer-based test interpretations of the MMPI.(In T. B. Gutkin & S. L. Wise (Eds.), *The computer and the decision-making process* (pp. 75—123). Hillsdale, NJ: Erlbaum.)

Farrell, A. D., Camplair, P. S. & McCullough, L. (1987). Identification of target complaints by computer interview: Evaluation of the Computerized Assessment System for Psychotherapy Evaluation and Research. *Journal of Consulting and Clinical Psychology, 55*, 691-700. PsycINFO

Finger, M. S. & Ones, D. S. (1999). Psychometric equivalence of the computer and booklet forms of the MMPI: A meta-analysis. *Psychological Assessment, 11*, 58-66. PsycINFO

First, M. B. (1994). Computer-assisted assessment of *DSM—III—R* diagnosis. *Psychiatric Annals, 24*, 25-29.

First, M. B., Opler, L. A., Hamilton, R. M., Linder, J., Linfield, L. S., Silver, J. M., Toshav, N. L., Kahn, D., Williams, J. B. W. & Spitzer, R. L. (1993). Evaluation in an inpatient setting of DTREE, a computer-assisted diagnostic assessment procedure. *Comprehensive Psychiatry, 34*, 171-175. PsycINFO

First, M. B., Williams, J. B. W. & Spitzer, R. L. (1989). *DTREE: The electronic DSM 2DIII 2 DR.* (Washington, DC: American Psychiatric Press)

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

Fray, P. J., Robbins, T. W. & Sahakian, B. J. (1996). Neuropsychological applications of CANTAB. *International Journal of Geriatric Psychiatry, 11*, 329-336. PsycINFO

French, C. C. & Beaumont, J. G. (1990). A clinical study of the automated assessment of intelligence by the Mill Hill Vocabulary Test and the Standard Progressive Matrices Test. *Journal of Clinical Psychology, 46*, 129-140. PsycINFO

Furnham, A. (1989). Personality and the acceptance of diagnostic feedback. *Personality and Individual Differences, 10*, 1121-1133. PsycINFO

Furnham, A. & Schofield, S. (1987). Accepting personality test feedback: A review of the Barnum effect. *Current Psychological Research and Reviews, 6*, 162-178. PsycINFO

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment.* (Washington, DC: American Psychological Association)

Garb, H. N. & Schramke, C. J. (1996). Judgement research and neuropsychological assessment: A narrative review and meta-analyses. *Psychological Bulletin, 120*, 140-153. PsycINFO

Golden, C. J. (1987). Computers in neuropsychology.(In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide.* New York: Basic Books.)

Greist, J. H., Klein, M. H., Erdman, H. P., Bires, J. K., Bass, S. M., Machtinger, P. E. & Kresge, D. G. (1987). Comparison of computer- and interviewer-administered versions of the Diagnostic Interview Schedule. *Hospital and Community Psychiatry, 38*, 1304-1310. PsycINFO

Grove, W. M. & Meehl, P. E. (1996). Comparative efficiency of information (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293-323. PsycINFO

Grove, W. M., Zald, D. H., Lebow, B., Smith, E. & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30. PsycINFO

Guastello, S. J., Guastello, D. D. & Craft, L. L. (1989). Assessment of the Barnum Effect in computer-based test interpretations. *The Journal of Psychology, 123*, 477-484. PsycINFO

Guastello, S. J. & Rieke, M. L. (1990). The Barnum effect and validity of computer-based test interpretations: The Human Resource Development Report. *Psychological Assessment, 2*, 186-190. PsycINFO

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

Helzer, J. E., Robins, L. N., McEvoy, L. T., Spitznagel, E. L., Stoltzman, R. K., Farmer, A. & Brockington, I. F. (1985). A comparison of clinical and

Diagnostic Interview Schedule diagnoses. *Archives of General Psychiatry, 42*, 657 PsycINFO

Hile, M. G. & Adkins, R. E. (1997). Do substance abuse and mental health clients prefer automated assessments? *Behavior Research Methods, Instruments, & Computers, 29*, 146-150. PsycINFO

Hirschfeld, R., Spitzer, R. I. & Miller, R. G. (1974). Computer diagnosis in psychiatry: A Bayes approach. *The Journal of Nervous and Mental Disease, 158*, 399-407. PsycINFO

Hofer, P. J. (1985). Developing standards for computerized psychological testing. *Computers in Human Behavior, 1*, 301-315. PsycINFO

Hofer, P. J. & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology, 53*, 826-838. PsycINFO

Holden, R. R. & Hickman, D. (1987). Computerized versus standard administration of the Jenkins Activity Survey (Form T). *Journal of Human Stress, 13*, 175-179. PsycINFO

Honaker, L. M. (1988). The equivalency of computerized and conventional MMPI administration: A critical review. *Clinical Psychology Review, 8*, 561-577. PsycINFO

Honaker, L. M., Harrell, T. H. & Buffaloe, J. D. (1988). Equivalency of Microtest computer MMPI administration for standard and special scales. *Computers in Human Behavior, 4*, 323-337. PsycINFO

Jemelka, R. P., Wiegand, G. A., Walker, E. A. & Trupin, E. W. (1992). Computerized offender assessment: Validation study. *Psychological Assessment, 4*, 138-144. PsycINFO

Knights, R. M. & Watson, P. (1968). The use of computerized test profiles in neuropsychological assessment. *Journal of Learning Disabilities, 1*, 6-19.

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

Lambert, M. E., Andrews, R. H., Rylee, K. & Skinner, J. (1987). Equivalence of computerized and traditional MMPI administration with substance abusers. *Computers in Human Behavior, 3*, 139-143. PsycINFO

Layne, C. (1979). The Barnum effect: Rationality versus gullibility? *Journal of Consulting and Clinical Psychology, 47*, 219-221. PsycINFO

Lees-Haley, P. R., Williams, C. W. & Brown, R. S. (1993). The Barnum effect and personal injury litigation. *American Journal of Forensic Psychology, 11*, 21-28.

Levitan, R. D., Blouin, A. G., Nvarro, J. R. & Hill, J. (1991). Validity of the compiterized DIS for diagnosing psychiatric inpatients. *Canadian Journal of Psychiatry, 36*, 728-731. PsycINFO

Maddux, C. D. & Johnson, L. (1998). Computer assisted assessment.(In H. B. Vance (Ed.), *Psychological assessment in children* (2nd ed.; pp. 87—105). New York: Wiley.)

Matarazzo, J. D. (1986). Computerized clinical psychological interpretations: Unvalidated plus all mean and no sigma. *American Psychologist, 41*, 14-24.

Mathisen, K. S., Evans, F. J. & Meyers, K. M. (1987). Evaluation of the computerized version of the Diagnostic Interview Schedule. *Hospital and Community Psychiatry, 38*, 1311-1315. PsycINFO

McKee, L. M. & Levinson, E. M. (1990). A review of the computerized version of the Self-Directed Search. *The Career Development Quarterly, 38*, 325-333. PsycINFO

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* (Minneapolis: University of Minnesota Press)

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

Merten, T. & Ruch, W. (1996). A comparison of computerized and conventional administration of the German versions of the Eysenck Personality Questionnaire and the Carroll Rating Scale for depression. *Personality & Individual Differences, 20*, 281-291. PsycINFO

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

Moreland, K. L. (1987). Computerized psychological assessment: What's available.(In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 64—86). New York: Basic Books.)

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

Moreland, K. L. & Onstad, J. A. (1989). Yes, our study could have been better: Reply to Cash, Mikulka, and Brown. *Journal of Consulting and Clinical Psychology, 57*, 313-314. PsycINFO

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

Pellegrino, J. W., Hunt, E. B., Abate, R. & Farr, S. (1987). A computer-based test battery for the assessment of static and dynamic spatial reasoning abilities. *Behavior Research Methods, Instruments, & Computers, 19*, 231-236. PsycINFO

Peters, L. & Andrews, G. (1995). Procedural validity of the computerized version of the Composite International Diagnostic Interview (CIDI-Auto) in the anxiety disorders. *Psychological Medicine, 25*, 1269-1280. PsycINFO

Pinsoneault, T. B. (1996). Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota Multiphasic Personality Inventory—2. *Computers in Human Behavior, 12*, 291-300. PsycINFO

Prince, R. J. & Guastello, S. J. (1990). The Barnum effect in a computerized Rorschach interpretation system. *Journal of Psychology, 124*, 217-222. PsycINFO

Reardon, R. & Loughead, T. (1988). A comparison of paper—pencil and computer versions of the Self-Directed Search. *Journal of Counseling and Development, 67*, 249-252. PsycINFO

Robins, L. N., Helzer, J. E., Croughan, J. & Ratchiff, K. (1981). National Institute of Mental Health Diagnostic Interview Schedule. *Archives of General Psychiatry, 38*, 381 PsycINFO

Roper, B. L., Ben-Porath, Y. S. & Butcher, J. N. (1995). Comparability and validity of computerized adaptive testing with the MMPI-2. *Journal of Personality Assessment, 65*, 358-371. PsycINFO

Ross, H. E., Swinson, R., Larkin, E. J. & Doumani, S. (1994). Diagnosing comorbidity in substance abusers: Computer assessment and clinical validation. *The Journal of Nervous and Mental Disease, 182*, 556-563. PsycINFO

Russell, E. W. (1995). The accuracy of automated and clinical detection of brain damage and lateralization in neuropsychology. *Neuropsychology Review, 5*, 1-68. PsycINFO

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178-200. PsycINFO

Schuldberg, D. (1988). The MMPI is less sensitive to the automated testing format than it is to repeated testing: Item and scale effects. *Computers in Human Behaviors, 4*, 285-298. PsycINFO

Sines, J. O. (1966). Actuarial methods in personality assessment.(Chapter in B. A. Maher (Ed.), *Progress in Experimental Personality Research* (Vol. 3; pp. 133—193). New York: Academic Press.)

Sips, H. J. W. A., Catsman-Berrevoets, C. E., van Dongen, H. R., van der Werff, P. J. J. & Brook, L. J. (1994). Measuring right-hemisphere dysfunction in children: Validity of two new computerized tests. *Developmental Medicine and Child Neurology, 36*, 57-63. PsycINFO

Snyder, C. R. & Newburg, C. L. (1981). The Barnum effect in a group setting. *Journal of Personality Assessment, 45*, 622-629. PsycINFO

Snyder, D. K., Widiger, T. A. & Hoover, D. W. (1990). Methodological considerations in validating computer-based test interpretations: Controlling for response bias. *Psychological Assessment, 2*, 470-477. PsycINFO

Styles, I. (1991). Clinical assessment and computerized testing. *International Journal of Man-Machine Studies, 35*, 133-150. PsycINFO

Sukigara, M. (1996). Equivalence between computer and booklet administrations of the new Japanese version of the MMPI. *Educational & Psychological Measurement, 56*, 570-584. PsycINFO

Tallent, N. (1958). On individualizing the psychologist's clinical evaluation. *Journal of Clinical Psychology, 14*, 243-245. PsycINFO

Wakefield, H. & Underwager, R. (1993). Misuse of psychological tests in forensic settings: Some horrible examples. *American Journal of Forensic Psychology, 11*, 55-75. PsycINFO

Watson, C. G., Juba, M., Anderson, P. E. & Manifold, V. (1990). What does the Keane et al. PTSD scale for the MMPI measure? *Journal of Clinical Psychology, 46*, 600-606. PsycINFO

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* (Reading, MA: Addison-Wesley)

Wilson, F. R., Genco, K. T. & Yager, G. G. (1985). Assessing the equivalence of paper-and-pencil vs. computerized tests: Demonstration of a promising methodology. *Computers in Human Behavior, 1*, 265-275. PsycINFO

Wilson, R., Omeltchenko, L. & Yager, G. G. (1991). Coping with stress: Microcomputer software for treatment of test anxiety. *Journal of Behavior Therapy and Experimental Psychiatry, 22*, 131-139.

Wyndowe, J. (1987). The microcomputerized Diagnostic Interview Schedule: Clinical use in an outpatient setting. *Canadian Journal of Psychiatry, 32*, 93-99. PsycINFO

Correspondence may be addressed to James N. Butcher, Elliott Hall, University of Minnesota, 75 East River Road, Minneapolis, Minnesota, 55455.
Electronic mail may be sent to butch001@tc.umn.edu

Received: July 10, 1998
Revised: March 22, 1999
Accepted: March 29, 1999

Table 1. Kappa Statistics for Various Types of Administration of the Diagnostic Interview Schedule (DIS)

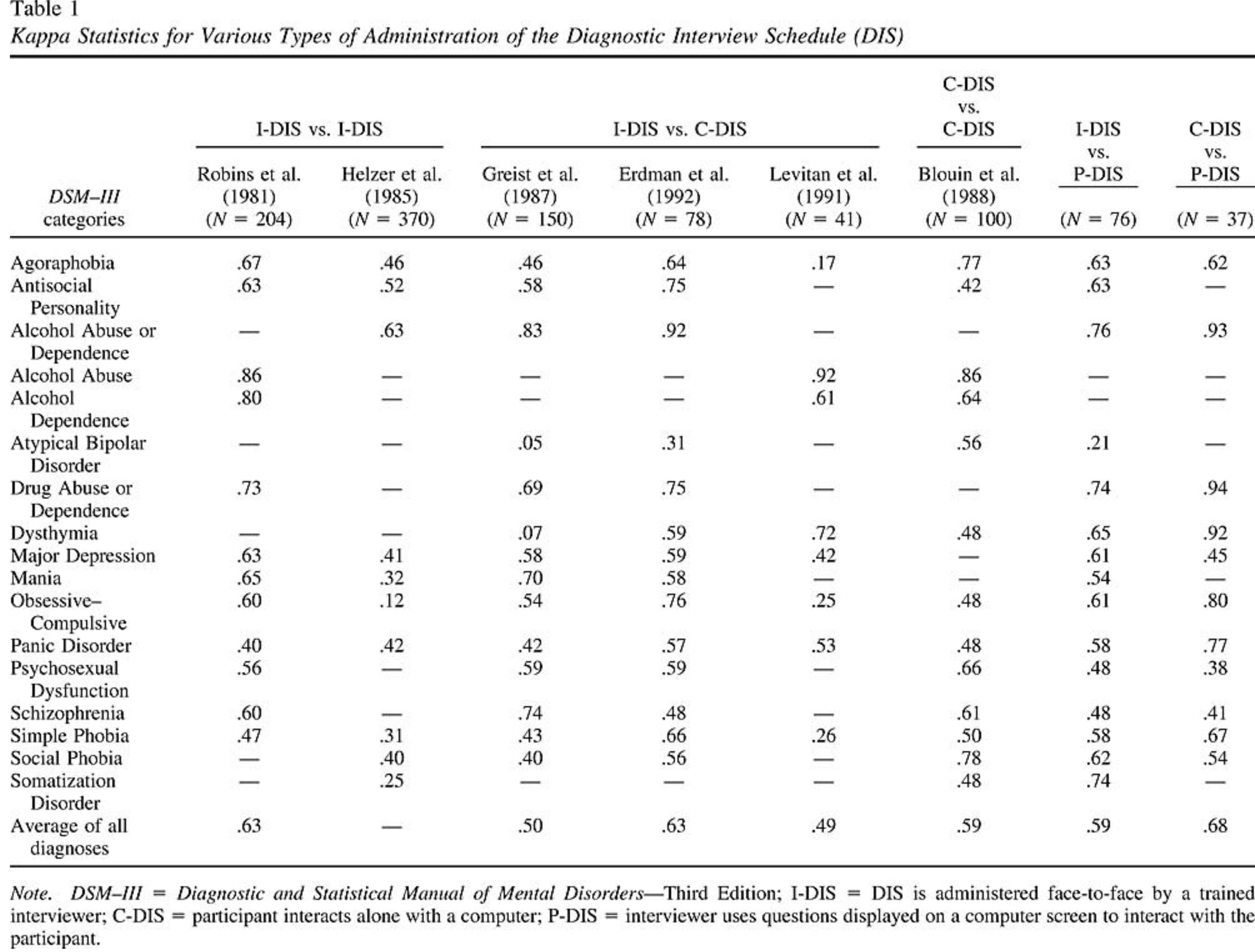

Table 1

Kappa Statistics for Various Types of Administration of the Diagnostic Interview Schedule (DIS)

	I-DIS vs. I-DIS		I-DIS vs. C-DIS			C-DIS vs. C-DIS	I-DIS vs. P-DIS	C-DIS vs. P-DIS
DSM–III categories	Robins et al. (1981) (*N* = 204)	Helzer et al. (1985) (*N* = 370)	Greist et al. (1987) (*N* = 150)	Erdman et al. (1992) (*N* = 78)	Levitan et al. (1991) (*N* = 41)	Blouin et al. (1988) (*N* = 100)	(*N* = 76)	(*N* = 37)
Agoraphobia	.67	.46	.46	.64	.17	.77	.63	.62
Antisocial Personality	.63	.52	.58	.75	—	.42	.63	—
Alcohol Abuse or Dependence	—	.63	.83	.92	—	—	.76	.93
Alcohol Abuse	.86	—	—	—	.92	.86	—	—
Alcohol Dependence	.80	—	—	—	.61	.64	—	—
Atypical Bipolar Disorder	—	—	.05	.31	—	.56	.21	—
Drug Abuse or Dependence	.73	—	.69	.75	—	—	.74	.94
Dysthymia	—	—	.07	.59	.72	.48	.65	.92
Major Depression	.63	.41	.58	.59	.42	—	.61	.45
Mania	.65	.32	.70	.58	—	—	.54	—
Obsessive–Compulsive	.60	.12	.54	.76	.25	.48	.61	.80
Panic Disorder	.40	.42	.42	.57	.53	.48	.58	.77
Psychosexual Dysfunction	.56	—	.59	.59	—	.66	.48	.38
Schizophrenia	.60	—	.74	.48	—	.61	.48	.41
Simple Phobia	.47	.31	.43	.66	.26	.50	.58	.67
Social Phobia	—	.40	.40	.56	—	.78	.62	.54
Somatization Disorder	—	.25	—	—	—	.48	.74	—
Average of all diagnoses	.63	—	.50	.63	.49	.59	.59	.68

Note. *DSM–III* = *Diagnostic and Statistical Manual of Mental Disorders*—Third Edition; I-DIS = DIS is administered face-to-face by a trained interviewer; C-DIS = participant interacts alone with a computer; P-DIS = interviewer uses questions displayed on a computer screen to interact with the participant.

Table 2. Studies of the Computerized Diagnostic Interview Schedule (C-DIS)

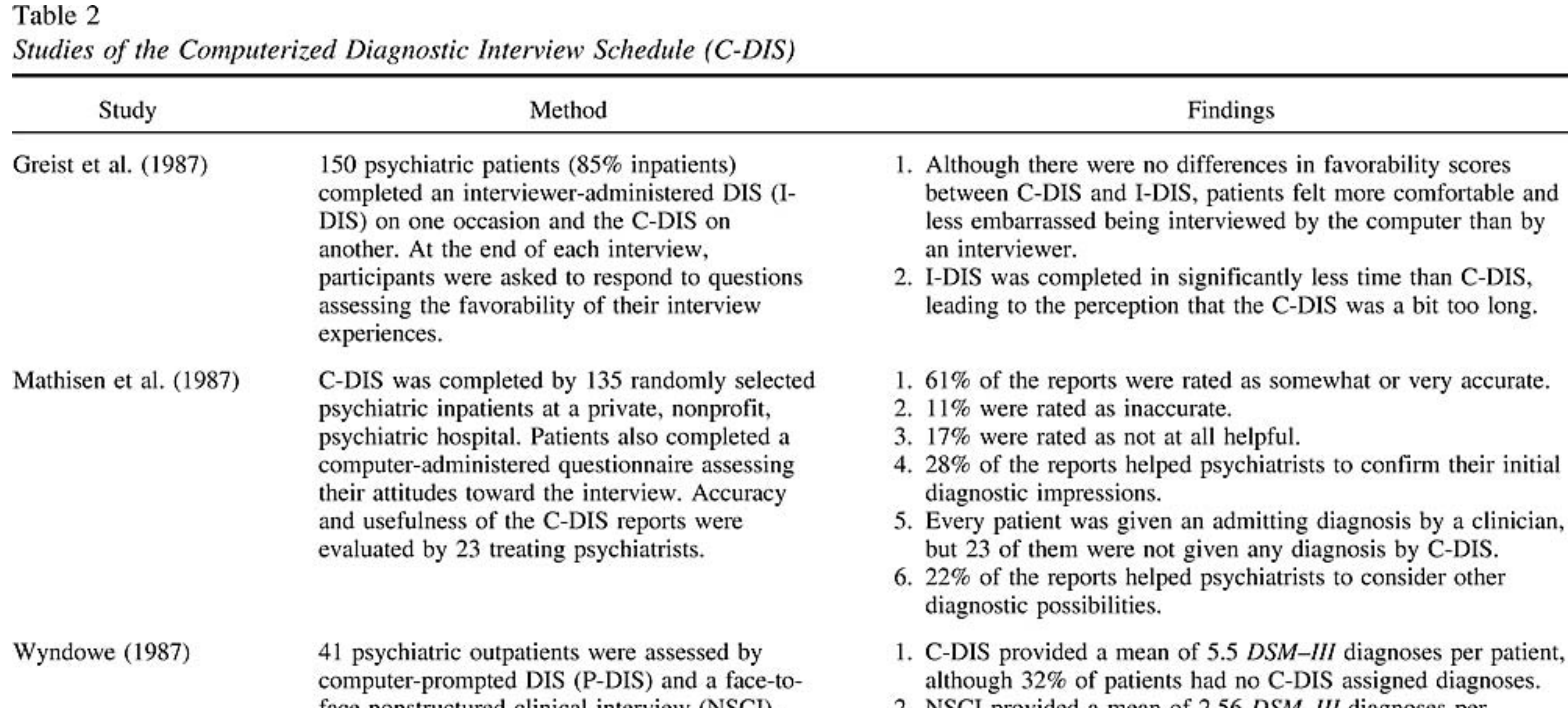

Table 2

Studies of the Computerized Diagnostic Interview Schedule (C-DIS)

Study	Method	Findings
Greist et al. (1987)	150 psychiatric patients (85% inpatients) completed an interviewer-administered DIS (I-DIS) on one occasion and the C-DIS on another. At the end of each interview, participants were asked to respond to questions assessing the favorability of their interview experiences.	1. Although there were no differences in favorability scores between C-DIS and I-DIS, patients felt more comfortable and less embarrassed being interviewed by the computer than by an interviewer. 2. I-DIS was completed in significantly less time than C-DIS, leading to the perception that the C-DIS was a bit too long.
Mathisen et al. (1987)	C-DIS was completed by 135 randomly selected psychiatric inpatients at a private, nonprofit, psychiatric hospital. Patients also completed a computer-administered questionnaire assessing their attitudes toward the interview. Accuracy and usefulness of the C-DIS reports were evaluated by 23 treating psychiatrists.	1. 61% of the reports were rated as somewhat or very accurate. 2. 11% were rated as inaccurate. 3. 17% were rated as not at all helpful. 4. 28% of the reports helped psychiatrists to confirm their initial diagnostic impressions. 5. Every patient was given an admitting diagnosis by a clinician, but 23 of them were not given any diagnosis by C-DIS. 6. 22% of the reports helped psychiatrists to consider other diagnostic possibilities.
Wyndowe (1987)	41 psychiatric outpatients were assessed by computer-prompted DIS (P-DIS) and a face-to-face nonstructured clinical interview (NSCI)	1. C-DIS provided a mean of 5.5 *DSM–III* diagnoses per patient, although 32% of patients had no C-DIS assigned diagnoses. 2. NSCI provided a mean of 2.56 *DSM–III* diagnoses per

	face nonstructured clinical interview (NSCI), about a week apart. After both interviews, patients answered seven questions about their reactions to P-DIS.	2. NSCI provided a mean of 2.56 *DSM–III* diagnoses per patient. 3. 9 (30%) had the same P-DIS and NSCI diagnosis. 4. 12 (29%) had primary NSCI diagnosis in the range of P-DIS but not supported by P-DIS. 5. 20 (49%) had primary NSCI diagnosis not supported by P-DIS. 6. Regarding endorsement of P-DIS items: Relevant = 74%; Interesting = 77%; Tiring = 23%; Difficult = 19%; Boring = 26%; Hard to understand = 16.9%; Left out important questions = 22%. 7. P-DIS resulted in longer interviewing time than that previously reported for I-DIS (i.e., about 1 hr).
Blouin et al. (1988)	80 psychiatric patients (50% inpatients) and 20 normal control subjects completed C-DIS on two occasions, about 1 week apart. Participants completed a Computer Attitude Scale (CAS) before the first C-DIS administration and after the second.	1. C-DIS provided a mean of 4.9 *DSM–III* diagnoses per patient. 2. Patients were significantly more intimidated and resentful of computers than normal participants. 3. The patients felt significantly more expert at using computers following second administration of C-DIS. 4. Normal participants felt significantly less intimidation and resentment following two administrations.
Levitan et al. (1991)	41 psychiatric inpatients were assessed by C-DIS after being administered a symptom checklist in the form of semi-structured *DSM–III* clinical interview.	1. C-DIS provided a mean of 3.9 *DSM–III* diagnoses per patient, as compared to 3.5 diagnoses from the symptom checklist.
Erdman et al. (1992)	117 patients (38% inpatient) completed I-DIS, C-DIS, and P-DIS. After completing the C-DIS, they responded to a paper-and-pencil evaluation of their experience. Each patient also completed measures of reading ability, deviant response bias, and social desirability response bias.	1. Patients who were more likely to give socially desirable responses reported fewer symptoms and received fewer diagnoses. 2. Patients who were more likely to give more deviant responses reported more symptoms and received more diagnoses. 3. There was no correlation between reading ability and diagnostic disagreement and interview length. 4. Participants did not feel that the computer interview was longer than a traditional interview. 5. Patients felt that they could better describe their ideas and feelings to a human, yet they found the computer to be less embarrassing.
Ross et al. (1994)	173 substance abusers were assessed by a revised C-DIS (on *DSM–III–R* diagnoses) followed by a Structured Clinical Interview for *DSM–III–R* (SCID). 1 to 2 weeks later, another SCID was administered by a different clinician, after which a consensus diagnoses was made based on all available information.	1. Diagnostic agreement between C-DIS and SCID was poor for all diagnoses other than psychoactive substance-use disorders. 2. Skipping out of disorders was a potential source of disagreement between positive C-DIS and negative SCID diagnoses.

Note. I-DIS = DIS is administered face to face by a trained interviewer; C-DIS = participant interacts alone with a computer; P-DIS = interviewer uses questions displayed on a computer screen to interact with the participant; *DSM–III* = *Diagnostic and Statistical Manual of Mental Disorders*—Third Edition.