1997

Errors in Scoring Objective Personality Tests

Gregory Allard

University of Rhode Island

Recommended Citation

Allard, Gregory, "Errors in Scoring Objective Personality Tests" (1997). *Open Access Master's Theses.* Paper 1105.
https://digitalcommons.uri.edu/theses/1105

ERRORS IN SCORING OBJECTIVE PERSONALITY TESTS

BY

GREGORY ALLARD

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTERS OF ARTS
IN
PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

1997

MASTER OF ARTS THESIS

OF

GREGORY ALLARD

Approved:

Thesis Committee

Major Professor David Faust

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

1997

Abstract

Scoring objective personality tests is considered clerical, and presumably, straightforward in nature. This may be the reason that few studies, if any, have investigated the impact of scoring error on widely used tests, such as the MMPI or BDI. Errors, even if infrequent (e.g. as few as 1% of tests), may adversely affect many hundreds or thousands of tests administered annually, however. In a study of three popular tests taken from three independent settings, this study found that the interpretation of popular tests are vulnerable even to small errors (i.e., 1 or 2 misscored items per test). This study explored the influence of two factors, scoring procedure complexity and commitment to scoring accuracy, hypothesized to be related to the occurence of scoring error with fewer errors occuring when higher commitment to accuracy and lower scoring procedure complexity are present. The scoring procedure complexity effect was predicted to be subordinate to the commitment to accuracy effect. Three popular tests were sampled from three different settings and rescored to check for accuracy. Twenty-one percent of tests scored with low commitment to accruacy were erroneous, while tests scored with full commitment to accuracy had 1% errors. Scoring procedure complexity, categorized as high and low, yielded 29% and 14% erroneous tests, respectively, in the less than full commitment to accuracy sample, and 0 and 4% in the full commitment to accuracy sample. The results provide strong support for the factors as major predictors of scoring error, as well as the interaction effect anticipated. Other risk factors, such as commercial computer scoring errors and lack of agreement on test scoring standards, were also found to distort scores. The frequency and severity of erroneous findings in this study, the author argues, are unlikely to be specific to this study, but instead more general. The author shows how awareness of the two factors, as well as other sources of error, can be used to reduce the risk of scoring error and offers practical recommendations to improve scoring accuracy.

Acknowledgments

I am grateful to my advisor, David Faust, for his contribution in completing this work. Through our collaboration, David has contributed more than mere content knowledge. Rather, he has imparted me with a new awareness of something basic: how to pose a question so that the answer can be worth knowing. This inquiry has stretched my perception, my thinking, increasing my momentum in the road that lies ahead.

I have been blessed with a tireless cheerleading squad: Beth, my wife, for her unending emotional support and love; Julian Butler for his rare insights and programmatic technical support; and Doreen Lawson for her talents in managing logistics.

Table of Contents

List of Tables

List of Figures

I. Scope of Study

Clinical interpretation cannot be better than the data upon which it depends. As with all psychological tests, certain sources of error are unavoidable, such as those stemming from limits in scientific knowledge and state-of-the-art measurement technology. Other errors are potentially avoidable, such as the failure to collect available but key sources of information and mechanical errors in scoring or tallying results on psychological tests. The design of some psychological tests, including objective personality instruments, virtually eliminates many types of errors. Nevertheless, some preventable errors, such as mechanical or clerical scoring errors, may still occur in the course of psychological testing.

This study aimed to determine whether error in scoring objective personality tests should concern the clinical community. The answer depends on the clinical significance and frequency of such errors. Although "clinical significance" is an open and value-laden construct, broad consensus is likely to be obtained in certain cases, such as those in which errors alter diagnoses or major treatment recommendations in a deleterious direction. Further, error rates on more popular tests, even if comparable to those found on more obscure tests, demand more immediate attention because of their greater overall adverse impact. In this inquiry, therefore, I focused on objective personality tests administered frequently nationwide.

A secondary focus of this inquiry was to explore whether scoring errors could be traced to systematic factors. Sources of systematic error include qualities of the scorers themselves, test settings, and tests' scoring procedures. One type of systematic error relating to scorer qualities might be level of training. For instance, Ph.D.'s may be less susceptible to scoring errors than non-doctorates. Such findings might suggest that Ph.D.'s should score objective personality tests over non-doctorates to reduce scoring error. In general, I attempted to evaluate the more "promising" sources of error and to consider possible corrective suggestions.

II. Justification and Significance of the Study

The primary justification of this study lies with the importance of accuracy in scoring frequently administered objective personality tests. Objective personality tests are used many thousands of times annually to aid clinicians in assessing individual psychological characteristics or maladies. In turn, these results may determine diagnosis, expert testimony in legal cases, or psychotherapeutic or psychopharmacological treatment recommendations. Thus, objective personality test results and interpretations can have a major impact on individual lives.

The clinical and scientific community has devoted little attention to possible scoring errors on objective personality tests, perhaps mistakenly. For frequently administered tests, even seemingly low or very low error rates can affect many individuals. For example, surveys conducted in the last two decades suggest that as many as four million people undergo psychological assessment across the US in a given year (Levine & Willner, 1976; Zilbergeld, 1983). If, say, the Minnesota Multiphasic Personality Inventory (MMPI), the most frequently used objective personality test (Piotrowski, Sherry, & Keller, 1985; Wade & Baker, 1977), is administered in 20% of those cases, this projects to 800,000 MMPI's administered annually. If scoring errors that result in clinically significant errors occurred at a seemingly low rate of 1 to 2% across MMPI's, then 8000 to 16,000 people in that year might be erroneously assessed on the MMPI due to potentially avoidable scoring error. From this standpoint, especially considering the feasibility of eliminating such errors almost entirely, what would appear to be a low scoring error rate is clearly unacceptable for tests administered so frequently. Obviously, if error rates are lower, mainly involve less popular tests, and rarely create meaningful changes in test profiles or interpretations, the problem may not merit much concern.

II.I Scoring Error on Cognitive Tests

Scoring accuracy has been scrutinized much more closely on cognitive tests as opposed to objective personality tests, with this research dating back at least 25 years (e.g., Miller, Chansky, & Gredler, 1970). Scoring and administrating cognitive tests requires considerable training, practice, skill, and subjective judgments, and thus scoring accuracy understandably has been of higher concern. With objective personality tests, interpretation is usually the primary concern; administration and scoring are considered merely clerical in nature.

Much of the literature on scoring accuracy focuses on the most frequently administered cognitive tests (Piotrowski & Keller, 1989; Piotrowski & Keller, 1992), such as the Wechsler Intelligence Scales and the Stanford-Binet.

Various studies have uncovered problems with scoring errors. For example, Warren and Brown (1973) rechecked 240 WISC's and Stanford-Binet's scored by 40 graduate students and found discrepant Full Scale IQ's in 37% of cases. Ryan, Prifitera, and Powers (1983) presented 19 psychologists and 20 graduate students with the same two WAIS-R protocols and later examined interrater scoring differences. About two-thirds of the test scores were not in agreement, and 23% of the differences exceeded one standard error of measurement. Because IQ scores are used to make academic, vocational, or other types on placement decisions, scoring errors can adversely affect the test taker's well-being or future opportunities.

Scoring cognitive tests can be a difficult task. Besides mechanical and clerical tasks, scoring sometimes requires sophisticated subjective judgments. Most literature has identified facets of scoring involving subjective judgment as a greater source of error than mechanical or arithmetic operations (Boehm, Duker, Haesloop, & White, 1974; Miller & Chansky, 1972; Slate & Chick, 1989; Slate & Jones, 1990; Slate, Jones, & Murray, 1991). Accordingly, corrective suggestions focus primarily on the subjective elements, such as practice or special training and instructor feedback programs designed to ensure more uniformity in scoring (Blakey, Fantuzzo, Gorsuch, & Moon, 1987; Boehm et al., 1974; Connor & Woodall, 1983; Slate et al., 1991).

The predominance of errors stemming from judgment factors does not mean that errors resulting from mechanical and clerical tasks are rare or insignificant. Such errors include the addition of subscale scores, table conversions, and calculation of chronological age, among others. Research suggests that such mechanical errors occur in anywhere from 1% to 50% of cases and can be of clinical significance (Beasley, Lobasher, Henley, & Smith, 1988; Boehm et al., 1974; Miller & Chansky, 1972; Miller et al., 1970; Sherretts, Gard, & Langner, 1979).

II.II Scoring Error Studies on Objective Personality Tests

Scoring objective personality tests entails mechanical and clerical tasks similar to those of cognitive tests. Currently, few published studies address the possible occurrence and impact of scoring error on objective personality tests. Allard, Butler, Shea, and Faust (1995) examined the accuracy with which individuals scored the Personality Diagnostic Questionnaire-Revised (PDQ-R). They found clerical errors in 53% of protocols, resulting in changed diagnostic classification in 19% of cases. Due to the PDQ-R's relatively low frequency of use and complexities involved in scoring it, Allard et al. conducted two additional exploratory analyses.

First, using data from the same setting but a different group of scorers, Allard et al. analyzed scoring accuracy for a more widely used measure, the Symptom Checklist-90, Revised (SCL-90R). T-score profile calculations at the study setting were performed by hand, wherein the scorer located the appropriate table and matched rounded raw scores within a T-score matrix. In a random sample of 35 protocols, the authors uncovered 85 hand-scoring errors (M = 2.43 errors per protocol), which altered T-score profiles in 29 cases, or 82.8% of the protocols. Second, the authors also contacted a half-dozen prominent consulting psychologists who, in the course of their practices, often check on the accuracy of psychological test scores. Each of these psychologists examines the work that other psychologists perform in the context of legal assessments and reviews cases from around the US. All indicated that they checked on the accuracy of objective personality test scores. Each psychologist also indicated that they found errors, although estimates of frequency varied from "not rare" to almost 50% of cases reviewed. All agreed that such errors could be highly significant. Although this small, informal "survey" obviously had serious methodological limits, the results, together with the analysis of the SCL-90R, lent further credence to the main findings of the PDQ-R study. Furthermore, these findings clearly raise the possibility that error in scoring objective personality tests represents a problem that may warrant concern.

II.III Review of Scoring Error Factors

In the search for factors associated with mechanical or clerical scoring errors, studies on cognitive tests have focused primarily on individual and setting variables. Individual variables have included demographics (e.g., educational level, gender) and type of test training and experience. These studies have shown small or contradictory effects. For instance, some studies on level of education (e.g., Ph.D. vs. graduate student) have shown a weak tendency toward students committing fewer errors than their mentors (Levenson, Golden-Scaduto, Aiosa-Karpas, & Ward, 1988; Ryan et al., 1983; Slate, Jones, Murray, & Coulter, 1993), but other investigations have yielded non-significant results (Oakland, Lee, & Axelrod, 1975; Sherretts et al., 1979). Other studies have investigated age and gender and have found minor, if any, effects (Oakland et al., 1975; Levenson et al., 1988). Further studies have explored the effect of training programs and practice. Except for one training program study (Boehm et al., 1974), most have shown some meaningful improvements in accuracy, but not in reducing mechanical or clerical error (Blakey et al., 1987; Connor & Woodall, 1983; Slate et al., 1991). A few studies have investigated differences between setting variables, such as metropolitan versus rural schools, or schools versus psychiatric clinics. Although small differences have sometimes been found between settings, error rates were found to be unacceptable across situations (Johnson & Candler, 1985; Sherretts et al., 1979).

The common element underlying many of these studies involving individual or setting variables is the lack of consistent or robust effects. This may be because such variables do not directly tap the most influential factors, and rather show weak probabilistic relations to underlying variables that exert more direct and powerful effects. One such underlying variable may be commitment to accuracy in scoring. In one study, metropolitan school psychologists were more accurate than rural school psychologists (Johnson & Candler, 1985). The researchers suggested that it was not the setting itself that directly accounted for the outcome, but rather that those in the metropolitan setting were more "conscientious" in their work, and therefore were more likely to score accurately compared to those in the predominantly itinerant, rural setting. Additionally, various researchers, who have studied scorer training programs have concluded that errors persist because of "carelessness," especially in clerical operations (Miller & Chansky, 1972; Slate & Chick, 1989; Slate, et al., 1991). Researchers, stymied by their efforts to rectify careless errors, have suggested using computer scoring programs (Johnson & Candler, 1985) or double-checking scoring (Miller & Chansky, 1972; Slate & Hunnicutt, Jr., 1988). These various findings, conclusions, and suggestions seem to converge on the same point: commitment to accuracy is a central determinant in scoring error.

Other potential variables associated with scoring error relate to the instruments themselves, in particular, the complexity involved in scoring them (Slate & Hunnicutt, 1988). Allard et al. (1995) found strong effects between the frequency of scoring error and scoring procedure complexity. When complexity of scoring operations increased, so, too, did scoring errors. The study revealed that items that are more difficult to score result in more errors, and scales that comprise higher quantities of heterogeneous scoring procedures or that require deeper cognitive processing are more prone to scoring error. In Allard et al.'s study, the effect of scoring procedure complexity was considerable, with the relationship between scoring error and scoring procedure complexity accounting for at least half of the total error variance. Other analyses on limited samples of more frequently administered objective personality tests, such as the MMPI, the Beck Depression Inventory, and the SCL-90R also seem to show error patterns that relate to scoring procedure complexity (Allard et al., 1995). The Beck Depression Inventory, which is simple to score, yielded much lower error rate than the SCL-90, which requires both addition and T-score profile conversions.

Thus, the relationship between test design and scoring error warrants further investigation.

III. Objectives, Variables under Study, and Clarification of Assumptions

The primary objective of the study was to determine whether scoring errors on objective tests should concern the clinical community. The study also attempted to examine two factors expected to relate to scoring accuracy: a) commitment to accuracy, and b) complexity of scoring procedures. Commitment to accuracy (CTA) was assessed by determining whether the scorer had taken certain actions in scoring a test. CTA was therefore conceptualized as a set of behaviors, rather than as a hypothetical construct. "Full" CTA was considered to be present when <u>all</u> operations of test scoring were either double-checked (i.e., scored twice) or optically scanned and computer scored; and "less-than-full" CTA was considered to be present when scoring operations consisted of unchecked keypunching or less than fully double-checked hand-scoring. It was expected that CTA would influence scoring accuracy such that tests scored with less-than-full CTA, unlike tests scored with full CTA, would yield problematic or unacceptable error rates. One way to determine the point at which error rate is unacceptable would be to survey the clinical community and solicit opinion on this matter.

Scoring procedure complexity (SPC) was also proposed to have strong effects on scoring accuracy. Scoring procedure complexity seems like a relatively straightforward concept and, for the purposes of this study, was defined as the number of procedures required to conduct scoring. Although fine distinctions in complexity may be challenging to assess, there are gross differences between the measures that were investigated in this study. For example, a test like the Beck Depression Inventory (BDI) merely requires the addition of one column of raw scores to attain a final score. The BDI has much lower scoring procedure complexity than, say, the hand-scored version of the MMPI, which requires not only addition, but a series of other procedures, such as correcting the number of raw scores on a number of scales by a different proportion of the score on another scale. It was expected that tests with lower SPC would yield fewer scoring errors than tests with higher SPC, in part depending on CTA. To test this, the study included tests that varied in SPC. Test SPC was thus expected to be a meaningful source of error only under conditions in which there was less-than-full CTA. Stated differently, even with complex protocols, tests scored with full CTA were expected to drastically reduce error.

A <u>precise</u> determination of the <u>ultimate impact</u> of scoring errors on some of the tests used in this study was not feasible. With tests like the BDI, where a distribution of errors can be easily converted into frequencies of change in classification, analyzing the impact of error is relatively straightforward. In contrast, tests like the MMPI pose certain difficulties that make a determination of impact a potentially formidable task. MMPI interpretation depends on the interrelation of 10 or more scale scores with different numbers of items that are coded in a variety of ways, and there is no obvious "population" of altered or misscored MMPI protocols.

What is known is that changes as little as one point on a single MMPI scale can alter the high two-point (scale) configuration, which is often considered the crux of MMPI interpretation. This phenomenon likely exists with other objective personality tests, too.

IV. Methodology

IV.I Sampling Domains

For the research questions posed here, sampling issues bear special attention. For one, examining whether scoring error should raise concern among clinicians requires directly tapping into the common tools of their practice; i.e., popular tests. Another reason for sampling popular tests is that tests even with seemingly low error rates might affect multitudes of tests administered nationally every year. Sampling all popular tests would likely represent most clinicians' testing armamentarium, but doing so would entail impractical burden. Sampling all popular tests would be more than sufficient to demonstrate that scoring errors should concern clinicians. Should unacceptable rates of error be found on a number of tests, this would suggest that the occurrence of error is not isolated to any one test. Furthermore, if errors are found on a variety of frequently administered tests, such results would raise concern, regardless of findings on other tests not sampled in this study.

As mentioned above, restricting the type of tests sampled in this study is a practical consideration, but restricting the number of tests sampled in this study also creates additional complications. To determine whether errors exist requires sampling enough tests to reveal the existence of error. Restricting the quantity of each test sampled increases the risk to the investigator falsely uncovering negative findings.

Additional caveats stem from the nature of the research design. Explicit constraints on collecting many test samples are imposed by the CTA factor related to scoring error, for its examination requires stratifying tests scored with full CTA and those with less-than-full CTA. Merely locating a few tests where full and less-than-full CTA samples are available would likely prove challenging; thus, finding all popular tests would not be practical. Another explicit constraint involves the implications of this research on those who participate; participants must be willing to undergo scrutiny that could reveal relevant and potentially damaging errors in patient records. Willing participants are thus unlikely to surface, thereby making it very difficult to sample all popular tests.

Still, tests must be chosen that are relevant to clinical practice and must also demonstrate the factors related to scoring error, CTA and SPC. Although sampling all popular tests is not feasible, choosing at least some popular tests seemed necessary to enhance clinical relevance and the potential for generalizing findings. Certainly, more than one test must be selected so as to expose whether scoring error is specific to one popular test, or rather, more general. Tests must also be chosen from sufficiently diverse settings. Note that the representation of diverse settings does not necessarily require sampling all types of clinical settings. For the purposes of this research, setting diversity is needed to discern whether error patterns discovered on tests sampled are isolated or maybe more general. Discerning whether scoring error patterns are specific to a particular test or setting is possible if all tests sampled are common to all settings. To explore factors associated with scoring error, types of tests chosen must vary in SPC levels. For all test types chosen, at least two test types must vary in SPC level to make SPC measurement possible. Lastly, measuring the CTA factor requires sampling tests scored with differing levels of CTA. To discern CTA effects from isolated effects of test design, common tests should be chosen for both CTA samples. Moreover, isolated effects attributed with particular settings can be controlled for by obtaining both CTA levels within each setting. Since full CTA test samples are expected to yield virtually no detectable scoring errors, however, collecting samples with full CTA from all participating settings would likely be redundant and thus unnecessary. Simply requiring only one setting to provide full CTA data for each test type sampled would seemingly be sufficient.

If the conditions as noted can be satisfied or even met roughly, the design of this study represents both a risky and specific test of whether scoring error should be of concern, and whether the factors in question have power in explaining the occurrence of error. For one, the examination of error at more than one setting allows for the disconfirmation of the assertion that error should be of general concern. Secondly, factors associated with scoring error can be examined to see whether they apply across settings, another risky test.

For the most part, the process of sampling settings and tests for the study went smoothly in that all settings queried agreed to participate. Three diverse settings elected to participate: a VA inpatient hospital, a VA outpatient clinic, and a private inpatient hospital. At each setting, many popular tests were available. Three popular tests were common across all three settings: the MMPI, the Beck Depression Inventory (BDI), and the Spielberger State Trait Inventory (STAI). Surveys show that all three tests chosen, particularly the MMPI and BDI, are considered among the most widely used in the clinical community (Piotrowski & Keller, 1989, 1992; Piotrowski & Lubin, 1990).

Each of these three test types when scored fully by hand also vary in SPC ratings. Table 1 shows the steps required to score each fully hand-scored test type and respective SPC rankings: low, medium, and high. Scoring procedures used in the settings that were sampled, however, reduced SPC ratings to two discernible categories, low and high. As noted, SPC was defined as the number of distinctly different cognitive or procedural operations required to arrive at an interpretable score. The BDI was rated as a low SPC test; it requires adding the raw item responses to derive a total score. The MMPI, if completely hand-scored, would have represented the highest SPC level among the three tests because it entails many steps; several subscales must be tallied and converted to T-scores on lookup tables. In practice, however, all three settings scored MMPI's with a computer program that only required keypunching item responses. This process reduced the SPC to one clerical task of low complexity.

Insert Table 1 about here

Table 2 reflects a revision of Table 1, showing the MMPI with the reduced number of scoring steps and corresponding reduced scoring complexity. Note that the STAI is administered to the patient in two parts, called the State (STAI_S) and Trait (STAI_T) forms. Because both parts were split during data collection, separate STAI_S and STAI_T samples were collected. Scoring both the State (STAI_S) and Trait (STAT_T) forms were considered to be of high complexity because some items must be reverse-coded before both scales are tallied (two separate steps per form). In summary, both STAI forms were rated as high SPC, and the MMPI and BDI were rated as low SPC.

Insert Table 2 about here

Participating settings were screened to assess their CTA. All three types of settings were to provide data scored with less-than-full CTA to determine its impact on accuracy. As noted, requesting all three settings to provide data with full CTA would likely have produced three error-free, and thus redundant, data sets. Therefore, choosing only one setting to provide such data would have been sufficient, particularly if that setting could have provided both types of data sets. In the event that no setting could supply both data types, I had intended to obtain an additional setting to satisfy the requirements of one setting with full CTA. Because none of the participating settings could furnish full CTA data sets and locating additional settings that could supply full CTA data became impractical, I created a simulated full CTA data set.

The final sampling consideration involves the number of specific tests from each setting. This number was set at 50 per test type at each setting. Besides test availability constraints in archives at the settings, 50 was considered to be a large enough sample to reveal relatively low frequencies of erroneous tests. Simulated data for the full CTA data set was derived from tests sampled at all three settings. Fifty MMPI's, 50 BDI's, 50 STAI_S's, and 50 STAI_T's were reproduced.

Besides sampling archived test data, I also conducted a survey of the clinical community to examine their perception of acceptable error rates. The sample chosen was comprised of randomly picked representatives of the American Psychological Association Clinical Psychology Fellows (Division 12). Nomination as a Fellow is intended to reflect outstanding and unusual professional contributions; thus, Fellows' opinions should carry some weight. The projected number of total survey participants was set at 50, and I sampled 25 in a pilot study to determine clinician attitudes, knowledge, and practices.

IV.II Procedure

Given the exposure of clinician practices and patient records this research entailed, sampling was carried out with strict regard for confidentiality. Despite the legal ramifications of placing clinical records under scrutiny, the settings were, thankfully, cooperative. I undertook three steps to provide assurances for legal and ethical concerns. First, the identities of participating settings were not and will not be disclosed in any publication or presentation. Second, each individual test was coded to ensure anonymity, and the lists containing the codes and names were stored in locked locations separate from the tests themselves. Lastly, I agreed to supply each participating setting general feedback on the findings pertaining to the specific setting.

The three participating settings provided access to test data for each of the three tests. I selected 50 of each test type (i.e., 50 BDI's, 50 MMPI's, 50 STAI_S's, and 50 STAI_T's) randomly from archives that were available. Note that because the STAI_S and STAI_T tests were sampled separately at each setting, I chose a random sample of 50 of each part. Added to the 50 BDI's and 50 MMPI's, each setting thus provided a total of 200 tests. In sampling test data, I obtained patients' raw data answer sheets, and, if applicable, derived summary score sheets or original keypunched patient responses. I assigned each test a unique ID number.

The resulting rescored tests were compared to the original hand-scored (or key-punched) portions of that test. To obtain accurate test data representation, the tests were independently rescored and double-checked electronically. All programming for this project was accomplished using Microsoft Excel 4.0 or 5.0 macros in PC and Macintosh environments (Microsoft, 1992, 1994). I recruited five high-grade point undergraduate assistants (rescorers) to rescore tests. Each rescorer used individualized scoring programs for each test type to keypunch patient responses and derived scale or summary scores. Tests were distributed such that every test was rescored by two independent rescorers. All scoring programs checked for previously entered test data to prevent each rescorer from scoring the same test twice. After all data were rescored twice, merging programs collected test data files and matched entries by ID numbers. The merging programs automatically compared raw data entries, scale scores, and T-scores for each test entry. These programs automatically identified discrepancies among re-keyed entries to facilitate accurate tracking of rescorer keypunching errors. In contrast to the BDI's and the STAI's, the MMPI's were originally computer scored from keypunched data. Unlike the BDI and STAI where discrepancies among the summary scores ultimately are the only indicators of scoring mistakes, the MMPI errors could be traced to mis-keyed items by comparing both patient item responses and corresponding keypunched responses. This level of detection required additional programming, but the added function enabled the detection of keypunching discrepancies in addition to resulting scale or T-score discrepancies.

After all discrepancies were rectified, rescoring programs automatically produced accurate summary scores or T-scores based on the verified raw data re-entries. The programs then compared the accurate summary and T-scores to those that were originally derived by the settings' scorers (or computer programs in the case of the MMPI's). This process provided the data for analyses that revealed discrepancies within the sampled tests.

Full CTA was simulated by rescoring patient data from each of the three settings using one of the previously described full CTA procedures, in this case, optical scanning and computer scoring. I created the scanning templates using National Computer Systems (NCS) ScanTools Software and scanned all tests using the NCS OpScan 5 optical scanner. Data from 50 of each test type were simulated for a total of 200 tests. The simulated MMPI data was taken from the private inpatient hospital sample because the patient test responses, coincidentally, were originally recorded on NCS scannable forms. The BDI and STAI data were not originally recorded on scannable forms at any of the settings; thus, these raw data had to be transcribed onto scannable sheets. The rescorers transcribed the BDI data from the VA outpatient clinic and STAI_S and STAI_T data from the VA inpatient hospital onto the NCS scannable forms. All scanned data were then compared to the verified double-checked rescored data sets mentioned in the previous paragraphs. Discrepancies between BDI and STAI rescored and scanned items revealed transcription errors. Transcription errors were rectified and the forms rescanned. Note that the MMPI scanned data set did not require rescanning because transcription was not necessary. Note also that the comparison of MMPI scanned data to the twice-rekeyed raw entries served as an additional check of optical scanning accuracy and double keypunching accuracy, both forms of full CTA.

All computer programs developed for rescoring tests reflect item construction, scale composition, norm groups, and scoring algorithms based on standards published in the literature or in test publishers' specifications. All participating settings used, scored, and interpreted the Beck Depression Inventory according to the most recent Beck Depression Inventory Manual (Beck & Steer, 1987). All settings used either the X or Y versions of the STAI_S and STAI_T reflecting item construction and scale composition as published in the Spielberger STAI Manual, Test, and Scoring Key (Spielberger, 1983). None of the sites specified interpretation protocols or norm groups. Instead, settings provided only unstandardized raw score totals. (As such, STAI SPC rankings in this study only included steps to score raw score totals.) MMPI scoring protocols at the participating settings were not fully specified, either. Although all settings endorsed using "recent" MMPI scoring programs, such programs were not identical across sites. Two settings identified NCS as the program manufacturer, but could not identify the software version. Another setting could not readily identify the software manufacturer. As such, scale composition and scoring algorithms could not be explicitly verified. All settings did, however, use adult male and female norms based on the K-corrected original Minnesota adult sample and item construction congruent with the NCS MMPI Manual for Administration and Scoring [NCS MMPI manual] (University of Minnesota, 1983). Rescoring programs developed for the current research used K-corrected original Minnesota norms, as well. Rescoring programs used item construction, scale composition, and scoring procedures in accordance with the accepted standard, An MMPI Handbook, Volume I (Dahlstrom, Welsh, & Dahlstrom, 1972). The NCS MMPI manual reflects the 1972 Dahlstrom et al. handbook, but corrects for round-off errors published in the Dahlstrom et al. T-score lookup values for K, Pd, Pa, Ma, and Si scales (an inadvertent discovery in the present study) (cf. University of Minnesota, 1983, pp. 19-20; Dahlstrom et al., 1972, pp. 380-383).

In classifying SPC levels, I presumed that scorers used procedures that were standardized according to test publishers' recommendations. Scorers may have used non-standardized scoring procedures, which can introduce inadvertent scoring complexities. Other assumptions regarding SPC require some level of subjective judgment about the demands placed on cognitive, motor, or even emotional facets of performance. In the present context, it was assumed that the act of counting was a "less complex" task than addition and subtraction or referring to the proper row and column of a T-score table.

Assessing CTA can be problematic. At settings where full CTA is not in force, it is not likely that scorers inadvertently employ full CTA procedures. Scorers are not likely to perform the extra effort required for double checking. To verify this assumption, I asked scorer supervisors to outline requirements, training programs, and incentives or policies for ensuring hand-scoring accuracy. No supervisors at any of the settings reported any procedures, policies, or behaviors that indicated scoring was performed with full CTA.

Some settings have designated test scorer positions; as few as three people may have scored tests from any particular setting. Thus, generalizing from any one setting may, in reality, only reflect the peculiarities of particular scorers. Because CTA and SPC are considered more meaningful predictors of error than demographic variables, such demographic variables were not considered for systematic study.

The clinician survey was performed as follows. A pilot study was conducted on a random selection of 25 APA Division 12 Fellows. Fellows received a survey questionnaire concerning aspects of objective personality test usage: MMPI scoring practices, experience in detecting MMPI scoring errors, questions about error rates that threaten clinical validity, and computer scoring program usage and associated errors. Appendix 1 contains a copy of the survey. The survey format was almost entirely objective, and it provided up to seven responses reflecting error rate range. Participants who did not respond within 60 days were sent remails. The survey was to be conducted in two stages, a pilot survey and a final survey with the purpose of attaining 50 responses. The pilot survey included 25 participants to approximate a return rate. The total number of final surveys to be sent was to be projected based on the pilot survey return rate.

V. Results

V.I Aggregated Error Rates on Sampled Test Data

Of the 600 tests sampled from all settings, 128 (21.33%) had scoring errors. All settings used less-than-full CTA scoring procedures. Of the 200 tests in the full CTA sample, two (1.00%) had scoring errors. SPC was assessed at two levels, low and high. In the less-than-full CTA sample, low SPC tests (the BDI's and MMPI's) evidenced about half of the proportion of errors found in the high complexity tests (the STAI_S's and STAI_T's). In the full CTA sample, both of the errors occurred with tests of low SPC (MMPI's). Table 3 shows the frequency of tests found with errors as a function of CTA by SPC. Table 4 shows the frequency of errors found on each test type as a function of CTA and

Insert Tables 3 and 4 about here

SPC. Analyses of the distribution of erroneous tests demonstrate strong support for the CTA and SPC factors. Tests scored with less-than-full CTA were erroneous in about a fifth of the sampled cases, whereas very few errors were discovered in the full CTA sample. As predicted, the effect of SPC on scoring error was dramatically different in the full CTA versus less-than-full CTA samples. The frequency of erroneous tests increased notably with increasing SPC in the less-than-full CTA sample, whereas tests scored with full CTA procedures virtually did not manifest errors at either SPC level. As predicted, full-CTA drastically reduces the occurrence of error.

Analyses of the frequency of scoring errors committed within a given test were only possible with the MMPI. Although most tests had no errors, six tests had at least five or more incorrectly keyed items, with 20 being the highest error count on a given test. Overall, 78 mis-keyed items were discovered in the 150 MMPI's sampled. In tests found with errors, 10 tests had one error, six tests had two to five errors, and six tests had six to 20 errors.

V.II. Error Rates for Each Test Type Disaggregated by Setting

The primary purpose of disaggregating results by setting was to examine whether error patterns were idiosyncratic to any one setting. Total error rates for each test type at each setting are shown in Table 5. The scoring error rate at the VA outpatient clinic was, notably, about six times greater than that found at

Insert Table 5 about here

the private inpatient hospital. Despite these differences in error frequencies among settings, each of the three settings produced higher frequencies of erroneous tests than the full CTA sample. Moreover, as shown in Table 6, error frequencies for each setting were concordant with SPC in all three cases; i.e., errors among high SPC tests occurred about twice as often as low SPC tests at each of the three settings. For each setting, as Table 5 demonstrates, the STAI_S and STAI_T error frequencies were in almost all cases notably higher than corresponding BDI or MMPI error frequencies.

Insert Table 6 about here

Figures 1 through 3, divided by setting, plot BDI total score discrepancies found in this study. BDI score discrepancies appear to manifest two patterns, one in which small numbers of items were mistallied in deriving total scores (e.g., a correct BDI score of 24 misscored as a 26), and another in which scores were off by about 21 points (e.g., a correct BDI score of 17 misscored as a 38). Note that solid black dots represent verified scores, whereas hollow black dots represent original hand-scores discrepant with verified scores.

Figures 4 through 9, divided by setting, display total score discrepancies for the STAI_S and STAI_T, respectively. For the private inpatient and VA inpatient hospitals, total raw score discrepancies ranged from 1 to 9 points when compared to correct scores. In the VA outpatient clinic, however, an additional pattern appeared in which total score discrepancies ranged from 20 to 30 points.

Insert Figures 1 through 9 about here

For the MMPI, hand-scorer errors were constrained to those created by mis-keying items. Despite 566 opportunities per questionnaire, tests with keypunching errors were 4% and 6% for VA and private inpatient settings. The errors found at these settings involved small numbers of mis-keyed items. For the third setting, however, 34% of tests showed keypunching errors. Six of these tests revealed 5 to 20 mis-keyed items each.

V.III. Errors in Test Interpretation for the BDI, STAI-S, and STAI-T

The frequency of erroneous tests in the aggregate sample was higher than anticipated. Data analyses revealed ample alterations in all three test types. Because the findings yielded so many errors altering clinical interpretations, describing the alterations was warranted.

To describe alterations requires reference to clinical interpretation standards. Because there appear to be few, if any, references regarding most popular or respected interpretation standards, such standards were chosen from those either frequently cited in the literature or in test publisher specifications. As noted earlier, I used interpretation standards based either on test publisher specifications or on frequently cited literature. To date there appear to be no formal surveys reporting the most popular interpretation standards, hence the qualification for the assumption. The BDI reference subdivides total scores into four ranges that signify minimal, mild, moderate, and severe depression in a clinically depressed outpatient population (Beck & Steer, 1987). Minimally depressed total scores range from 0 to 9; mildly depressed from 10 to 16; moderately depressed from 17 to 29; and severely depressed from 30 to 63. According to these standards, two cases from the VA outpatient clinic produced errors that altered classification. In two cases, clients who should have been classified as moderately depressed (scores of 19 and 26) received hand-scored totals indicating severe depression (scores of 40 and 47, respectively). In another case, a score indicating mild depression (14) was miscored as severe depression (35).

STAI data collected from the settings only included raw total scores. Understanding the implications of scoring errors on interpretation requires the assignment of norm groups and respective standardized scores. The STAI manual (Spielberger, 1983) provides norm groups and T-score conversions for inferring alterations in interpretation. In using an inpatient psychiatric reference group (Spielberger, 1983, pp. 25-26), several erroneous hand-scores resulted in 10 to 20 point T-score discrepancies. Due to the large frequency of such discrepancies, only a few will be highlighted. Table 7 shows a sample of STAI_S and STAI_T raw scores where scoring errors misrepresent high situational anxiety scores as normal situational anxiety scores. Note that most dramatic STAI score discrepancies were found in the VA outpatient clinic data.

Insert Table 7 about here

V.IV. MMPI Interpretation Errors and Unanticipated Sources of Error

In the case of the more popular MMPI, demonstrating definitive links between scoring errors and alterations in interpretation proved challenging for the data sampled in this study. Measuring alteration in clinical interpretation requires reference to an interpretation standard. Since commercial MMPI computer programs embed scoring algorithms and interpretation protocols, discrepancies can result from either or both sources. None of the contacts at the settings could readily identify scale compositions, scoring algorithms, or interpretation standards used in their MMPI scoring programs and computerized interpretations. First pass analyses revealed discrepancies between keypunched and verified raw scale scores and T-scores, even in instances where no keypunching errors appeared. Understanding the source of error in T-score profile comparisons thus required further analysis. Results of these analyses showed that T-score profile discrepancies stemmed from three phenomena: 1) some subscales comprised items that differed from those assumed in the NCS MMPI standard (used in this study); 2) roundoff errors were discovered in the T-score lookup tables published in the Dahlstrom manual (Dahlstrom et al., 1972); and, 3) T-score ceiling values were found to differ from the NCS MMPI standard used in this study. If the effects of these three sources of error found in the private inpatient hospital scoring program were included, 48 of the 50 MMPI's sampled at that setting would have evidenced profile alterations, even though only two of those 50 tests had keypunching errors.

A major thrust of this study was to understand the significance of interpretation errors that result from clerical scoring errors, rather than those resulting from computer scoring program errors or vague interpretive standards. To adhere to this objective, I decided to impose a scoring and interpretation standard on the MMPI's. This required rescoring the originally keypunched data using some reference as an interpretation standard. I chose the 1983 NCS MMPI scoring manual (as noted in the previous section) because of its widespread use and because two of the three settings sampled use MMPI scoring programs purportedly published by NCS. This rescoring process employed the same computer scoring program used in the rekeying effort. This rescored version was then compared to the double-rekeyed effort from the rescorers. To summarize, the same interpretation standard was used to compare scores generated from the patient's raw item entries to those generated from original keypunched item entries as reflected in the patient's original computer printout. By using this approach, discrepancies that appeared between the two resultant profile interpretations could be attributed specifically to keypunching errors.

Although 22 of the 150 MMPI's (14.7%) sampled in this study had keypunch errors, 12 (8.0%) were found with keypunching errors that produced discrepant profiles across the 10 clinical scales and 3 validity scales (see Figures 10 through 21). Perhaps the most common practice for interpreting

Insert Figures 10 through 21 about here

profiles uses the two highest subscale T-scores that exceed a T-score of 70 (or 65 on the MMPI-2) (cf. Dahlstrom et al., 1972; Greene, 1991), although others exist. For instance, a protocol with the two highest scores on the 2 and 4 subscale would be labeled a 2-4 codetype. Most interpretation manuals present a set of descriptors or associated features for common codetypes. This study revealed an instance in which a 2-7 was altered to a 2-4 codetype (see Figure12). If Dahlstrom's work is used major shifts in the interpretive test result. Excerpts from the codetype descriptors are included to demonstrate this difference.

2-4 codetype:

In psychiatric populations this pattern is likely to be found in a psychopathic person who is in trouble and appears at a medical center. Alcoholism, addiction, and legal difficulties are frequent in the patterns of these cases. Although the distress of these persons seems genuine it does not reflect internal conflicts that they may be suffering so much as situational pressures from legal confinement, psychiatric commitment, or close supervision and scrutiny. While the insight these persons show at this time may be good and their verbal protestations of resolve to do better may seem genuine, long-range prognosis is poor. Recurrences of acting out and subsequent exaggerated guilt are common (Dahlstrom et al., 1972, pp. 259-260).

2-7 codetype:

The prominent feature of this group in presenting complaints...is depression, with tenseness and nervousness as frequent accompaniments. Many of these patients also suffer from anxiety, insomnia, and undue sensitiveness. For both sexes, these authors reported a modal diagnosis of reactive depression, with obsessive-compulsive neurosis a close second, but mixed psychoneuroses and conversion reactions are unlikely (Dahlstrom et al., 1972, pp. 260-262).

The protocol associated with the above example had 20 keypunching errors, the most found within the sample. However, two-point codetype alterations can occur with just one keypunching error. For instance, if one patient had originally responded to one more item on the Scale 9, the sole keypunching error found on that test would have shifted a 2-9 codetype to a 2-4 codetype (see Figure 13).

V.V. Clinical Community Scoring Error Survey

Results of a pilot survey submitted to the APA Division 12 Fellows indicated a response bias. Eighteen of the 25 sampled (72%) returned the survey, but 12 (48%) indicated that they were not qualified to answer the scoring error survey. Of the six who responded to the survey, only three (12% of sample) provided complete responses. However, by this time, initial analysis of scoring error had negated the original rationale for obtaining clinicians' opinions about scoring error.

The survey was terminated after the pilot study. Much of the reason behind soliciting the Fellows' opinions was to characterize their views about the relevance of errors on interpreting tests, especially if such errors were perceived to be relatively infrequent. Considering the frequency and magnitude of errors found in the present study and the implications of related alterations in clinical interpretation, the need for soliciting opinion on this matter appears moot.

VI. Discussion

The primary objective of the present study was to address whether scoring errors on objective personality tests should concern the clinical community. The answer to this question depends upon the clinical significance and frequency of such errors and the extent of test usage. Tests chosen for this study are administered frequently nationwide—tens or even hundreds of thousands of times annually. Thus, even infrequent errors can have implications for many individuals. The results of this study provide strong evidence that the three commonly used tests, the MMPI, BDI, and STAI are vulnerable to scoring error. The frequency of erroneous tests ranged from 2% to 56% across all samples at each setting; thus, all settings produced errors in each test sample. Of course, these results are not necessarily representative of the population of scoring errors and may be, for example, over- or under-estimates of the frequency of error in other settings. It is, however, unlikely that scoring errors of the frequency and magnitude discovered in this study are exclusive to the tests sampled from the settings that elected to participate in the current study. Thus, it quite likely that other tests in other settings are also scored erroneously and that some rates in some settings are also "alarming." The results of the study also clearly revealed that scoring errors can change the interpretation of test results, which, for example, could alter whether a patient is prescribed needed medications. Although clinicians may argue that clinical decisions are not made on the basis of an isolated test score, salient information, even that stemming from a single variable (e.g. a test score), can predominate judgments (Faust, 1984). Just how often, and to what extent, judgment and treatment decisions are altered by scoring error is a question beyond the scope of this study.

A second objective was to examine possible factors that prior research (e.g., Allard et al., 1995) has suggested are related to scoring error—CTA and SPC. Results indeed suggest that the factors analyzed, SPC and CTA, are associated with the occurrence of scoring errors. The research design did pose methodological limits on measuring the SPC and CTA factors as fully intended, however. Additionally, peculiarities in error patterns at each setting also revealed limits to the CTA and SPC constructs as defined in this study. Typically, peculiarities threaten the applicability or implications of a study's findings. Interestingly, the peculiarities found in this study do not undermine support for the factors related to scoring error; moreover, the peculiarities in some ways increase concern for addressing quality assurance problems in scoring popular objective personality tests. A discussion of limits and peculiarities as well as their implications follows.

Analyses of both aggregated and disaggregated data show that SPC rank, when conceptualized as the number of distinctly different operations needed to produce summary scores, shows a strong positive relation to error rate. Quite simply, the STAI_S and STAI_T tests, which required two steps to derive summary scores, evidenced discernibly higher error rates than either the BDI or the keypunched MMPI, which required only one step to derive final scores. It is important to note that SPC was only measured at two ranked levels, although the research design originally called for three (recall that the MMPI's sampled were keypunched at all three settings). If three separate levels had been measured, then the findings might have provided much stronger support for the notion that the SPC ranking scheme used in this study was indeed associated with increasing scoring error. The SPC construct, thus, was not fully tested as intended. Although more work is needed to boost confidence in the SPC ranking scheme as defined in this study, the findings are consistent with previous research on one objective personality test with subscales that contained several different SPC levels (cf. Allard et al., 1995). Still, the findings support the notion that tests requiring more complicated scoring procedures are associated with increased frequency of scoring error. Although this point seems obvious in hindsight, it is of interest that few prior studies have examined complexity as a factor in error rate.

As the results show, CTA proved to be an even stronger factor in predicting the presence of scoring error. The findings possibly call into question the definition of "full" CTA used in this study. The optically scanned full CTA sample did not, after all, yield error free results on the MMPI. The optical scanning process, the NCS OpScan 5 in this study, was not flawless in distinguishing between erased items and marked ones, as two tests had one item miscored each. Even so, aggregated error frequencies representing each of the less-than-full CTA settings substantially exceeded aggregated error frequencies in the full CTA sample. Although full CTA practices virtually eliminated scoring errors in comparison to tests scored with less-than-full CTA, full CTA was not measured to the extent specified above. The full CTA sample consisted exclusively of optically scanned computer scoring. Recall that the full CTA construct was defined as optically scanned computer scoring or fully double-checked scoring. Obtaining full CTA samples proved more difficult than originally anticipated; no settings could be readily identified where practices even included double-checking. This research, thus, did not fully test the levels as defined in the CTA construct. If attainable, such research requires the examination of double-checked test data to determine whether such a process out-performs less-than-full CTA procedures. Support for the notion that double-checking increases accuracy was obtained informally during the data entry process by the rescorers in the present study. Fortunately, no errors in double checking were discovered when MMPI patient item responses were compared against respective keypunched item answers. This finding was, however, incidental in that double-checking accuracy was not formally tracked and the double-checking process was completed by the research team.

Despite limits in measuring full CTA, the full CTA procedures still clearly reduced errors in comparison to less-than-full CTA procedures. Barring any changes future research holds for the full CTA distinction, examination of the idiosyncrasies in error patterns among the settings also suggest that less-than-full CTA might be better represented by two subcategories, partial CTA and low CTA. Clarifying this intended recategorization requires the explication of "large" versus "small" errors. Here, "small" scoring errors mean either small magnitude errors or small total number of errors per test. Small errors were evident in all test types for all settings. Small errors, as seen in the exemplars, can be misleading and hazardous. The conventions used in interpreting the MMPI and BDI place alteration of interpretation at risk when small errors are committed. Recall the case (Figure 13) where one item error out of 566 questions almost changed a 2-9 profile to a 2-4 profile. This mistake might result in denying the potentially manic patient a lithium prescription. The BDI interpretation could change from minimal depression to mild depression with just one counting error.

Stark differences in error frequencies among the three participating settings emerged when scoring error analyses were performed on disaggregated data. The VA outpatient clinic exhibited gross error frequencies (between 17% and 56%, i.e., "large" scoring errors) on the MMPI and STAI compared to the VA and private inpatient settings. Both VA settings exhibited high BDI error rates compared to the private inpatient setting. The less-than-full CTA category reflected notable heterogeneity in that error rates for specific test types differed dramatically among settings. These gross error rates suggested the presence of a lower CTA level than the less-than-full CTA category and warrant further consideration within this study.

One explanation accounting for such divergent error rates among less-than-full CTA levels concerns the use of non-standard test forms. Upon scrutinizing the BDI's taken from the VA settings, two test forms were discovered. Note that the BDI is a 21-item test. On one of the forms, the item responses ranged from 0 to 3, in accordance with the test publisher's specifications. The second form had item responses ranging from 1 to 4. Upon scrutinizing individual tests with 21-point discrepancies, the source of the problem indeed appeared to be in item format of the test. According to the BDI manual (Beck & Steer, 1987), item responses should be numbered 0 to 3 such that the scorer can simply add all response values to derive the final score. On tests numbered 1 to 4, adding item values on these tests yields a total score that is 21 points higher than the intended score. For these tests, final scores required subtraction by 21, which apparently did not happen in some cases, thereby creating the 21-point discrepancies. The existence of both test forms at a setting may thus create confusion. Additionally, the existence of these improperly coded test forms may have effectively increased SPC via the addition of the total score adjustment step. Also, it is conceivable that some scorers use an alternate and, unfortunately, more complex adjustment method by subtracting 1 from each item as the items are tallied. These 21-point errors account for the large discrepancies shown in Figures 1 and 2. Such large errors are not at all unlikely to have deleterious implications. For instance, one patient was classified as severely depressed when the verified score indicated mild depression. Such errors could result in mis-prescribing anti-depressant medications or rendering unwarranted services, such as suicide prevention, ECT, inpatient hospitalization, or other treatments with serious implications. The converse is possible, as well; patients could be classified as minimally or mildly depressed when actually severely depressed, possibly leading to negligence of treatment. In either case, the morbidity that can flow from gross misscoring of tests, or the failure to do something so basic as coding scores properly on a scale of 0 to 3, could easily lead to lawsuits.

The STAI tests were subject to large errors in the VA outpatient clinic, too. Unlike the BDI, there were no apparent test form item value errors that could account for the magnitude of such discrepancies. Instead, these errors may be the result of neglecting the addition of a single factor, approximately 27 points. An error of this kind can occur if the reverse coded items are tallied in a separate pass and the total adjusted with the addition of a constant. In some cases, scorers may have neglected the addition of the constant. This could account for the large discrepancies reflected in Figures 4 and 7. Such large errors on the STAI tests can shift the category or interpretation of anxiety levels. If pathological Trait anxiety (STAI_T) is misclassified as normal anxiety, then therapy may mistakenly be directed towards other concerns, when ironically, gross elevations in anxiety impede the outcome of all other therapeutic efforts.

The scorers in the VA outpatient clinic made far more errors on the MMPI than those in the private and VA inpatient hospitals. On the MMPI, error pattern analysis revealed that scorers misaligned item-response columns when keypunching. One such MMPI with 20 mis-keyed items changed a 2-7 profile to a 2-4 profile (Figure 12). Because the mistaken profile suggests anti-social personality disorder instead of anxiety or depression, this alteration could affect whether the patient even receives therapy. Informal surveys of the setting test scoring environment revealed that the VA outpatient clinic MMPI data entry terminal was located in the reception area where phone calls and patients perpetually disrupt the scoring process. Other settings appeared to have quiet locations for keypunching the MMPI. Environmental factors could perhaps contribute to the quality of the MMPI scoring process.

Another possible explanation for such noticeable differences among each setting's test type error rates may concern an aspect of CTA not assessed by the behavioral criteria used in this study. Based on informal discussions with scorer supervisors, two of the three settings had hired full-time test scorers. The setting that did not have such professionals relied mainly on temporary workers to score test data; this setting was the VA outpatient clinic. It is possible that full-time workers produce higher quality test scoring than transient or temporary workers. Johnson and Chandler (1985) noted this phenomenon between two samples of cognitive test scorers, one consisting of full-time psychologists and the other consisting of transient psychologists. CTA thus may be variably affected by a variable such as “commitment to job.”

The preceding analysis, although post hoc, suggests the need for exploring further distinctions in CTA levels. Less-than-full CTA certainly comprises unchecked keypunching or less than fully checked hand-scoring. Where some combination of non-standard test forms, non-standard scoring practices, frequent interruptions, or transient scorers exist, large errors are likely to appear. Settings that exhibit these practices warrant the label of "low" CTA. Barring methodological limits, less-than-full CTA should be divided into "low" and "partial" CTA, or some continuous measure. Partial CTA procedures, as demonstrated in this study, still result in potentially deleterious misclassifications. If considered the mode on frequently administered tests, as it quite possibly could be, many individuals can be affected adversely. In settings where low CTA procedures are present, and it would seem almost certain that the two settings we studied in which the label appears justified are not the sole instances in the country or world, deleterious misclassifications are likely ubiquitous. This inference is, however, based on the limited findings within the present study. Given the potential for widespread negligence, further research is needed to establish the generality of low and partial CTA practices and to make improvements as soon as possible.

VII. Recommendations

Recently, Moreland, Eyde, Robertson, Primoff, and Most (1995) published test user qualifications. Among the 12 minimum test user competencies listed, the authors cited first: "Avoiding errors in scoring and recording." The findings of this study suggest a clear need to avoid scoring errors.

Scoring errors appeared on commonly administered objective personality tests at all three settings in this study. Both large and small errors were discovered on all three tests. The data in this study provided rich examples of errors that lead to distortions in interpretation. Such distortions possibly reduce clinical efficacy and hamper the appropriateness of assessment based interventions. At worst, they can lead to serious, if not potentially fatal, errors. For researchers, distorted test findings inflate error terms and likely decrease the likelihood of detecting meaningful differences in the data.

Full CTA procedures appear to virtually eliminate scoring error but unfortunately do not come without barriers, costs, and new pitfalls. Manually rescoring tests requires more than twice the labor or time to deliver highly accurate test scoring, adding to the current burden on dwindling resources. Automated rescoring may entail investment in computer and optical scanner technology, both of which can be expensive. However, return on investment can be realized in labor savings alone in as little as two years.

Moreover, as this study has demonstrated, scanning technology is fallible and computer scoring programs can contain programming errors or use inconsistent test standards. In short, the clinicians cannot or should not be expected to trust the veracity of complex scoring programs if publisher's scoring standards are not clearly communicated.

This study intended to investigate a particular source of scoring error, that generated by the process of manual scoring. Unexpectedly, however, the study showed that scoring errors are not the only source of erroneous test scores. Three additional sources of error were uncovered: errors in computer scoring programs, lack of standard references for scoring tests, and optical scanner errors. Given that error sources resided in what were presumed to be processes ensuring high scoring accuracy, recommendations to merely use full CTA, as defined in this study, do not completely address the problem at hand. In fact, regardless of scoring error rates, scoring program errors, whether it be roundoff, scale composition, or ceiling values errors, can alter profiles dramatically. Two classes of recommendations, if followed, can go a long way towards addressing the problems uncovered in this study: changing scoring practices and changing the way test scores are interpreted.

At a bare minimum, tests that require more than one step to score should be scored with full CTA procedures. To alleviate preventable sources of error, though, not only demands eliminating hand scoring and keypunching without double verification, but must also address other threats to reliability and sound interpretive practices. These include using verified scoring programs and eliminating optical scanner errors. Verifying the accuracy of computer scoring programs imposes the clinician with a frustrating onus. Although tests can be hand scored against computer scored output, the process defeats the purpose of automation. Secondly, verifying hand scores against computer scores does not necessarily expose computer scoring program bugs. Programming bugs can produce obvious or consistent scoring discrepancies or subtle and sporadic ones. The onus belongs on the scoring program manufacturer, who should publish the method by which their scoring program was verified. For instance, scale composition and item membership in the MMPI scoring programs used in this study was verified by using a "jack-knifing" program that reproduced the composition tables (in the NCS MMPI manual) by scoring all combinations of tests with only one item pathologically endorsed in each case. The resulting lookup table was matched against the NCS manual scale composition lookup table (University of Minnesota, 1983, pp. 19-21).

Optical scanning errors may prove difficult to prevent. On occasion, the test taker partially erases or accidentally smudges the items, which produces ambiguity for the scanner. Optical scanners may improve scoring accuracy, as demonstrated in this study, but may not ascertain scoring accuracy. Scorers can check answer sheets to remove smudge marks and improve scanner performance. This practice could greatly alleviate scanning errors, yet some smudges will continue to elude the scanner. Ascertaining scoring accuracy may ultimately require eliminating processes between the test-taker and the scoring program. Administering the test on computer may be the most effective way to ensure such accuracy.

Scanner accuracy notwithstanding, verifying scoring programs necessarily entails a standard for scoring and interpretation. Although this problem was cited almost 30 years ago in the literature (Fowler, 1968), surprisingly, no such standard ostensibly exists for the MMPI, making it difficult to discern among computer program scoring errors and errors in misinterpreting various standards. In the present study, 96% of the private inpatient hospital profiles were discrepant with profiles generated from the scoring protocol published in the NCS MMPI manual. These errors were due to a combination of roundoff, ceiling score discrepancies, and item composition discrepancies.

Another change in the scoring interpretation process could drastically reduce interpretation discrepancies by mitigating the effects of small errors. Tests, such as the BDI and MMPI, that are interpreted using point scores, thresholds, and cutoffs, are particularly susceptible to the effects of not only large, but also small errors. These scoring structures create vulnerabilities for cases where scale scores "sit near the fence." It is therefore important to be conscious of the interaction between measurement practices and the things being measured. An interpretation strategy that starts with an awareness of the probabilistic versus deterministic nature of test scores would usually neutralize the potential effects of small errors. For example, test scores could be reported with certainty estimates, such as standard error of measurement (SEM). Using SEM's, small scoring errors may slightly alter presumed interpretation profiles, but are much less likely to result in the type of categorical shifts that easily result from point score and threshold structure. The MMPI T-scores could, for instance, be represented by error bands rather than points. With improvements in scoring accuracy, combined with interpretation systems that emphasize the use of SEM's, the problem could be virtually eliminated, most likely sparing hundreds, if not thousands, of individuals' needless suffering.

Table 1: Tests selected for current study, respective SPC rankings, and number of steps required to obtain total scores

Test:	BDI	STAI	MMPI
SPC Ranking:	low	medium	high
Scoring Steps:	1) total item values	1) locate reverse-coded items 2) transform reverse-coded values 3) total all item values	1) locate correct gender template 2) locate correct subscale template 3) total all marked items for each subscale 4) plot totaled scores on T-score lookup table 5) record T-score for each subscale

Table 2: Tests selected for the current study, with revised respective SPC rankings, and revised number of steps required to obtain total scores

Test:	BDI	STAI	MMPI
SPC Ranking:	low	medium	high
Scoring Steps:	1) total item values	1) locate reverse-coded items 2) transform reverse-coded values 3) total all item values	1) keypunch all item values

Table 3: Frequency and percentage of tests found with errors as a function of CTA and SPC

		CTA		
		Full	Less-than-full	SPC Totals
SPC Rank	High	0 0.0% n=100	86 28.7% n=300	86 21.5% n=400
	Low	2 2.0% n=100	42 14.0% n=300	44 11.0% n=400
CTA	Totals	2 1.0% n=200	128 21.3% n=600	130 16.3% n=800

Table 4: Frequency and percentage of errors found on each test type as a function of CTA and SPC

			CTA	
		Test Type	Full	Less-than-full
SPC Rank	High	STAI_S	0 0.0% n=50	44 29.3% n=150
	High	STAI_T	0 0.0% n=50	42 28.0% n=150
	Low	BDI	0 0.0% n=50	20 13.3% n=150
	Low	MMPI	2 4.0% n=50	22 14.7% n=150
		CTA Totals	2 1.0% n=200	128 21.3% n=600

Table 5: Frequency and percentage of tests found with errors as a function of setting and test type

	Setting		
Test Type	VA Outpatient	VA Inpatient	Private Inpatient
STAI_S	23 46.0% n=50	15 30.0% n=50	6 12.0% n=50
STAI_T	28 56.0% n=50	10 20.0% n=50	4 8.0% n=50
BDI	9 18.0% n=50	10 20.0% n=50	1 2.0% n=50
MMPI	17 34.0% n=50	3 6.0% n=50	2 4.0% n=50
Setting Totals	77 38.5% n=200	38 19.0% n=20	13 6.5% n=200

Table 6: Frequency and perceintage of tests found with errors as a function of setting and SPC

			Setting	
		VA Outpatient	VA Inpatient	Private Inpatient
		51	25	10
	High	51.0%	25.0%	10.0%
		n=100	n=100	n=100
SPC		26	13	3
	Low	26.0%	13.0%	3.0%
		n=100	n=100	n=100

Table 7: Sample of STAI cases where discrepancies between hand and verified raw scores would have produced relevant T-score alterations in interpretation

Test Type	Raw Hand Score	Raw Hand Score	T-Score based on Raw Hand Score	T-Score based on Raw Verified Score
STAI_S	50	80	52	72
STAI_S	42	68	46	64
STAI_S	28	45	36	48
STAI_T	40	63	45	62
STAI_T	52	73	55	72
STAI_T	59	80	60	72

Figure 1. Comparisons between 50 hand-scored and verified BDI scores sampled from the VA outpatient clinic.

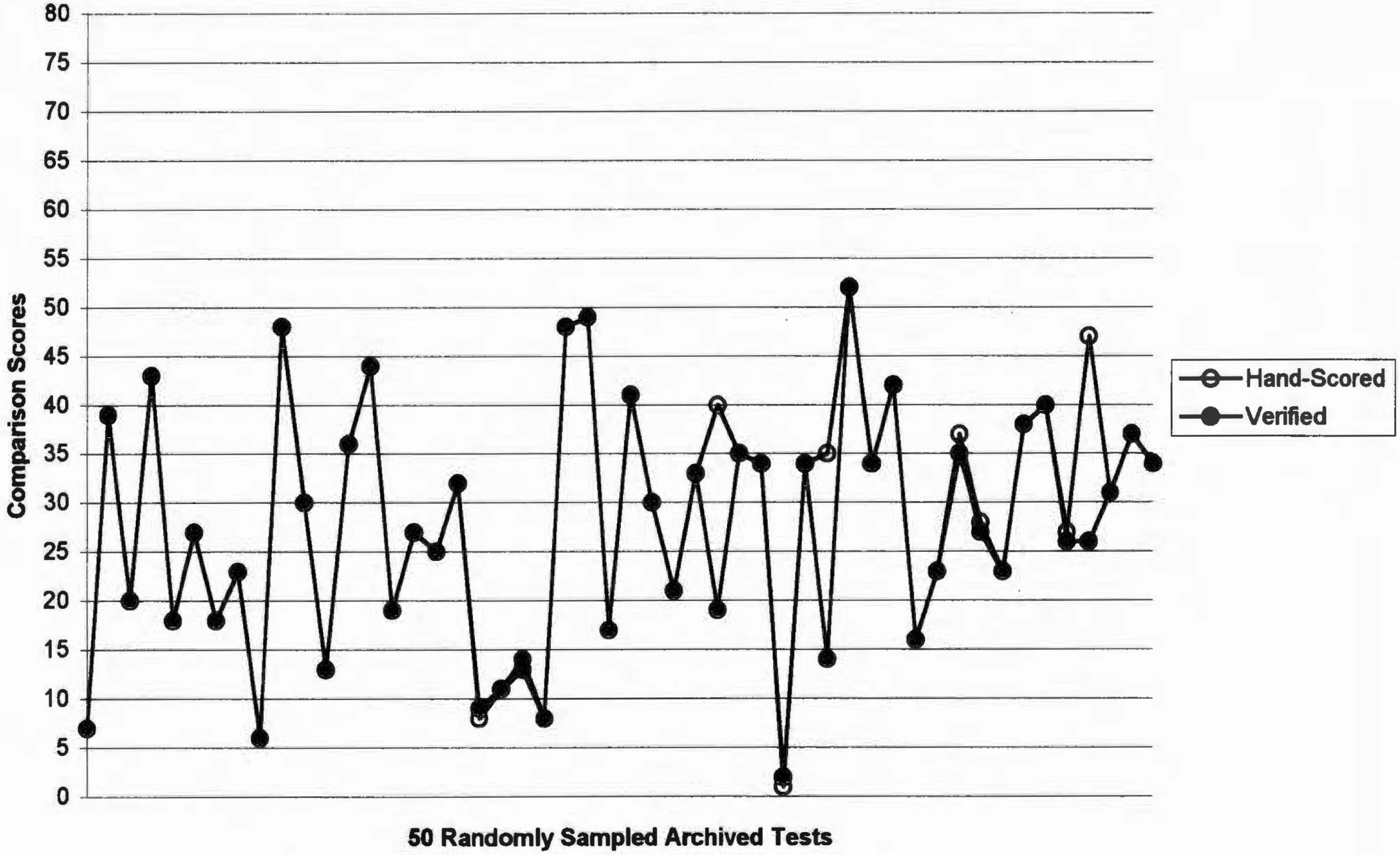

80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Comparison Scores
50 Randomly Sampled Archived Tests
Hand-Scored
Verified

Figure 2. Comparisons between 50 hand-scored and verified BDI scores sampled from the VA inpatient hospital.

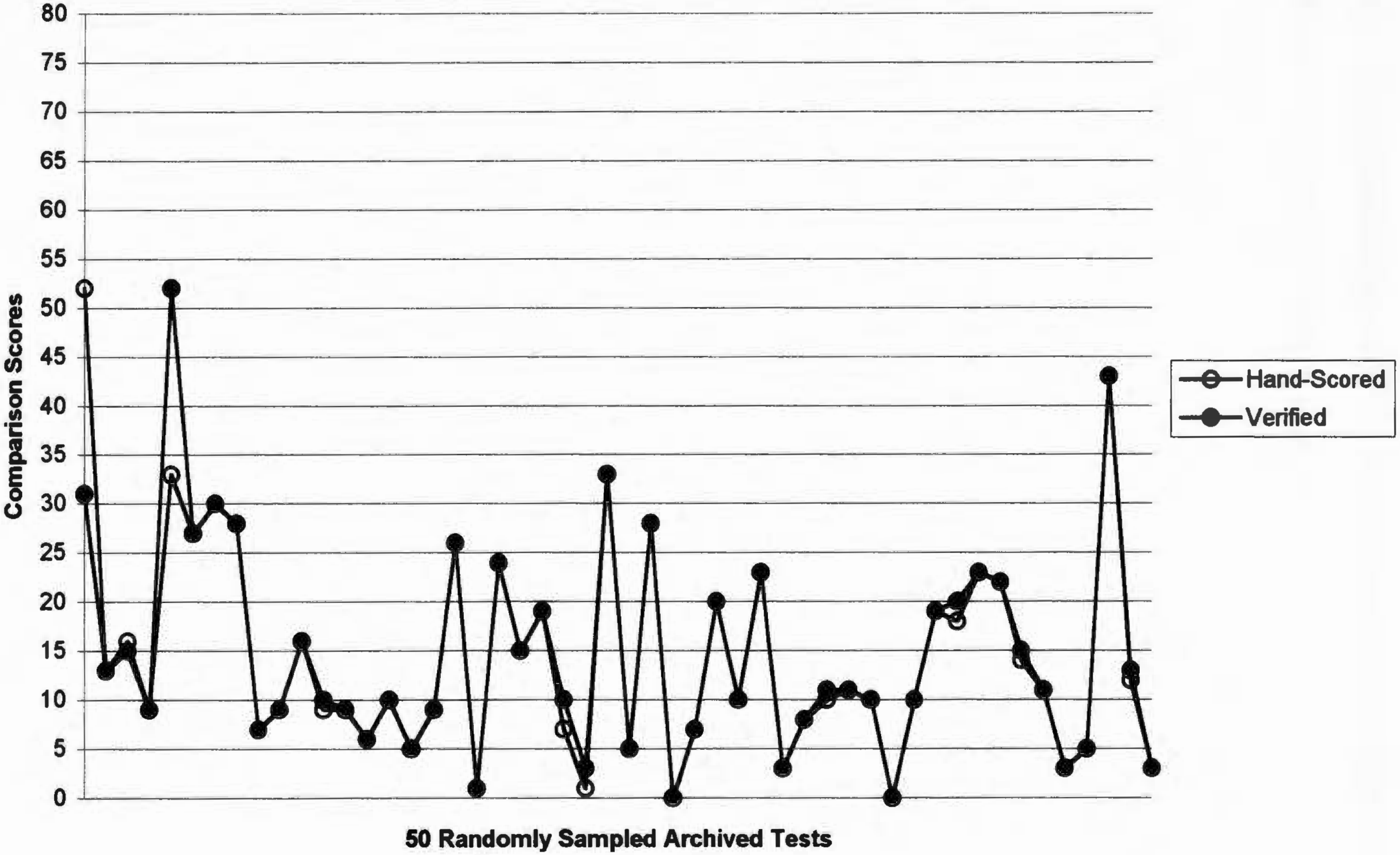

80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Comparison Scores
50 Randomly Sampled Archived Tests
Hand-Scored
Verified

Figure 3. Comparisons between 50 hand-scored and verified BDI scores sampled from the private inpatient hospital.

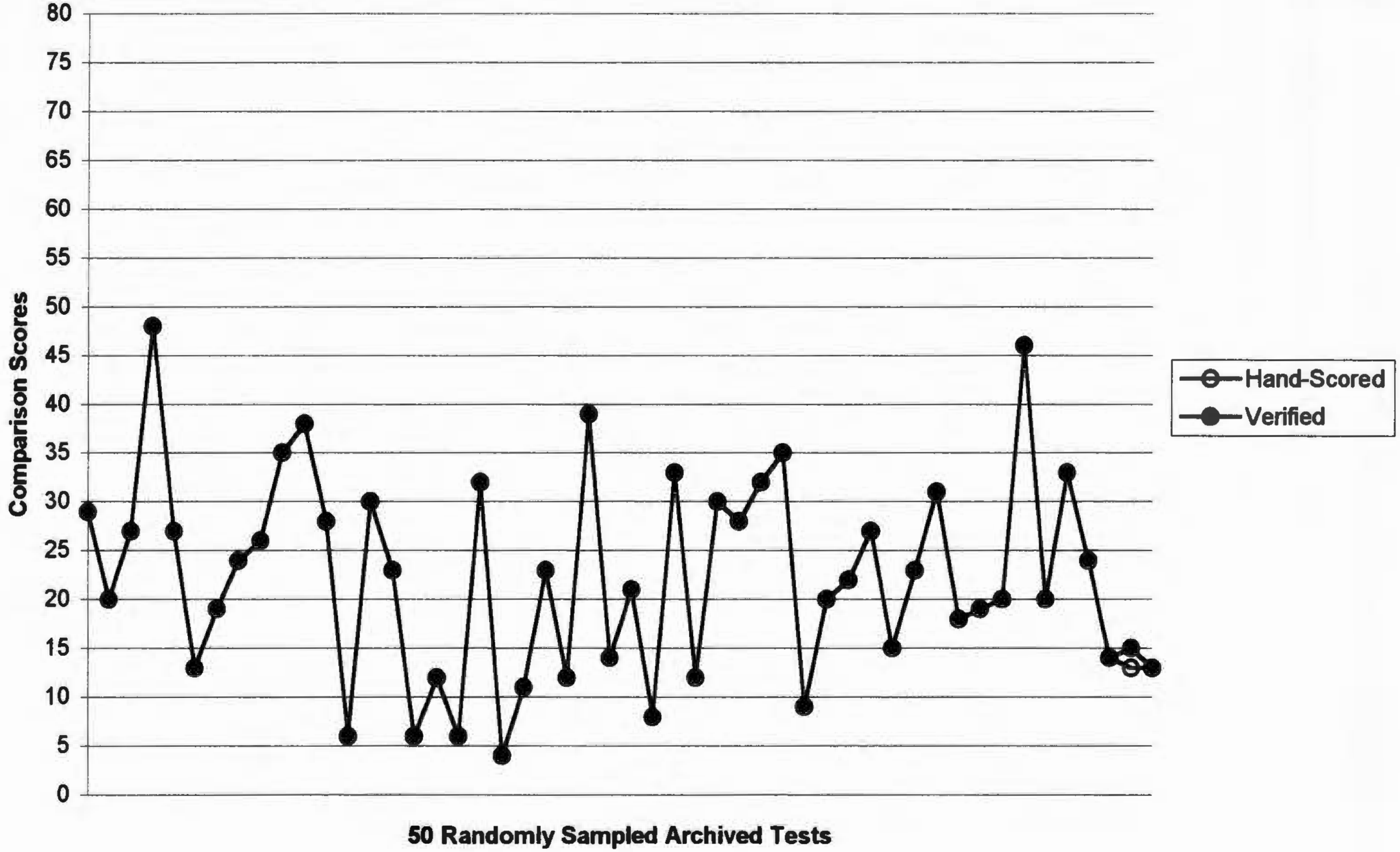

Comparison Scores
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
50 Randomly Sampled Archived Tests
Hand-Scored
Verified

Figure 4. Comparisons between 50 hand-scored and verified STAI_S raw scores sampled from the VA outpatient clinic.

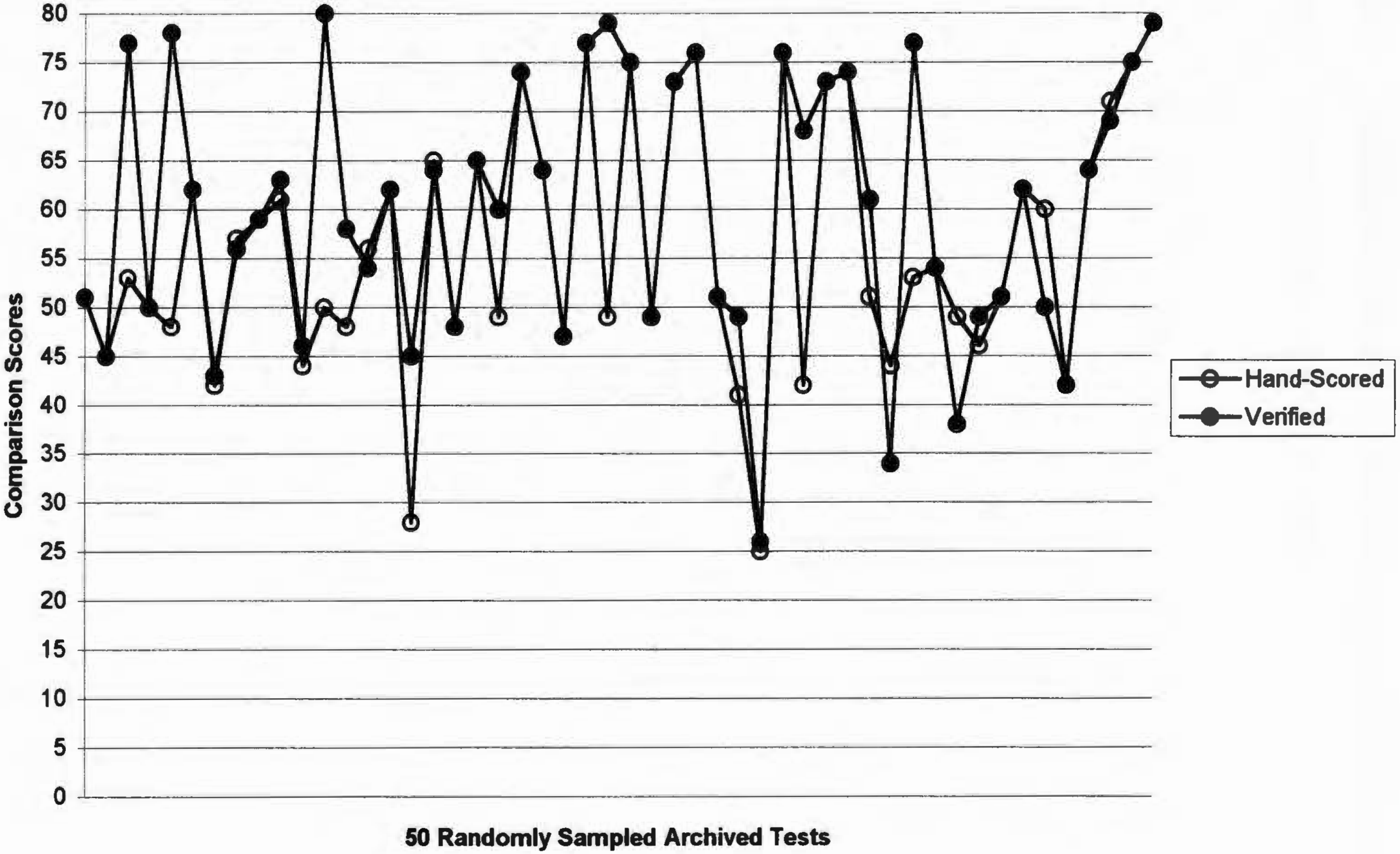

80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Comparison Scores
50 Randomly Sampled Archived Tests
Hand-Scored
Verified

Figure 5. Comparisons between 50 hand-scored and verified STAI_S raw scores sampled from the VA inpatient hospital.

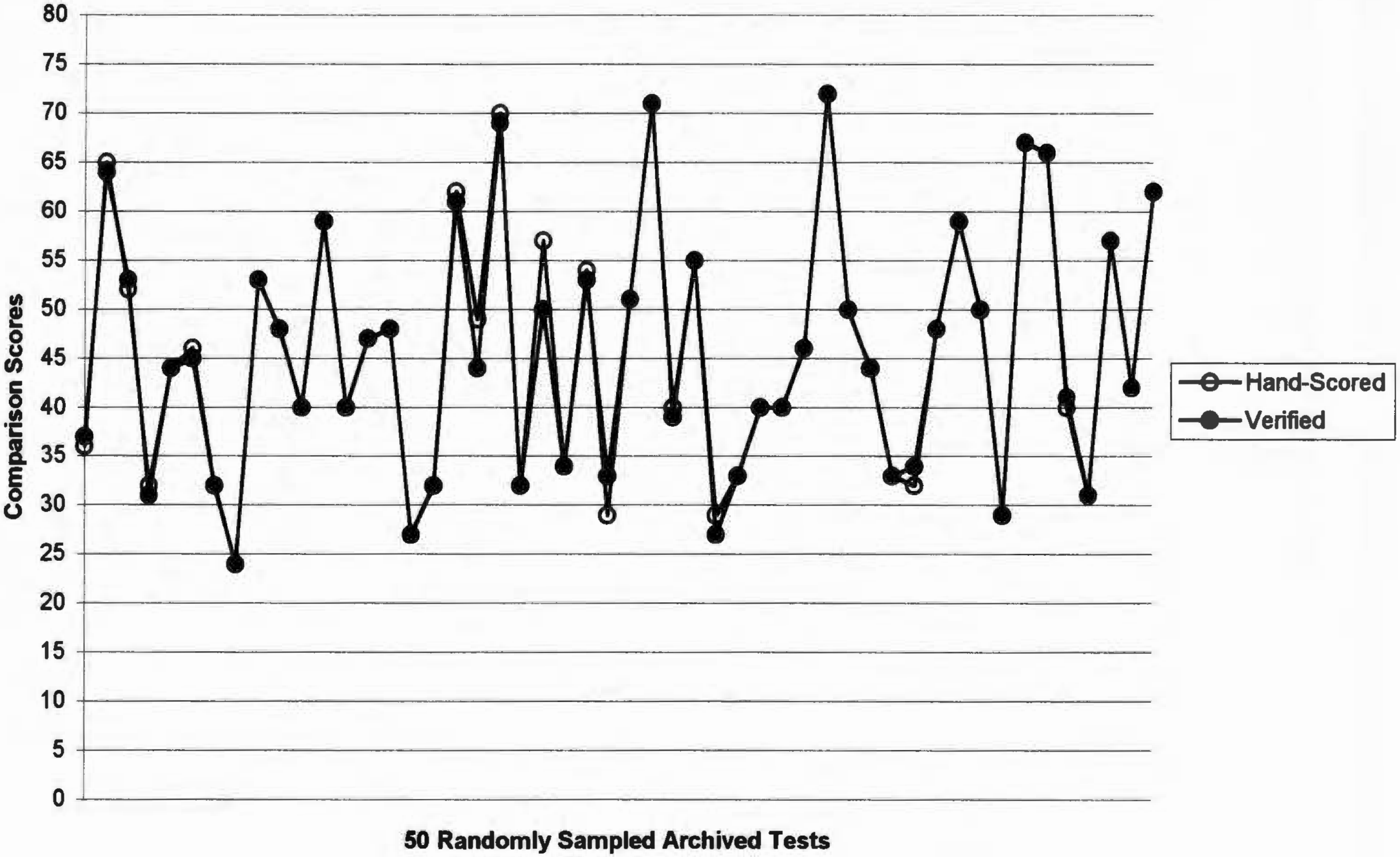

Comparison Scores
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
50 Randomly Sampled Archived Tests
Hand-Scored
Verified

Figure 6. Comparisons between 50 hand-scored and verified STAI_S raw scores sampled from the private inpatient hospital.

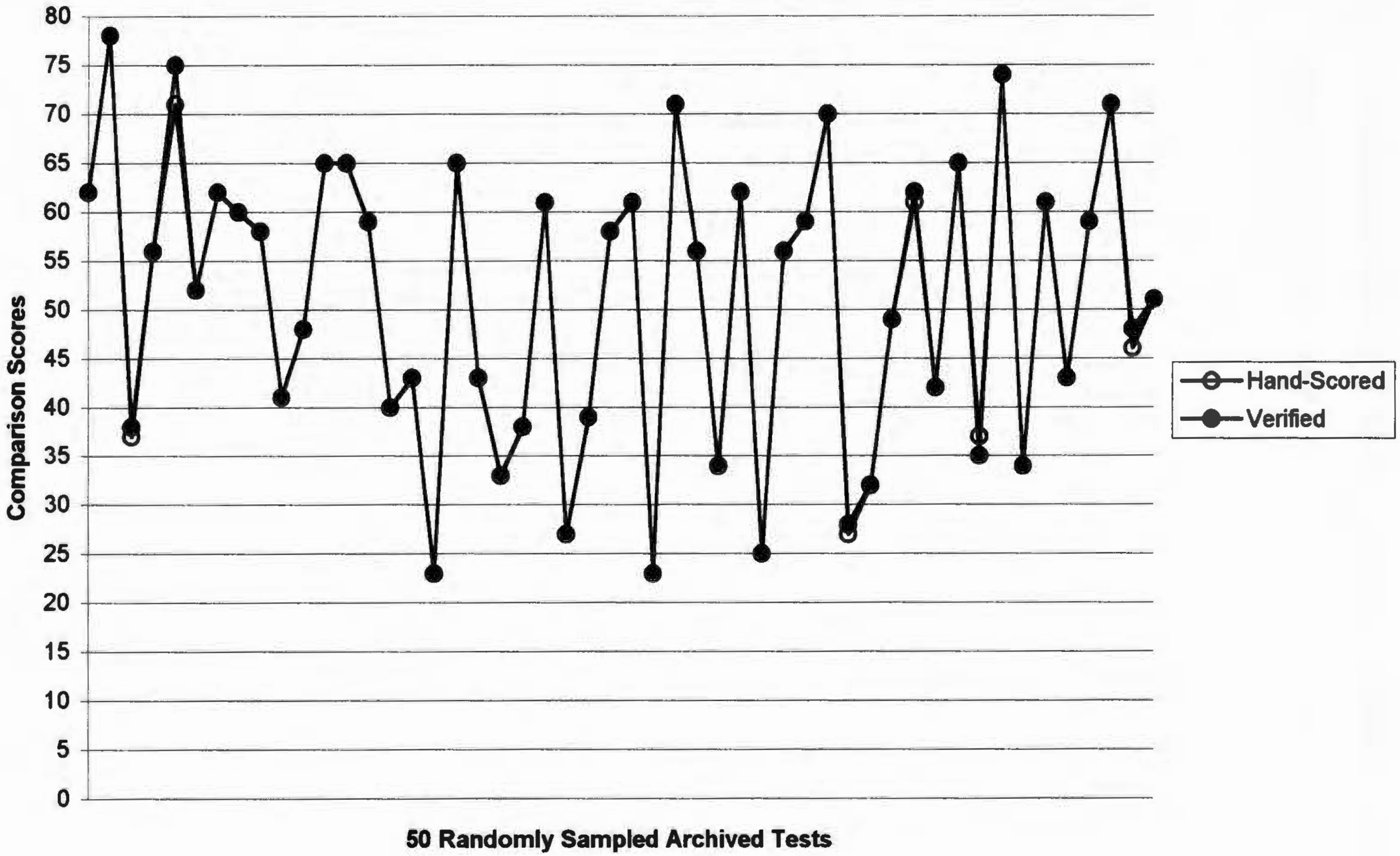

Comparison Scores
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
50 Randomly Sampled Archived Tests
Hand-Scored
Verified

Figure 7. Comparisons between 50 hand-scored and verified STAI_T raw scores sampled from the VA outpatient clinic.

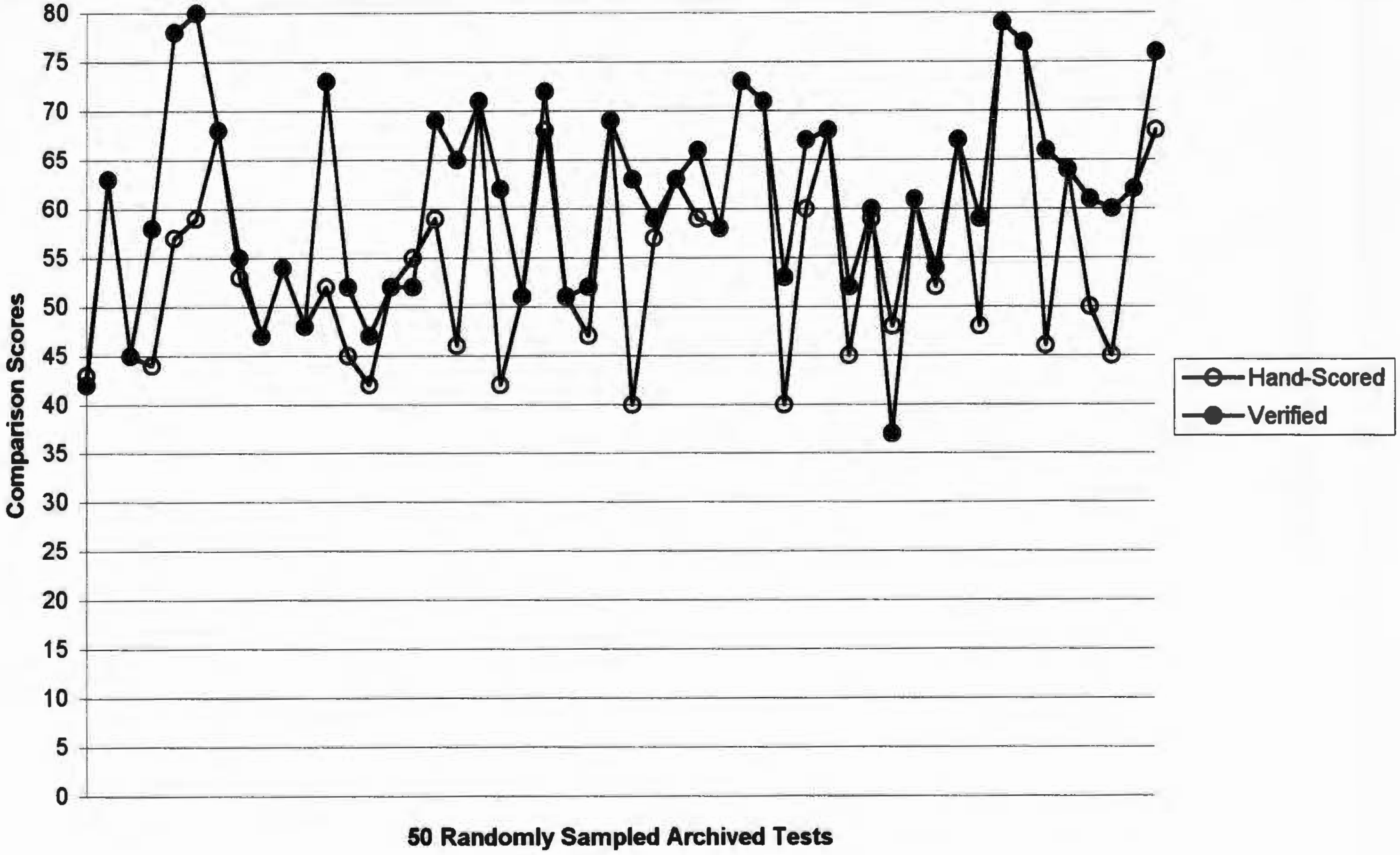

80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Comparison Scores
50 Randomly Sampled Archived Tests
Hand-Scored
Verified

Figure 8. Comparisons between 50 hand-scored and verified STAI_T raw scores sampled from the VA inpatient hospital.

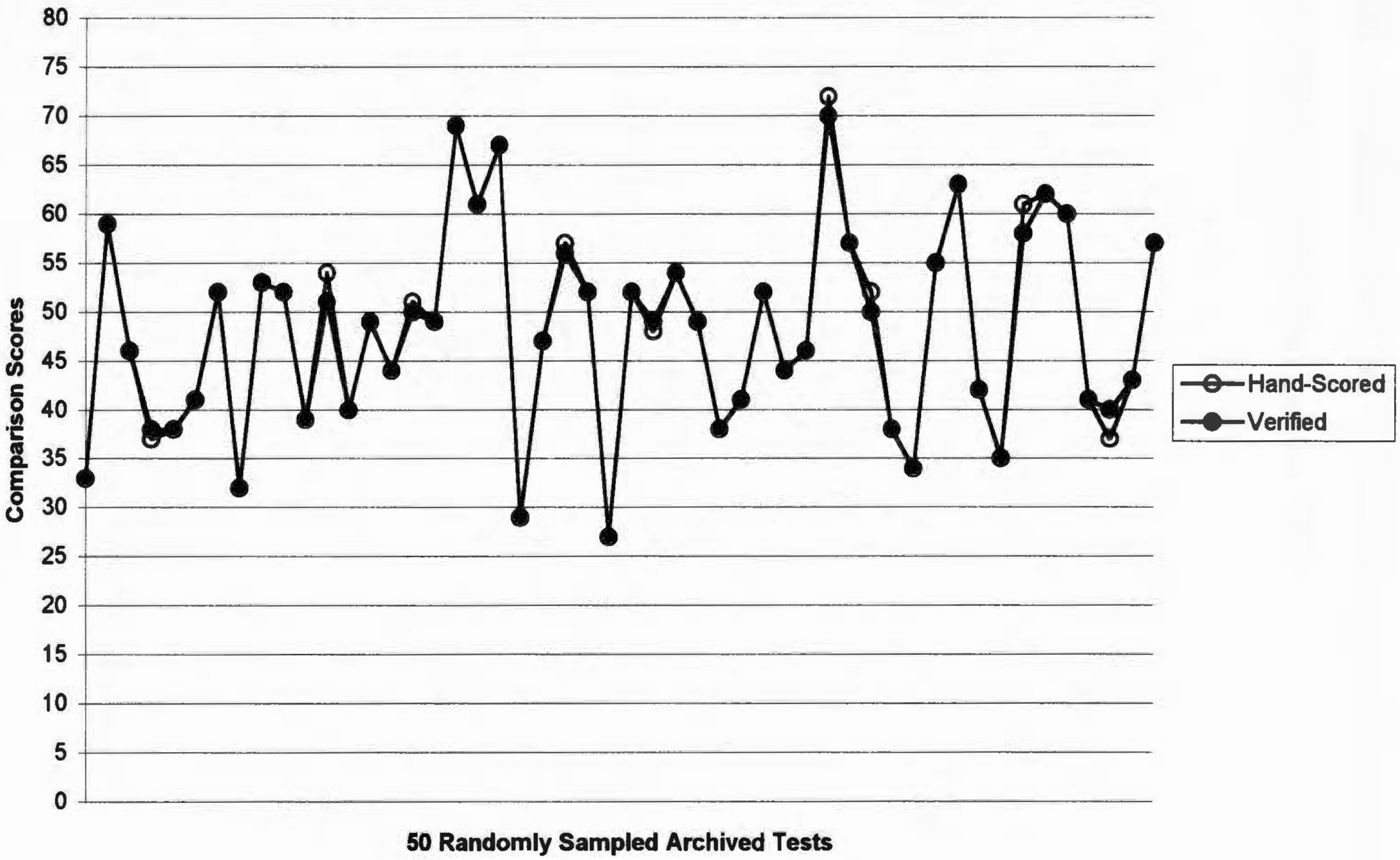
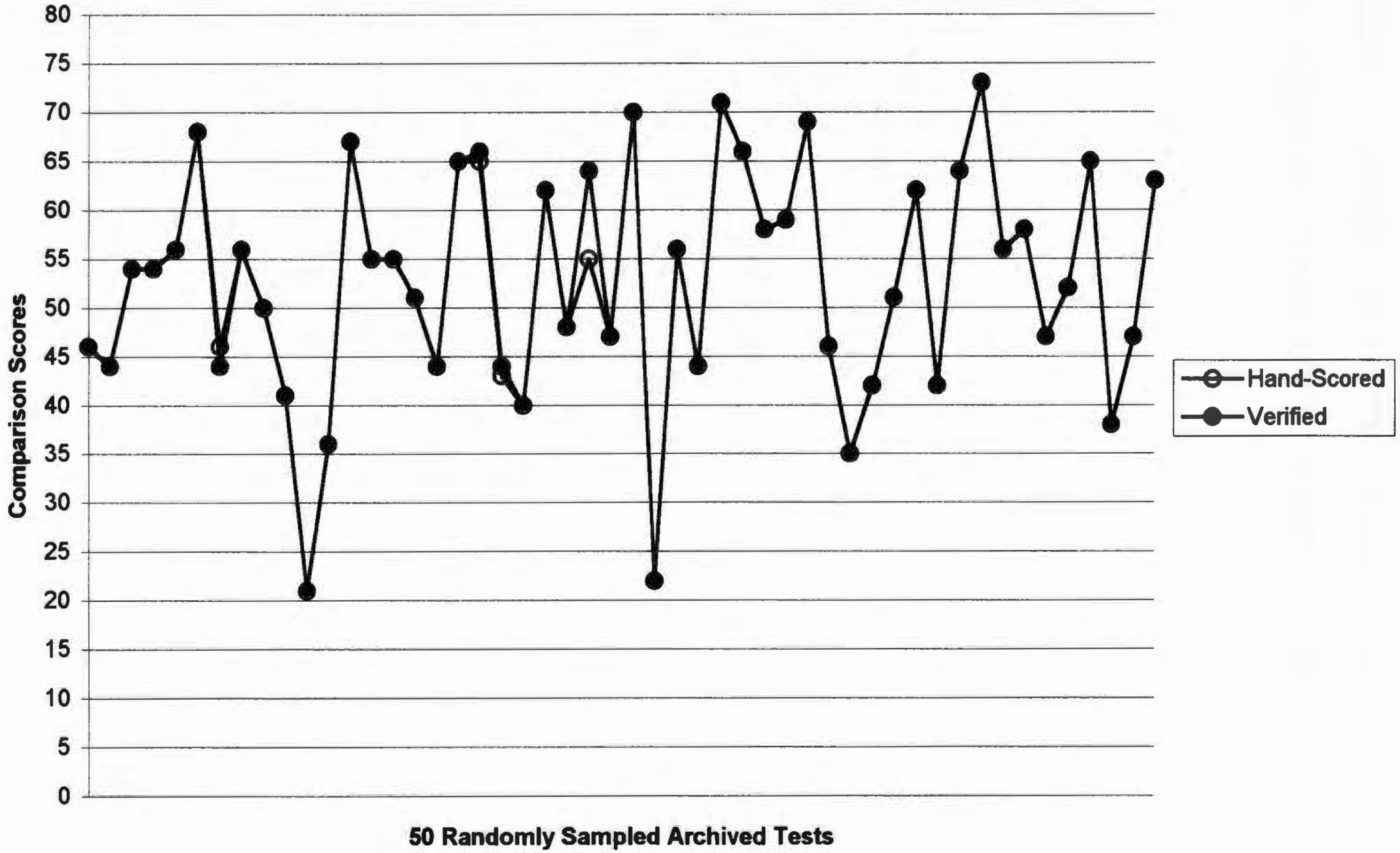
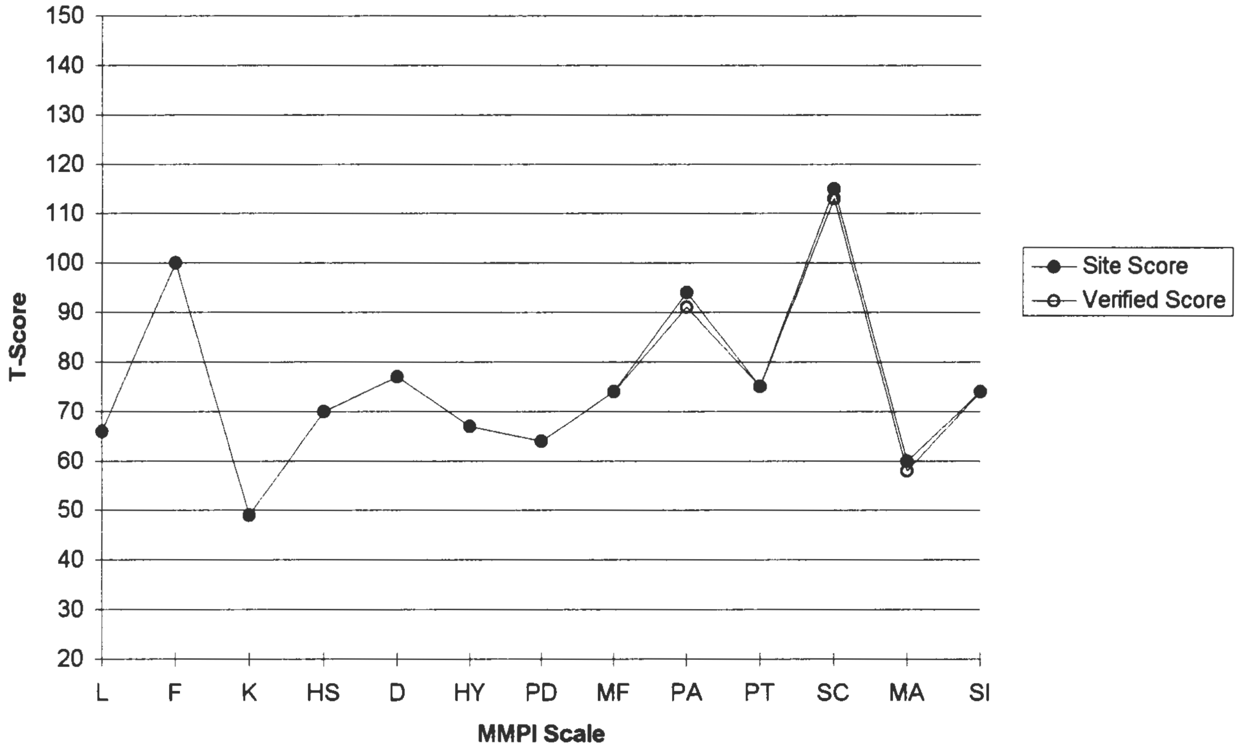

Comparison Scores
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
50 Randomly Sampled Archived Tests
Hand-Scored
Verified

Figure 9. Comparisons between 50 hand-scored and verified STAI_T raw scores sampled from the private inpatient hospital.


80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Comparison Scores
50 Randomly Sampled Archived Tests
Hand-Scored
Verified

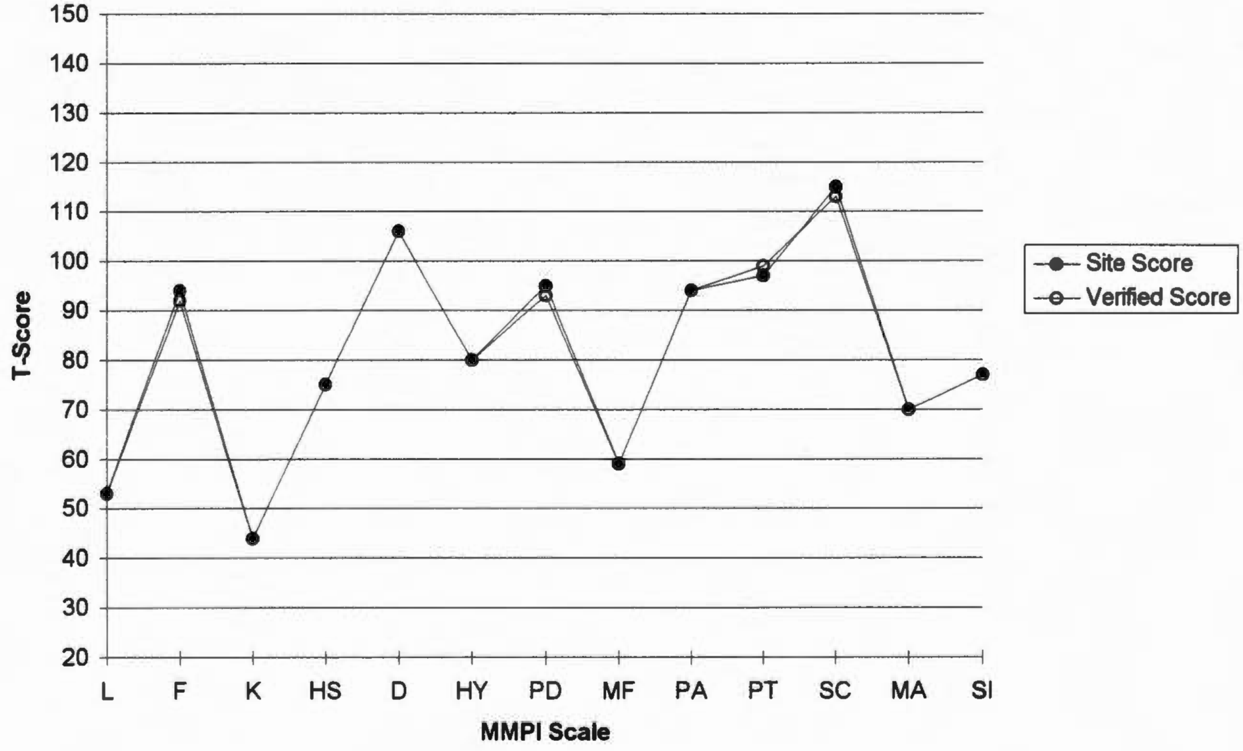
Figures 10. Case 2966. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



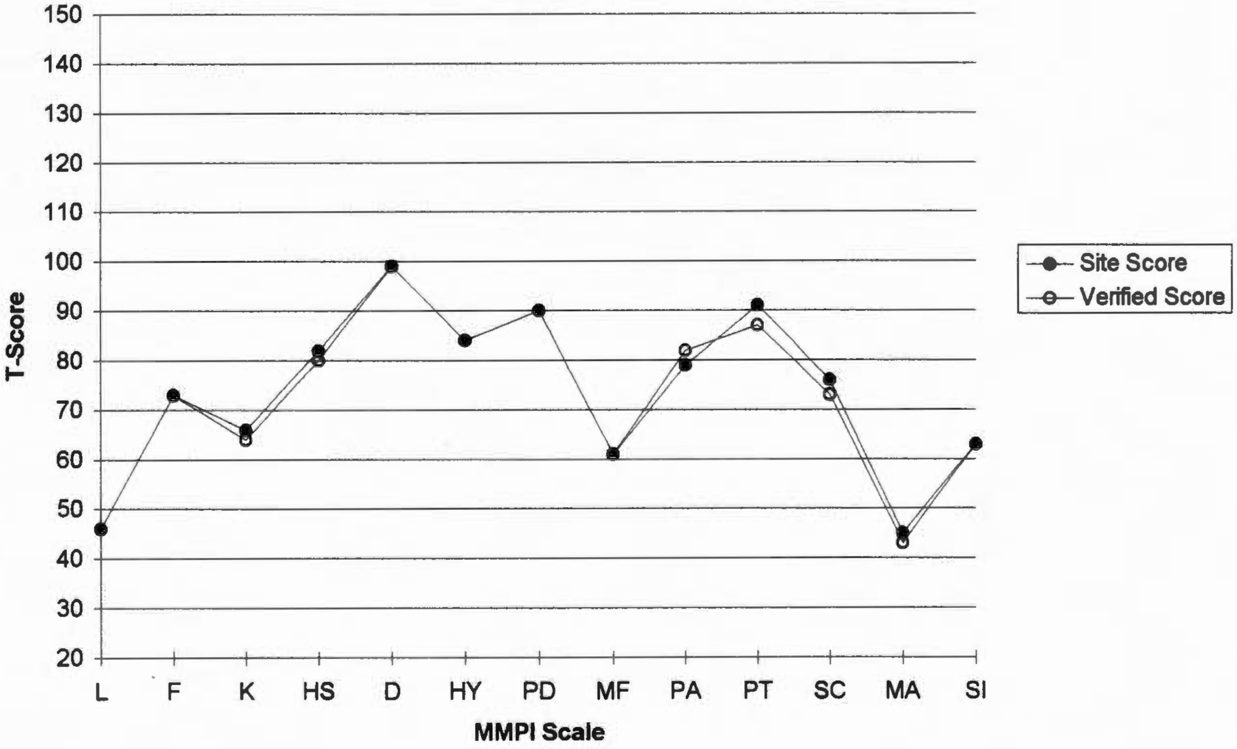
Figures 11. Case 2988. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



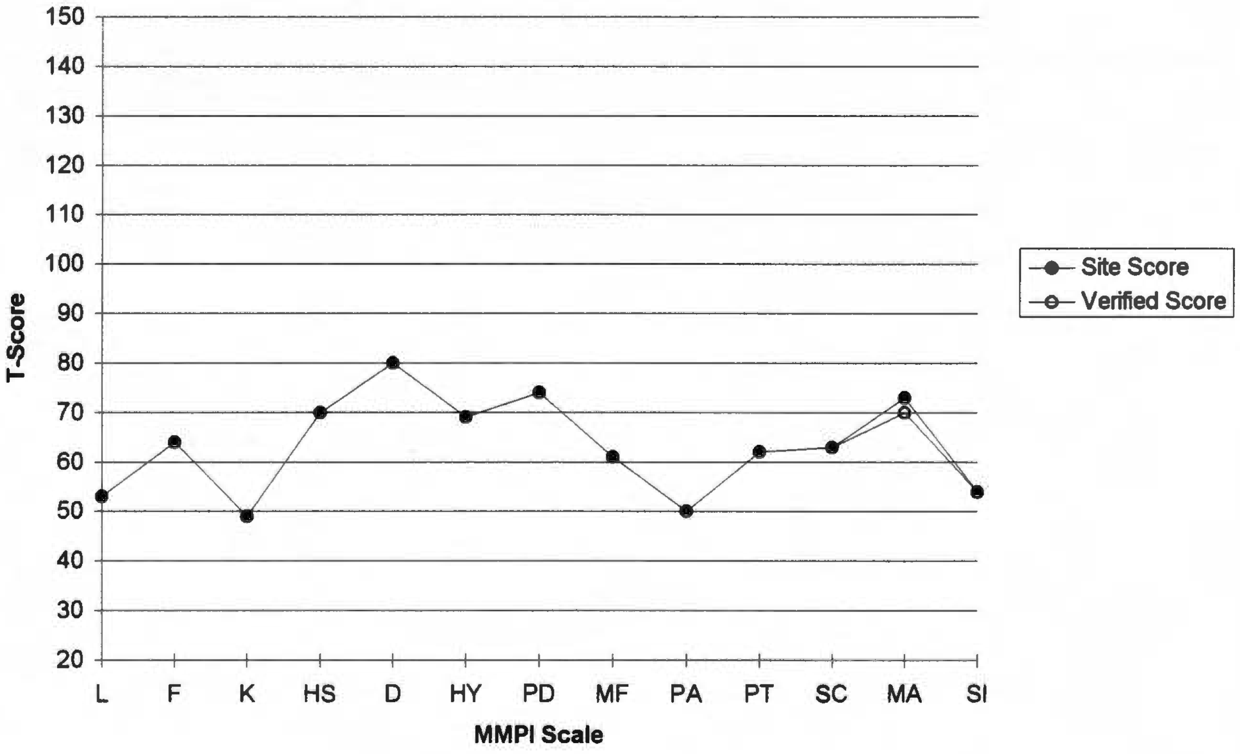
Figures 12. Case 6189. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



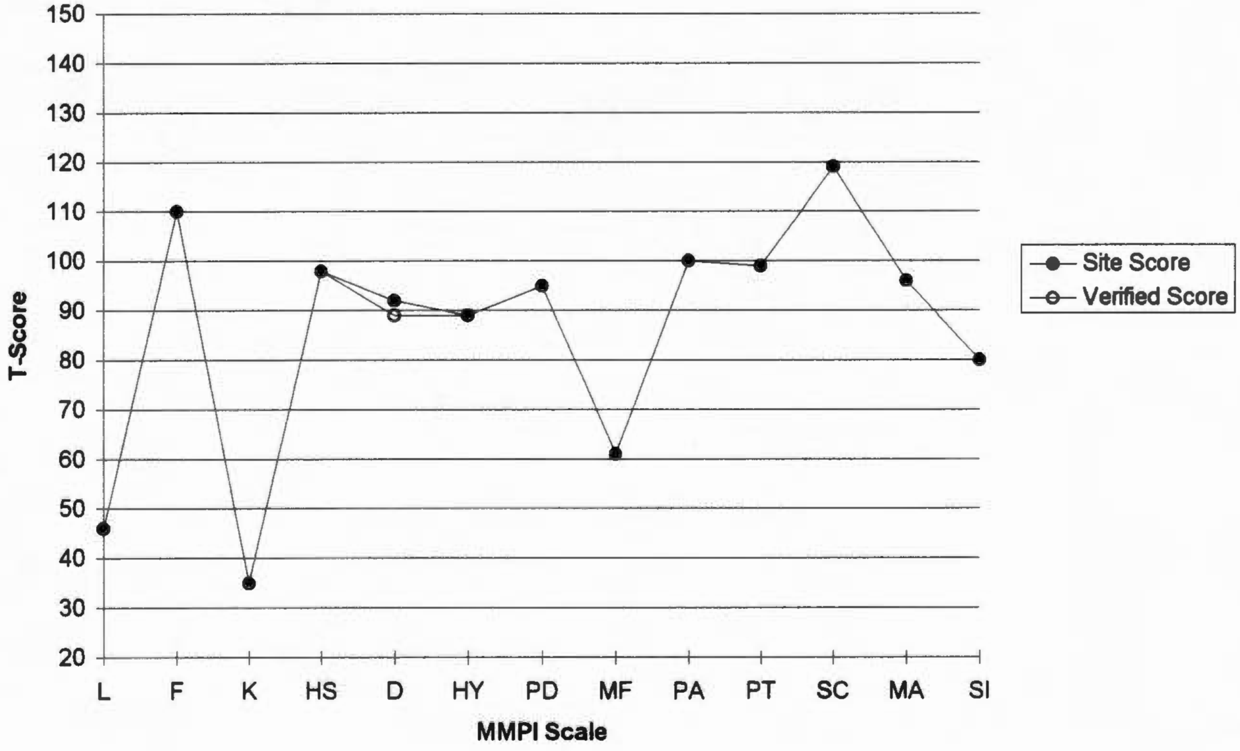
Figures 13. Case 6233. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



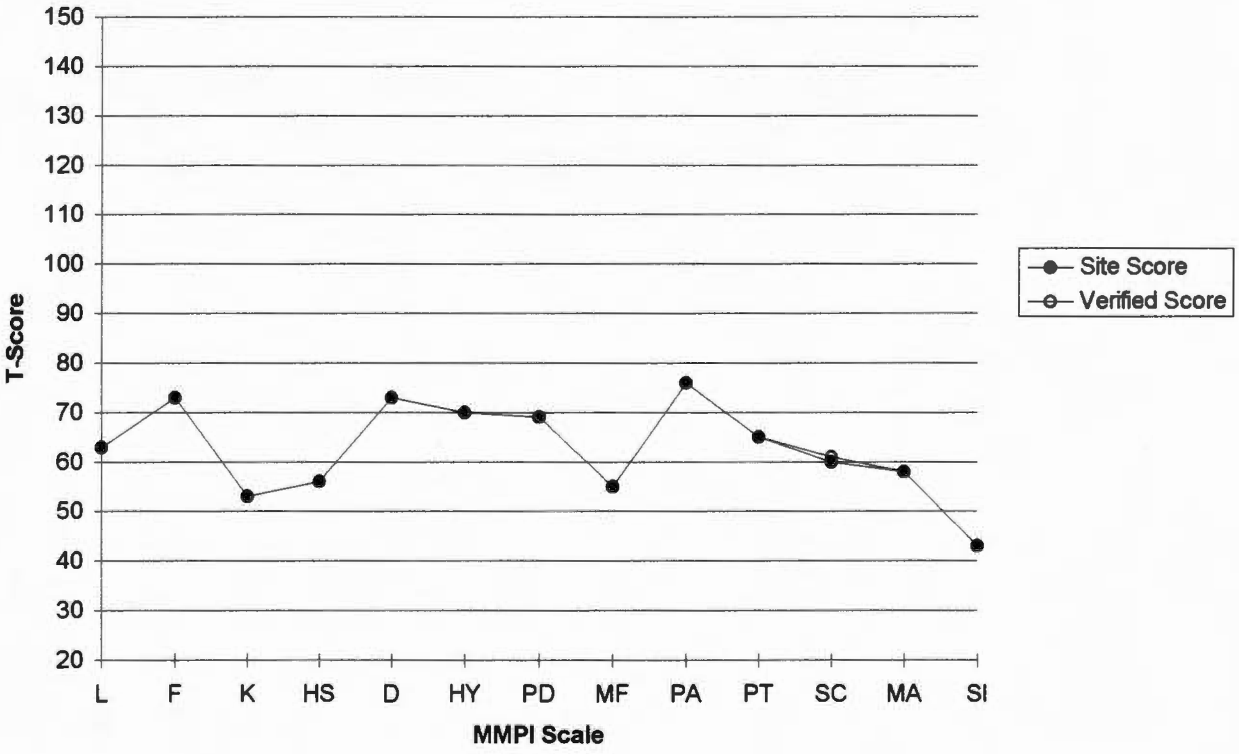
Figures 14. Case 8832. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



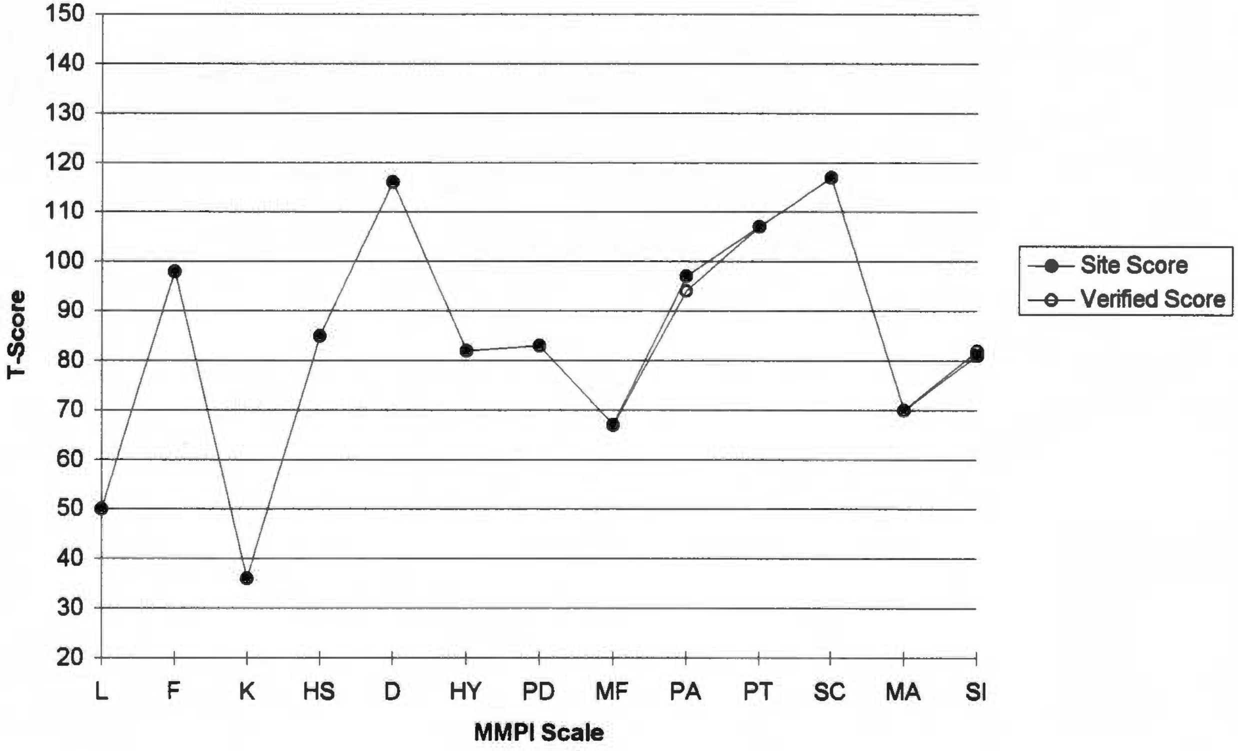
Figures 15. Case 3-45339. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



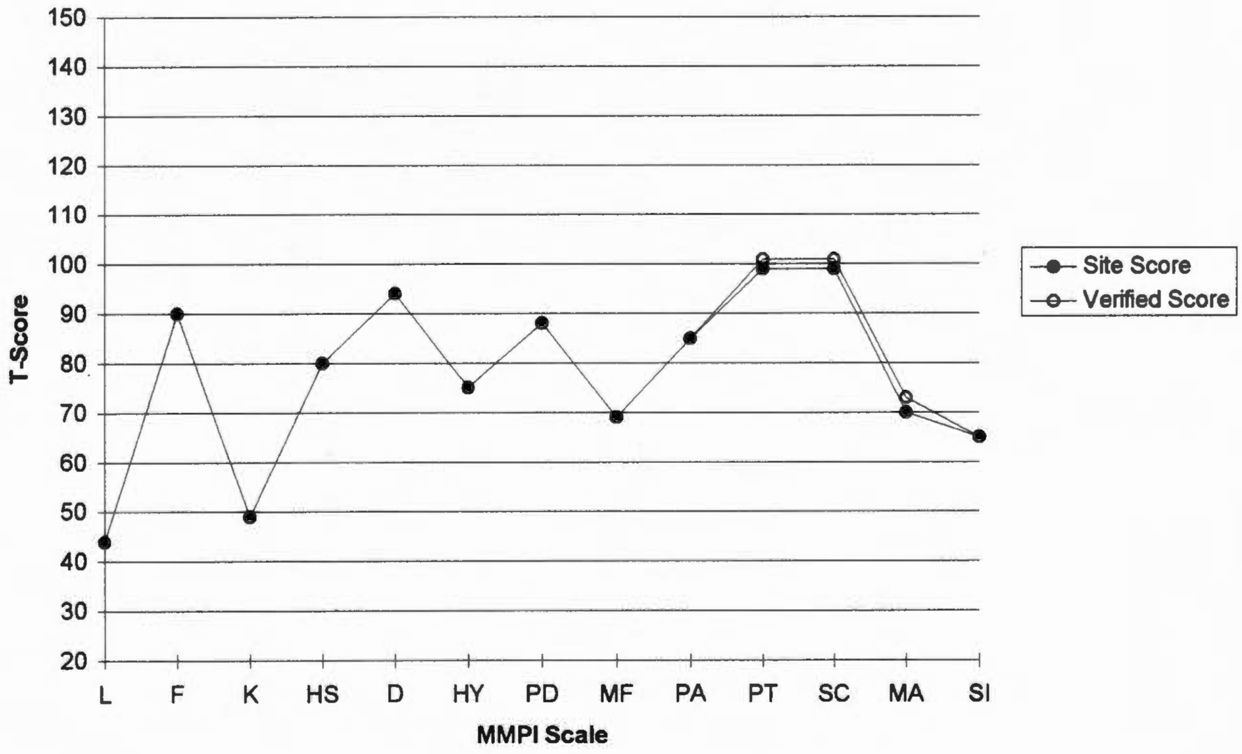
Figures 16. Case 5670-0773. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



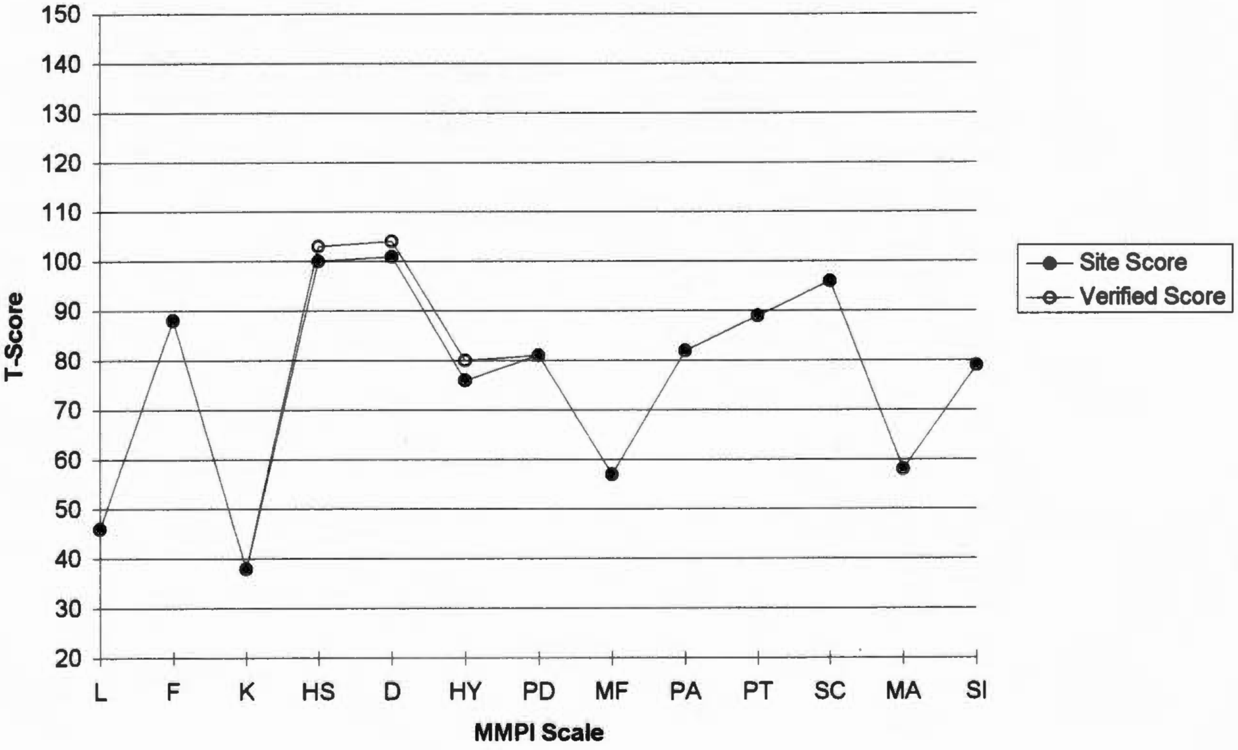
Figures 17. Case 6789-174. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



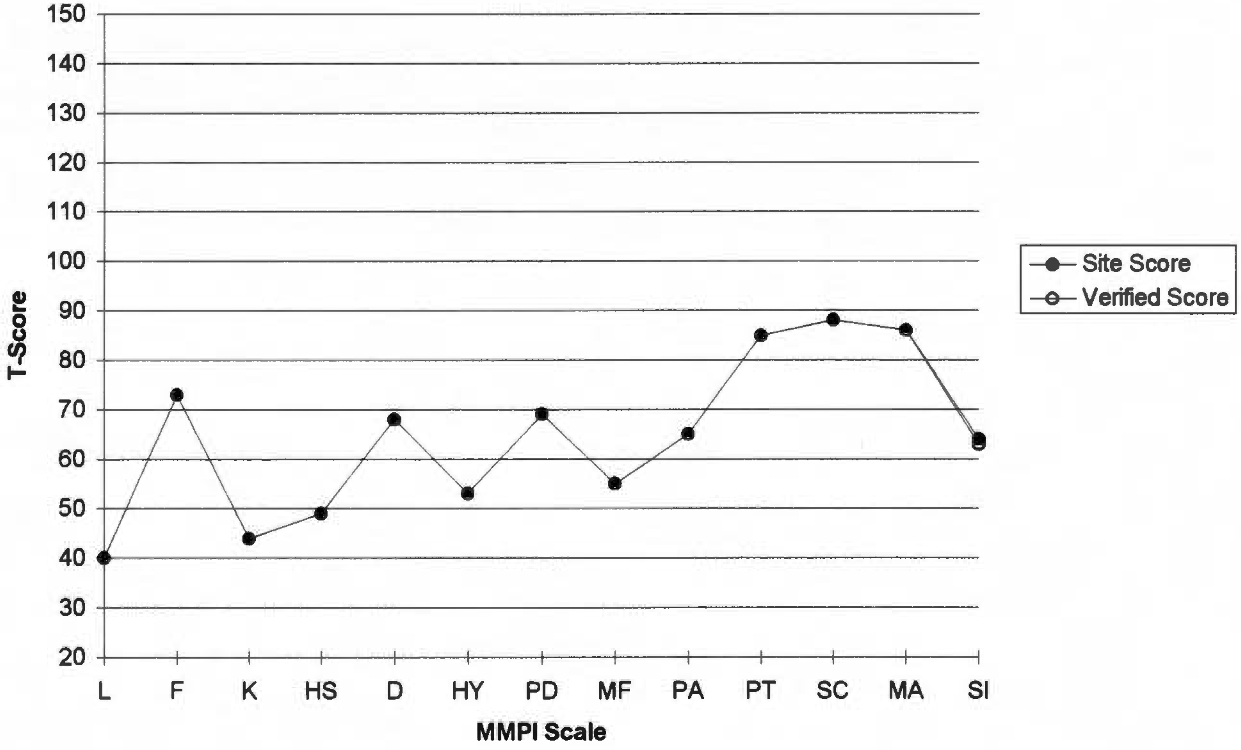
Figures 18. Case 6871-0789. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



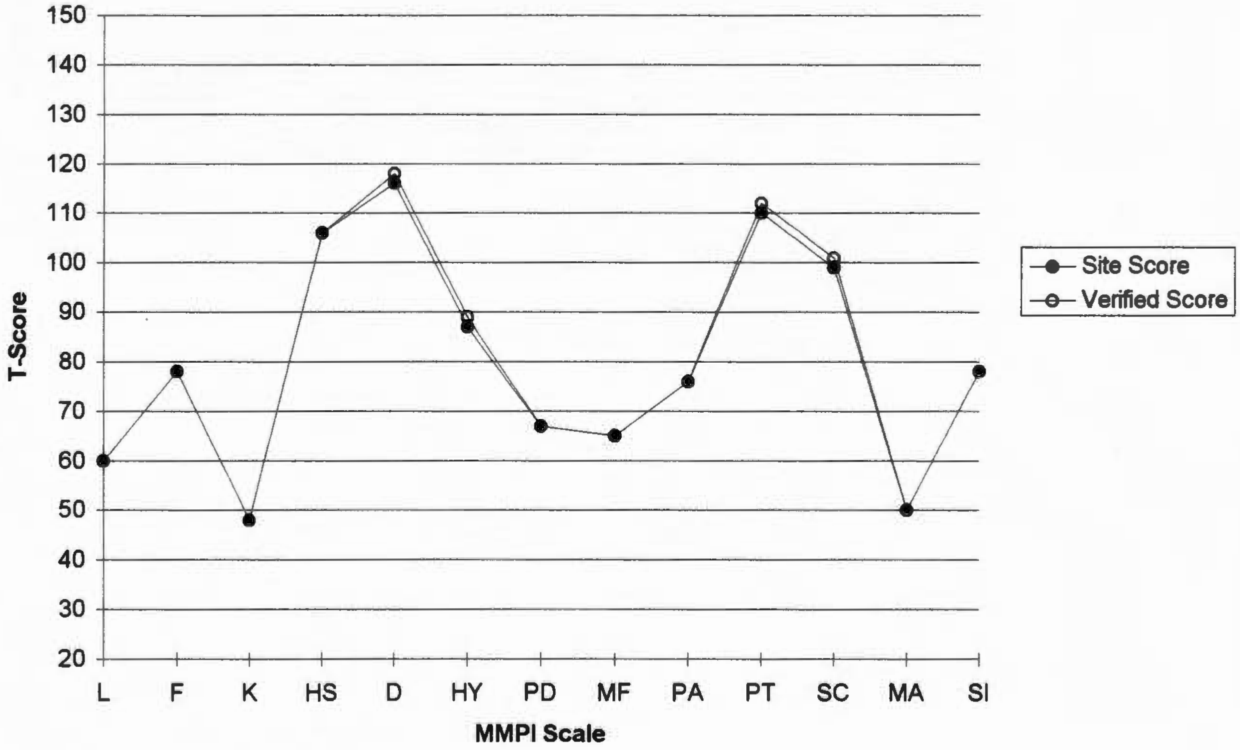
Figures 19. Case 7967-0145. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



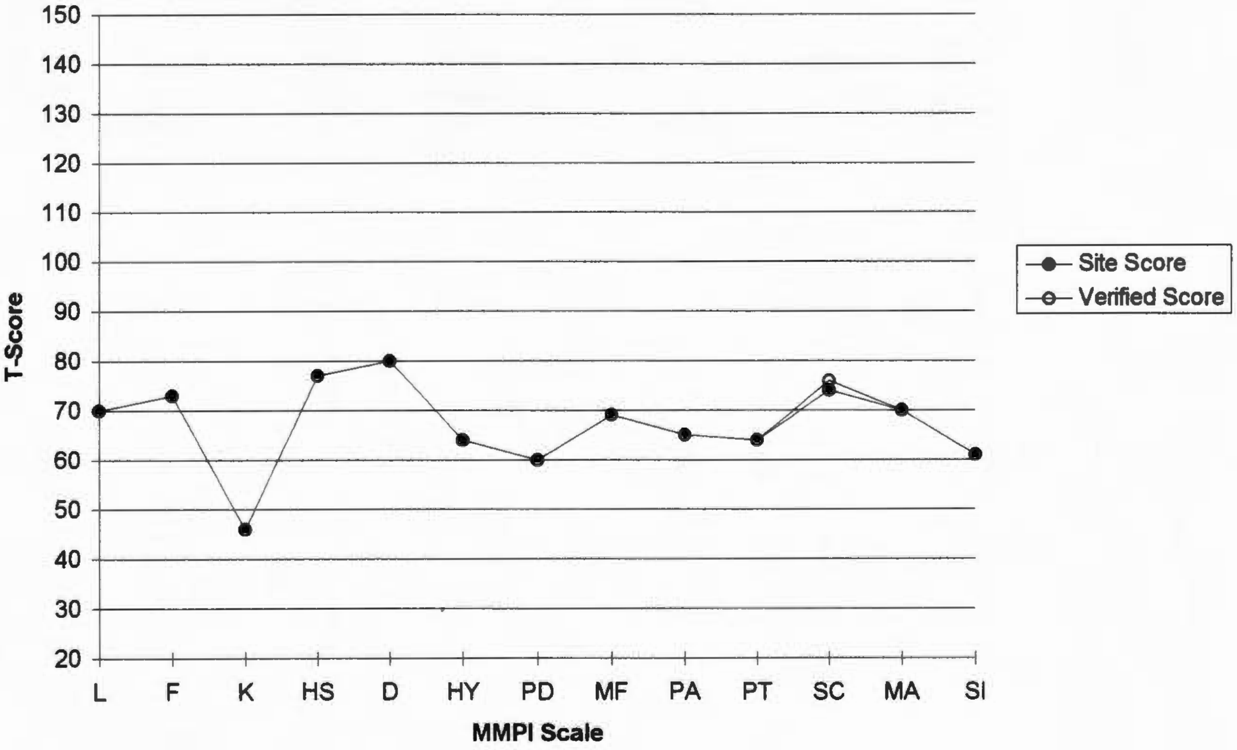
Figures 20. Case 8252-0761. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



Figures 21. Case 9890-0805. MMPI profile discrepancy recalculated using NCS MMPI scoring standard (University of Minnesota, 1983).

MMPI Site Scored vs. Verified Profiles
Based on 1983 NCS MMPI Scoring Standard



Appendix 1

Scoring error coverletter and survey submitted to APA Division 12 Fellows

This survey pertains only to the use of objective personality tests for clinical purposes. Please answer each question accordingly.

1.) Please use the 5-point scale below to rank frequency of use for each of the following tests.

Usage Rank:
5 = Always, 4 = Frequently, 3 = Sometimes,
2 = Rarely, 1 = Never

Name of Instrument	Rank
MMPI (Original version)	______
MMPI-2	______
Beck Depression Inventory	______
Strong-Campbell Interest Inventory	______
16 PF Questionnaire	______
MCMI (Original version)	______
MCMI-II	______
MCMI-III	______
California Psychological Inventory	______
SCL-90R	______
Other:____________________	______
Other:____________________	______
Other:____________________	______

The remaining questions concern the MMPI. All questions refer only to the three standard validity scales and the ten standard clinical scales.

2.) If you score the MMPI by hand; e.g., using templates, in what percentage of cases do you score the test a second time for accuracy?

a) Not applicable; I don't score by hand
b) Never
c) 1-10% rescored
d) 11-25% rescored
e) 26-50% rescored
f) 51-75% rescored
g) 75-100% rescored

3.) If you rescore MMPIs, how often do you find protocols with small errors, i.e., errors that alter scale T-scores by less than 5 points?

a) Not applicable; I don't rescore MMPIs
b) 0% of protocols
c) 1-5% of protocols
d) 6-10% of protocols
e) 11-20% of protocols
f) 21-35% of protocols
g) 36-50% of protocols
h) 51-75% of protocols
i) 76-100% of protocols

4.) If you rescore MMPIs, how often do you find protocols with larger errors, i.e., errors that alter scale T-scores by 5 or more points?

a) Not applicable; I don't rescore MMPIs
b) 0% of protocols
c) 1-5% of protocols
d) 6-10% of protocols
e) 11-20% of protocols
f) 21-35% of protocols
g) 36-50% of protocols
h) 51-75% of protocols
i) 76-100% of protocols

5.) For scoring errors on the MMPI limited to less than 5 T-score points, what frequency of scoring errors would you deem clinically acceptable per protocol?

a) No more than 5 errors per protocol
b) No more than 2-4 errors per protocol
c) No more than 1 error per protocol
d) No more than 1 error per 2-4 protocols
e) No more than 1 error per 5-9 protocols
f) No more than 1 error per 10 protocols
g) Less than 1 error per 10 protocols, but something more than no errors at all
h) No errors at all

6.) For scoring errors on the MMPI ranging from 5 to 10 T-score points, what frequency of scoring errors would you deem clinically acceptable per protocol?

a) No more than 5 errors per protocol
b) No more than 2-4 errors per protocol
c) No more than 1 error per protocol
d) No more than 1 error per 2-4 protocols
e) No more than 1 error per 5-9 protocols
f) No more than 1 error per 10 protocols
g) Less than 1 error per 10 protocols, but something more than no errors at all
h) No errors at all

7.) For scoring errors on the MMPI greater than 10 T-score points, what frequency of scoring errors would you deem clinically acceptable per protocol?

a) No more than 5 errors per protocol
b) No more than 2-4 errors per protocol
c) No more than 1 error per protocol
d) No more than 1 error per 2-4 protocols
e) No more than 1 error per 5-9 protocols
f) No more than 1 error per 10 protocols
g) Less than 1 error per 10 protocols, but something more than no errors at all
h) No errors at all

8.) Of all MMPI results you have reviewed that have been scored by *others*, what percentage of cases have you obtained the raw data *and* rescored the tests?

a) Not applicable; I don't review others' MMPI results
b) 0% of cases
c) 1-5% of cases
d) 6-10% of cases
e) 11-20% of cases
f) 21-35% of cases
g) 36-50% of cases
h) 51-75% of cases
i) 76-100% of cases

The remaining questions pertain solely to MMPI computer scoring, not computer interpretation.

9.) How often you use computer programs or services in scoring the MMPI?

a) Never
b) 1-5% of protocols
c) 6-10% of protocols
d) 11-20% of protocols
e) 21-35% of protocols
f) 36-50% of protocols
g) 51-75% of protocols
h) 76-100% of protocols

10.) If you use computer scoring programs or services, how often do you check on the accuracy of the computer scoring?

a) Not applicable; I don't use computer programs or services
b) I never check computer scoring
c) 1-5% of protocols
d) 6-10% of protocols
e) 11-20% of protocols
f) 21-35% of protocols
g) 36-50% of protocols
h) 51-75% of protocols
i) 76-100% of protocols

11.) When you check on computer scoring, how often do you find scoring errors?

a) Not applicable; I don't use computer programs or services
b) 0% of protocols
c) 1-5% of protocols
d) 6-10% of protocols
e) 11-20% of protocols
f) 21-35% of protocols
g) 36-50% of protocols
h) 51-75% of protocols
i) 76-100% of protocols

12.) If you use computer scoring, which choice best describes the protocol you follow?

a) Not applicable; I don't use computer scoring
b) I send tests to computer scoring services for scoring
c) I have an onsite facility for keypunching raw data into a computer
d) I have an onsite facility for keypunching raw data into a computer, and it has a double entry system for accuracy
e) I use an optical scanner to enter data for a computer scoring program
f) Other:________________________

VIII. Bibliography

Allard, G. A., Butler, J., Shea, M. T., & Faust, D. (1995). Errors in hand-scoring objective personality tests: The case of the Personality Diagnostic Questionnaire--Revised (PDQ-R). Professional Psychology: Research and Practice, 26(3), 304-308.

Beasley, M. G., Lobasher, M., Henley, S., & Smith, I. (1988). Errors in computation of WISC and WISC-R intelligence quotients from raw scores. Journal of Child Psychology and Psychiatry, 29(1), 101-104.

Beck, A. T., & Steer, R. A. (1987). Beck Depression Inventory manual. San Antonio, TX: The Psychological Corporation.

Blakey, W. A., Fantuzzo, J. W., Gorsuch, R. L., & Moon, G. W. (1987). A peer-mediated, competency-based training package for administering and scoring the WAIS-R. Professional Psychology: Research and Practice, 18(1), 17-20.

Boehm, A. E., Duker, J., Haesloop, M. D., & White, M. A. (1974). Behavioral objectives in training for competence in the administration of individual intelligence tests. Journal of School Psychology, 12(2), 150-157.

Connor, R., & Woodall, F. E. (1983). The effects of experience and structured feedback on WISC-R error rates made by student-examiners. Psychology in the Schools, 20, 376-379.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). An MMPI handbook: Volume 1: Clinical interpretation.

Faust, D. (1984). The limits of scientific reasoning. Minneapolis, MN: University of Minnesota Press.

Fowler, R. D. & Coyle, F. A. (1968). Scoring error on the MMPI. Journal of Clinical Psychology, 25, 62-63.

Greene, R. L. (1991). The MMPI-2/MMPI: An interpretive handbook. Boston, MA: Allyn and Bacon.

Johnson, D. L., & Candler, A. C. (1985). Using a small computer for test score conversion. College Student Journal, 19(1), 102-106.

Levine, D. S., & Willner, S. G. (1976). The cost of mental illness, 1974. In Mental Health Statistical Note No. 125 (pp. 1-7). Washington, D. C.: National Institute of Mental Health.

Microsoft Excel 4.0 [Computer program]. (1992). Redmond, WA: Microsoft Corporation.

Microsoft Excel 5.0 [Computer program]. (1994). Redmond, WA: Microsoft Corporation.

Miller, C. K., & Chansky, N. M. (1972). Psychologists' scoring of WISC protocols. Psychology in the Schools, 9, 144-152.

Miller, C. K., Chansky, N. M., & Gredler, G. R. (1970). Rater agreement on WISC protocols. Psychology in the Schools, 7, 190-193.

Moreland, K. L., Eyde, L. D., Robertson, G. J., Primoff, E. S., & Most, R. B. (1995). Assessment of test user qualifications: A research-based measurement procedure. American Psychologists, 50, 14-23.

Oakland, T., Lee, S. W., & Axelrod, K. M. (1975). Examiner differences on actual WISC protocols. Journal of School Psychology, 13(3), 227-233.

Piotrowski, C., & Keller, J. W. (1989). Psychological testing in outpatient mental health facilities: A national study. Professional Psychology: Research and Practice, 20, 423-425.

Piotrowski, C., & Keller, J. W. (1992). Psychological testing in applied settings: A literature review from 1982-1992. The Journal of Training & Practice in Professional Psychology, 6(2), 74-82.

Piotrowski, C., & Lubin, B. (1990). Assessment practices of health psychologists: Survey of APA Division 38 Clinicians. Professional Psychology: Research and Practice, 21(2), 99-106.

Piotrowski, C., Sherry, D., & Keller, J. W. (1985). Psychodiagnostic test usage: A survey of the Society for Personality Assessment. Journal of Personality Assessment, 49, 115-119.

Levenson, Jr., R. L. Golden-Scaduto, C. J., Aiosa-Karpas, C. J., & Ward, A. W. (1988). Effects of examiners' education and sex on presence and type of clerical errors made on WISC-R protocols. Psychological Reports, 62, 659-664.

Ryan, J. J., Prifitera, A., & Powers, L. (1983). Scoring reliability on the WAIS-R. Journal of Consulting and Clinical Psychology, 51(1), 149-150.

Sherretts, S., Gard, G., & Langner, H. (1979). Frequency of clerical errors on WISC protocols. Psychology in the Schools, 16(4), 495-496.

Slate, J. R., & Chick, D. (1989). WISC-R examiner errors: Cause for concern. Psychology in the Schools, 26, 78-84.

Slate, J. R., & Jones, C. H. (1990). Identifying students' errors in administering the WAIS-R. Psychology in the Schools, 27, 83-87.

Slate, J. R., Jones, C. H., & Murray, R. A. (1991). Teaching administration and scoring of the Wechsler Adult Intelligence Scale-Revised: An empirical evaluation of practice administrations. Professional Psychology: Research and Practice, 22(5), 375-379.

Slate, J. R., Jones, C. H., Murray, R. A., & Coulter, C. (1993). Evidence that practitioners err in administering and scoring the WAIS-R. Measurement and Evaluation in Counseling and Development, 25, 156-161.

Slate, J. R., & Hunnicutt, Jr., L. C. (1988). Examiner errors on the Wechsler scales. Journal of Psychoeducational Assessment, 6, 280-288.

Spielberger, C. D. (1983). State-Trait Anxiety Inventory for adults: Sampler set, manual, test, scoring key. Palo Alto, CA: Consulting Psychologists Press, Inc.

University of Minnesota. (1983). Minnesota Multiphasic Personality Inventory manual for administration and scoring. Minnesota, MN: University of Minnesota Press.

Wade, T. C., & Baker, T. B. (1977). Opinions and use of psychological testing: A survey of clinical psychologists. American Psychologist, 32, 874-882.

Warren, S. A., & Brown, Jr., W. G. (1973). Examiner scoring errors on individual intelligence tests. Psychology in the Schools, 10, 118-122.

Zilbergeld, B. (1983). The Shrinking of America. Boston: Little, Brown.