Cross-validation of indicators of malingering
A comparison of nine neuropsychological tests, four tests of malingering, and behavioral observations☆

Tina Hanlon Inmana, David T.R. Berryb,*

aAIDS Neurological Center, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA
bDepartment of Psychology, University of Kentucky, 115 Kastle Hall, Lexington, KY 40506-0044, USA

Accepted 10 July 2000

Abstract

Few studies to date have cross-validated indicators of malingering that have been suggested on various neuropsychological tests. This study presents data cross-validating several indicators of malingering on neuropsychological tests, as well as on tests of malingering and via behavioral observations. It incorporates methodological recommendations by Rogers [Researching dissimulation. In: R. Rogers (Ed.), Clinical assessment of malingering and deception (pp. 309–327). New York: Guilford Press.] resulting in an ecologically valid design utilizing college students with a history of mild head injury as analog malingerers. Results indicated that the Letter Memory Test (LMT) and the Digit Memory Test (DMT) attained the highest hit rates for the detection of malingering, while the sensitivity of many other measures declined on cross-validation. © 2001 National Academy of Neuropsychology. Published by Elsevier Science Ltd.

Keywords: Malingering; Cross-validation; Letter memory test

The evaluation of malingering in neuropsychological assessment has become increasingly important in recent years with the growth of forensic evaluations. In many cases, neuropsychological test data provide the only potentially objective evidence of deficits. This is especially true in cases of mild head injury where neuroimaging tests are often negative and neurological signs are often absent. The ability of neuropsychologists to judge the motivation

☆ The study reported in this article was derived from the dissertation of Tina Hanlon Inman.
* Corresponding author. Tel.: +1-859-257-5451; fax: +1-859-323-1979.
E-mail addresses: theinmans@mindspring.com (T.H. Inman), dtrb@pop.uky.edu (D.T.R. Berry).
of patients during testing has been criticized in the past (Faust, Hart, & Guilmette, 1988; Faust, Hart, Guilmette, & Arkes, 1988; Heaton, Smith, Lehman, & Vogt, 1978), however, several promising indicators of malingering have been developed in recent years. Previously published indicators of malingering include behavioral rating scales, scores on standard neuropsychological tests, and tests designed specifically for the detection of malingering.

Research on the use of behavioral rating scales has been limited. Several researchers have suggested the use of behavioral rating scales for the determination of cooperation in neuropsychological assessment (Allen, Lewis, Wyman, & Coyne, 1989; Frederick, Sarfaty, Johnston, & Powel, 1994; Snow, Tierney, Zorzitto, Fisher, & Reid, 1990). The study by Frederick et al. (1994) is the only study to date that has used a behavioral rating scale in an experimental design including possible malingerers. Their results suggested that behavioral rating scales may be useful in detecting malingering, however, further research is needed with more clearly defined groups of possible or simulated malingerers.

Numerous studies have been done examining various indicators of malingering on standard neuropsychological tests. Studies have included all or parts of the following tests: the Benton Visual Retention Test (Benton & Spreen, 1961), the Bender–Gestalt Test (Bruhn & Reed, 1975; Schretlen & Arkowitz, 1990), the Rey–Osterrieth Complex Figure Test (Bernard, 1990; Bernard, Houston, & Natoli, 1993), the Wechsler Memory Scales — Revised (Bernard, 1990; Bernard et al., 1993; Greiffenstein, Baker, & Gola, 1994; Iverson & Franzen, 1996; Mittenberg, Azrin, Millsaps, & Heilbronner, 1993; Trueblood & Schmidt, 1993), the Memory Assessment Scales (Beetar & Williams, 1995), the Rey Auditory Verbal Learning (AVLT) Test (Bernard, 1990, 1991; Bernard et al., 1993; Binder, Villaneuva, Howieson, & Moore, 1993; Greiffenstein et al., 1994; Greiffenstein, Gola, & Baker, 1995; Hiscock, Branham, & Hiscock, 1994), the California Verbal Learning Test (Frederick et al., 1994; Rose, 1993; Trueblood, 1994; Trueblood & Schmidt, 1993), the Recognition Memory Test (Iverson & Franzen 1994; Millis, 1992, 1994), the Halstead–Reitan Neuropsychological Battery (Goebel, 1983; Heaton et al., 1978; Hiscock et al., 1994; Mittenberg, Rotherol, Russell, & Heilbronner, 1996; Trueblood & Schmidt, 1993), the Luria–Nebraska Neuropsychological Battery (Mensch & Woods, 1986), the Wechsler Adult Intelligence Scales — Revised (WAIS-R) (Bernard et al., 1993; Greiffenstein et al., 1994, 1995; Heaton et al., 1978; Iverson & Franzen, 1994, 1996; Martin, Hayes, & Gouvier, 1996; Mittenberg et al., 1993; Trueblood & Schmidt, 1993), the Symbol Digit Modalities Test (Hiscock et al., 1994), and the Wisconsin Card Sorting Test (WCST) (Bernard, McGrath, & Houston, 1996).

Many of the above studies have shown promising results for the use of indicators of malingering on neuropsychological tests. Some of the most well-supported measures are the Digit Span subtest from the WAIS-R (Greiffenstein et al., 1994, 1995; Iverson & Franzen, 1994, 1996; Martin et al., 1996; Trueblood, 1994; Trueblood & Schmidt, 1993), the Rey AVLT (Greiffenstein et al., 1994, 1995; Hiscock et al., 1994), the Recognition Memory Test (Iverson & Franzen, 1994; Millis, 1992, 1994), and the Halstead–Reitan Neuropsychological Battery (Mittenberg et al., 1996; Trueblood & Schmidt, 1993). Although there have been many studies examining these indicators, the indicators are rarely cross-validated, they lack consistent cutting scores, and they are rarely directly compared with other methods of assessing motivation. Advantages of these indicators are they add no additional testing to the...
battery and they provide potentially useful information regarding brain dysfunction. A disadvantage is they are prone to false-positive diagnoses of malingering due to inadequate validation of cutting scores and discriminant functions.

The third method of evaluating motivation during neuropsychological testing involves using tests specifically designed for the detection of malingering. Several tests have been designed for this purpose, including the Rey’s 15-Item Test (FIT) (Rey, 1964, cited in Lezak, 1995), Rey’s 15-Item Word List (Rey, 1941, cited in Lezak, 1995), Rey’s Dot Counting Test (Rey, 1941, cited in Lezak, 1995), Symptom Validity Testing (Pankrantz, 1979), the Portland Digit Recognition Test (Binder & Willis, 1991), the Digit Memory Test (DMT) (Hiscock & Hiscock, 1989), the 21-Item Word Test (Iverson, Franzen, & McCracken, 1991), the Forced Choice Test of Nonverbal Ability (Frederick & Foster, 1991), the Letter Memory Test (LMT) (Inman et al., 1998), the Test of Memory Malingering (Rees, Tombaugh, Gansler, & Moczynski, 1998), and the 16-Item Test (Paul, Franzen, Cohen, & Fremouw, 1992).

According to a recent meta-analysis reviewing five tests of malingering including the DMT, the Portland Digit Recognition Test, the 21-Item Test, the Dot Counting Test, and the Rey’s FIT (Vickery, Berry, Inman, Harris, & Orey, in press), the DMT had the largest mean effect size at $d = 1.95$ (CI 1.75–2.16). It was significantly greater than all tests, but the 21-Item Test that had a mean effect size of 1.26 (CI 1.04–1.48). The Portland Digit Recognition Test ranked third, followed by the Dot Counting Test and the FIT. When comparing individual classification rates, the DMT ranked highest with a mean sensitivity of 83.4% and specificity of 95.1%. The 21-Item Test ranked lowest, with a mean sensitivity of 22% and a specificity of 100%. Rey’s 15-Item Word List, Symptom Validity Testing, the Forced Choice Test of Nonverbal Ability, the LMT, the Test of Memory Malingering, and the 16-Item Test were not included in this meta-analysis due to an insufficient number of studies or inconsistent cutting scores.

In general, tests designed for the detection of malingering are intended to have a low true difficulty level, but a high face difficulty level, thereby enticing malingers to perform poorly while allowing motivated test-takers to perform well regardless of deficits. As such, they tend to have a low incidence of false-positive diagnoses of malingering. However, more sophisticated malingers may “see through” the high face difficulty of malingering tests, and escape detection — especially if they are coached by unscrupulous attorneys aware of the true difficulty level of the tests. Another disadvantage of incorporating tests of malingering into a neuropsychological test battery is that they take between 5 and 45 min to administer, and generally only yield information regarding the patient’s level of motivation, not brain dysfunction.

Studies validating indices of malingering on both neuropsychological tests and tests of malingering have suffered from similar weaknesses: In many cases, they have not been adequately cross-validated; they often fail to include ecologically valid analog malingering groups; they often fail to provide adequate incentives for malingering; they rarely evaluate the effectiveness of tests of malingering within a neuropsychological battery; and they rarely directly compare the effectiveness of individual tests with one another and in combination. In order to develop effective indicators of malingering, indicators must be validated in methodologically sound experiments. Simulation designs have the advantage of a high degree of experimental control, however, they tend to be less ecologically valid.
than known group designs (Rogers, 1988). In order to increase the generalizability of simulation designs, a possible alternative would be to include head-injured subjects as analog malingerers. This would begin to examine some of the more complex questions, such as: (1) Is it possible for individuals with documented injuries to increase their level of impairment? and (2) What would this pattern of performance look like? Two studies have examined this question to date: Palmer, Boone, Allman, and Castro (1995) presented a case study in which an individual with documented brain lesions appeared to be exaggerating his deficits, and Rees et al. (1998) employed a sample of individuals with primarily mild head injuries as analog malingerers in a validation study of the Test of Memory Malingering.

In the present study, data are presented which examine the performance of individuals with mild head injuries who were asked to exaggerate their deficits on neuropsychological testing. Their performance was compared to a control group of head-injured individuals, as well as nonhead-injured analog malingerers and controls in an empirical design. Both groups were administered a battery of tests including traditional neuropsychological tests and malingering tests.

It is felt that this study addresses the methodological weaknesses identified by Rogers (1988), while maximizing the level of generalizability. This study included detailed instructions, a manipulation check, an incentive, a control group, multiple indices, appropriate statistical analyses, and test administrators who were blind to test condition of each participant. This study did not include a coached group, but thorough instructions were provided in order to maximize a subject’s ability to successfully malinger. Generalizability was enhanced by including head-injured malingerers and controls, applying previously published cutting scores and discriminant function analyses, and using a battery of tests that approximates the experience of a standard neuropsychological assessment.

1. Methods

1.1. Participants

A total of 108 participants were initially recruited for this study. These participants were not included in any previously reported study (Inman et al., 1998; Orey, Cragar, & Berry, 2000). Fifty-five participants were selected from a pool of subjects enrolled in an introductory psychology course at the University of Kentucky because they had a history of head injury of at least mild severity. Head-injury severity was determined by self-reported loss of consciousness following head injury and self-reported length of post-traumatic amnesia. Head-injured participants were randomly assigned to the head-injured control (HIC) group or the head-injured malingering (HIM) group. Head-injured participants were excluded if they were currently seeking compensation for their injury, if their injury had occurred more than 5 years prior to the present study, or if they had been diagnosed with a severe mental illness, such as major depression or schizophrenia.

Fifty-three non-head-injured participants were also obtained from the pool of students enrolled in introductory psychology. These participants were matched to those in the head-
injured groups on the demographic variables of age, sex, years of education, and race. These participants were randomly assigned to either a normal control (NC) group or a normal malingering (NM) group. These individuals were excluded if they had suffered a blow to the head, which resulted in a loss of consciousness, or if they had been diagnosed with a severe mental illness, such as major depression or schizophrenia.

1.2. Materials

After participants completed the informed consent, they completed a questionnaire that requested basic demographic information, history of head injury, substance use, mental illness, current medications, and legal status in regard to head injury. Participants then read the appropriate instructions for their assigned groups. Instructions stated that the experimenter did not know participants’ group membership and that they should not reveal their instructions. Instructions for the honest-responding groups stated that they would be taking a battery of neuropsychological tests, and they should take them to the best of their ability. The instructions stressed the importance of putting maximum effort and attention while taking the tests. Instructions for the HIM group stated:

Think about the types of problems that you have had since your head injury. In this experiment, your job is to make your symptoms appear worse on a battery of neuropsychological tests. Imagine that you are having an evaluation which will determine the amount of a settlement which you will receive for your injury. While you don’t want to be dishonest, you want to reflect the severity of the problems that you have experienced. Therefore, you want to perform worse on the tests than you may ideally be able to do. As you go through the tests, try to make your problems appear more severe, but be careful not to be too obvious, as you would not want the examiner to know that you are exaggerating.

Instructions for the NM group stated:

Imagine that you have had a head injury. In this experiment, your job is to make your symptoms that you have experienced since your injury appear worse on a battery of neuropsychological tests. Imagine that you are having an evaluation which will determine the amount of a settlement which you will receive for your injury. While you don’t want to be dishonest, you want to reflect the severity of the problems that you have experienced. Therefore, you want to perform worse on the tests than you may ideally be able to do. As you go through the tests, try to make your problems appear more severe, but be careful not to be too obvious, as you would not want the examiner to know that you are exaggerating.

After the participants finished reading the instructions, the test battery was administered. The test battery contained tests specifically designed to assess motivation and malingering, as well as tests that measured neuropsychological functioning. Malingering tests that were given included a computerized version of the 21-Item Test (Iverson et al., 1991), a computerized version of the DMT (Hiscock & Hiscock, 1989), the FIT (Rey, 1964), and the LMT (Inman et al., 1998). Neuropsychological tests included the Digit Span, Information, and Digit Symbol (DSY) subtests from the Wechsler Adult Intelligence Test — Revised (Wechsler, 1981), Seashore Rhythm Test, Controlled Oral Word Association (COWA) Test, the Rey AVLT, Grip Strength, Grooved Pegboard, and the WCST.
A computerized version of the 21-Item Test was used (Inman et al., 1998; Iverson et al., 1991). Participants were presented with 21 words, one at a time, on the computer screen. They were allowed to study each word for 5 s before the next word appeared. After the 21 words were presented, participants were asked to write as many of the words as they could remember. The forced-choice portion followed, in which both the target word and a foil were displayed on the screen. Participants chose the correct word by pressing the number on the keyboard that corresponded to the number shown on the screen above their choice.

A computerized, abbreviated version of the DMT was used (Hiscock & Hiscock, 1989; Inman et al., 1998). Participants were presented with a five-digit number that they were instructed to remember. After a delay, they identified the correct number from two choices. This version of the DMT included three blocks of 12 items each with delays of 2.5, 5, and 10 s, respectively. No digits were shared between the target and the foil on any given trial.

The FIT was also included (Rey, 1964, cited in Lezak, 1995). Participants were asked to memorize 15 items and then recall them. In reality, it was fairly simple to group the items, so that test-takers only had to memorize five “chunks.” The presentation of the items was in five rows with three characters per line: A, B, C; 1, 2, 3; a, b, c; a circle, a square, a triangle; and Roman numerals I, II, and III.

The final test of malingering was the LMT (Inman et al., 1998). The LMT was a computerized test in which participants were presented a three-, four-, or five-letter stimulus, for 5 s. The stimuli were constructed from the first 10 consonants in the alphabet (B, C, D, F, G, H, J, K, L, M). Participants were instructed to remember the stimulus, and, after a 5-s delay, they were asked to choose the target set from among two, three, or four choices. Targets appeared equally in every position, and choices were displayed horizontally across the middle of the screen. Participants chose the correct set by pressing the number on the keyboard that appeared on the screen above the letter set they wished to choose. There were five trials of each combination of length of stimulus and number of foils. For example, there were five trials of a three-letter stimulus with four foils and five trials of a five-letter stimulus with two foils. There were a total of 45 trials with 45 different target stimuli. The combinations of length and foils were presented sequentially starting with three letters with two foils, then three letters with three foils, then progressing through the four- and five-letter combinations, and finally ending with five letters with four foils.

Neuropsychological tests were chosen, which sampled different domains of cognitive functioning and, when possible, had previous research on indicators of malingering. Digit Span and the Seashore Rhythm Test sampled attention and concentration. A second area of cognitive functioning, intellectual ability, was represented by the Information subtest of the WAIS-R (Wechsler, 1981). The area of language abilities was measured by the COWA test, which is a test of verbal fluency (Benton & Hamsher, 1976). The area of memory abilities was represented by the Rey AVLT (Rey, 1964; Taylor, 1959, as cited in Lezak, 1995). Sensation and perception were measured by the Seashore Rhythm Test, which served as a measure of auditory perception (Seashore, Lewis, & Saetveit, 1960). Motor skills were sampled using two tests, the Grooved Pegboard Test (PEG) and the Grip Strength Test. Visual–Motor Integration was measured by the Digit Symbol subtest of the WAIS-R. Frontal/executive functioning was measured by two tests, the COWA and the WCST (Heaton, 1981).
1.3. Procedure

Head-injured participants were randomly assigned to the malingering (HIM) or honest-responding groups (HIC). Matched normal controls and NM participants were also randomly assigned to the malingering (NM) or the honest-responding groups (NC). Before the manipulation, participants completed the informed consent form. After participants signed the consent form and all questions were answered, they filled out the demographics questionnaire.

To guarantee that experimenters were blind to participants’ conditions, this study employed the following procedure, developed by Bernard (1990).

All participants were given a sealed envelope that contained the appropriate instructions for their assigned group. The experimenter, who was blind to participants group status, then left the room for 10 min to allow participants to read the instructions and prepare their strategies.

In order to induce motivation on this task, participants in the malingering groups were told that they would obtain a prize of US$20.00 if they were able to successfully fool the examiner. For ethical reasons, after participants in this study were tested, all participants in the malingering groups were given this US$20.00-prize regardless of test performance.

As stated above, all participants were given 10 min to prepare their dissimulation strategy. Participants in the honest-responding groups were told to wait quietly for 10 min for the assessment to begin. After the preparation period had passed, participants were told to place their instructions back into the envelope. The experimenter then began administering the battery of tests.

The first test in the battery was always a test of malingering; however, the order of the malingering tests was counterbalanced. It was important to present a malingering test first because some have suggested that malingerers may appear more impaired on the first test of the battery, as they have no means of comparing the difficulty level of the first test with other procedures (Bernard, 1990).

After the participant had finished all of the measures, he/she was asked to complete a feedback questionnaire that also served as a manipulation check. The participant was asked to summarize his/her instructions and to provide ratings of his/her compliance with instructions, level of attention, level of effort, and perceived success on the task.

At the end of each administration, the examiner completed a rating of the participant’s motivation based on his/her behavioral observations.

2. Results

2.1. Participants

In order to ensure internal validity, participants were screened for their ability to recount the experimental instructions accurately and on the basis of ratings given on the post-testing debriefing questionnaire. Two variables were chosen from the questionnaire that were thought to relate significantly to test scores. These were the participants’ ratings of their understanding
of the instructions and their level of effort to comply with the instructions. Each of these variables was rated by participants on a score from 0 to 10, with 0 being the lowest rating (e.g., *no effort*) and 10 being the highest rating (e.g., *full effort*). Participants who rated themselves as less than 4 on either of these variables were eliminated. This was intended to exclude subjects who were unable or unwilling to cooperate with instructions. Other subjects were eliminated for compensation seeking following head injury and for a head injury occurring greater than 5 years ago. Below is a group-by-group discussion of the implementation of these criteria.

The HIM group consisted of 21 participants, after one participant was excluded because of compensation seeking, one participant was excluded for a head injury occurring more than 5 years ago, one participant was excluded because of an inadequate recounting of instructions, and four participants were excluded because of self-reported inadequate effort to follow the instructions. All participants in the head-injured groups had suffered a closed head injury of at least mild severity. Participants reported their head injury severity by selecting categories for length of loss of consciousness and time since injury along with other severity indicators, such as presence of post-traumatic amnesia and hospitalization. The median and the mode for loss of consciousness for the HIM group fell at 1–5 min (range from 1–5 min to 24–48 h). Time since injury was bimodal at less than 1 year and 1–2 years, but the median fell at 1–2 years (range from less than 1 year to 3–5 years). Three of the participants in the HIM group (14%) reported a period of retrograde amnesia, while two reported a period of post-traumatic amnesia (10%). Twelve participants (57%) went to the hospital following their injury, but only five (24%) were admitted. Two participants (10%) reported a second head injury with a loss of consciousness.

There were 24 participants in the HIC group, after one was excluded for compensation seeking, one was excluded for a head injury occurring more than 5 years prior to testing, and one was excluded for poor rating of expended effort. All participants in the HIC group reported a head injury with a loss of consciousness. Both the median and the mode for loss of consciousness fell at 1–5 min (range from 1–5 min to 24–48 h). The time since injury had a median of 2–3 years and a mode of 3–5 years (range from less than 1 year to 3–5 years). Five participants reported a period of retrograde amnesia (21%) and two (8%) reported a period of post-traumatic amnesia. Fifteen (63%) participants went to the hospital following their injury, but only one (4%) was admitted. Five (21%) participants reported a second head injury with a loss of consciousness.

The NM group contained 23 participants, after two were excluded for a low rated understanding of instructions and one was excluded for a low rating of effort to comply with instructions. No participants reported a history of head injury.

Twenty-four participants were included in the NC group, after one was excluded for an inadequate recounting of instructions and two were excluded for low effort to comply with instructions. No participants in the NC group reported a history of head injury.

2.2. Demographic data

Because neuropsychological tests may be sensitive to demographic and other characteristics (Larrabee, 1990), the four groups were first compared on these variables to determine if
any reliable differences existed. Demographic data were analyzed using analysis of variance (ANOVA) on continuous variables and chi-square analyses on categorical variables. Head-injury severity data were compared only between the two head-injured groups using $t$ tests. Major demographic variables are presented in Table 1.

It can be seen that the groups did not differ reliably on any of the major demographic variables. Participants were also compared on a wide variety of background variables including alcohol use, drug use, medication use, and psychiatric treatment. Variables, for which there were no significant differences, included drug use (amount or type of drug), alcohol use, medication use, current psychiatric treatment, and psychiatric diagnosis. A significant difference was found among groups in terms of the number of participants who reported previous psychiatric treatment. This effect was followed by pairwise $t$ tests using Tukey’s Honestly Significant Difference to correct for an inflated type I error rate. Significantly more participants in the HIM reported previous psychiatric treatment than did participants in the NM and NC groups. The frequency of previous psychiatric treatment in the HIC group was not reliably different from any other group. Follow-up regression and discriminant function analyses revealed that the incidence of previous psychiatric treatment did not predict a significant amount of variance in malingering group status when compared with other neuropsychological and malingering test scores.

Head-injury severity variables were compared only between the HIC and HIM groups using chi-square analyses, as no participants in the NC or NM groups had any history of head injury. The head-injured groups were not significantly different on any measure of head-injury severity.

2.3. Behavioral observations

The test protocol was administered by research assistants who were blind to participants’ group membership. At the end of each testing session, the research assistant rated a participant’s effort, honesty, accuracy of performance, and classified him/her as malingering or honest. Ratings of level of effort, level of honesty, and accuracy of performance were given on a Likert scale spanning from 0 (e.g., poor effort) to 10 (e.g., full effort). A summary of these ratings is provided in Table 2.
Experimenters rated participants in the malingering conditions as significantly lower in level of effort and level of honesty than participants in the honest conditions. However, when forced to categorize participants as either malingering or honest, they correctly classified significantly fewer participants in the malingering groups as compared to the honest groups. Thus, although they had a high degree of specificity (92%, equal for the HIC and NC groups), misclassifying few honest participants, they had only a moderate level of sensitivity (63%).

2.4. Neuropsychological tests

For all of the following analyses the significance level was set at \( P < .01 \) in order to control for alpha inflation in the multiple comparisons. All scores were compared among groups using a MANOVA model, with head-injury status and malingering status as factors. On no test was there an effect of head-injured status or an interaction of head-injured status with malingering condition. The results presented below show only the differences found between the malingering and nonmalingering groups.

2.5. Information, Grip Strength, and COWA

Scores on the Information subtest from the WAIS-R and the Grip Strength test were calculated using age, education, and sex-corrected normative T-scores (Heaton, 1992; Heaton, Grant, & Matthews, 1991). Percentile rankings for the COWA were calculated using age, education, and sex-corrected normative scores (Ruff, Light, & Parker, 1996). There were no
significant differences found among groups on the Information subtest of the WAIS-R, the Grip Strength test, or the COWA Test. Group means are presented in Table 3.

2.6. Digit Span

Digit Span scores were analyzed using Heaton’s (1992) normative tables for the WAIS-R with age, education, and sex-corrected T-scores. Digit Span scores were also analyzed using the Reliable Digit Span, which was calculated according to the instructions in Greiffenstein et al. (1994). Results are presented in Table 4. Participants in the HIM and NM groups scored significantly lower than those in the honest condition on both the Digit Span T-score and the Reliable Digit Span score. Hit rates were calculated based on a cutting score of Reliable Digit Span less than eight as suggested by Greiffenstein et al. (1994). This cutting score resulted in a specificity of 100% and a sensitivity of 27%, for an overall hit rate of 65%.

2.7. Seashore Rhythm Test

Seashore Rhythm Test scores were analyzed using Heaton et al.’s (1991) age, education, and sex-corrected T-scores. The number of errors on the Seashore Rhythm Test was also analyzed in order to cross-validate a cutting score suggested by Trueblood and Schmidt (1993). Results are summarized in Table 5. Participants in the malingering conditions scored significantly lower than participants in the honest conditions on both the T-score and the number of errors on the Seashore Rhythm Test. Trueblood and Schmidt (1993) suggested a

Table 3
Scores on Information, Grip Strength, and COWA for malingering and honest participants

<table>
<thead>
<tr>
<th></th>
<th>HIC</th>
<th>NC</th>
<th>HIM</th>
<th>NM</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information T-score</td>
<td>M</td>
<td>46.13</td>
<td>47.67</td>
<td>45.24</td>
<td>41.74</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>9.28</td>
<td>9.32</td>
<td>8.68</td>
<td>10.78</td>
</tr>
<tr>
<td>Grip dominant hand T-scores</td>
<td>M</td>
<td>48.71</td>
<td>42.29</td>
<td>41.24</td>
<td>41.65</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>9.09</td>
<td>6.31</td>
<td>8.60</td>
<td>8.10</td>
</tr>
<tr>
<td>Grip nondominant hand T-scores</td>
<td>M</td>
<td>47.50</td>
<td>44.04</td>
<td>43.29</td>
<td>42.30</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>8.70</td>
<td>5.79</td>
<td>6.62</td>
<td>9.36</td>
</tr>
<tr>
<td>COWA percentile</td>
<td>M</td>
<td>56.17</td>
<td>59.38</td>
<td>53.81</td>
<td>41.83</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>30.36</td>
<td>25.73</td>
<td>32.72</td>
<td>28.35</td>
</tr>
</tbody>
</table>

Means with same letter subscripts are not significantly different at \( P < .01 \).

Table 4
Digit Span scores for malingering and honest participants

<table>
<thead>
<tr>
<th></th>
<th>HIC</th>
<th>NC</th>
<th>HIM</th>
<th>NM</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-score</td>
<td>M</td>
<td>57.96</td>
<td>55.58</td>
<td>47.24</td>
<td>40.30</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>9.21</td>
<td>10.07</td>
<td>11.14</td>
<td>13.51</td>
</tr>
<tr>
<td>Reliable Digit Span</td>
<td>M</td>
<td>11.58</td>
<td>10.83</td>
<td>9.19</td>
<td>8.09</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>2.39</td>
<td>2.16</td>
<td>2.27</td>
<td>2.41</td>
</tr>
</tbody>
</table>

* \( P < .01 \)
A cutting score of more than eight errors to indicate malingering. In this sample, that cutting score resulted in a specificity of 98% and a sensitivity of 27%, for an overall hit rate of 64%.

### 2.8. Rey AVLT

Scores on the AVLT were analyzed using the raw scores. Analyses were conducted to examine differences among trials on recall, recognition, and the serial position effect. Results are presented in Table 6.

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Seashore Rhythm Test scores among malingering and honest participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HIC</td>
</tr>
<tr>
<td>T-score</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>Errors</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
</tbody>
</table>

Means with same letter subscripts are not significantly different at P<.01.

* P<.01.

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Scores on the AVLT for the malingering and honest groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HIC</td>
</tr>
<tr>
<td>Trial 1</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>Trial 2</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>Trial 3</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>Trial 4</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>Trial 5</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>Mean number of words recalled per trial</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>Total recall across trials</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>List B</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>Immediate recall</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>Delayed recall</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
<tr>
<td>Recognition</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
</tr>
</tbody>
</table>

Means with same letter subscripts are not significantly different at P<.01.

* P<.01.
The malingering groups recalled significantly fewer words than the honest groups on all of
the acquisition trials, except for trial 1. The groups did not reliably differ on recall of the
interference list, List B. The malingering groups again scored significantly lower than the honest
groups on both the immediate and the delayed recall trials, as well as the word recognition trial.

Bernard (1991) suggested that the serial position effect of word recall on the AVLT might
be a useful indicator of malingering. Specifically, he found that the participants in an analog
malingering condition had a significantly reduced primacy effect as compared to the
participants with closed head injuries. Closed head-injury patients and analog malingerers
did not differ in the recency effect. Primacy was defined as the mean recall of the first five
words in the list, and recency was defined as the mean recall of the last five words in the list.
Bernard’s (1991) hypothesis was tested in this sample, but no difference was found in the
primacy and recency effect between the malingering and honest conditions. The participants
in the malingering conditions recalled significantly fewer words overall, but showed a similar
serial position effect to those in the honest conditions. Fig. 1 graphically illustrates the pattern
of word recall.

Greiffenstein et al. (1994) suggested a cutting score of less than eight correct on AVLT
recognition to indicate malingering. When this cutting score is applied to this sample, it
results in a specificity of 100%, sensitivity of 9%, and an overall hit rate of 57%. Hiscock et
al. (1994) also derived a cutting score based on the mean number of words recalled across the
five acquisition trials. They found that a cutting score of less than 7.2 words recalled
effectively discriminated between the honest and the malingering groups. In this sample, the
cutting score resulted in a specificity of 100%, sensitivity of 16%, and a hit rate of 60%.

2.9. Grooved Pegboard

Scores for the Grooved Pegboard were analyzed using both the dominant and nondominant
hand scores, which were transformed to T-scores using the Heaton et al. (1991) normative
sample. Results are presented in Table 7.
2.10. Digit Symbol

Scores from the Digit Symbol subtest of the WAIS-R were calculated using Heaton’s (1992) age, sex, and education corrected T-scores. Results are presented in Table 8. Digit Symbol scores for the HIM and NM groups were significantly below scores in the HIC and NC groups. Trueblood (1994) suggested a cutting score of a scaled score of less than five to indicate malingering. This cutting score was cross-validated in this sample resulting in a specificity of 100%, sensitivity of 2%, and an overall hit rate of 53%.

2.11. Wisconsin Card Sorting Test

Scores from the WCST were analyzed using the T-score calculated from Heaton et al.’s (1991) norms based on the number of perseverative responses. The raw scores of the number of categories completed, number of perseverative responses, and the number of perseverative errors are also presented. The results are presented in Table 9. Scores in the malingering groups were significantly lower than scores in the honest groups on all measures from the WCST. A discriminant function derived by Bernard et al. (1996) was cross-validated in this sample. This discriminant function included the raw scores of the number of perseverative errors and the number of categories completed. In Bernard et al.’s (1996) study, it achieved a specificity of 94–97% and a sensitivity of 86–100% in two samples discriminating analog

Table 7
T-scores on the Grooved Pegboard Test for malingering and honest groups

<table>
<thead>
<tr>
<th></th>
<th>HIC</th>
<th>NC</th>
<th>HIM</th>
<th>NM</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dominant hand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>43.67</td>
<td>45.33</td>
<td>38.24</td>
<td>32.17</td>
<td>F(1,88) = 14.90*</td>
</tr>
<tr>
<td>S.D.</td>
<td>9.87</td>
<td>11.30</td>
<td>11.90</td>
<td>12.95</td>
<td></td>
</tr>
<tr>
<td>Nondominant hand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>42.04</td>
<td>45.13</td>
<td>40.05</td>
<td>34.91</td>
<td>F(1,88) = 9.08*</td>
</tr>
<tr>
<td>S.D.</td>
<td>8.59</td>
<td>10.89</td>
<td>9.10</td>
<td>10.00</td>
<td></td>
</tr>
</tbody>
</table>

Means with same letter subscripts are not significantly different at \( P < .01 \).

* \( P < .01 \).

The malingering groups scored significantly lower than the honest groups on both the dominant and nondominant hands of the Grooved Pegboard Test.

Table 8
Digit Symbol scores for honest and malingering conditions

<table>
<thead>
<tr>
<th></th>
<th>HIC</th>
<th>NC</th>
<th>HIM</th>
<th>NM</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>57.38</td>
<td>56.58</td>
<td>50.00</td>
<td>44.44</td>
<td>F(1,88) = 20.17*</td>
</tr>
<tr>
<td>S.D.</td>
<td>6.41</td>
<td>9.95</td>
<td>12.33</td>
<td>12.20</td>
<td></td>
</tr>
<tr>
<td>Scaled score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>11.88</td>
<td>11.75</td>
<td>10.10</td>
<td>9.04</td>
<td>F(1,88) = 17.39*</td>
</tr>
<tr>
<td>S.D.</td>
<td>1.45</td>
<td>2.69</td>
<td>3.10</td>
<td>2.84</td>
<td></td>
</tr>
</tbody>
</table>

Means with same letter subscripts are not significantly different at \( P < .01 \).

* \( P < .01 \).
malingerers from closed head-injury patients. In this sample, the specificity was 100%, but the sensitivity was only 9% resulting in a low overall hit rate of 58%.

2.12. Malingering tests

2.12.1. 15-Item Test

Scores on the FIT were analyzed using the raw score, which was the number of items correctly recalled regardless of their position in the matrix. Results are presented in Table 10. Participants in the malingering groups scored significantly lower than those in the honest conditions. Several different cutting scores have been suggested, but the most consistent is a cutting score of less than nine items recalled (Arnett, Hammeke, & Schwartz, 1995; Frederick et al., 1994; Goldberg & Miller, 1986; Greiffenstein, Baker, & Gola, 1996; Lee, Loring, & Martin, 1992; Millis & Kler, 1995; Morgan, 1991). This cutting score resulted in a specificity of 100%, a sensitivity of 2%, and a hit rate of 53%.

2.12.2. 21-Item Test

Scores on the 21-Item Test were analyzed using the number of words recalled on the free recall portion and the percentage correct on the recognition portion. Results are presented in Table 11. Participants in the malingering conditions scored significantly lower than those in the honest conditions on both the recall and recognition portions. A cutting score of less than three words recalled in the free recall trial has been established in several previous studies (Frederick et al., 1994; Inman et al., 1998; Iverson et al., 1991). When applied to this sample,
a cutting score of less than three words recalled resulted in a specificity of 100%, a sensitivity of 0%, and a hit rate of 50.59%. A cutting score of less than nine words identified on the recognition portion (43% correct) has also been commonly used (Inman et al., 1998; Iverson & Franzen, 1996; Iverson et al., 1991; Iverson, Franzen, & McCracken, 1994; Rose, 1993). This cutting score resulted in a specificity of 100%, a sensitivity of 5%, and a hit rate of 54%.

2.13. Digit Memory Test

Scores on the DMT were analyzed using the total percent correct. Results are presented in Table 12. Participants in the malingering conditions scored significantly lower than the participants in the honest groups. A cutting score of less than 90% correct was cross-validated in this sample (Guilmette, Hart, & Guiliano, 1993; Guilmette, Hart, Guiliano, & Leininger, 1994; Inman et al., 1998). This cutting score produced a specificity of 100%, a sensitivity of 64%, and an overall hit rate of 82%.

2.14. Letter Memory Test

Scores on the LMT were analyzed using the total percent correct. Results are presented in Table 13. Participants in the malingering conditions scored significantly lower than did those in the honest conditions. A cutting score of less than 93% correct was cross-validated in this sample (Inman et al., 1998). It resulted in a specificity of 100%, a sensitivity of 73%, and an overall hit rate of 87%.

There are two manipulations of face difficulty included in the LMT. The first is the increasing length of the stimulus to be remembered (three, four, or five letters). The second is the increasing number of choices from which the stimulus must be identified (two, three, or
A MANOVA with malingering and head-injured status as between-subjects variables and length of stimulus and number of choices as within-subjects variables was conducted in order to explore the effect of this manipulation. The results indicated that the malingering groups overall obtained uniformly lower scores on the LMT regardless of increments of length and of number of choices \(F(1,87) = 58.89, P < .01\). The increasing length of the stimulus did not have a significant effect on test scores \(F(2,174) = 3.74, P > .01\). However, with the increase in the number of choices test scores did show a significant change \(F(2,174) = 11.36, P < .01\). There was a significant interaction between malingering group status and an increasing number of choices, suggesting that participants in the malingering groups missed more items as the number of choices increased \(F(2,174) = 8.51, P < .01\). This effect was analyzed further and it was found that participants in the malingering conditions obtained significantly lower scores with three and four choices vs. two choices; however, there was not a significant difference in scores between three and four choices. These results are graphically illustrated in Fig. 2.

### 2.15. Combined malingering tests

Scores on the malingering tests were considered in combination by calculating the sensitivity and specificity based on the number of impaired scores on any one of the four choices. A MANOVA with malingering and head-injured status as between-subjects variables and length of stimulus and number of choices as within-subjects variables was conducted in order to explore the effect of this manipulation. The results indicated that the malingering groups overall obtained uniformly lower scores on the LMT regardless of increments of length and of number of choices \(F(1,87) = 58.89, P < .01\). The increasing length of the stimulus did not have a significant effect on test scores \(F(2,174) = 3.74, P > .01\). However, with the increase in the number of choices test scores did show a significant change \(F(2,174) = 11.36, P < .01\). There was a significant interaction between malingering group status and an increasing number of choices, suggesting that participants in the malingering groups missed more items as the number of choices increased \(F(2,174) = 8.51, P < .01\). This effect was analyzed further and it was found that participants in the malingering conditions obtained significantly lower scores with three and four choices vs. two choices; however, there was not a significant difference in scores between three and four choices. These results are graphically illustrated in Fig. 2.

#### Table 13

<table>
<thead>
<tr>
<th></th>
<th>HIC</th>
<th>NC</th>
<th>HIM</th>
<th>NM</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score % correct</td>
<td>M</td>
<td>99.30&lt;sub&gt;a&lt;/sub&gt;</td>
<td>99.50&lt;sub&gt;a&lt;/sub&gt;</td>
<td>80.40&lt;sub&gt;b&lt;/sub&gt;</td>
<td>71.10&lt;sub&gt;b&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>1.70</td>
<td>0.90</td>
<td>15.20</td>
<td>25.20</td>
</tr>
</tbody>
</table>

Means with same letter subscripts are not significantly different at \(P < .01\).

\* \(P < .01\).

![Fig. 2. Effect of increasing number of choices on LMT scores for malingering and honest conditions.](https://academic.oup.com/acn/article-abstract/17/1/1/2053)
four malingering tests. The number of participants who scored below each malingering test’s cutting score was calculated (LMT total < 93, DMT total < 90, 21-Item total < 43, 15-Item total < 9). The number of impaired malingering tests for each group is presented in Table 14. When the four malingering tests were considered in combination, one impaired score on either the DMT or the LMT resulted in a specificity of 100%, a sensitivity of 77%, and an overall hit rate of 89%, which was slightly higher than the hit rate found by using the LMT alone (87%). The addition of scores from the 21-Item Test and the FIT did not improve the hit rate.

2.16. Nonparametric analyses

Iverson and Franzen (1996) suggested that the highly divergent scores on tests that are often obtained between honest and malingering groups do not meet the level of homogeneity required for an ANOVA. Each of the neuropsychological and malingering test scores were examined in order to determine if the assumption of homogeneity of variance was met using the Bartlett–Box F test of homogeneity. Scores on five tests failed to meet that assumption of homogeneity of variance, the LMT, DMT, 21-Item Test, Digit Symbol, and AVLT. For each of these tests, follow-up analyses using a nonparametric procedure, the Kruskal–Wallis one-way ANOVA, were conducted. For each of these analyses, the significant differences between the malingering and honest groups were confirmed. Results are shown in Table 15.
3. Discussion

In recent years, the importance of developing indicators of malingering for neuropsychological testing has been recognized. Several tests have been developed for the purpose of detecting malingering. Although some of these tests have been adequately validated in analog and known-groups designs, few studies have compared malingering tests with one another and with standard neuropsychological tests. Furthermore, past studies have often been poorly controlled (i.e., not matched by demographic variables, not controlling for litigation) or have used analog malingerers who are naive to the effects of head injury, and seldom have they been cross-validated. Since scores are inherently more variable on standard neuropsychological tests, as compared to those designed specifically for the detection of malingering, cross-validation is particularly important. This empirical void has made it impossible to determine how a measure will perform as part of a neuropsychological battery. The present study has endeavored to fill this empirical void. It incorporated methodological guidelines by Rogers (1988), employing a battery of tests that mimicked a “real-world” neuropsychological evaluation, using real and simulated mild head-injury patients, employing rigorous screening criteria, matching participants on demographic variables, employing a manipulation check, and, most importantly, allowing for the cross-validation of previously published cutting scores.

A total of nine neuropsychological tests were included in the test battery. There were no significant differences among groups based on a history of head injury. This suggests that the mildly HICs were functioning at a level equivalent to the non-HICs. As such, specificity data should be interpreted with caution since it may be lower in a more severely head-injured population. With regard to the malingering manipulation, there were no significant differences among groups on the Grip Strength, Information, and COWA tests. Participants in the malingering groups scored significantly lower than those in the honest-responding groups on the Digit Span, Seashore Rhythm, Digit Symbol, Grooved Pegboard, WCST, and A VLT tests. There were no reliable main effects of head injury or interactions of head injury and malingering group status. Previously published cutting scores and discriminant functions were cross-validated resulting in good specificity (100%), but poor to moderate sensitivity (2–27%).

Perhaps the depressed sensitivity, using indicators on standard neuropsychological tests, is due to the fact that the true difficulty of these tests results in a larger range of scores from honest-responding participants and may force the cutting score to be set at a low level in order to maximize specificity. In this sample, it is notable that among 48, presumably cognitively intact, honest-responding, undergraduates, over 50% had at least one impaired score on the neuropsychological tests, as defined as a score less than or equal to the fifth percentile based on the norms described in Section 3. No participants in the control conditions scored within the impaired range on the Grip Strength, Digit Symbol, WCST, or Digit Span tests. However, on the Seashore Rhythm, Information, Grooved Pegboard, COWA, and A VLT, 4–21% of the honest participants scored within the impaired range.

Of the tests designed for the detection of malingering, the LMT and the DMT had the highest levels of specificity and sensitivity. Both the LMT and the DMT had a specificity of 100%. The LMT had a slightly higher sensitivity (73%) than the DMT (64%). Abnormal
scores on either the LMT or the DMT resulted in a specificity of 100%, and a sensitivity of 77%. Use of the 21-Item Test and the FIT contributed nothing to the overall hit rate.

When considering the utility of the DMT and the LMT to detect malingering in an individual case in clinical practice, one must compute the positive (PPP) and negative predictive power (NPP), which takes into account the base rate of malingering in the population. In previous studies, base rates have been suggested ranging from 15% to 48% of referrals in various settings (Binder et al., 1993; Guilmette et al., 1994; Trueblood & Schmidt, 1993; Youngjohn, Burrows, & Erdal, 1995). PPP and NPP were computed based on the minimum (15%), maximum (48%), and median (23%) base rates reported in the literature, and the sensitivity (77%) and specificity (100%) of one impaired score on either the DMT or the LMT. The PPP at each of these base rates was constant at 100%. The NPP declined as the base rate increased, from 96%, at a base rate of 15%; to 93%, at a base rate of 23%; to 83%, at a base rate of 48%; but remained fairly strong even at the highest base rate.

It is important that new tests of malingering are developed in order to combat coaching attempts by unscrupulous attorneys (Youngjohn, 1995). These tests must be well normed, such that they include psychiatric patients, individuals with cognitive deficits, and individuals with a wide range of educational levels and socioeconomic backgrounds. This is especially true when considering indicators of malingering on traditional neuropsychological tests that may be prone to false-positives. The LMT and the DMT have begun to accrue this type of normative data, although further research is needed (Guilmette et al., 1994; Inman et al., 1998). As the area of malingering research continues to grow, it becomes increasingly important to address the theoretical issues in the area. Research expanding theories of malingering (Rogers, 1990; Rogers, Sewell, & Goldstein, 1994) and creating more valid criteria for the diagnosis of malingering will be very important to the future of the research area.

References


