The Effects of Coaching on the Sensitivity and Specificity of Malingering Measures

Julie A. Suhr and John Gunstad

Ohio University

This study sought to identify patterns of performance indicative of malingering on the Auditory Verbal Learning Test (AVLT). Participants were randomly assigned to perform normally, simulate head injury, or simulate head injury with warning that there might be attempts to detect malingering. Participants completed an expanded AVLT and a forced-choice task, in addition to several other memory tests. The warned simulators performed worse than normals on the forced-choice task, but better than those simulating head injury without a warning, suggesting that the warned subjects recognized forced choice as a malingering test. A combination of AVLT indices was able to predict group status for both naïve and warned malingerers (73.6% for naïve malingerers, 84.8% for warned, no false positives). The forced-choice measure detected only 31.6% of the naïve malingerers when specificity was maximized, and detected only 6.5% of the warned malingerers, a significant drop in detection rate. Findings suggest that pattern of performance indices are useful in detecting malingering, even when subjects are aware of attempts to detect malingering. © 2000 National Academy of Neuropsychology. Published by Elsevier Science Ltd

Keywords: head injury, malingering detection

Memory impairment is a common and well-known symptom of brain injury. Eighty-two percent of the general public are aware that a concussion can result in memory impairment (Gouvier, Prestholdt, & Warner, 1988). Further, memory impairment is not specific to head injury. Claims of memory impairment are common among those with neurological injury, but are also common among those feigning neurological injury, personal injury litigation, and other nonneurological patients (Brandt, 1988). In particular, individuals who attempt to malinger head trauma symptoms often report a variety of memory problems (Mittenberg, Azrin, Millsaps, & Heilbronner, 1993) and perform poorly on memory tests. Thus, in any clinical situation, it is important for neuropsychologists to estimate the likelihood of non-brain injury factors (such as poor motivation or lack of effort) affecting cognitive performance.

A widely used approach to detection of dissimulation is the symptom validity technique, in which each item has a 50% probability of obtaining a correct response without...
any knowledge of the correct response. In theory, given enough items, a person scoring below chance is most likely malingering. A popular version of this technique is the Portland Digit Recognition Test (PDRT; Binder, 1993; Binder & Willis, 1991). Studies using the PDRT have demonstrated that brain-damaged individuals get about 78–85% correct, while malingerers average about 56–76% correct (Binder, 1993; Binder & Willis, 1991; Greiffenstein, Baker, & Gola, 1994; Rose, Hall, & Szalda-Petree, 1998).

However, there are some drawbacks to this technique. An effective forced-choice procedure must not only look difficult, but must have numerous items, making the test monotonous and time-consuming. A person who realizes that the task is easy may then become annoyed with the lengthy task, stop attending, and perform poorly (Lezak, 1995). On the other hand, a distressed or depressed person may become overwhelmed by the lengthy and seemingly difficult task and perform poorly. Thus, poor performance may not reflect conscious malingering, but other factors as well. A further limitation is that the task requires that the participant believe it is a difficult task. To the extent that the malingerer does not realize that the task is quite easy, he or she will perform more poorly than those with true brain injury. However, a well-informed or coached patient may recognize the attempt at malingering detection and not perform poorly on the task. In this context, it is important to note that 47–48% of layers believe they should provide specific information about psychological tests to their clients, including information about validity measures (Wetter & Corrigan, 1995).

Is it possible to detect malingering without adding length to an already lengthy neuropsychological evaluation? Early approaches to the assessment of malingering focused on the use of standardized neuropsychological measures. It is well-known that confirmed and suspected malingerers tend to perform worse on standard attention and memory tests than do normal controls and individuals with brain damage (Hayward, Hall, Hunt, & Zubrick, 1987; Heaton, Smith, Lehman, & Vogt, 1978; Iverson & Franzen, 1996; Mittenberg et al., 1993). However, some studies have demonstrated that malingerers cannot always be distinguished from those with actual brain injuries using the cutoffs of standard tests (Goebel, 1983; Leininger, Gramling, Farrell, Kreutzer, & Peck, 1990). While cutoff scores for standardized tests can be generated to distinguish malingered performance from normal effort, it is more difficult to find cutoff scores that give adequate sensitivity to malingering, while remaining specific to malingering (i.e., not accidentally labeling those with true brain injury as malingerers).

However, it may be possible to develop more complex indicators within existing neuropsychological instruments. For example, researchers have detected particular patterns of performance within and among neuropsychological tasks that are able to discriminate malingerers from those with true neurologic damage (Barrash, Suhr, & Manzel, 1998; Bernard, 1991; Bernard, Houston, & Natoli, 1993; Mittenberg, Rothold, Russell, & Heilbronner, 1996; Suhr, Tranel, Wefel, & Barrash, 1997). For example, while it may be relatively easy to lower the number of words one recalls on a memory test below a certain level, faking a convincing pattern of impaired test performance may prove more difficult. Furthermore, it would be more difficult for a sophisticated or coached malingerer to keep track of and alter patterns of cognitive performance in a way that is believable. For these reasons, indices based on patterns of performance may be less vulnerable to the effects of coaching.

Bernard (1991) demonstrated that malingerers do not demonstrate the serial position effect seen on list learning tasks (those with normal memory, closed head injury, and dementia remember items better from the beginning and end of a list relative to those in the middle) (Bigler, Rozca, Schutz, & Hall, 1989). Undergraduates asked to mangle on a word list did not show the serial position effect, and appeared to suppress recall of words from the beginning of the list. This serial position effect was not replicated in a
follow-up study (Bernard et al., 1993); however, Suhr et al. (1997) demonstrated that pa-
tients who are strongly suspected of malingering do not show a normal serial position ef-
fect. Probable malingerers in their sample suppressed both the beginning and middle por-
tions of the list compared to those with head injury and those with psychiatric disorders.
Mittenberg et al. (1993) found that head trauma patients do better on attention tasks rel-
ative to memory tasks, while malingerers do worse on attention relative to their perfor-
ance on memory. A difference score based on the Wechsler Memory Scale General
Memory versus Attention/Concentration subscales accurately distinguished patients
with head injury from controls asked to simulate head injury.

Suhr et al. (1997) found that a combination of cutoff scores on various memory indi-
ces (very low scores on Auditory Verbal Learning Test [AVLT] recognition, very low
Benton Visual Retention Test scores, very low Digit Span scores, and failure at 30-
minute recognition to recognize words that had been recalled at least three times during
the learning trials on the Auditory Verbal Learning Test) was sensitive and specific to
malingering in a clinical sample. Having at least one of these four indicators detected
38% of the probable malingerers, with a specificity of 85% or above for all control
groups (mild head injury, moderate to severe head injury, depression, and somatiza-
tion). Having at least two of these indicators detected 25% of the probable malingerers,
with 100% specificity. The use of multiple cutoff scores on standardized tests increased
sensitivity to a great degree, without harming specificity. The final malingering indicator
was not a standard cutoff score, but a more subtle indicator of malingering. This indica-
tor (failure at 30 minutes to recognize words that had been recalled at least 3 times dur-
ing the learning trials of the AVLT) was developed based on observations that patients
with brain damage usually recognized words that they had been able to freely recall dur-
ing most learning trials, while malingerers often showed no recall or recognition of these
words despite having recalled them during most learning trials.

An interesting approach to the detection of malingering was taken by Barrash et al.
(1998), who used standard neuropsychological measures and clinical samples. Patients
who were referred for evaluation of memory complaints but who were judged to be ma-
lingering were compared to patients with brain damage from various etiologies and to
patients with psychiatric disorders. It was hypothesized that malingerers would be more
likely to demonstrate inconsistencies in their memory over time, relative to those with
brain injury. A combination of performance errors on the AVLT (worse recall at 60-
minute than 30-minute delay, worse recognition at 60-minute than 30-minute, and fail-
ure to recognize words at 60 minutes that had been recalled at 30 minutes) identified
48% of a sample of malingerers, with false-positive rates ranging from 2.5–4% (brain-
damaged and psychiatric patients, respectively).

One way to look at patterns of performance is to use discriminant function analysis to
determine combinations of test performance among multiple tests that are best able to
classify malingerers and nonmalingerers, a method strongly recommended by Rogers,
a pattern of performance on the Wechsler Memory Scale Figural Memory and Visual
Reproduction subscales that accurately classified 74% of simulated malingerers, and a
pattern of performance on AVLT Trial 1, AVLT 30-minute recognition and Complex
Figure Test recall that accurately identified 77% of malingerers. Reliability of these pat-
terns was supported in a follow-up study of simulated malingering (Bernard et al., 1993);
the Wechsler Memory Scale pattern classified 88% of the malingerers with no false pos-
itives, and the AVLT/Complex Figure Test pattern classified 86% of the malingerers
with no false positives. However, these discriminant function analyses require empirical
examination within a clinical population. The AVLT/Complex Figure Test pattern is
consistent with other observations (Wiggins & Brandt, 1988) that malingerers perform worse on recognition tasks than those with true brain injury.

Using a simulated malingering paradigm, the present study compared several AVLT-related malingering indices to a standard forced-choice procedure in the detection of malingering. We hypothesized that warning or coaching could influence the classification accuracy of a more obvious malingering detection device (such as forced choice), but would not be detrimental to the sensitivity and specificity of more subtle and complex malingering indices. We tested the ability of these tests to detect malingering in two situations: the relatively naïve malingerer, and the malingerer who is warned that attempts to detect malingering would occur during testing. We predicted that individuals asked to simulate head injury would do worse on all malingering indices than individuals asked to perform with full effort. We further hypothesized that warning about malingering detection would adversely affect obvious malingering indices (such as forced choice) but not more subtle indices (e.g., AVLT indices).

**METHOD**

**Participants**

Participants were 108 undergraduates randomly assigned to one of three groups: Effort ($n = 33$), Naïve Simulators ($n = 40$), and Simulate Head Injury with Information and Warning (hereafter called Warned) ($n = 35$). The Naïve group consisted of 20 subjects who were completely naïve about head injury and 20 subjects who received some information about head injury. Preliminary analyses found no differences between these groups on demographic or memory measures, and so they were combined in all later analyses. Two subjects (one from Naïve and one from the Warned group) were excluded from the study based on self-report of lack of compliance with experimental instructions (see below). Another two subjects (one from Naïve and one from Warned) did not complete one of the two tasks of interest, and so were excluded from analyses. Thus, the final sample sizes were Effort, $n = 33$, Naïve $n = 38$; and Warned $n = 35$. The three groups were not different in age, education, handedness, or gender distribution (see Table 1).

**Procedure**

Each participant received an envelope with instructions prior to beginning the testing. The Effort group was told to perform their best. The Naïve group was given a scenario about an accident in which they received a head injury, and they were asked to act as though they had deficits resulting from that accident (see Appendix). Fifty percent of these subjects were also given very nonspecific information about head injury symptoms. The Warned group read the same scenario, but also received a nonspecific warning that

<table>
<thead>
<tr>
<th>Demographic Information for the Three Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic Variable</td>
</tr>
<tr>
<td>Mean age (SD)</td>
</tr>
<tr>
<td>Mean years education (SD)</td>
</tr>
<tr>
<td>% Right-handed</td>
</tr>
<tr>
<td>% Female</td>
</tr>
</tbody>
</table>
malingering detection may occur during testing (see Appendix). All three groups were instructed not to reveal the details of their instructions to their examiner.

Trained examiners (who were unaware of group status) then administered a neuropsychological battery to each participant. The battery included an expanded AVLT (Barrash et al., 1998; Rey, 1964), the PDRT (Binder, 1993), the Complex Figure Test (Rey, 1941), Warrington Recognition Memory Test (Warrington, 1984), Benton Visual Retention Test (Sivan, 1992), and the Digit Span subtest from the Wechsler Adult Intelligence Test-III (Wechsler, 1997). The expanded AVLT includes a 60-minute delayed recall and recognition to create the Barrash et al. (1998) malingering indices.

After completion of the neuropsychological tests, the participants were asked to complete a form to judge recall of and compliance with the instructions they were given at the beginning of the experiment. All participants recalled their instructions accurately. Two subjects reported lack of compliance with the instructions (see above) and were excluded from further analyses.

**Malingering Measures**

**Forced choice.** We used a computerized version of the PDRT (Binder, 1993; Binder & Willis, 1991) consisting of 72 trials in which a five-digit number is seen on a computer screen, one digit at a time (as per Rose et al., 1995, 1998). Thus the modification was the presentation of the number sequence visually, rather than hearing the numbers spoken aloud. After either a 5, 15, or 30-second delay, the five-digit number appears in its entirety, together with a different five-digit number serving as a foil. The participant must watch the numbers appear, count backward during the delay, and then choose the five-digit number that was presented. The score utilized in the present study was the total percent correct on all three trials (5, 15, and 30-second delays).

**Attention versus memory.** Prior research (Mittenberg et al., 1993) suggests that a difference between basic attention and memory might distinguish malingerers from true brain injury. To test this hypothesis, we derived an Attention versus Memory Index composed of Digit Span Forward span length minus the number of words recalled on AVLT learning trial 1.

**Learning versus recognition.** As per Suhr et al. (1997), we tested whether failure at 30 min recognition to recognize words learned at least three times in the learning trials of the AVLT would effectively discriminate the groups. The score utilized was the total number of words meeting this criterion.

**Memory inconsistencies.** We tested the three indices identified by Barrash et al. (1998) and described above (30–60-minute recall, 30–60-minute recognition, and failure to recognize words at 60 minutes that had been recalled at 30 minutes).

**Recognition versus recall.** Given that differences in recognition versus recall performance have been described in malingering groups (Bernard et al., 1993; Wiggins & Brandt, 1988), we derived a Recognition vs. Recall Index composed of the percent of words recognized during the 30-minute delayed recognition of the AVLT minus the percent of words recalled during the 30-minute delayed recall of the AVLT.

**Learning Span.** The final index, Learning Span, was derived from the number of words recalled on Trial 5 of the AVLT minus the number of words recalled on Trial 1. This in-
dex was based on the idea that persons who malinger may not show as much learning over trials relative to those with true brain injury. In a sample of 20 mild to moderately head-injured patients, it was shown that one could expect an average of 5.8 words learned from Trial 1 to Trial 5 ($SD = 2.6$), while a group of 31 patients identified as probable malingerers had an average of only four words learned from Trial 1 to Trial 5 ($SD = 2.7$), a significant difference at $p < .05$ (Suhr, unpublished data).

**RESULTS**

**Group Differences**

Means and standard deviations of each group on the malingering indices are presented in Table 2. Analysis of variance (ANOVA) revealed that several of the potential discriminating variables were reliably different among groups, including Forced Choice, Learning vs. Recognition, 30–60-minute Recall, Recognition vs. Recall, and Learning Span. Attention vs. Memory and the other Barrash et al. (1998) indices did not significantly distinguish between groups and were not included in further analyses.

Posttests (Bonferroni corrected) revealed that both malingering groups performed significantly worse than the Effort group on Learning vs. Recognition, Recognition vs. Recall, and Learning Span. For the most part, the warning did not alter performance on the AVLT indices. An exception was 30–60-minute Recall, where the Naïve group performed worse than the Effort group, but there was no difference between the Warned group and the Effort group. For Forced Choice, both malingering groups performed significantly worse than the Effort group, but the Naïve group also performed significantly worse than the Warned group. Thus, the general warning about malingering detection did alter performance on this index.

**Creation of the Performance Pattern Index**

We decided to combine the pattern of performance indices (Learning vs. Recognition, Recognition vs. Recall, Learning Span, and 30–60-minute Recall) to create a Performance Pattern Index. As a first step, we confirmed that the four individual pattern of performance indices were not highly correlated, to confirm that each would be offering

<table>
<thead>
<tr>
<th>Variable</th>
<th>Effort</th>
<th>Simulate Head Injury</th>
<th>Simulate Head Injury-Warned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forced Choice**</td>
<td>94.0 (9.9)</td>
<td>62.1 (17.5)</td>
<td>77.9 (13.7)</td>
</tr>
<tr>
<td>Attention vs. Memory</td>
<td>−0.3 (1.8)</td>
<td>−1.0 (2.4)</td>
<td>−0.2 (1.6)</td>
</tr>
<tr>
<td>Learning vs. Recognition**</td>
<td>0.3 (0.6)</td>
<td>1.4 (1.2)</td>
<td>1.1 (1.1)</td>
</tr>
<tr>
<td>Memory Inconsistencies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVLT 30 – 60-minute recall*</td>
<td>0.0 (0.3)</td>
<td>0.6 (0.9)</td>
<td>0.4 (0.7)</td>
</tr>
<tr>
<td>AVLT 30 – 60 minute recognition</td>
<td>−0.6 (1.5)</td>
<td>−0.5 (1.2)</td>
<td>0.2 (1.5)</td>
</tr>
<tr>
<td>No. of words not recognized at</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60-minutes and recalled at 30-minute delay on AVLT</td>
<td>−0.6 (1.20)</td>
<td>−0.5 (1.2)</td>
<td>0.0 (1.4)</td>
</tr>
<tr>
<td>Recognition vs. Recall**</td>
<td>0.3 (0.6)</td>
<td>1.4 (1.2)</td>
<td>1.1 (1.1)</td>
</tr>
<tr>
<td>Learning Span**</td>
<td>6.4 (1.9)</td>
<td>4.2 (3.1)</td>
<td>3.9 (2.7)</td>
</tr>
</tbody>
</table>

AVLT = Auditory Verbal Learning Test.

*p < .005.

**p < .001.
independent information. Table 3 shows the correlations among the indices used in the study. The highest correlations were between Forced Choice and each of the other indices. The correlations among the pattern of performance indices were not very high (although some were statistically significant), suggesting that there should not be problems with multicollinearity when combining these indices.

We then chose cutoffs for each qualitative index in order to maximize specificity for each (i.e., there would be no false positives among the normal Effort controls). The cutoffs were: Learning vs. Recognition score greater than 1, Recognition vs. Recall greater than 47, Learning Span less than 3, and 30–60-minute Recall greater than 0. Each of these was then coded as a categorical variable (i.e., above or below the cutoff) for each subject. We then compared the malingering detection rates for the Naïve group to the detection rates for the Warned group. In the Naïve group, having at least one of the four indicators detected 73.6% of the malingerers. In the Warned group, only 6.5% of the malingerers were detected. The difference between detection of the Naïve and Warned malingerers was significant ($\chi^2 = 7.27, p < .01$).

Performance of the PDRT

To determine how well forced choice could distinguish among the groups, we again chose a cutoff score that would maximize specificity for the total forced-choice score (53%). We then compared the malingering detection rates for the Naïve group to the detection rates for the Warned group. In the Naïve group, this cutoff detected 31.6% of the malingerers. In the Warned group, only 6.5% of the malingerers were detected. The difference between detection of the Naïve and Warned malingerers was significant ($\chi^2 = 7.27, p < .01$).

**DISCUSSION**

Results suggest that patterns of performance on the AVLT were useful in distinguishing potential malingerers from controls. Warnings about the presence of malingering detection during testing lowered the sensitivity of the forced-choice procedure in detecting malingerers among the Warned group. However, the pattern of performance variables remained good predictors of malingering even in the presence of warning.

The Warned group did significantly better than the Naïve group on overall perfor-
mance on the forced choice procedure (see Table 2), which likely explains its drop in sensitivity to malingering following a warning. However, the warning about malingering used in the study was very generic (see Appendix) and should not have led participants to believe any specific test was a malingering detection test. Table 4 shows the results of $t$ tests between the Naïve and Warned groups on various memory measures used in the study. Although many of the differences were not different, they were all in the expected direction (i.e., the Warned participants did not perform as poorly as the Naïve participants). Thus, it appears that warning caused individuals to perform in a less impaired way on obvious memory indicators, but did not alter their general pattern of performance on memory tests, allowing the AVLT indices to remain sensitive to malingering.

It is important to note that the students simulating head injury were provided with no instructions regarding how malingering tests work, or how to avoid detection. Further, they were provided with only generic information regarding the effects of head injury (see Appendix). Thus, the degree to which they were “coached” to perform in a believable manner was minimal. It could be hypothesized that, given more information, the more “sophisticated” malingerer would be even more likely to escape detection by well-known and more face-valid malingering detection devices. This suggests a need for more subtle and complex malingering detection strategies, such as those identified in the present study.

The present findings are limited by the use of a simulated malingering design. It is unknown whether the identified malingering indices would be sensitive to exaggeration in a sample of suspected malingerers, or whether they would inappropriately classify patients with true brain injury as malingerers. The present results need to be replicated, and future studies should include patients with brain damage as a comparison group. However, the findings do suggest that the use of patterns of performance on memory tests can be useful in detecting malingering. We caution that the patterns of performance identified in the present study be used only in combination with other important clinical information (e.g., medical/legal records, clinical interviews, and other well-established malingering measures), to make a well-informed decision about the likelihood of malingering in any given case.

### Table 4
Comparisons Between the Simulate Head Injury and Warned Groups on Memory Variables

<table>
<thead>
<tr>
<th>Test</th>
<th>Simulate Head Injury</th>
<th>Warned</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digits Forward (raw score)</td>
<td>4.8 (1.8)</td>
<td>5.1 (1.7)</td>
<td>-0.64</td>
<td>ns</td>
</tr>
<tr>
<td>Digits Backward (raw score)</td>
<td>3.5 (1.3)</td>
<td>4.0 (1.2)</td>
<td>-1.77</td>
<td>&lt;.10</td>
</tr>
<tr>
<td>AVLT 30-minute recall</td>
<td>5.7 (2.3)</td>
<td>6.2 (4.0)</td>
<td>-0.68</td>
<td>ns</td>
</tr>
<tr>
<td>Recognition Memory Test Words (raw score)</td>
<td>36.7 (8.3)</td>
<td>40.5 (8.5)</td>
<td>-1.96</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Recognition Memory Test Faces (raw score)</td>
<td>34.2 (6.2)</td>
<td>36.8 (7.0)</td>
<td>-1.67</td>
<td>&lt;.10</td>
</tr>
<tr>
<td>Visual Retention Test Number correct</td>
<td>4.4 (2.2)</td>
<td>4.7 (2.8)</td>
<td>-0.59</td>
<td>ns</td>
</tr>
<tr>
<td>Visual Retention Test Number errors</td>
<td>9.5 (6.4)</td>
<td>8.2 (5.6)</td>
<td>0.9</td>
<td>ns</td>
</tr>
</tbody>
</table>

AVLT = Auditory Verbal Learning Test.
REFERENCES


APPENDIX

Scenarios Used in the Study

Simulate Head Injury—Not Warned

Imagine that you were in a car accident in which another driver hit your car. You were knocked unconscious, and woke up in the hospital. You were kept overnight for observation. The doctors told you that you experienced a concussion. Try to imagine that a year after the accident, you are involved in a lawsuit against the driver of the other car. If you are found to have experienced significant injuries as a result of the accident, you are likely to receive a bigger settlement. You have decided to fake or exaggerate symptoms of a brain injury in order to increase the settlement you will receive. As a part of the lawsuit, you are required to undergo cognitive testing to determine whether or not you have experienced a brain injury. If you can successfully convince the examiner that you have experienced significant brain damage, you are likely to get a better settlement. If the examiner detects that you are faking, you are likely to lose the lawsuit.

You are about to take a series of cognitive tests that would be used in such a situation. I would like you to simulate brain damage, but in a believable way, such that your examiner cannot tell that you are attempting to fake a brain injury.

[1/2 the sample also had the following additional instruction: Below is a list of common problems following brain injury, which may help you in your simulation of head injury: frequent headaches, being easily fatigued, problems with memory, difficulty attending and concentrating, slowed responses, irritability, anxiety, and depression.]

Simulate Head Injury—Warned

This group had the same instructions as above, in addition to the following instruction: At least one of the tests you will be given is designed to catch you faking, because it’s easier than it looks. Be careful.