ERRORS IN SCORING OBJECTIVE PERSONALITY TESTS

Gregory Allard
David Faust
University of Rhode Island

Given the paucity of previous research, we examined the occurrence of scoring error on widely used objective personality tests and examined its possible relation to two factors: scoring procedure complexity (SPC) and commitment to accuracy (CTA). We double-checked the scoring of three tests (MMPI, Beck Depression Inventory, Spielberger State/Trait Anxiety Inventory) across three settings. Each of the tests were misscored at a surprisingly high rate in at least one setting, and some such errors altered major interpretive implications. Tests of higher SPC showed greater error rates, but high CTA greatly reduced the occurrence of error across levels of SPC. Unexpected sources of error were also uncovered, such as commercial computer scoring errors and disagreement in scoring standards among test publishers. Practical suggestions for improving scoring accuracy are offered.

Keywords: Objective personality tests, test administration, scoring error, quality assurance, computerized test scoring, MMPI, MMPI-2, BDI, STAI

Objective personality tests are used many thousands of times annually to aid clinicians in assessing psychological characteristics or maladies. The results may greatly influence diagnosis, expert testimony in legal cases, or treatment recommendations. Therefore, objective personality test results can have a major impact on individual lives.

Considerable research has demonstrated the vulnerability of other types of psychological measures to scoring error, such as the Wechsler Intelligence Scales (e.g., Blakey, Fantuzzo, Gorsuch, & Moon, 1987; Boehm, Duker, Haesloop, & White, 1974; Connor & Woodall, 1983; Miller & Chansky, 1972; Slate & Chick, 1989; Slate & Jones, 1990; Slate, Jones, & Murray, 1991). Errors often are made on items requiring examiner judgment, but also in the performance of simple clerical operations (e.g., Beasley, Lobasher, Henley, & Smith, 1988; Miller & Chansky, 1972; Miller, Chansky, & Gredler, 1970; Sherretts, Gard, & Langner, 1979).

In contrast, little attention has been directed to possible scoring errors on objective personality tests, perhaps because the accuracy of scoring has almost been taken for granted. Ironically, “avoiding errors in scoring and recording” was cited as
the first of 12 minimum test user competencies in a recent publication by the Test User Qualifications Group, a subcommittee of the Joint Committee on Testing Practices (Moreland, Eyde, Robertson, Primoff, & Most, 1995). At least one study (Allard, Butler, Shea, & Faust, 1995) suggested, in fact, that errors in scoring objective personality tests may not be rare or benign. Allard et al. found clerical errors in 53% of protocols of one test, resulting in changed diagnostic classification in 19% of cases. These same authors conducted an exploratory survey of consultants who review objectively scored personality tests on a national level, the results of which also suggested problematic error rates on other measures.

Concerns about error in scoring objective personality tests are heightened by the frequency with which these tests are administered, and by the potential impact of even minor errors, in light of the general preference for interpretation based on scale score patterns and diagnostic thresholds. For example, based on projections of the number of annual administrations of a widely used measure such as the MMPI (Piotrowski & Keller, 1989; Wade & Baker, 1977), an error rate as low as 1% per test could affect hundreds or thousands of protocols and inevitably would alter major interpretive implications in some cases. Finally, it is of pragmatic utility to determine the frequency and magnitude of scoring error because such knowledge can guide intervention, if needed, and can be used to identify effective corrective steps.

The primary aim of this study was to extend Allard et al.’s (1995) earlier research by examining the frequency of scoring error on objective personality tests across multiple sites and tests to help clarify whether scoring accuracy on such measures merits attention. In addition, if unacceptable error rates were uncovered, an additional aim was to examine factors that might explain differences in error rates in order to better understand and potentially correct the problem. Research with cognitive measures suggests that demographic variables (e.g., educational level, gender, training in test administration, and clinical experience) show minimal relation with the occurrence of mechanical or clerical error (cf. Levenson, Golden-Scaduto, Aiosa-Karpas, & Ward, 1988; Oakland, Lee, & Axelrod, 1975; Ryan, Prifitera, & Powers, 1983; Sherrretts et al., 1979; Slate, Jones, Murray, & Coulter, 1993). Two factors, however, may show strong relations to error rate: (a) commitment to scoring accuracy (Miller & Chansky, 1972; Slate & Chick, 1989; Slate et al., 1991), and (b) the complexity of the measures themselves, or the steps need to score them (Johnson & Candler, 1985; Miller & Chansky, 1972; Slate & Hunnicutt, 1988). Allard et al. (1995) found strong associations, accounting for at least half of the variance, between the frequency of scoring error and scoring procedure complexity. When complexity of scoring operations increased, so, too, did scoring errors. Thus, in addition to examining the rate of scoring error, we also analyzed the relation between its occurrence, commitment to accuracy, and scoring complexity.

Method

Settings

Obtaining particular sites in a study such as this is not easy, as the sites potentially expose themselves to legal jeopardy. Furthermore, we wished to obtain a diversity of sites so that some reasonable conclusions could be drawn about whether errors, if uncovered, were idiosyncratic. Additionally, the sites needed to use common measures. Fortunately, we were able to obtain cooperation from all three sites we initially targeted based on their seeming fit with our research aims: a VA inpatient hospital, a VA outpatient clinic, and a private inpatient hospital.

Tests

We wished to select tests that are used frequently (in order to maximize the potential relevance of the study), vary in scoring complexity, (given the wish to study the effects of this variable), and were common across participating sites. Three tests met these requirements: the Minnesota Multiphasic Personality Inventory (MMPI; University of Minnesota, 1983), the Beck Depression Inventory (BDI; Beck & Steer, 1987), and the State-Trait Anxiety Inventory (STAI; Spielberger, 1983). Surveys show that these tests, particularly the
Errors in Scoring Objective Tests

MMPI and BDI, are among the most widely used clinical measures (Piotrowski & Keller, 1989, 1992; Piotrowski & Lubin, 1990).

The STAI is administered in two parts, the State (STAI_S) and the Trait (STAI_T). Although initial analyses were conducted on each part of the test separately, results were nearly identical. Hence, only the combined results are reported.

Measures of Commitment to Accuracy and Scoring Complexity

Commitment to accuracy (CTA) was conceptualized as a set of behaviors employed in scoring a test rather than as a hypothetical construct. We considered “full” CTA (F-CTA) to be present when all operations of test scoring were either double-checked (i.e., scored twice) or optically scanned and computer scored, and “less-than-full” CTA (L-CTA) to be present when scoring operations consisted of unchecked keyboarding or less than fully double-checked hand scoring. We hypothesized that tests scored with L-CTA, unlike tests scored with F-CTA, would yield problematic or unacceptable error rates.

Scoring procedure complexity (SPC) was defined as the number of steps required to score a protocol. We realize that this is an overly simple definition, but it proved to be a powerful discriminator in the Allard et al. (1995) study. Tests with lower versus higher SPC were expected to yield fewer scoring errors, in part depending on CTA; that is, SPC was expected to lead to problematic error rates only with L-CTA. Stated differently, even with complex protocols, F-CTA was expected to nearly eliminate error. Thus, measures were selected in part to reflect different levels of scoring complexity.

Tests scored with L-CTA were drawn from all three settings in order to examine whether error frequencies or patterns were idiosyncratic or more general. Tests scored with F-CTA versus L-CTA were expected to reveal far lower error rates, and consequently F-CTA tests were sampled from only one setting.

The MMPI, BDI, and STAI vary in SPC ratings when scored fully by hand. Only two SPC ratings were used, however, because at all three sites the MMPI scoring process was limited to keyboarding item responses into a computer scoring program. This process reduced the SPC for the MMPI to one clerical task of low complexity. The BDI, which requires the addition of item responses to arrive at a total score, was also considered to be low complexity. Scoring the STAI was considered to be of high complexity because one must identify a subset of items that are reverse-coded before both scales are tallied (two separate steps per part).

Procedure

We initially contacted the three sites to solicit their participation. All three sites agreed to participate and were screened to assess their CTA. We obtained data scored with L-CTA at each site. We had hoped to locate at least one setting that demonstrated F-CTA, but none of these three settings, which otherwise met our research requirements quite well, satisfied this condition. After careful consideration and exploration of other potential sites, we decided to create a simulated F-CTA data set.

Fifty tests of each type were chosen at each setting. Based on the original division of the STAI into two parts and subsequent pooling of results, the sample sizes for the STAI are double those for the MMPI and BDI. All three settings hand-scored the STAI and BDI. All three settings hand-keyed MMPI raw data into MMPI scoring programs.

F-CTA was simulated by rescoring patient data from each of the three settings using optical scanning and computer scoring. The scanning templates were created using National Computer Systems (NCS) ScanTools Software (1990), and all tests were scanned using the NCS OpScan 5 optical scanner. Data from 50 of each test type (including each part of the STAI) were simulated for a total of 200 tests. Data for the STAI and BDI protocols had to be transcribed onto scannable forms. Transcription accuracy was verified against the double-checked original raw data.

1At the time of data collection, all three settings had just begun administering the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). We chose to collect MMPI data primarily because of its availability across three settings. While the items, norms, and profiles of the two tests are different, the focus of this study concerns the scoring process which is similar across both older and newer editions.
again using electronic verification. MMPI data had already been entered onto scannable forms at one setting, eliminating the need for transcription. All scanned data were then compared to the verified double-checked rescored data sets previously described.

For each test sampled, patient's raw data answer sheets, and, if applicable, derived summary score sheets or original keyboarded patient responses were compared against corresponding rescored raw data. Raw data were keyboarded into computer programs, which then calculated summary scores, the verified scores. Raw data and original scores were also keyboarded into computer programs to provide a representation of original scores. To ensure accuracy of electronic representation of both original and verified data and scores, the computer program was designed for double-entry by two independent keyboarders. Discrepancies and comparisons were all identified with computer programs. Testing for keyboarding errors on the MMPI required further comparisons which necessitated both the keying of raw data from the patient's answer sheet and corresponding keyboarded raw data.

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 BDI 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 State-Trait Anxiety Inventory Manual, Test, and Scoring Key (Spielberger, 1983). None of the sites specified interpretation protocols or norm groups. Instead, settings provided only non-standardized raw score totals. (As such, STAI SPC rankings in this study only included steps to score raw score totals.) Although all settings used MMPI computer scoring programs with K-corrected original Minnesota norms, none of the settings could readily provide software brand or version information that would allow for verifying scale composition or scoring algorithms. To bypass this potential source of variance, MMPI raw data was rescored using a scoring program based on the most recent NCS MMPI Manual for Administration and Scoring (NCS scoring standard; University of Minnesota, 1983).

Once our analyses were completed, we contacted each site to provide detailed information about their respective results.

**Results**

**Aggregated Error Rates on Sampled Test Data**

In Table 1, results are subdivided by level of CTA and high versus low SPC. Across the 600 tests sampled from the three settings that used L-CTA scoring procedures, 78 (21%) had scoring errors. The low SPC tests (BDI and MMPI) showed about half the error rate as the high SPC test (STAI). Of the 200 tests in the F-CTA condition, only two (1%) had scoring errors, and the overall error rate was very low, regardless of test complexity. Although both errors occurred with a low SPC test (MMPI), the small number of errors within this condition and across conditions indicates that F-CTA greatly attenuated scoring errors, even with higher SPC. Thus, as expected, error rates showed clear association with scoring complexity and commitment to accuracy.

|       | CTA |       |   |
|-------|-----|-------|---|
|       | L-CTA |       | F-CTA |
| SPC rank | Error rate | n | Error rate | n |
| High SPC | 29% | 300 | 0% | 100 |
| Low SPC | 14% | 300 | 2% | 100 |
| CTA Totals | 21% | 600 | 1% | 200 |

*Note. L-CTA = Low commitment to accuracy; F-CTA = Full commitment to accuracy.*

*aIncludes the combined results for the State-Trait Anxiety Inventory, State Form and Trait form. bIncludes the combined results for the Beck Depression Inventory and the Minnesota Multiphasic Personality Inventory.*
Analyses of the frequency of scoring errors committed within a given test were only possible with the MMPI. Overall, 78 miskeyed items were discovered in the 150 MMPI protocols sampled. In those tests with errors, 10 tests had one error, 6 tests had two to five errors, and six tests had 6 to 20 errors.

**Error Rates for Tests Across Settings**

Total error rates for each test type at each setting are shown in Table 2. The scoring error rate at the VA outpatient clinic was notably more than five times greater than that found at the private inpatient hospital. Despite these differences in error frequencies among settings, each setting produced higher frequencies of erroneous tests than the F-CTA sample. Moreover, as shown in Table 3, error frequencies for each setting were generally concordant with SPC; errors on the high SPC test occurred at about twice the rate as that seen on the two low SPC tests combined.

**Types of Errors**

We examined the original test materials to look for possible sources and patterns of scoring error. BDI score discrepancies appeared to show two patterns. Either small numbers of items were mistaillied in deriving total scores, a pattern common to all three settings, or, at two of the settings, some scores were off by about 21 points, the likely result of a disparity in test forms. These two settings used two different forms, one of which assigned scores of 0 to 3 on the 21 items (corresponding to the standard BDI test forms), and another of which mistakenly assigned scores of 1 to 4, the latter of which likely led to the 21-point scoring errors.

STAI score discrepancies also manifested two patterns. For two of the three settings, total raw score

| Test  | Setting     | VA outpatient | VA inpatient | Private inpatient |
|-------|-------------|---------------|--------------|-------------------|
|       |             | n  | Error rate | n  | Error rate | n  | Error rate |
| STAI  | VA outpatient | 100 | 51%     | 100 | 25%     | 100 | 10%     |
| BDI   | VA inpatient  | 50  | 18%     | 50  | 20%     | 50  | 2%      |
| MMPI  | Private inpatient | 50  | 34%     | 50  | 6%      | 50  | 4%      |
| Setting Totals |          | 200 | 39%     | 200 | 19%     | 200 | 7%      |

*Note. STAI = State-Trait Anxiety Inventory, BDI = Beck Depression Inventory, MMPI = Minnesota Multiphasic Personality Inventory.*

| SPC rank | Setting  | VA outpatient | VA inpatient | Private inpatient |
|----------|----------|---------------|--------------|-------------------|
| High SPCa | VA outpatient | 51%*          | 25%          | 10%              |
| Low SPCb  | VA inpatient  | 26%           | 13%          | 3%               |

*aIncludes the combined results for the State-Trait Anxiety Inventory, State and Trait forms.
bIncludes the combined results for the Beck Depression Inventory and Minnesota Multiphasic Personality Inventory. *n = 100 for each cell.
errors ranged from 1 to 9 points. One of three settings also demonstrated another error pattern in which total score discrepancies ranged from 20 to 30 points, the likely result of a non-standard scoring routine which required adding 27 points to final scores.

For the MMPI, hand-scorer errors were constrained by the scoring format, that is, keyboarding. Despite 566 opportunities for error per questionnaire, two of the settings made errors in only a small percentage of protocols, consisting of one or a few miskeyed items. For the third setting, however, 34% of tests showed keyboarding errors. Six of these tests revealed 5 to 20 miskeyed items. Unlike the other two settings, MMPIs in this setting were all scored in a distracting receptionist area.

**Potential Impact on Test Interpretation**

Some examples of alterations in test results due to scoring error are presented here for illustrative purposes. In the case of the BDI, scoring errors of 21 points that occurred in a number of cases had the potential for grossly altering the classification of results. For example, in one case, a score indicating mild depression (14) was misscored as indicating severe depression (35).

For the STAI, errors frequently resulted in 10 to 20 point changes in T scores. For example, a STAI_S score representing high situational anxiety (72) was misscored as indicating normal situational anxiety (52) (Spielberger, 1983, pp. 25-26).

With the MMPI, interpretation often focuses on the two highest scale scores that exceed a T of 70 (or 65 on the MMPI-2) (e.g., Dahlstrom, Welsh, & Dahlstrom, 1972; Greene, 1991). In one example with 20 keyboarding errors (the most found within the sample), a 2-7 profile was altered to a 2-4 codetype. However, two-point codetype alterations can occur with just one keyboarding error. For instance, one client’s 2-9 codetype could have shifted to a 2-4 codetype with just one such error.2

The MMPI results described to this point only include human mechanical errors, such as keyboarding errors. Bear in mind that because each site could not readily identify the brand or exact scoring protocol employed in their scoring programs, we chose to eliminate a potential source of variance by recalculating scale scores and T scores using a single standard (the 1983 NCS scoring standard) for our analyses. Had we decided to include the scores produced by the settings’ scoring programs, analyses would reveal some striking unanticipated discrepancies that are worthy of further mention. After eliminating effects due to keyboarding errors, the T-score profiles generated with the 1983 NCS scoring standard compared against those based on scoring programs from each setting would have produced discrepant profiles in almost all (96%) of the cases sampled. Further analyses uncovered three sources for these discrepancies: (a) some subscales used different items from those used in the 1983 NCS scoring standard, (b) programs that used T-score lookup tables based on Dahlstrom et al. (1972) inherited corresponding rounding errors, and (c) T-score ceiling values were found to differ from the 1983 NCS scoring standard used in this study.

**Discussion**

The answer to the primary question addressed in this study—whether scoring errors on objective personality tests merit concern—appears resoundingly affirmative. Errors were found in the scoring of all three tests at all three sites, with rates ranging from a low of 2% to a high of 56%! Given these rates and common interpretive practices, it is a virtual certainty that a number of these errors would have (or did) substantially alter test interpretation. Furthermore, considering the number of times objective personality tests are administered, the consistency of error across tests and settings in this study, and the extremely low likelihood that these results are idiosyncratic to the study settings, the potential problem of error in scoring objective personality tests hardly seems trivial. At the same time, recognition of this

---

2We recognize that with the MMPI the use of “defined code types” may mitigate the impact of isolated scoring errors. Many MMPI profiles do not achieve this criterion, however, and many MMPI users may not apply this type of analysis. Moreover, many tests other than the MMPI do not have such a potential safeguard.
potential for error can serve constructive purposes because these problems are almost entirely correctable, as the analysis of high commitment to accuracy suggests.

There is no claim or expectation here that the error rates found in this study are representative of particular settings nationally, and it was not our intent to produce such numbers. Rather, what can be said is that it is very unlikely, given the Allard et al. (1995) comparable results and the diversity of tests and settings in the present study, that scoring errors of the frequency and magnitude we uncovered are exclusive to the current sample. Thus, it is quite likely that other tests in other settings are also scored erroneously and that some rates of error in some settings are also alarming.

The exact impact of errors on clinical interpretation is difficult to determine. It seems likely that the gross errors that sometimes occur, which can result in dramatic shifts in a BDI score or an MMPI code type, could greatly affect clinical judgment. Although it may be argued that clinical decisions should not rest on isolated test scores, salient information, even that stemming from a single variable (e.g., a test score), can predominate judgments (Faust, 1984). Furthermore, in some settings, it is not unusual to limit psychometric evaluation to a single measure, such as the MMPI. Attempting to delineate the precise consequences of various types and magnitudes of scoring error on clinical decision making is a considerable undertaking, and we think time and effort might be better directed towards solving the problem of erroneous scoring, because it is relatively simple to alleviate.

A second objective of the study was to examine factors that prior research (e.g., Allard et al., 1995) suggests are related to scoring error – CTA and SPC. Analyses of both aggregated and disaggregated data demonstrated a strong positive relation between SPC and error rate. The STAI, which required two steps to derive summary scores, evidenced discernibly higher error rates than either the BDI or the keyboarded MMPI, which each required only one step to derive final scores. The strength of the association between SPC and error rate may have been stronger still had the range of complexity been extended beyond the two, somewhat truncated, levels examined in this study. In fact, Allard et al. (1995) included a test with more complex scoring than any of the procedures examined here and found very strong associations with error rate. In any case, both studies support the seemingly obvious notion that tests requiring more complicated scoring procedures are associated with increased frequency of scoring error. Considering the strength of this factor, it is of interest that so few prior studies on scoring error have measured or analyzed complexity explicitly.

CTA proved to be an even stronger factor than SPC in determining scoring error. Whereas all L-CTA samples revealed errors, the F-CTA sample revealed very few mistakes. The SPC factor, which was clearly influential in the L-CTA sample, had little or no impact in the F-CTA sample. Despite the clear advantage of F-CTA, errors did surface with the optical scanner. The NCS OpScan 5 used in this study was not flawless in distinguishing between erased items and marked ones, as two tests had one item misscored each.

As part of this study, various test protocols were double-checked, allowing an exploratory analysis of the effectiveness of double-checking in reducing errors. The accuracy of rescorers’ double-checking efforts was assessed by comparing MMPI raw test data to the corresponding keyboarded test data, both of which were double-checked. No discrepancies surfaced due to flaws in the double-entry process. This limited analysis is encouraging but far from conclusive, and the inefficiency of double-checking is an obvious drawback. However, for other types of test scoring methods, double-checking is presently the only alternative available, and its effectiveness merits further investigation.

Gross discrepancies, such as 21-point errors on the BDI or 27-point errors on the STAI discovered at some settings (one in particular), were likely the result of unorthodox scoring procedures and test forms or distracting working conditions (such as keyboarding in a receptionist area). Such errors can have serious consequences. For instance, misidentifying a patient as showing severe (versus mild or unremarkable) depression on the BDI
could result in prescribing unnecessary antidepressant medications or rendering unwarranted services, such as suicide prevention, electroconvulsive therapy, 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 negligent treatment. In either case, the morbidity that can flow from gross miscoring 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.

**Additional Sources of Error**

Although this study’s primary focus concerned risk for error related to manual scoring processes, analyses also uncovered additional sources of scoring error, particularly for the MMPI: computer program errors and lack of scoring standards. These problems stem from such factors as lack of agreement about items that compose certain subscales and about T-scores ceiling values for certain subscales, and rounding errors in T-score lookup tables published in primary sources (e.g., Dahlstrom et al., 1972). Just how much these discrepancies can influence interpretation is beyond the scope of this study. It is remarkable to note, however, that these sources of error altered MMPI profiles in 48 of 50 (96%) cases in one setting, as well as large percentages of profiles in the other settings. More remarkable is that discrepancies in MMPI computer scoring were first reported more than 30 years ago (Fowler & Coyle, 1968), yet lack of agreement persists to this day! It is unlikely that the scoring discrepancies found at all three sites are the only ones that exist in the population of MMPI scoring programs in current use. At the risk of repeating old concerns, findings of this magnitude still deserve immediate follow through.

**Recommendations for Test Users and Publishers**

Results of this study lead to recommendations or guides for test users and for test and software publishers, which will be presented in turn.

**Guides for Sites/Test Users**

Popular tests, as currently designed and scored, are generally vulnerable to (i.e., can easily be influenced by) even minor scoring errors. Sites/test users should therefore implement F-CTA scoring procedures to alleviate the potential for errors. Optical scanning is one option, although scoring errors can still occur occasionally with this technique, as demonstrated in this study. There are two sources of error, both of which are correctable by the user; the scanner scoring profiles may be incorrectly programmed, or the scanner may miscode eraser smudge marks created by the test taker. Errors due to eraser smudge marks can be largely prevented if scorers visually scan answer sheets to remove such marks before scanning. Double-checking hand-scoring processes would likely reduce error rates greatly, although further research is needed on this procedure and the potential for systematic error is not eliminated. For example, scorers may learn problematic scoring practices from one another or the same trainer, such as scoring the STAI using an unorthodox, although systematic, scoring procedure and neglecting to add 27 points to final scores, leading each scorer to repeat the same mistake. The ultimate method for achieving F-CTA would require eliminating processes between the test-taker and the scoring program. Administering the test on computer would accomplish this, provided computer scoring errors are not present within such a scoring program.

Sites/users that employ a single instrument, such as the MMPI alone (or any other instrument), should strongly consider employing F-CTA procedures for two reasons. When administered in isolation, tests are likely to be at greater risk for errors going undetected than tests that are administered as part of a battery. Other tests scores within the battery provide a backdrop for detecting a potentially discrepant test score. Secondly, errors in isolated test scores are at higher risk for adversely influencing interpretations given that clinical judgements can be predominated by a single test score.
Errors in Scoring Objective Tests

Sites may balk at implementing F-CTA procedures if the benefits are not carefully considered. Perceived time or budgetary constraints may be typical causes for such reluctance. Manually rescoring tests doubles the labor or time, 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 clerical staff labor savings alone in as little as 2 years (based on scoring 10 MMPI’s per week, for example), and even less where tests are administered in high volume.

Individual test users can (and of course often do) use commercial services, although the accuracy of such services should not be assumed and periodic hand-checks are needed. Based on our and others’ clinical experience, it is suspected that one widely used commercial service often miscores MMPIs, and another uses idiosyncratic norms. Discrepancies can be double-checked to determine the source of error and the reliability of the commercial service.

Sites can perform a fairly straightforward audit of their scoring practices to determine risk for gross errors. Where some combination of non-standard test forms, non-standard scoring practices, and frequent interruptions are present, large magnitude errors are likely to occur. These additional handicaps, in effect, boost the scoring complexity of tests. Sites can train scorers about the importance of adhering to the scoring methods specified by the test manuals, as well as the pitfalls of using colloquial scoring practices, such as those uncovered in the site that produced tests with gross errors (e.g., 27-point errors on the STA1). Sites should also ensure that consistent scoring forms are used throughout, otherwise large magnitude errors can result (e.g., 21-point errors on the BDI). One way to ensure consistency in scoring procedures and forms is to insist on using the test publishers’ scoring procedures, scoring forms, and response sheets. Sites can also try to relocate scoring areas from noisy or distracting environments.

From a research perspective, identifying scoring practices that result in gross error should be the first priority, since the risk for misclassification and harm is substantial. Future studies might reclassify the L-CTA distinction as “low” and “partial” CTA, where “low” CTA represents scoring practices that could result in gross discrepancies.

**Recommendations for Test and Software Publishers**

If, among all recommendations presented herein, only one can be pursued, it should concern the clarification of scoring standards. Verifying scoring programs necessarily entails a standard for scoring and interpretation. Until this problem is resolved and tests such as the MMPI are scored with the same procedures, the validity of test results will remain in question, regardless of improvements made in the accuracy of the scoring process.

Verifying the accuracy of computer scoring programs imposes clinicians and settings with a frustrating onus. Outside of purchasing scoring programs that claim to be based on test manufacturers standards, verifying the accuracy of scoring programs is not easy. Programming bugs can either produce obvious or consistent scoring discrepancies or subtle and sporadic ones. The responsibility really belongs to the scoring program manufacturer, who should publish the method by which their scoring program was verified.

Test designers could also promulgate a new scoring interpretation process that 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 small errors. These scoring structures create vulnerabilities for cases where scale scores “sit near the fence,” especially when the pattern or interrelationship of scale scores (e.g., high two-point code) is the primary unit of analysis. It is therefore important to recognize 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 will usually reduce or mitigate the potential
effects of small errors. For example, test scores could be reported with certainty estimates, such as standard error of measurement (SEM). Using SEMs, small scoring errors may slightly alter presumed interpretive correlates or the probability of a finding, 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 SEMs, the problem of scoring errors on objective personality tests could be virtually eliminated.

References

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-Revise (PDQ-R). Professional Psychology: Research and Practice, 26, 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, 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, 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, 150-157.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemper, B. (1989). Minnesota Multiphasic Personality Inventory-2: Manual for administration and scoring. Minneapolis: University of Minnesota Press.

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. Minneapolis: University of Minnesota Press.

Faust, D. (1984). The limits of scientific reasoning. Minneapolis: 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: Allyn and Bacon.

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

Levenson, R. L., Jr., 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.

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., & Greder, 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 Psychologist, 50, 14-23.

Oakland, T., Lee, S. W., & Axelrod, K. M. (1975). Examiner differences on actual WISC protocols. Journal of School Psychology, 13, 227-239.

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

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, 74-82.

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

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

ScanTools Release 2.4A [Computer program]. (1990). Minneapolis, MN: National Computer Systems.

Sherrett, S., Gard, G., & Langner, H. (1979). Frequency of clerical errors on WISC protocols. Psychology in the Schools, 16, 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., & Hunnicutt, L. C., Jr., (1988). Examiner errors on the Wechsler scales. Journal of Psychoeducational Assessment, 6, 280-288.

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, 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.

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

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

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