Development of a Novel Malingering Detection Method Involving Multiple Detection Strategies

John J. Neudecker*, Reid L. Skeel

Department of Psychology, Central Michigan University, Mount Pleasant, MI, USA

Accepted 3 December 2008

Abstract

This study utilized a novel computer-administered measure designed to assess multiple malingering detection strategies, as well as measure genuine impairment in a brain-injured sample. Ninety-four neurologically normal subjects comprised control and simulator groups. Twenty individuals with moderate to severe head injuries comprised a clinical sample. The summary score from the measure yielded a sensitivity of .81 and a specificity of .89. Total completion time and learning curves were utilized as genuine impairment indicators. Sensitivity (.95) and specificity (.95) were quite high for these dimensions. Results suggest there is utility to combining detection techniques, including indicators of genuine impairment.

Keywords: Neuropsychological assessment; Traumatic brain injury; Malingering

Introduction

Effective dissimulation of neuropsychological deficits relies on potential feigners obtaining information concerning the affect of brain injury as well as assessment techniques. Information regarding traumatic injury has been available to interested parties for some time. A greater, and perhaps more pressing danger, is the ability of present day potential malingerers to now find information regarding the exact mechanisms utilized by tests. Internet search engines yield websites containing relevant information while journal articles and books discuss techniques in depth (Bauer & McCaffrey, 2006). Research regarding malingering detection techniques may help malingerers to drastically improve their performance on previously sound tests. For example, if a malingerer was to discover that in order to “pass” a certain test, they must simply obtain a 100% score, then it is unlikely that any suspect effort would be detected.

Therefore, a key to the evaluation of malingering test development will be to devise detection strategies that are too complex for an individual to feign, even if exact detection mechanisms are made known. Combining established malingering techniques may be more resistant to test specific education. Further, measures may be more robust if they are also sensitive to genuine characteristics of human responding, such as consistency over time or natural learning curves.

In addition, it is important to note that because of the nature of malingering, it is difficult to conclusively determine how common the phenomenon is. Base rate estimates of malingering brain injuries range from 1% to 50% (Binder & Rohling, 1996; Miller, 1996; Reynolds, 1998). The monumental Slick, Sherman, and Iverson (1999) study suggests that the probability of malingering may be even higher. The combined rate of probable and definite malingering of neurocognitive dysfunction in their analysis was 54.3%. Larrabee (2007) has reported that the probability of malingering generally increases as actual symptom severity decreases. Considering these rates, the variability of estimates, and the serious implications of clinical statements regarding individual patients, there is a substantial need for aggregation of indicators of effort in each clinical case.

In summary, the main purpose of this study was to evaluate the usefulness of a multi-pronged detection approach that may also be sensitive to the most common type of genuine impairment within a single measure. A brief review of the detection strategies employed is provided, followed by limited overview of “embedded” techniques, and consideration of other similar approaches.

* Corresponding author at: Central Michigan University, 720 W. Wackerly Street, Suite 12, Midland, MI 48640, USA. Tel.: +1-989-839-6565; fax: +1-989-839-5794.
E-mail address: jneudecker@yahoo.com (J.J. Neudecker).

© The Author 2009. Published by Oxford University Press on behalf of the National Academy of Neuropsychology. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org.
doi:10.1093/arclin/acp008 Advance Access publication on 11 March 2009
The floor effect is the most commonly employed detection strategy, perhaps because it is simple to design and generally produces fairly robust findings. The floor effect occurs when a measure is so easy that virtually all individuals (regardless of neurological status) can respond correctly to all prompts with adequate effort. The distribution of scores is negatively skewed, so that low scores are particularly rare. In malingering measures that use the floor effect, low scoring outliers are thought to be performing poorly on purpose (Rogers, Harrell, & Liff, 1993). There are many examples of tests utilizing this strategy, including the 15-Item Test (Rey, 1964), the Test of Memory Malingering (TOMM; Tombaugh, 1996), the Word Completion Memory Test (WCMT; Hilsabeck & Gouvier, 2005), the Digit Memory Test (Hiscock & Hiscock, 1989), and the Portland Digit Recognition Test (Binder & Willis, 1991).

Performance curve analysis as a method of detection has received less attention in the literature (Rogers et al., 1993; Wogar, van den Broek, Bradow, & Szabadi, 1998). This method was originally identified and developed to evaluate differentiations in an individual’s performance across trials of varying complexity (Gudjonsson & Shackleton, 1986). The first assumption is that individuals will tend to respond correctly to easier items, and incorrectly to more difficult items. Secondly, it is assumed that individuals should remain consistent in their ability to solve certain types of problems across trials, and thus should reflect this consistency in their responses. One example of a test using this paradigm is the Word Memory Test (Green, Allen, & Astner, 1996).

The measurement of response time or latency is also a developing area. Bolan, Foster, Schmand, and Bolan (2002) conducted an analysis of the Amsterdam Short Term Memory Test (ASTM; Schmand, de Sterke, & Lindeboom, 1999), the TOMM, and a modified digit recognition test. Researchers found that across all three tests, malingers had the longest response delay, suggesting a measurable difference in their approach to the task, not just in their actual responses. Wogar et al. (1998) also utilized response time as a detection tool, theorizing that effortful test-takers would require more time to complete more challenging items compared with their time required to complete easier items, while malingers would fail to consider this phenomenon. In essence, these investigators were considering that naturally evoked learning curves could help differentiate between genuine and contrived performances.

The Use of Non-malingering Neuropsychological Tests

In addition to stand-alone malingering tests, researchers and clinicians have utilized “embedded” malingering indicators in more traditional neuropsychological measures. Embedded indicator refers to any component or analysis of a symptom-specific test (such as a list-learning task designed to assess memory) that suggests malingering or inadequate effort. Embedded measures offer the advantage of assessing effort within the context of a face valid test. In addition, implication by an embedded measure allows the clinician to make statements about that specific measure, rather than inferring suspect effort identified on a stand-alone malingering measure to a clinical measure administered separately.

The California Verbal Learning Test (CVLT; Delis, Kramer, Kaplan, & Ober, 1987) and the Rey Auditory Verbal Learning Test (RAVLT; Rey, 1964) have been utilized in such studies (Baker, Donders, & Thompson, 2000; Barrash, Suhr, & Manzel, 2004; Curtis, Greve, Bianchini, & Brennan, 2006; Millis & Putnam, 1997; Millis et al., 1995; Silverberg & Barrash, 2005), which is not surprising considering the heavy reliance on memory assessment by most malingering measures (Baker et al., 2000; King, Gfeller, & Davis, 1998). Several studies evaluated the utility of embedded detection with the Wechsler Memory Scale, Third Edition (WMS-III; Wechsler, 1997) as well (Killgore & DellaPietra, 2000; Langeluddecke & Lucas, 2003). Finally, another popular and well-researched embedded approach involves utilization of the Digit Span subtest from Wechsler tests (Greiffenstein, Baker, & Gola, 1994). Referred to as Reliable Digit Span, procedures essentially involve consideration of minimally acceptable performances (via floor effect). This approach has been demonstrated to be useful for distinguishing malingerers from clinical samples of varying severity (Heinly, Greve, Bianchini, Love, & Brennan, 2005), although other authors have indicated that the use of Digit Span alone is insufficient (Axelrod, Fichtenberg, Millis, & Wertheimer, 2006). In general, findings communicate support for the use of embedded measures. Embedded measures and other indicators of malingering are particularly useful when they agree, markedly increasing the probability of accurate identification as the number of indicators increases (Larrabee, 2007). Readers interested in a more thorough review are referred to the study of Larrabee (2007).

Another, more recent, example of multiple detection strategies that has demonstrated preliminary success is the Medical Symptom Validity Test (MSVT; Green, 2005). Essentially, this measure utilizes the floor effect and a form of performance curve analysis. With a fairly extensive database of clinical profiles (including dementias), clinical comparisons can be made between individual performances and referral-specific norms.

The Current Approach

The current study utilized a novel computer-administered multiple choice measure involving visuospatial pattern completion. It is comprised of seven-item categories with eight items in each category. The measure relies on four principles,
initially derived from literature review and theoretical conceptualization. These principles were subsequently pilot-tested in a study that included 35 control group subjects, 35 subjects asked to fake impairment, and nine subjects with documented traumatic brain injury (TBI) (none of whom were used in the current study). The pilot project was similar in design to the current study, and will thus be referred to as the “pilot” study for the rest of this manuscript.

The first principle is that individuals who are genuine in their approach will show relative consistency across problem categories that are similar in nature to their counterpart on a previous trial. In other words, if subjects respond correctly on one item of a category, they should respond correctly on subsequent problems that utilize the same principle. This first assumption is essentially a performance curve analysis. Items are ordered from the most to least difficult within item categories so as to maximally elicit consistency of performance in non-simulating subjects.

The second principle is response time. As subjects repeatedly answer questions utilizing the same underlying principle, their response time for solving similar problems will decrease. This second assumption is basically a utilization of normal learning curves. The detection of malingering occurs as these two principles interact. Item categories were organized from least to most difficult in presentation order. In general, effortful respondents should require more time per item category as the test progresses, while completion time within a category should decrease over trials. The result should be an upward traveling “sawtooth” graph of response time, reflecting within category response time decreases and overall measure response time increases.

The third principle is based on the floor effect. Utility of this effect has ample support in the literature, as discussed previously. Three of the seven categories are designed specifically to elicit this effect.

The fourth principle involves the time required to complete the entire measure. In a pilot study, individuals with impairment required nearly twice as much time to complete the measure as both simulators and controls, who were not significantly different. This principle is included as a possible indicator of genuine impairment, noting that decreased processing speed is probably the most common deficit in TBI (Langeluddecke & Lucas, 2003; Lezak, Howieson, & Loring, 2004). It was hypothesized that a combination of these four principles would be superior in discriminating between individuals feigning impairment than any of the principles individually. Increasing accuracy in detection through aggregation of multiple indicators is strongly suggested in recent research (Larrabee, 2007).

**Materials and Methods**

Participants included 20 individuals who had sustained a moderate-to-severe brain injury and 94 individuals without a history of brain injury. The clinical group included individuals who had sustained a documented moderate-to-severe brain injury. These individuals were invited to participate because of the severity of their injury (non-mild). Clinical records were reviewed with permission at three Michigan TBI rehabilitation facilities and one Michigan private practice neuropsychology clinic. Criteria for participation included a documented loss of consciousness (via medical records) greater than 30 min (Levin, Eisenberg, & Benton, 1989) and/or positive indication of brain damage detected via neurodiagnostic assessment. In addition, the injury must not have occurred prior to the individual’s eighteenth birthday. None of the clinical group members were currently involved in litigation. Former neuropsychological reports were available for review of most clinical group subjects. None of these reports identified suspect effort for any clinical group member. Clinical group subjects were compensated $25 for participation. Undergraduate students attending a large Midwestern University were recruited from psychology classes to participate on a voluntary basis to serve as the control and simulating samples. Subjects were compensated with extra-credit in their respective course. Subjects were recruited and randomly assigned to the control group or the simulator group. Subjects in all the three groups provided signed consent, consistent with IRB review requirements. Additional written permission and oversight were obtained from TBI rehabilitation centers for clinical subjects who resided there at the time of the study.

**Materials**

This study utilized ePrime for construction and implementation of the figural reasoning task. The task consisted of seven problem categories with eight items in each category. All problems were composed of figural sequences in which the subject is provided the first three figures and instructed to select the fourth from four alternatives to complete the pattern. The seven problem categories, in order of presentation, include sizing, dice, pendulum, migration, order, and analogies. A sizing problem consists of three frames in which a figure grows larger and the individual must select the next logical size from four choices. Dice problems involve counting in the forward or reverse direction depending on the number shown on the first three dice. Pendulum problems depict a pendulum in one of the three positions, with action lines suggesting movement. The subject must pick the terminal location of the pendulum. Migration problems utilize a 4 × 2 grid with a dot in one of the eight possible positions. The dot is “moved” in each of the first three frames, and the subject must pick the terminal location. Rotation problems involve rotating objects, occasionally with multiple parts rotating in opposite directions. Subjects must
identify terminal locations. Finally, analogies problems involve geometric shapes in which the first shape relates to the second in a meaningful way. Subjects must pick the fourth shape that best goes with the third shape, based on the relationship between shapes one and two. The entire task was administered by computer in private cubicles. Individuals were tested in groups no larger than four. Response time, response given, and the accuracy of the response were recorded by the computer.

A questionnaire was developed and implemented as a manipulation check to determine if participants who received injury-specific information actually learned the material. The questionnaire contained four simple true/false questions pertaining to head injury sequelae. The Education Questionnaire was not completed by the clinical group, as they were not exposed to the information regarding head injury sequelae.

Procedure

Non-clinical subjects were randomly assigned to one of the two groups. Group one was the control group and consisted of individuals who received no indication as to the actual purpose of the measure. The second group was designated as educated malingerers, and were instructed to feign impairment that results from brain injury, consistent with the brief statement read to them. Subjects in the control and simulator groups were read a statement regarding the effects of moderate head injury. They were asked to read along while the administrator read aloud. This information was adapted from the National Institute on Deafness and Other Communication Disorders (NIDCD) Health Information website (http://www.nidcd.nih.gov/health/voice/tbrain.asp) and is available in the Appendix.

Subjects in control and simulator groups were given a statement regarding how they should approach the measure. Control group participants were asked to help evaluate a new measure by trying to do their best. Their statement indicated that the measure was to be used as an index of brain injury, and that it was important that they try to do well. Subjects in the simulator group were given a scenario in which they were to pretend that they had been involved in a motor vehicle accident in which they had sustained moderate physical (non-brain) injuries. Despite their injuries, they were informed that their insurance company was denying their injury claims, and in order to obtain the required compensation, they were asked to feign cognitive impairment. The clinical sample received a brief oral description of the measure. They were told that the device was a measure of problem-solving and asked to do their best.

After completing all measures, subjects were thanked, debriefed in private, and compensated. All subjects were also provided with contact information should further questions arise.

Results

Demographic Variables

The sample of 94 participants without a history of TBI consisted of significantly more women \(n = 68\) than men \(n = 26\), \(\chi^2 (1, n = 94) = 18.8, p < .01\), while the clinical group was comprised primarily of men \(n = 15\), \(\chi^2 (1, n = 20) = 5.0, p < .05\). Analysis of gender differences within each of the three groups revealed no significant gender differences for items correct, total time to complete the measure, response time slopes within item categories, or across-categories response-time slopes. The groups did not differ on the majority of demographic variables. However, pooled simulators and controls were significantly younger than members of the clinical group, \(t(1,112) = 14.86, p < .001\). Table 1 provides additional information regarding age, gender, ethnicity, and educational level of participants. Subjects from all three groups overwhelmingly rated their effort (pertaining to experimental instructions) as an 8, 9, or 10 on a Likert Scale ranging from 1 to 10, with no effort ratings obtained below 5. Subjects who read the information regarding what to expect from a brain injury (control and simulator groups) were asked to complete a simple, 4-item true/false questionnaire to assess their attendance to the statement read to them. Subjects overwhelmingly obtained four correct responses. No subject responded incorrectly to two or fewer items. There were no subjects removed from either group owing to an unacceptable performance on the validity check.

Setting Cut-scores

Multiple variables were investigated as hypothesized for their ability to differentiate between controls and simulators. Cut-scores were determined by plotting individual scores by groups and then determining optimal points to dichotomize the data into probable malingering versus effortful test-taking. A bookmark cut-score method was used, with a slight preference for superior specificity in cases of ambiguous cut-score points. The bookmark method involves rank-ordering individual variables and making a visual cut in the data (across all groups) in order to optimize predictive power.
Item Analysis and Reliability

Individual item analysis was conducted to identify items that contributed to and detracted from overall reliability of the measure. Responses obtained from control and clinical group members were used for analysis as inclusion of simulator responses would not be meaningful. Mean total number of items correct for each problem category are provided by group in Table 2.

The total reliability for the entire measure (alpha) is .85 based on 67 cases. Category reliabilities were much lower, based on the limited number of items in each category. As expected, reliability for the two easiest categories (sizing and dice) was particularly poor because of the intended floor effect. There was little variability in responding, resulting in limited differentiation between respondents. Removal of these two categories in total measure reliability analysis did not meaningfully improve the overall reliability. Table 2 also contains alpha coefficients for each category including the total reliability.

Performance Curve Analysis

The utility of a performance curve analysis was analyzed by computing inconsistency in responding across items of the same category. Any instance in which an individual missed an item of a certain category immediately after correctly responding to that item category was recorded. The sum of these instances was used as an indication of inconsistency. Only incorrect responses immediately following correct responses were used in an attempt to avoid implicating individuals who respond correctly once by chance, and incorrectly to all like-category items to follow. Inconsistency among items for four of seven

Table 2. Mean total number of correct responses and inter-item scale reliability; group membership by problem category

<table>
<thead>
<tr>
<th>Problem Category</th>
<th>Control Group</th>
<th>Simulator Group</th>
<th>Clinical Group</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sizing</td>
<td>7.5 (0.8)</td>
<td>5.2 (2.2)</td>
<td>7.3 (0.8)</td>
<td>.19</td>
</tr>
<tr>
<td>Dice</td>
<td>7.8 (0.4)</td>
<td>5.8 (2.2)</td>
<td>7.8 (0.7)</td>
<td>.39</td>
</tr>
<tr>
<td>Pendulum</td>
<td>6.2 (1.5)</td>
<td>4.2 (2.3)</td>
<td>5.1 (1.8)</td>
<td>.52</td>
</tr>
<tr>
<td>Migration</td>
<td>6.0 (1.3)</td>
<td>4.6 (2.0)</td>
<td>5.0 (2.1)</td>
<td>.53</td>
</tr>
<tr>
<td>Order</td>
<td>4.9 (2.5)</td>
<td>3.3 (2.4)</td>
<td>4.5 (2.3)</td>
<td>.77</td>
</tr>
<tr>
<td>Analogies</td>
<td>6.4 (1.2)</td>
<td>4.4 (2.1)</td>
<td>5.0 (2.1)</td>
<td>.56</td>
</tr>
<tr>
<td>Rotation</td>
<td>5.0 (1.7)</td>
<td>3.5 (1.6)</td>
<td>4.7 (2.4)</td>
<td>.59</td>
</tr>
<tr>
<td>Total</td>
<td>43.8 (5.9)</td>
<td>31.0 (11.3)</td>
<td>39.2 (8.7)</td>
<td>.85</td>
</tr>
</tbody>
</table>

Notes: Standard deviations in parentheses, number of items per problem category = 8, total items = 56. Reliability configured using control and clinical group members (n = 67).

Performance Curve Analysis

The utility of a performance curve analysis was analyzed by computing inconsistency in responding across items of the same category. Any instance in which an individual missed an item of a certain category immediately after correctly responding to that item category was recorded. The sum of these instances was used as an indication of inconsistency. Only incorrect responses immediately following correct responses were used in an attempt to avoid implicating individuals who respond correctly once by chance, and incorrectly to all like-category items to follow. Inconsistency among items for four of seven
categories contributed to differentiation between simulators and controls. A p-value of .01 as opposed to .05 was adopted as the level of significance to help compensate for the large number of comparisons. Independent sample t-tests indicated significant differences between simulator and control groups for these variables at \( p < .001 \) (Table 3). Effect sizes for the following categories were substantial: sizing \( (d = 1.26) \); dice \( (d = 1.27) \); migration \( (d = .54) \); and analogies \( (d = .81) \).

A cut-score was set for inconsistency for each category that showed a difference between groups based on the bookmark method. Two or more instances of inconsistent responding optimized classification for sizing, migration, and analogies categories, while the dice category required only one instance of inconsistency. One point was added to the Malingering Index for each of these four criteria depending on the subject’s performance.

**Response–time Analysis**

Response time was hypothesized to decrease for all of the subject groups within each of the categories. Slopes across the eight items of each problem category were calculated using the slope of the best-fit line for each subject. The formula for the slope of a line was used \( y = mx + b \), where \( m \) = slope. The slope of the decrease was hypothesized to be steeper for control and clinical group members than for simulators. However, average aggregate learning curve slopes were not significantly different between simulator and control groups, \( t(92) = .160, p = .873 \). The clinical group, however, demonstrated on average much steeper learning curves within categories, \( F(3,111) = 48.4, p < .001 \), compared with both simulator and control groups. This observation will be discussed in a later section.

The problem categories were also ordered based on difficulty, with the easiest categories being placed at the beginning, and the most difficult categories being placed at the end of the administration. Via pilot-testing, sizing was determined to be the most simplistic category, followed by dice, pendulum, migration, order, analogies, and rotation categories, respectively. Effortful participants (control and clinical groups) were expected to require more and more time to complete items, on average, as the measure progressed. Paired with within-category response times, a line graph of response times for an effortful participant should result in a “sawtooth” pattern. The total response time slope (across all 56 items) was computed for each participant. The total response time slope was effective in differentiating between simulators and controls, \( t(92) = 4.75, p < .05 \). A graphical display of average response times by group across all 56 items is available in Fig. 1. Again, the bookmark method was used to set an optimal cut-point. After implementing this method it was determined that the \( m < 90 \) indicated malingering.

**Symptom Validity Analysis/Floor Effect**

Two categories were identified because they contained items that were very simple in nature, and thus, all items should be answered correctly by both clinical and control group members. The sizing category was minimally changed from the pilot study, whereas the dice scale was a new addition. The dice category was designed to be overly simplistic a priori, although it has not been previously investigated. The total items missed in each of these two categories were used separately in the overall Malingering Index. For the sizing category, the optimal cut-point was made at five or less correct (of eight total), whereas the optimal cut-score for the dice category was two or more incorrect responses.

**Malingering Index**

The Malingering Index is considered the overall “score” for the measure. It is comprised of seven observations of data for each subject, including four possible instances of inconsistent responding, two possible instances of failing floor-effect scales.

**Table 3. Independent sample t-test results for inconsistency variables**

<table>
<thead>
<tr>
<th>Category</th>
<th>Effortful Mean (SD)</th>
<th>Simulator Mean (SD)</th>
<th>t-Value</th>
<th>Effect size (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sizing</td>
<td>0.22 (0.52)</td>
<td>1.15 (0.91)</td>
<td>6.9**</td>
<td>1.26</td>
</tr>
<tr>
<td>Dice</td>
<td>0.12 (0.37)</td>
<td>1.04 (1.02)</td>
<td>6.8**</td>
<td>1.27</td>
</tr>
<tr>
<td>Pendulum</td>
<td>0.99 (0.77)</td>
<td>1.26 (0.97)</td>
<td>1.7</td>
<td>.34</td>
</tr>
<tr>
<td>Migration</td>
<td>1.00 (0.76)</td>
<td>1.45 (0.78)</td>
<td>3.1*</td>
<td>.54</td>
</tr>
<tr>
<td>Order</td>
<td>.90 (0.72)</td>
<td>1.17 (0.82)</td>
<td>1.9</td>
<td>.39</td>
</tr>
<tr>
<td>Analogies</td>
<td>.70 (0.72)</td>
<td>1.32 (0.78)</td>
<td>4.6**</td>
<td>.81</td>
</tr>
<tr>
<td>Rotation</td>
<td>1.24 (0.76)</td>
<td>1.47 (0.86)</td>
<td>1.5</td>
<td>.27</td>
</tr>
</tbody>
</table>

Notes: Effortful group \( (n = 67) \) consists of normal controls and clinical group members; Simulator \( (n = 47) \); variables reported in means and (standard deviations); figures represent instances of inconsistent responding as described in the text; \( df = 112 \) for all comparisons.

*Significant at \( p < .01 \).

**Significant at \( p < .001 \).
and consideration of the slope of response time across the measure. The Malingering Index thus ranged from 0 to 7. Individuals received a point for each criterion described earlier on which they performed below the cut-point. Malingering Index scores were subjected to the cut-off method and it was determined that the optimal cut-point was 2 or above. Individuals receiving an index score of 1 or 0 were identified as effortful test-takers, whereas individuals receiving index scores of 2 or greater were identified as malingerers. Using this criterion, 38 of 47 simulators were correctly classified as malingering while 5 of 47 controls were incorrectly classified. The Malingering Index incorrectly classified 2 of 20 genuinely impaired individuals as malingering. The overall sensitivity of the measure was .81, whereas the overall specificity was .89. Table 4 contains cut-score information. The clinical group and the control group were used for calculating these coefficients, making the base rate of malingering in the sample 42%. The resultant positive predictive power is equal to .84. Predictive power is expected to decrease as the base rate of malingerers decrease. The expected positive predictive power for a base rate of 20% would be equal to .66, whereas positive predictive power for a base rate of 5% would be equal to .32. Similarly, negative predictive power for the current sample was .97. With a base rate of 20%, negative predictive power would be .95, while with a base rate of 5% negative predictive power would be .99. A receiver-operating characteristic curve was equal to .873.


Indicators of Impairment

It was hypothesized that the total time to complete the measure would be a useful indicator of impairment in the present study. An analysis of variance determined that the time required to complete the measure by the clinical group (1,118,228 ms; under 20 min) was significantly greater than the time required by both the simulator (400,555 ms) and control (366,938 ms) groups: \(F(2, 111) = 78.04, p < .0001\). There was no difference between the control and simulator groups. A cut-score was set for time to complete the measure to determine the measure’s ability to detect true impairment in addition to malingering.

As mentioned previously, a malingering measure that is also capable of demonstrating impairment would be the optimal device for assessment of effort. In addition to total time to complete the measure, learning curves were evaluated. Average response time aggregates within item categories were computed for each individual (available as Table 5). Initial analysis suggested a strong difference between clinical group members versus simulators and controls for average response time slopes across the entire measure: \(F(1, 112) = 97.69, p < .0001\). Qualitatively, this finding suggests that on average, individuals with impairment required more time to complete initial items within categories, but decreased their response times at a faster rate than those without impairment. Essentially, sample individuals with TBI had more room for improvement. In order to simulate this effect, a simulator would need to not only start each category with extreme patience, but he/she would then need to reliably increase response times through the end of the category.

A cut-score was set for this variable. The average learning curve variable was then used in conjunction with total time to complete to evaluate the measure’s ability to detect impairment. If either cut-score implicated a subject, he/she was determined to be impaired. Using this as cut criteria, sensitivity (.95) and specificity (.95) were computed. Considering the sample base rate (17.5% impaired), the sample-specific positive predictive power is equal to .79. However, fluctuations in base rate affect predictive power. If a 5% base rate is utilized, positive predictive power declines to .49, whereas a base rate of 40% results in positive predictive power of .92. Similarly, negative predictive power for the current sample was .99. With a base rate of 5%, negative predictive power would remain at .99, whereas with a base rate of 40% negative predictive power would be
Qualitatively, only one simulator was implicated as impaired. However, that same individual scored a four out of seven on the overall malingering scale. Four control group members were implicated as impaired, though none of these four were implicated as malingering. Cut-score information is provided in Table 6.

There are four possible decision outcomes based on index scores, as represented in Fig. 2. These include malingering with evidence of impairment, malingering with no evidence of impairment, effortful test-taking with evidence of impairment, and effortful test-taking with no evidence of impairment. Essentially, simulators should be identified as belonging to the second group, clinical subjects should fit best with the third group, while controls should fit best with the fourth group. In fact, most subjects were assigned to the most appropriate group, as demonstrated in the figure. It is most telling to explore the designated status of individuals in the simulator group. Of these 47 individuals, only one was able to demonstrate impairment. This subject, however, was also implicated as malingering. Nine of the 47 simulators were not detected as malingering, although they also failed to demonstrate impairment. Finally, the majority of simulators (37 of 47) were identified as not impaired and malingering. The clinical group also fared well regarding designated status. An overwhelming 19 of 20 clinical subjects were identified as impaired and not malingering. The one subject who failed to demonstrate any impairment was also not identified as malingering. This finding is, in fact, similar to the vast majority of cases involving malingering in which clinical

---

**Table 4. Cut-score results for the malingering index**

<table>
<thead>
<tr>
<th></th>
<th>Simulators (n = 47)</th>
<th>Non-simulators (n = 67)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified as Malingering</td>
<td>38 True-positives</td>
<td>7 False-positives</td>
</tr>
<tr>
<td>Identified as Effortful</td>
<td>9 False-negatives</td>
<td>60 True-negatives</td>
</tr>
</tbody>
</table>

95% Confidence interval

| Sensitivity = .81 | .66 | .90 |
| Specificity = .89 | .79 | .95 |
| Positive predictive power = .84 | | |
| Negative predictive power = .87 | | |

**Notes:** The non-simulators group described in the table is comprised of clinical and control group members. The malingering index is comprised of seven data points, with the cut-score set at two violations or greater for implication as malingering.

**Table 5. Mean response time; group membership by problem category**

<table>
<thead>
<tr>
<th></th>
<th>Control, mean (SD)</th>
<th>Simulator, mean (SD)</th>
<th>Clinical, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sizing</td>
<td>21.95 (8.11)</td>
<td>37.85 (16.00)</td>
<td>85.23 (75.09)</td>
</tr>
<tr>
<td>Dice</td>
<td>20.69 (3.59)</td>
<td>38.52 (18.88)</td>
<td>71.169 (65.87)</td>
</tr>
<tr>
<td>Pendulum</td>
<td>39.26 (12.79)</td>
<td>49.86 (16.03)</td>
<td>135.62 (95.38)</td>
</tr>
<tr>
<td>Migration</td>
<td>44.84 (13.83)</td>
<td>55.03 (17.71)</td>
<td>127.28 (62.65)</td>
</tr>
<tr>
<td>Order</td>
<td>81.21 (41.84)</td>
<td>74.90 (33.42)</td>
<td>242.98 (112.24)</td>
</tr>
<tr>
<td>Analogies</td>
<td>67.43 (20.78)</td>
<td>71.86 (22.91)</td>
<td>185.21 (91.15)</td>
</tr>
<tr>
<td>Rotation</td>
<td>91.53 (37.46)</td>
<td>72.51 (32.09)</td>
<td>270.72 (149.90)</td>
</tr>
<tr>
<td>Total</td>
<td>366.93 (108.08)</td>
<td>400.55 (121.10)</td>
<td>1,118.22 (519.78)</td>
</tr>
</tbody>
</table>

**Note:** Response times displayed in seconds.

**Table 6. Cut-score results for impairment indicators**

<table>
<thead>
<tr>
<th></th>
<th>History of TBI (n = 20)</th>
<th>No history of TBI (n = 94)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified as impaired</td>
<td>19 True-positives</td>
<td>5 False-positives</td>
</tr>
<tr>
<td>Identified as normal</td>
<td>1 False-negatives</td>
<td>89 True-negatives</td>
</tr>
</tbody>
</table>

95% Confidence interval

| Sensitivity = .95 | .73 | .99 |
| Specificity = .95 | .87 | .98 |
| Positive predictive power = .79 | | |
| Negative predictive power = .99 | | |

**Notes:** History of traumatic brain injury (TBI) is comprised of clinical group members, whereas no history of TBI is comprised of control and simulator subjects. The Impairment Index is composed of two criteria, with a cut set at one or greater implicating genuine impairment.
measures identify what appears to be impairment while malingering measures (embedded and stand-alone) identify suspect effort. In these cases, overwhelmingly, clinical test results are regarded as inconclusive because of suspect effort. In essence, the presence of suspect effort “trumps” clinical tests that show impairment. The power of malingering indicators relative to clinical indicators also rises substantially as indications of suspect effort increase in number for the individual patient (Larrabee, 2007; Slick et al., 1999).

Discussion

Summary and Integration of Findings

The purpose of the study was to evaluate the ability of a novel measure to identify malingerers in a sample of simulators, controls, and genuinely impaired individuals. The first hypothesis posited that the simulator group would show greater inconsistency in responding. Four of the seven categories demonstrated the ability to discriminate between groups based on inconsistent responding. It is unclear why the three other categories were less successful, though each showed a trend towards significance. This observation is likely because of the differential difficulty of items within each category. Items were ordered based on the results of a pilot study, and may need some revision in terms of their sequential placement.

Presenting items within categories from the most to the least difficult appears to have been quite effective in capturing the importance of consistency in responding. Green, Lees-Haley, and Allen (2002) also demonstrated a successful utilization of response consistency through their validation of the word memory test (WMT). Simulators were expected to demonstrate a discrepancy between words recalled in the initial recall stage and the delayed recall stage, whereas non-simulators were expected to demonstrate superior consistency. The current measure utilizes similar items thought to tap the same construct, while the WMT allows for consistency comparisons between items that are the same. Repeating items could be added to future revisions to augment the performance curve analysis utilized in the current study.

The second hypothesis stated that simulators would demonstrate less-steep response times over the entire measure when compared with clinical and control group members. The total slope of response times (across all 56 items) was significantly different for the simulator group (vs. clinical and control groups; Fig. 1), as they did not produce a substantially disproportionate amount of time to complete simple items compared with more difficult items. Wogar et al. (1998) successfully demonstrated increased latencies in response time across items of increasing difficulty for control subjects and subjects with impairment. Simulators, however, were unable to demonstrate a proportionate increase in response latency. In the present study, a different format for stimuli was used, and results were consistent with Wogar et al.’s findings, suggesting that response time is a less-intuitive detection method that is harder to avoid when other demands are present.

The third hypothesis also involved response time. However, the level of analysis was within-item categories across like items. These within-category slopes can be thought of as learning curves, as they should decrease with time because of decreasing difficulty of the items and familiarity with the item category by an effortful test-taker. As mentioned earlier, there was a trend towards significance for each category, but the differences were not large enough to reliably discriminate between groups. It is unclear why learning curves were not as predicted. Means were graphed for clinical and control group members, resulting

![Decisional tree for indices used together. MI = Malingering Index, II = Impairment Index; figures in terminal boxes represent raw data based on group assignment, where ratios represent the number of individuals meeting criteria versus the sample size for the group in question; italics represent optimal group placement.](https://academic.oup.com/acn/article-abstract/24/1/59/3736)
in generally smooth learning curves. Simulators also demonstrated these curves, though perhaps for slightly different reasons. In order to purposefully select an incorrect response, simulators would need to first identify the correct response to avoid. If a simulator did this for every item, their learning curves should remain intact, and appear normal, as item difficulty decreased within-item categories. However, simulators did on average provide correct responses (60% correct response rate), with no clear preference for differential responding early or late within-item categories.

Hypothesis four stated that individuals in the simulator group would incorrectly answer items that were rarely missed by the clinical and control groups. This proved to be one of the more robust findings. Similar to results commonly obtained on the TOMM (Tombaugh, 1996), simulators incorrectly answered several of the 16 items identified as “easy.”

The current analysis was hindered to some degree in that only 16 items were deemed too easy to miss. The TOMM, for example, contains 50 items per trial, with a total of three trials. However, a strength of the current measure is that it is not entirely composed of floor effect items. These easy items are presented among items that are frequently responded to incorrectly by control group members. Therefore, unlike the TOMM or the WMT, the current measure does not elicit an overall floor effect. Educated malingerers must evaluate item difficulty when deciding how to respond. It would be suspect for a test-taker to incorrectly respond to several of the floor effect items, and then later correctly respond to a disproportionate number of more difficult items.

The fifth hypothesis involved the ability of the measure to detect impairment. The total time required to complete the measure was a significant predictor of group membership. In the present study, it was hypothesized that clinical group members would require more time in total to complete all 56 items. The notion behind this prediction is that individuals who have sustained a TBI are more likely to have deficits in processing speed than any other domain. Clinical group members demonstrated this tendency by requiring more time to respond to each item. If malingerers are made explicitly aware of this tendency, they may attempt to simulate it. However, as they vary the amount of time needed to respond to items, analyses such as the slope of response time curves across categories or over the entire measure may be more sensitive to unusual patterns of responding. The total time to complete the measure was combined with within-category slopes to compute a total impairment scale. When clinical group members were compared with simulators and controls (subjects without impairment), learning curves were significantly steeper for genuinely impaired individuals. In essence, individuals with genuine impairment required much more time for early (more difficult) items within-item categories, but quickly improved over time, thus establishing much steeper response time slopes. The rationale for the early delay in responding is that individuals with impairment likely have difficulty transitioning from one item category to the next. Learning curves were then incorporated with the total time to complete the measure in order to create a brief scale of impairment. While there are only two indicators, the ‘scale’ was able to differentiate between groups quite well. Sensitivity and specificity for this scale were quite high, as total time to complete was bolstered by including learning curves. This data is promising because malingering measures typically cannot also detect genuine impairment.

The primary hypothesis for the study was supported as it indicated that the combination of detection approaches would correctly identify the majority of simulators and non-simulators. This combination index demonstrated a much stronger sensitivity than any single hurdle alone without sacrificing specificity. This supports the hypothesis that the combination of multiple approaches has more utility for the detection of malingering.

The measure is unique because of its apparent abilities to measure effortful test-taking in addition to deficits (primarily in processing speed). The Malingering Index and the impairment index cannot be meaningfully combined into one index, as cut-score criteria are not applicable to three criterion groups. However, an individual’s performance on both indices can be interpreted by the clinician to aid in decision-making.

Overall, the current measure shows considerable promise as a clinical device. In order to prepare it for clinical use, the measure will benefit from item revision, minor changes in item presentation order, and validation with additional clinical groups. Establishment of concurrent validity will also be essential. This is particularly important with regards to the clinical assessment of malingering. Considering the limited amount of time available for the neuropsychological assessment, clinicians need a measure that can permit more confident decisions. False accusations as well as incorrect detection both bear significant ramifications. Thus, the standard for malingering detection measures needs to be quite high.

Analysis furthered the notion that this measure may not only give the clinician an indication of effort, but may also provide evaluation of the severity of genuine symptoms on the same measure for those not identified as possibly malingering. Individuals with a history of TBI have consistent deficits in speed of processing. The current measure proved to be sensitive to these deficits in a clinical sample.

Limitations and Future Directions

The current evaluation was limited by several factors, including limited sample size, population differences, and no measure of concurrent validity. Control and simulator samples are limited in that they are predominantly composed of college students.
The use of students is appropriate for this stage of research, but less useful for drawing conclusions to compare with the general population. In addition, simulators in this study may not adequately represent genuine malingerers based on the difference in value afforded by study participation versus worker’s compensation. Although Haines and Norris (2001) reported that student simulators may actually be more difficult to detect, markedly greater compensation may change the responding patterns of simulators significantly.

The clinical group was composed of individuals with a moderate-to-severe TBI. Validation with clinical groups including individuals who have sustained mild TBI will be important, given the increased base rate of malingering in groups of individuals with a history of mild TBI. A final population concern is the difference between college-based samples and the clinical group. Although education was similar between the groups, age differences may make comparisons less clear-cut. Additional criterion groups validation will address this concern.

As mentioned previously, concurrent validity needs to be established during the next stage of research with this measure. Evaluation of the measure’s concurrent validity should include comparison with both a well-established malingering measure in addition to a measure of clinical impairment. For example, a validation study may include the TOMM or WMT in addition to measures of processing speed, visuoperceptive skills, and visuospatial problem-solving.

The rarely missed index was one of the more useful single techniques for identification. However, it was hampered by the limited number of available observations. Addition of more easy items could permit for greater differentiation between groups.

The cut-scores determined for the current study resulted in optimal classification rates. After augmenting the measure, cut-scores determined from those data should be applied to a second group of subjects. This procedure will help to determine if the phenomena that cut-scores are based on are exclusive to the group they were created from or if they are observable across individuals. In addition, special attention should be given to the differential occurrence of malingering among TBI severity classes and other factors, as cut-score statistics are very much dependent on sample base rates (Meehl & Rosen, 1955).

Study limitations notwithstanding, it appears as though there is promise in measure development for multifaceted malingering detection devices, particularly as face-validity and information availability concerns increase. In addition, measures sensitive to impairment as well as effort appear to be particularly valuable. They permit an individual to evaluate one sample of behavior in terms of effort and ability as opposed to separate measurements of effort and impairment, which are by current standards samples of separate instances of behavior.

**Funding**

This study is a continuation of a dissertation product and there is no outside funding source to acknowledge.

**Conflict of Interest**

None declared.

**Appendix**

**Statement Regarding Head Injury Sequelae**

This statement was adapted from the NIDCD Health Information website for TBI (http://www.nidcd.nih.gov/health/voice/tbrain.asp).

TBI can be a significant cause of distress for the patient and family. In general, the common symptoms include disturbances in arousal, attention, and concentration. Memory impairments can occur, either due directly to memory function damage, or owing to poor attention and concentration in the encoding stage. Individuals often have problems with higher order or executive functions, including poor planning, sequencing, and judgment. This means that they have difficulty doing tasks with multiple steps or in cases where they must organize and implement a plan for future actions. Individuals with TBI may make errors due to impulsivity, and find that they have trouble in shifting between tasks. While the severity of certain deficits is associated with the degree of damage to different areas of the brain, an individual with a TBI is expected to show a range of deficits. Also, the individual’s overall cognitive ability, or intelligence level, will likely be decreased. Difficulties may not be obvious, but when an intellectually challenging task is presented, the individual is expected to have great difficulty.

**References**
