The Question of Validity of Neuropsychological Test Scores Among Head-Injured Litigants: Development of a Dissimulation Index

Ralph M. Reitan and Deborah Wolfson
Reitan Neuropsychology Laboratory, Tucson, AZ

Two groups of 20 head-injured subjects were studied. One group was involved in litigation related to the head injury and the other group was not involved in litigation. Every subject had been tested twice, with the intertest mean interval for the two groups ranging from 12.00 to 14.45 months, a statistically insignificant difference. Two indexes of consistency of test–retest performance, which have been reported previously (Reitan & Wolfson, 1995, in press), were combined to form the Dissimulation Index. Comparison of the distributions for the two groups on the Dissimulation Index revealed no overlap; in fact, a gap of several points separated the groups. Every litigation subject performed with less consistency than any nonlitigation subject. Using an approach based on intra-individual test comparisons shows promise of alleviating the vexing problem of invalidity of neuropsychological test results.

The problem of malingering, dissimulation, or even nondeliberate production of invalid performances on neuropsychological tests cuts at the heart of the scientific basis of clinical practice in neuropsychology. The Diagnostic and Statistical Manual of Mental Disorders (DSM–IV) (American Psychiatric Association, 1994) cautions that malingering should be strongly suspected if an examination is done in a medicolegal context (e.g., the client was referred by an attorney for an evaluation). Binder (1990), in his review of forensic cases, stated that malingering should be considered a possibility whenever the test results may be related to an opportunity for financial gain.

The obvious need for a valid index of malingering or dissimulation has led to many studies, most of which used normal subjects instructed to pretend that they were brain-injured while taking various neuropsychological tests (see Nies & Sweet, 1994, for a review of this literature). Several investigations reported statistically significant differences between feigned malingerers and groups not suspected of malingering but two major problems plague these research efforts.
First, it is naive to assume that normal subjects instructed to fake the type of deficits that would result from a head injury would actually be able to simulate the neuropsychological impairment caused by brain damage. The subjects — who are usually normal, unimpaired, and relatively young — typically know very little about the behaviors they are instructed to simulate (e.g., neuropsychological deficits). In addition, normal subjects pretending to be brain-damaged share none of the stresses, anxieties, guilt, depression, and desperation experienced by many litigants whose future financial stability may depend on the outcome of the neuropsychological examination.

Many plaintiffs may not even realize that they are not putting forth their best possible effort when taking the tests. In such cases, the real-life situation is far different from the laboratory, and laboratory results are open to serious question. Because few subjects ever reveal that they have malingered, no one has assembled a representative group to perform even the first valid comparison of malingers and nonmalingers. How can the unique characteristics of malingers be identified if we cannot assemble an appropriate group to study?

The second major problem of current research studies is that the results reach, at best, only certain levels of statistical significance. The investigation of malingering requires methods that “diagnose” each subject correctly rather than methods that yield only a probability statement about the likelihood that intergroup differences are due to chance! Guesswork, based on probability, is not an adequate solution to the problem of identifying the individual subject who is malingering. As with neurosurgeons diagnosing brain tumors, the validity of the procedure narrows down to the individual case.

This problem might at first seem to be beyond solution, given the fact that intergroup overlap occurs in essentially all distributions of scaled psychological performances. The problem derives primarily from the wide range of performances represented by the normal probability distribution. A bright person who is malingering might perform better than a less able nonmalinger, and presumably 50% of nonmalingers fall below the median.

Nevertheless, neuropsychologists have devised a number of ingenious techniques and approaches that are helpful in some cases. Tests have been devised that are actually very easy but are presented to the subject as being very difficult (Binder, 1993; Lee, Loring, & Martin, 1992).

Some subjects appear to “fall for the bait” (i.e., believe that the test is difficult), perform well below a chance level, and thereby produce results suggesting that the subjects have not made a legitimate effort to produce the best performances of which they are capable. Clinical observations of some severely brain-damaged persons, however, indicate that significant neuropsychological impairment can also cause a subject to perform below chance levels.

Reitan and Wolfson (1993a) as well as others have demonstrated that subjects involved in litigation sometimes perform exceedingly poorly on some tests and do well on other tests, presenting combinations of scores that occur rarely, if ever, clinically. Pankratz (1988) and his associates developed symptom validity testing, a method of evaluating sensory and memory complaints that can be used to identify malingers. This technique uses a forced-choice paradigm in which a 50% rate of correct responses represents a chance performance. The chance probability of scoring below the 50% level may readily be calculated, indicating the likelihood that any particular poor level of performance would occur by chance.

Depending on the scores, each of these approaches may yield strong presumptive evidence of malingering in the case of certain subjects. There is no assurance, however, that all instances of invalid test results are detected.

Cullum, Heaton, and Grant (1991) used a method for detecting invalid results that they believed had previously been unstudied. They elected to use each of three subjects as his/her own control and compare results on repeated testings. These authors suggested that “examination of performance reliability across testings may be a powerful means by which
the neuropsychologist can detect patients who are not consistently putting forth adequate effort on the examination” (p. 168).

As demonstrated both in neurological (DeMyer, 1994) and neuropsychological (Reitan & Wolfson, 1993b) evaluations, using the subject as his/her own control has many advantages. This method is frequently used clinically to compare the subject’s performances on the same test on the two sides of the body. Many subjects, particularly those involved in litigation, are tested two or more times before the case comes to trial or is settled. Our hypothesis was that there would be a great advantage in comparing the subject’s test results obtained on two testings rather than comparing the subject’s performances with the rest of the population. If the subject performed to the best of his/her ability on both examinations, one would presume that the scores on the second testing would be at least comparable or possibly better (because of practice-effects). If the subject was not putting maximum effort, the scores on the two testings would be more variable, and perhaps even worse on the second testing, because of the need to “prove” one’s impairment as the time for a judgment came closer.

Reitan and Wolfson (in press) compared the test results of two groups of 20 adult head-injured subjects. One group was involved in litigation related to the head injury; the other group was not. The subjects in each group had been tested twice using the same instruments, with the interval between the testings not differing significantly. The tests used were the Comprehension, Picture Arrangement, and Digit Symbol subtests from the Wechsler Scale and the Category Test, Part B of the Trail Making Test, and Tactual Performance Test–Localization component from the Halstead–Reitan Battery. These tests were selected because prior research has generally shown that they are sensitive to cerebral damage and appropriate for evaluating subjects with a head injury (Reitan & Wolfson, 1986, 1988, 1994). Table 1 presents mean scores and SDs for both groups and Wechsler results used in an additional study by Reitan & Wolfson (1995).

The mean scores on the first testing for both groups are generally consistent with the possible effects of head injury and are analyzed in more detail in another publication (Reitan & Wolfson, in press). Because some subjects had been given the WAIS and other subjects had been given the WAIS–R, it was not possible to draw statistical inferences about the probability that intergroup mean differences for the Wechsler data were due to chance. In some instances, the same tests (Category Test, Part B of the Trail Making Test, and the Tactual Performance Test–Localization) had been given to each group on Testing I, and comparisons of the distributions yielded the t ratios and probability levels shown in Table 1. Note that on the first testing, the litigation group

<table>
<thead>
<tr>
<th>TABLE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Means and SDs on Testings I and II for Nonlitigation and Litigation Groups, and t Ratios Comparing Means on the First Test for Those Tests That Were Identical for the Two Groups</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>Inf</th>
<th>Comp</th>
<th>Arith</th>
<th>Simil</th>
<th>Vocab</th>
<th>PA</th>
<th>BD</th>
<th>Digit Symbol</th>
<th>Category</th>
<th>Trails B</th>
<th>TPT-Loc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonlitigation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing I M</td>
<td>10.20</td>
<td>10.45</td>
<td>9.10</td>
<td>11.60</td>
<td>10.10</td>
<td>9.00</td>
<td>11.40</td>
<td>8.75</td>
<td>45.95</td>
<td>109.00</td>
<td>3.25</td>
</tr>
<tr>
<td>SD</td>
<td>3.27</td>
<td>3.65</td>
<td>4.91</td>
<td>4.22</td>
<td>5.37</td>
<td>3.51</td>
<td>4.68</td>
<td>3.42</td>
<td>27.88</td>
<td>63.91</td>
<td>1.65</td>
</tr>
<tr>
<td>Testing II M</td>
<td>10.45</td>
<td>11.30</td>
<td>9.95</td>
<td>11.70</td>
<td>10.20</td>
<td>11.00</td>
<td>11.70</td>
<td>12.05</td>
<td>35.90</td>
<td>89.40</td>
<td>4.35</td>
</tr>
<tr>
<td>SD</td>
<td>3.27</td>
<td>3.80</td>
<td>4.48</td>
<td>4.66</td>
<td>3.93</td>
<td>3.77</td>
<td>3.60</td>
<td>3.50</td>
<td>27.45</td>
<td>53.76</td>
<td>2.52</td>
</tr>
<tr>
<td>t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Litigation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing I M</td>
<td>9.75</td>
<td>9.65</td>
<td>8.10</td>
<td>9.25</td>
<td>10.20</td>
<td>9.00</td>
<td>7.55</td>
<td>7.15</td>
<td>65.80</td>
<td>118.25</td>
<td>2.60</td>
</tr>
<tr>
<td>SD</td>
<td>2.07</td>
<td>2.70</td>
<td>2.77</td>
<td>2.84</td>
<td>2.42</td>
<td>2.38</td>
<td>2.16</td>
<td>2.25</td>
<td>27.13</td>
<td>64.55</td>
<td>1.90</td>
</tr>
<tr>
<td>Testing II M</td>
<td>8.85</td>
<td>8.70</td>
<td>7.65</td>
<td>8.35</td>
<td>9.20</td>
<td>7.50</td>
<td>7.15</td>
<td>5.20</td>
<td>80.20</td>
<td>153.50</td>
<td>1.75</td>
</tr>
<tr>
<td>SD</td>
<td>2.50</td>
<td>2.39</td>
<td>3.30</td>
<td>2.62</td>
<td>2.14</td>
<td>2.74</td>
<td>2.80</td>
<td>1.38</td>
<td>33.57</td>
<td>79.94</td>
<td>1.25</td>
</tr>
</tbody>
</table>
had a poorer mean score than the nonlitigation group on 9 of the 11 tests, even though the mean scores for the two groups were fairly similar on most tests. We have no way to determine whether this trend represents a lower ability level or suboptimal effort by the litigants. The fact that the litigation group performed more poorly, in absolute terms, on every test on the second as compared with the first testing would tend to implicate motivational factors.

Table 1 was presented here to show that neither group was grossly impaired. Further aspects of these results are presented in Reitan and Wolfson (in press). Because all analyses were based entirely on intra-individual difference scores obtained from the two testings, none of these intergroup comparisons are directly relevant to either the results of the present study or to the prior report (Reitan & Wolfson, in press). A 5-point scale was developed for converting raw-score differences to scaled scores. The litigation group showed more test–retest variability at highly significant levels on each test. A total score based on the 5-point scale (the Retest Consistency Index) also differentiated the groups at highly significant levels. The Retest Consistency Index separated 90% of the litigation group from 95% of the nonlitigation group using an optimal cutting point.

Reitan and Wolfson (1995) used these same groups of subjects to perform a response consistency evaluation in which specific responses were compared on the two testings. Tests that require rather specific and discrete responses to individual items were necessary to facilitate test–retest comparisons, and the Information, Comprehension, Arithmetic, Similarities, and Vocabulary subtests of the Wechsler Scale were selected. Intra-individual test scores were again used and no tests for intergroup mean differences for level of performance were performed. Data was generated by comparing responses on the two testings. If a subject’s response to an item earned the same score on both testings, no points were assigned to that item. An improved response on the second testing earned 1 point and a poorer response on the second testing received a score of 2 points. A raw-score distribution was generated for each group, and these raw scores were transformed into scaled scores using a 5-point scale. The mean scaled scores were less consistent for the litigation group than for the nonlitigation group, and the mean differences in scaled scores were highly significant on each test as well as for the total score (Response Consistency Index). The best cutoff point correctly classified 100% of the nonlitigants and 90% of the litigants.

The present study reports results using the Dissimulation Index, a summarical combination of the Retest Consistency Index and the Response Consistency Index.

METHOD

Subjects

In this study, we used the same two groups, each composed of 20 head-injured subjects, that had been used in a previous study (Reitan & Wolfson, 1995). The nonlitigation group was composed of 17 men and 3 women who had sustained a head injury but who had no litigation pending or contemplated. These subjects had volunteered to participate in a research project. The litigation group consisted of 14 men and 6 women, all of whom were involved in litigation to recover damages allegedly resulting from a head injury.

The nonlitigation group had a mean age of 29.65 years ($SD = 14.91$) and a mean education of 12.30 years ($SD = 2.23$). The litigation group had a mean age of 41.85 years ($SD = 10.19$) and a mean education of 13.95 years ($SD = 2.84$). The groups differed significantly on these variables ($p < 0.01$ for age and $p = 0.05$ for education), but because only intra-individual difference scores were used in the data analyses, these differences may not have had much significance.
It would be useful to have information about specific aspects of the subjects’ head injuries, such as loss of consciousness, duration of coma, post-traumatic amnesia, results of clinical neurological examinations, and evidence of brain tissue damage as revealed by specialized neurodiagnostic tests. This kind of information was being disputed in every case involving litigation. However, we did confirm that no subject in either group had sustained a penetrating head injury or a depressed skull fracture, and none of the subjects had had surgical intervention for the closed head injury.

The mean time between the first and second neuropsychological examination was 12.00 months for the nonlitigation group and 14.45 months for the litigation group. The intergroup difference was not statistically significant (p < 0.20).

Procedure

Several constraints were significant in determining the procedure used in this study. First, the test data on many of the litigants was variable, and subject selection was determined in part by the need to have data on the same tests across all cases on two testings. Second, it was desirable for the two testings to have been given within each group at comparable intervals and cases were selected accordingly. As noted, we also wished to include brain-sensitive tests as a basis for the Retest Consistency Index and tests that required specific responses for the Response Consistency Index.

The Dissimulation Index represented the sum of the Retest Consistency Index score (Reitan & Wolfson, in press) and the Response Consistency Index score (Reitan & Wolfson, 1995). The Retest Consistency Index was computed by obtaining difference scores, based on two testings of each subject, on the Comprehension, Picture Arrangement, and Digit Symbol subtests of the Wechsler Scales and the Category Test, Part B of the Trail Making Test, and the Localization component of the Tactual Performance Test. Differences among achievement and error scores were treated so that the difference scores consistently reflected poorer versus better performances. As noted, these difference-score distributions for each test were converted to a 5-point scaled-score distribution with a score of 5 representing the 20% of the scores that were most variable and a score of 1 representing the 20% that were least variable.

The Response Consistency Index compared responses on two testings for individual items on the Information, Comprehension, Arithmetic, Similarities, and Vocabulary subtests of the Wechsler Scale. Raw scores were generated for each subject on each test and a total score was determined, with larger scores reflecting greater variability on the two testings. These distributions were converted to a 5-point scale, with scores of 5 representing the least consistent performances on the two testings.

The two groups were given exactly the same tests from the Halstead–Reitan Battery on the two testings. However, some subjects had been given the WAIS on both testings and other subjects had been given the WAIS–R on both examinations. In the nonlitigation group, 18 subjects were given the WAIS and 2 subjects were given the WAIS–R. In the litigation group, 13 subjects had taken the WAIS on both testings and 7 subjects had been given the WAIS–R. A chi-square test of these frequency differences, applied with Yates’ correction for continuity because of small expected frequencies in two of the four cells, yielded a value of 2.30 (p < 0.20).

The Dissimulation Index, obtained by adding the scores for the Retest Consistency Index and the Response Consistency Index, represented the sum of 11 scores, with each score ranging from 1 to 5. For example, a subject who was inconsistent on each test of the two testings could receive a maximal score of 55. A subject who performed consistently could obtain a score of 11 (the most consistent score possible).
The Dissimulation Index was computed for each of the 40 subjects. Means and SDs were determined for each group and t ratios were computed to compare the distributions. The degree of overlap of the distributions was used to determine the best cutoff scores.

RESULTS

Table 2 presents means, SDs, and data representing inferential comparisons of the litigation and nonlitigation groups on the Retest Consistency Index, the Response Consistency Index, and the Dissimulation Index. It is apparent that the groups were significantly differentiated on each of the three variables.

The mean Dissimulation Index was 42.05 (SD = 6.14) for the litigation group and 23.50 (SD = 3.00) for the nonlitigation group, yielding a t ratio of 12.13 (p < 0.001). These results indicate that inconsistencies between the first and second testings were far more likely to be demonstrated by litigants than by nonlitigants.

The distributions of the two groups showed that Dissimulation Indexes ranged from 33 to 55 points for the litigants and from 18 to 29 points for the nonlitigants. The frequency distributions are presented in Table 3.

Dissimulation Indexes for the two groups did not overlap; in fact, none of the 40 subjects received scores of 30, 31, or 32, thus revealing a gap between the two distributions. On the basis of these samples, we judged that the best cutoff point would be 31/32.

DISCUSSION

The results of this study demonstrated that retest performances were far less consistent in litigants than in nonlitigants. As noted, there was a gap in the scores between the two distributions of Dissimulation Indexes. A cutoff point that separated the groups perfectly was identified, although this study, as with any first report, requires cross-validation. Also note that specific characteristics of litigants or factors that might influence the test-taking behavior of litigants may eventually be identified. In addition, the method of test–retest consistency should be studied among other groups in whom the possibility of malingering may be likely.

Implementation of the suggestion by Cullum, Heaton, and Grant (1991) to study the consistency of test performances, rather than depend on level of performance on a single examination, clearly appears to be worthwhile. The results of this study strongly suggest that using the subject as his/her own control has significant advantages over interpreting the subject’s score with relation to the rest of the population.

TABLE 2

Summary of Means, SDs, t Ratios, and Probability Levels for Head-Injured Subjects Involved in Litigation Versus Subjects Not Involved in Litigation on the Retest Consistency Index, the Response Consistency Index, and the Dissimulation Index

<table>
<thead>
<tr>
<th>Group</th>
<th>Retest Consistency Index</th>
<th>Response Consistency Index</th>
<th>Dissimulation Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonlitigation</td>
<td>M 13.10</td>
<td>10.40</td>
<td>23.50</td>
</tr>
<tr>
<td></td>
<td>SD 2.67</td>
<td>2.62</td>
<td>3.00</td>
</tr>
<tr>
<td>Litigation</td>
<td>M 22.50</td>
<td>19.55</td>
<td>42.05</td>
</tr>
<tr>
<td></td>
<td>SD 4.12</td>
<td>2.82</td>
<td>6.14</td>
</tr>
<tr>
<td></td>
<td>t 8.55</td>
<td>10.64</td>
<td>12.13</td>
</tr>
<tr>
<td></td>
<td>p &lt; 0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>
TABLE 3
Distributions of Dissimulation Indexes for Two Head-Injured Groups, One Involved in Litigation and the Other Not Involved in Litigation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonlitigation group</td>
<td>4</td>
<td>11</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Litigation group</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

The approach used to develop the Dissimulation Index has the advantage of bearing no direct relation to the subject's level of performance. Of course, if a subject scored 0 on every item, there would be no evidence of inconsistency. Consistently poor performances might also lower the Dissimulation Index. However, subjects in litigation generally earn fairly adequate scores, with means that are not consistently significantly lower than the means earned by subjects not in litigation (Reitan & Wolfson, in press), even though in this study the Dissimulation Index perfectly differentiated the litigants from the nonlitigants.

Note that this study was not directly concerned with malingering; it compared persons involved in litigation with persons not involved in litigation. We knew of no reason why the nonlitigation subjects would not put forth their best effort when taking the tests.

It is possible, though, that subjects involved in litigation might deliberately, or because of the stress and pressure of the existing circumstances, fail to do their best on a second examination. In any case, the results of this study suggest that these two groups, in terms of the Dissimulation Index, represent quite different samples and raise the probability of eventually identifying test results that are not completely valid.

Limitations of this study principally center around the small sample size and the resulting question of whether the samples are representative of the populations from which they were drawn. One would certainly expect that some subjects involved in litigation would fall in the score range of nonlitigants, even though this did not occur in the present study. We should also emphasize the fact that, in this study, all the nonlitigants were examined under carefully controlled testing conditions in one laboratory whereas the litigants were evaluated by practicing neuropsychologists, some hired by the plaintiff and some hired by the defense, and may be a more typical representation of a neuropsychologist's clinical practice.

Finally, we should point out that because the characteristics and/or behavior of litigation samples may vary, future investigations may not obtain the same results as this study. None of the subjects in our litigation group were aware that comparisons might be made between the results of Testing I and Testing II and that they may be identified as litigants if they did not consistently put forth their best effort during the testing. There may well have been examiner differences (see Reitan & Wolfson, in press) in testing and in motivating subjects in litigation. If future litigants became aware that they must consistently do their best on every testing, it is likely that their scores on the Dissimulation Index will tend to fall in the range occupied exclusively in this study by the nonlitigants. If such an effect were actually to occur, the validity of psychological testing in the legal context would almost certainly improve.

REFERENCES


