Detecting simulated memory impairment: Further validation of the Word Completion Memory Test (WCMT)

Robin C. Hilsabeck*, Wm. Drew Gouvier

Louisiana State University, USA

Accepted 25 May 2005

Abstract

The Word Completion Memory Test (WCMT) was developed to detect sophisticated attempts at simulating memory impairment. The primary purpose of the present study was to provide additional validity and reliability information about the WCMT. Seventy-one participants were recruited for this study: 30 normal volunteers and 11 memory-disordered patients instructed to perform their best, and 30 normal volunteers instructed to fake memory impairment. Normal volunteers were administered five tests of neuropsychological functioning and five tests of simulation to explore the convergent and divergent validity of the WCMT. Two weeks later, these participants completed all 10 measures a second time. Memory-disordered patients were administered the WCMT and two additional simulation measures as part of a comprehensive neuropsychological evaluation. The WCMT successfully discriminated simulators from nonsimulators with an overall classification accuracy of 97% and demonstrated good psychometric properties. In conclusion, the WCMT continues to show promise as a measure of simulated memory impairment.

Keywords: Malingering; Cognitive; Neuropsychological; Faking bad; Implicit memory

Estimates of the number of cases in civil litigation involving some component of simulated neuropsychological deficits range from 15–64% depending upon the patient population sampled (e.g., workers’ compensation cases, outpatients) and the criteria used to identify simulation (Binder & Willis, 1991; Greiffenstein, Baker, & Gola, 1994; Guilmette, Sparadeo, 2005).
Whelihan, & Buongiomo, 1994; Trueblood & Schmidt, 1993; Youngjohn, 1991). Since the likelihood of simulators confessing their deceit is very low, researchers in this area have used differential prevalence and simulation designs to examine two primary detection methods: identification of aberrant performances on existing neuropsychological tests and development of tests specifically designed to detect simulation.

Research investigating differences in neuropsychological test performance patterns between simulators and truly impaired individuals is important because these tests are administered as part of a neuropsychological evaluation anyway; thus, administration is cost effective and time efficient, and the tests are not easily recognized as simulation measures (Hayes, Hilsabeck, & Gouvier, 1999). Unfortunately, this method is limited because performance patterns of truly impaired individuals, as well as those of persons with no history of neurological problems, have not been clearly elucidated and sometimes may overlap (Ashendorf, O’Bryant, & McCaffrey, 2003; Lu, Boone, Cozolino, & Mitchell, 2003; O’Bryant, Hilsabeck, Fisher, & McCaffrey, 2003). Therefore, identification of simulated performance patterns on existing neuropsychological tests cannot be established fully until those of the criterion groups are more clearly identified.

The second method used to increase neuropsychologists’ abilities to detect simulated cognitive deficits is development of tests specifically designed for that purpose (i.e., domain-specific tests). There are four categories into which most domain-specific simulation tests are placed based on their underlying principles: floor effect, response bias/inconsistency, symptom validity/forced-choice procedures, and priming/implicit memory (Hayes et al., 1999). The floor effect refers to failing tasks that even severely impaired individuals are able to perform with some success. Therefore, tests relying on the floor effect often possess high specificity rates because few reasons exist for failing such simple tasks other than faking bad. The primary limitation of these tests is poor sensitivity as they are often viewed as too easy or too obvious to be faked (Haines & Norris, 1995; Millis & Kler, 1995). Therefore, only unsophisticated attempts at simulation are reliably detected with these measures. Another important drawback is that very poor performances are sometimes found among truly impaired persons, so even gross failures on these tests cannot be confidently equated with simulation (Lee, Loring, & Martin, 1992).

The hypothesis behind measures of response bias and/or response inconsistency is that some simulators may answer test items in a random or inconsistent manner. Random and bizarre responding have been indicated as specific simulation strategies by analogue simulators (Iverson, 1995). The Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) is a well-known test that includes measures of response bias and response inconsistency (e.g., L, F, K, TRIN, VRIN scales); however, the MMPI-2 was designed as a measure of psychopathology and may not be appropriate as a measure of cognitive dysfunction (McCaffrey, O’Bryant, Ashendorf, & Fisher, 2003; Mittenberg, Tremont, & Rays, 1996). Moreover, research has shown that the MMPI-2 is less sensitive than domain-specific measures of cognitive dysfunction (Berry et al., 1995; Greiffenstein, Gola, & Baker, 1995; Lamb, Berry, Wetter, & Baer, 1994; Wetter & Deitsch, 1996).

Currently, symptom validity tests or forced-choice procedures are the most widely used and researched measures of simulated cognitive deficits (Gervais, Rohling, Green, & Ford, 2004). Forced-choice procedures typically are based on the binomial distribution theorem; that is, on a test consisting of items with two possible answers, examinees should be able to obtain the
correct answer 50% of the time by chance alone. Advantages of the forced-choice procedure are its adaptability to any neuropsychological function where a two-alternative response format is possible (Frederick et al., 1995), the ease with which it can be computer-administered (Niccolls & Boller, 1991; Rose, Hall, & Szalda-Petree, 1995), and lack of viable reasons an examinee would score at a less-than-chance level (Rogers, Harrell, & Liff, 1993).

The primary disadvantage of symptom validity tests, whether administration is manual or computerized, is their low level of sensitivity for detecting simulation when using the original criterion of below chance performance. Although subsequent research has generated more sensitive cut-off scores (Binder, 2002; Millis, 2002), the frequency at which these tests have been employed in litigation cases and simulation research likely has resulted in reduced sensitivity due to coaching (Youngjohn, 1995). Symptom validity tests may be susceptible to coaching because they can be singled out easily from other measures by virtue of typically being computerized and involving a multiple-choice response format. Thus, an attorney can "warn" their client to beware of computerized tests and/or memory tests requiring recognition of previously presented items from an array of two or more.

The fourth type of domain-specific simulation test relies on the principles of priming and implicit memory. Priming refers to the facilitation of performance as the result of having previously viewed a target stimulus (Graf & Schacter, 1985). Implicit memory is revealed when task performance is facilitated by previously presented information without reference to that information, as opposed to explicit memory, which is revealed when task performance requires conscious recollection of the previously presented stimuli (Roediger, 1990). The reason priming tests might be useful in detecting simulation is because performances of amnesics on these tasks are counterintuitive to what laypersons might expect (cf., Graf, Squire, & Mandler, 1984). Simulation measures employing priming/implicit memory principles have been the least studied of the four types of simulating tasks (Davis, King, Bloodworth, Spring, & Klebe, 1997; Davis et al., 1997; Feldstein, Durham, Keller, Klebe, & Davis, 2000).

The fourth type of domain-specific simulation test relies on the principles of priming and implicit memory. Priming refers to the facilitation of performance as the result of having previously viewed a target stimulus (Graf & Schacter, 1985). Implicit memory is revealed when task performance is facilitated by previously presented information without reference to that information, as opposed to explicit memory, which is revealed when task performance requires conscious recollection of the previously presented stimuli (Roediger, 1990). The reason priming tests might be useful in detecting simulation is because performances of amnesics on these tasks are counterintuitive to what laypersons might expect (cf., Graf, Squire, & Mandler, 1984). Simulation measures employing priming/implicit memory principles have been the least studied of the four types of simulating tasks (Davis, King, Bloodworth, Spring, & Klebe, 1997; Davis et al., 1997; Feldstein, Durham, Keller, Klebe, & Davis, 2000).

The Word Completion Memory Test (WCMT; Hilsabeck & Lecompte, 1997) is a priming/implicit memory measure developed to answer calls from researchers for domain-specific simulation tests aimed at detecting more “sophisticated” simulators. For example, Guilmette, Hart, Giuliano, and Leininger (1994) noted, “the development of techniques resistant to ‘coaching’ or to sophisticated subjects remains an important area of inquiry” (p. 293). With this in mind, the WCMT was developed using well-coached analogue participants as the criterion group rather than analogue simulators simply instructed to fake “believable” memory impairment. The methodology used to coach analogue participants for development of the WCMT involved: (1) providing a detailed scenario of emotional, occupational, and cognitive difficulties (i.e., anterograde versus retrograde memory impairment) experienced by a claimant, (2) allowing participants to practice their roles by completing a questionnaire about the experiences of the claimant, and (3) excluding participants who did not demonstrate an accuracy rate of 80% or better on questions about problems experienced by the person in the scenario. An accuracy rate of less than 80% was believed to indicate poor attention to detail, inadequate understanding of the role participants were being asked to play, and/or poor motivation, which would be very unlikely in sophisticated, real-world simulators.
Recent studies have provided preliminary evidence of the validity of the WCMT. Hilsabeck, LeCompte, Marks, and Grafman (2001) investigated the classification accuracy of the WCMT using 69 control participants, 58 well-coached analogue simulators, and 14 memory-impaired patients. Results showed that 100% of control and memory-impaired participants and 93% of analogue simulators were correctly classified, resulting in an overall correct classification rate of 97%. A study using the German version of the WCMT found an overall correct classification rate of 100% of controls and well-coached analogue simulators (Merten, Henry, & Hilsabeck, 2004). The Word Memory Test (Green, 2003) and Amsterdam Short-Term Memory Test (Schagen, Schmand, de Sterke, & Lindeboom, 1997) also were investigated and found to have overall classification accuracy rates of 100 and 95%, respectively. The WCMT was viewed as more difficult than either of these measures, however, as well as Part A of the Trail Making Test (Reitan & Wolfson, 1993) and the copy trial of the Rey Complex Figure Test (Rey, 1941), suggesting the WCMT achieved its goal of appearing more sophisticated and more like a “true” neuropsychological test. Thus, initial validity data appear promising, but additional validation studies of the WCMT are needed.

The primary purposes of this study were to replicate the findings of these prior studies and to examine the psychometric properties of the WCMT. Four hypotheses were tested: (1) WCMT performances of participants instructed to fake memory problems would be significantly worse than WCMT performances of participants instructed to perform their best, (2) the WCMT would demonstrate acceptable convergent and discriminant validity, (3) the WCMT would display 2-week test-retest reliabilities >0.69, and (4) the WCMT would show classification accuracy superior to other measures of simulation. Secondary goals of this study were to investigate the psychometric properties of existing simulation measures, as this area of research has been neglected (Hayes et al., 1999; Nelson et al., 2003), and to explore differences, if any, between performances of undergraduate and community volunteers since generalizability of research with undergraduate populations has been questioned due to homogeneity of demographic characteristics (Arnett, Hammeke, & Schwartz, 1995; Rose et al., 1995).

1. Method

1.1. Participants

Seventy-one participants were recruited for the present study. Participants were placed into one of five groups: (1) community controls (N=15), (2) undergraduate controls (N=15), (3) community simulators (N=15), (4) undergraduate simulators (N=15), and (5) memory-impaired patients (N=11). Undergraduate controls (UC) and undergraduate simulators (US) were recruited from a large southern university campus and received extra credit in psychology courses in exchange for participation. Community controls (CC) and community simulators (CS) were recruited via undergraduate students who received extra credit in psychology courses for recruiting community participants. Undergraduate and community participants had no known histories of closed head injury, neurologic problems, substance abuse, or psychiatric illness as indicated by self-report.
The UC group was composed of 13 females and 2 males. Thirteen were Caucasian, one was African-American, and one was Asian-American. Average age was 19.60 years (S.D. = 1.45), ranging from 18 to 23 years. Average education was 13.20 years (S.D. = 1.08), with a range of 12–15 years. The US group consisted of 12 females and 3 males. Twelve were Caucasian, two were Asian-American, and one did not indicate race. Average age was 19.40 years (S.D. = 1.24), ranging from 18 to 21 years, and average education was 13.07 years (S.D. = 1.16), with a range of 12–15 years.

In an effort to ensure demographic differences between the undergraduate and community samples and to more closely approximate demographic characteristics of persons involved in litigation resulting from traumatic brain injury (Greiffenstein et al., 1994; Trueblood & Schmidt, 1993), community participants between the ages of 24 and 56 only were recruited for the study. The CC group was composed of 11 females and 4 males. Eleven were Caucasian, three were African-American, and one was Asian-American. Average age was 31.33 years (S.D. = 6.79), ranging from 24 to 46 years. Average education was 14.00 years (S.D. = 1.31), with a range of 12–16 years. Twelve of the CC participants were employed, one in a professional occupation, three as managers/supervisors, seven as service workers/clerical staff, and one as an unskilled laborer. The CS group consisted of 11 females and four males. Twelve were Caucasian, two were African-American, and one was Hispanic. Average age was 30.67 years (S.D. = 8.71), ranging from 24 to 53 years, and average education was 14.67 years (S.D. = 1.63), with a range of 12–18 years. Similar to the CC group, 12 CS participants were in the labor force, two in professional occupations, three as managers/supervisors, and seven as service workers/clerical staff.

Memory-impaired (MI) participants were patients who underwent comprehensive neuropsychological evaluations at the request of a physician or vocational rehabilitation counselor. Inclusion of this group was important to more clearly elucidate how patients with true memory impairment perform on the WCMT. All MI participants demonstrated impaired performances on tests of memory ability while retaining intellectual abilities in the low average range or above (mean WAIS-R Full Scale IQ = 97.00, S.D. = 16.37). MI participants were seven males and four females. Nine were Caucasian and two were African-American. Average age was 38.64 years (S.D. = 20.19), ranging from 17 to 69 years, and average education was 15 years (S.D. = 4.47), ranging from 11 to >20 years. Most of these patients were not employed at the time of their evaluations (N = 7). The remaining four patients worked, one as accountant, one as a business owner, one as a wildlife officer, and one as an office manager; however, their abilities to perform their job duties were raised as part of the referral question. Eight of the MI participants were referred by physicians, and three were referred by vocational rehabilitation counselors. Referral problems were closed head injury (N = 4), probable Alzheimer’s disease (N = 2), hypoxia (N = 2), brain tumor (N = 1), toxic exposure (N = 1), and cerebral palsy (N = 1). None of these participants was involved in litigation, and there was no obvious evidence of other secondary gain.

1.2. Materials

1.2.1. Screening questionnaire

A screening questionnaire was administered to undergraduate and community participants to obtain information about demographic characteristics, past and current medical and psy-
chological problems, including history of head injury or other neurologic problem, litigation status, and current medication and substance use.

1.2. Simulation measures

As noted above, the primary simulation measure of interest in this study was the WCMT (Hilsabeck & LeCompte, 1997). The WCMT consists of two subtests, Inclusion and Exclusion, and yields three scores: the Inclusion score (I score), the Exclusion score (E score), and the R score, which is the difference between the I and E scores. To establish convergent validity and to compare classification accuracies, four additional measures of simulation were investigated: the Memory for 15 Items Test (MFIT; Rey, 1964), Dot Counting Test (DCT; Rey, 1941, 1964), Multi-Digit Memory Test (MDMT; Niccolls & Bolter, 1991), and Recognition Memory Test (RMT; Warrington, 1984). The MFIT and DCT were chosen because they are two of the most widely used and well-researched domain-specific simulation tests available, and they represent the principles of floor effect (i.e., MFIT) and response bias/inconsistency (i.e., DCT). In addition, they are brief and easy to administer and score. Although several scoring criteria and methods have been suggested for detecting cognitive simulation with these measures, only the total number of items correctly recalled on the MFIT and total number of errors on the DCT were examined.

The remaining two simulation measures, the MDMT and RMT, were chosen because they are forced-choice recognition memory tests. Although the MDMT has not received as much attention from researchers as some other computerized digit recognition tests, its procedure is nearly identical, and there is no reason to believe its discriminative ability is significantly different from more widely researched measures (e.g., Hiscock Forced Choice Procedure, Hiscock & Hiscock, 1989; Portland Digit Recognition Test, Binder, 1993). The RMT was included in this study because it is a well-known, brief memory test that was not developed originally as a simulation-specific measure but has since been used as a potential measure of simulation due to its forced-choice recognition format (Iverson & Franzen, 1998; Millis, 1992; Millis & Putnam, 1994). For the purpose of this study, total correct was the variable examined for the MDMT, as well as RMT Words and Faces.

1.2.3. Nonsimulation measures

Five nonsimulation measures were utilized to establish divergent validity of the WCMT: Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990), Finger Tapping Test (FTT; Reitan & Wolfson, 1993), Grip Strength (GS; Reitan & Wolfson, 1993), Grooved Pegboard (GP; Klove, 1963; Matthews & Klove, 1964), and Boston Naming Test (BNT: Kaplan, Goodglass & Weintraub, 1983). The KBIT was chosen for inclusion in the study because it is a well-known, brief memory test that was not developed originally as a simulation-specific measure but has since been used as a potential measure of simulation due to its forced-choice recognition format (Iverson & Franzen, 1998; Millis, 1992; Millis & Putnam, 1994). For the purpose of this study, total correct was the variable examined for the MDMT, as well as RMT Words and Faces.
1.3. Procedure

After obtaining informed consent, undergraduate and community participants were asked to complete the screening questionnaire. Participants who met exclusion criteria (i.e., history of closed head injury, neurologic problem, psychiatric illness, or significant substance use) were debriefed, given the appropriate extra credit, and thanked for their participation. Participants who did not meet exclusion criteria were asked to schedule an individual 2-h testing session at a convenient date and time. Those who agreed were randomly assigned to either the control or simulating condition (UC, US, CC, and CS groups).

Undergraduate and community participants returning for the individual testing sessions again provided written informed consent. Participants assigned to the simulating condition (US and CS groups) were instructed to take on the role of a person who had suffered a mild closed head injury in an automobile accident for which they were seeking compensation. US and CS participants were provided with a detailed scenario of this person’s memory problems and the motivations which might lead the person to feign or exaggerate memory difficulties. The scenario was read aloud to simulating participants to ensure exposure to the information, and they were allowed to refer back to the scenario throughout the testing session in an effort to aid their abilities to undertake the role. Simulating participants were specifically instructed to fake memory difficulties only, in an effort to increase the sophistication of their approach to feigning to more closely simulate real world simulators (pilot data confirmed that these instructions were effective, as participants reported on a post-experiment questionnaire that they attempted to simulate memory deficits only, just as instructed).

Next, US and CS participants completed a questionnaire designed to assess their understanding of the role they were asked to play (that of a simulator) and to give them an opportunity to practice that role. Because real world simulators are very likely to have a clear understanding of their roles and are well practiced at playing them, only data from US and CS participants achieving 80% correct on the first 10 items of the questionnaire were included in the analyses to more closely approximate the level of understanding of a real world simulator. No simulating participant failed to meet this criterion. Participants assigned to the control condition (UC and CC groups) were instructed to do their best on all tests.

After the UC, CC, US, and CS groups were given the appropriate instructions for their experimental conditions, each participant was administered the above 10 tests in the following fixed order: WCMT, BNT, MFIT, GP, MDMT, GS, RMT, K-BIT, DCT, and FTT. The order of tests was fixed so potential interference of similar test content (e.g., word lists) could be minimized, and so tests measuring neuropsychological abilities could be alternated with those assessing simulation. However, the starting point in the battery was counterbalanced across participants to help control for order and fatigue effects. Upon completion of these tests, the US and CS groups were administered a post-experiment questionnaire to assess compliance with instructions and to elicit strategies used to simulate memory impairment. Two weeks later, UC, CC, US, and CS participants completed all 10 measures a second time, in an order different from their first administration, and the US and CS groups were again administered a post-experiment questionnaire.

Participants in the MI group were administered the WCMT, MFIT, and DCT as part of a comprehensive neuropsychological evaluation after obtaining written informed consent. These
participants were not given instructions other than those for the standard administration of these measures.

2. Results

2.1. Comparisons between undergraduate and community participants

As expected, UC and CC groups differed significantly in age \( t(28) = -6.55, P < .001 \), as did US and CS groups \( t(28) = -4.96, P < .001 \). In both cases, community participants were significantly older than undergraduate participants. No significant group difference in education was found between the UC and CC groups, but there was a significant educational difference between the US and CS groups \( t(28) = -3.09, P = .004 \). The CS group had significantly more education than the US group, although this difference was not clinically meaningful (i.e., 14.7 vs. 13.1 years, respectively).

Two-tailed independent samples \( t \)-tests were conducted on each of the 10 measures to determine if undergraduate and community participants could be combined to form one control group and one simulator group regardless of these significant differences in age and education. For the control participants, there was only one significant difference out of 18 variables, nondominant hand Grip Strength \( t(28) = -2.06; P = .05 \). On this task, CC participants showed a significantly stronger nondominant hand Grip Strength than UC participants (35.0 kg vs. 28.2 kg, respectively), which may have been due, in part, to the greater number of males in the former group (4 vs. 2, respectively). There also was only one significant group difference out of 18 variables for the simulating participants, Dot Counting Test total errors \( t(28) = -2.11; P = .05 \). CS participants made significantly more errors on this task than US participants (3.4 vs. 1.6, respectively). Given the large number of variables examined (i.e., 36), at least two significant differences would have been expected by chance alone. Therefore, these two differences were disregarded, and UC and CC participants were combined to form one control group (CON; \( N = 30 \)), and US and CS participants were combined to form one simulation group (SIM; \( N = 30 \)). Test means and standard deviations of the combined groups (CON and SIM), as well as the MI group, where applicable, are presented in Tables 1 and 2.

2.2. Comparisons among control, simulator, and memory-impaired groups

A one-way analysis of variance (ANOVA) and was conducted to investigate whether group differences in age and education existed between the MI group and the CON and SIM groups. Results showed that significant group differences were found for age \( F(2) = 7.26, P = .001 \) but not for education \( F(2) = 1.69, P > .05 \). A Tukey HSD post hoc test revealed the MI group was significantly older (i.e., 38.64 years) than both the CON and SIM groups (i.e., 25.47 and 25.03 years, respectively), which did not differ significantly from one another. In addition, a Kruskal–Wallis revealed significant group differences for gender \( \chi^2(2) = 8.06, P = .02 \), and Mann–Whitney \( U \)-tests showed that the MI group consisted of significantly more males than both the CON and SIM groups (\( U = 93.00 \) and 98.50, respectively, \( P < .05 \)). The CON and SIM groups did not differ significantly from each another (\( U = 435.00, P > .05 \)).
Table 1
Means (standard deviations) of nonsimulation measures by group

<table>
<thead>
<tr>
<th></th>
<th>CON (N=30)</th>
<th>SIM (N=30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-BIT Composite IQ</td>
<td>101.80 (7.67)</td>
<td>96.37 (14.07)</td>
</tr>
<tr>
<td>K-BIT Vocabulary</td>
<td>102.57 (8.28)</td>
<td>96.93 (13.84)</td>
</tr>
<tr>
<td>K-BIT Matrices</td>
<td>100.77 (9.58)</td>
<td>96.53 (16.56)</td>
</tr>
<tr>
<td>FTT DH</td>
<td>45.55 (7.02)</td>
<td>40.23 (10.27)</td>
</tr>
<tr>
<td>FTT NDH</td>
<td>40.80 (5.63)</td>
<td>37.48 (8.11)</td>
</tr>
<tr>
<td>GS DH</td>
<td>34.05 (9.00)</td>
<td>34.03 (10.44)</td>
</tr>
<tr>
<td>GS NDH</td>
<td>31.62 (9.57)</td>
<td>31.97 (10.50)</td>
</tr>
<tr>
<td>GP DH</td>
<td>62.97 (10.15)</td>
<td>72.50 (19.11)</td>
</tr>
<tr>
<td>GP NDH</td>
<td>68.30 (11.43)</td>
<td>76.57 (19.69)</td>
</tr>
<tr>
<td>BNT</td>
<td>53.37 (4.33)</td>
<td>50.10 (9.79)</td>
</tr>
</tbody>
</table>

Note: CON = controls; SIM = simulators; K-BIT = Kaufman Brief Intelligence Test; FTT = Finger Tapping Test; DH = dominant hand; NDH = nondominant hand; GS = Grip Strength; GP = Grooved Pegboard; BNT = Boston Naming Test.

A Kruskal–Wallis one-way analysis of variance of ranks was utilized to examine group differences on the WCMT variables. The Kruskal–Wallis analysis indicated a significant main effect of group for I, E, and R scores ($\chi^2 (2) = 36.33, 42.74, and 46.28$, respectively, $P < .001$). Because there were significant group differences in age and gender, a multivariate analysis of variance (MANOVA) employing age and gender as covariates was conducted. It was found that the group differences for WCMT I, E, and R scores remained significant ($F (2) = 70.12, 34.20, and 79.56$, respectively, $P < .001$). Follow-up Mann–Whitney U-tests revealed that the CON group obtained significantly higher WCMT I and R scores than both the SIM group ($U = 38.00$ and 29.00, respectively, $P < .001$) and the MI group ($U = 99.00$ and 84.00, respectively, $P < .05$). Results also showed that the MI group obtained significantly higher WCMT I and R scores than the SIM group ($U = 27.50$ and 15.50, respectively, $P < .001$). With regard to WCMT E scores, the CON and MI groups obtained significantly lower scores than the SIM group.

Table 2
Means (standard deviations) of simulation measures by group

<table>
<thead>
<tr>
<th></th>
<th>CON (N=30)</th>
<th>SIM (N=30)</th>
<th>MI (N=11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCMT I score</td>
<td>23.97 (2.65)$^a$</td>
<td>12.10 (5.44)$^b$</td>
<td>21.55 (3.75)$^c$</td>
</tr>
<tr>
<td>WCMT E score</td>
<td>1.93 (2.16)$^a$</td>
<td>12.20 (6.86)$^b$</td>
<td>3.36 (3.56)$^c$</td>
</tr>
<tr>
<td>WCMT R score</td>
<td>22.03 (3.56)$^a$</td>
<td>–0.10 (9.60)$^b$</td>
<td>18.18 (4.26)$^c$</td>
</tr>
<tr>
<td>MFIT total correct</td>
<td>14.83 (5.94)$^a$</td>
<td>10.80 (3.83)$^b$</td>
<td>13.45 (2.16)$^c$</td>
</tr>
<tr>
<td>DCT total errors</td>
<td>1.27 (1.14)$^a$</td>
<td>2.5 (2.47)$^b$</td>
<td>1.27 (1.62)$^c$</td>
</tr>
<tr>
<td>MDMT total correct</td>
<td>70.93 (1.74)$^a$</td>
<td>41.93 (18.05)$^b$</td>
<td>NA</td>
</tr>
<tr>
<td>RMT Words</td>
<td>47.97 (2.36)$^a$</td>
<td>27.13 (12.15)$^b$</td>
<td>NA</td>
</tr>
<tr>
<td>RMT Faces</td>
<td>40.73 (4.23)$^a$</td>
<td>27.57 (8.82)$^b$</td>
<td>NA</td>
</tr>
</tbody>
</table>

Note: CON = control; SIM = simulators; MI = memory-impaired; WCMT = Word Completion Memory Test; MFIT = Memory for 15 Items Test; DCT = Dot Counting Test; MDMT = Multi-Digit Memory Test; RMT = Recognition Memory Test. Superscript letters indicate means that differ significantly from other means designated with a different letter.
Additional Kruskal–Wallis analysis of variance of ranks found significant group differences on MFIT total correct \( \chi^2 (2) = 24.03, P < .001 \). A MANOVA also was conducted with age and gender entered as covariates, and results remained significant \( F (2) = 18.11, P < .001 \). Mann–Whitney U-tests revealed that the CON and MI groups performed significantly better than the SIM group \( (U=168.00\text{ and } 96.00,\text{ respectively, } P<.001\text{ and }<.05\text{, respectively}), \) but did not differ significantly from each other \( (U=116.50, P=.16\text{; see Table 2}). \) With regard to DCT total errors, a Kruskal–Wallis failed to find significant group differences \( \chi^2 (2) = 4.21, P = .12 \). However, a MANOVA using age and gender as covariates revealed significant group differences \( F (2) = 3.84, P = .03 \), and Tukey’s HSD post hoc analyses showed that SIM participants made significantly more errors than CON and MI participants, who did not differ significantly from each other (see Table 2).

Results of one-tailed independent samples t-tests indicated significant differences between the CON and SIM groups on the two remaining simulation measures, MDMT and RMT (note that MI participants were not administered these two measures). Findings revealed that SIM participants obtained significantly lower total correct scores on the MDMT \( t (58) = 8.75, P < .001 \), as well as RMT Words and Faces \( t (58) = 9.22 \text{ and } 7.37,\text{ respectively, } P < .001 \text{; see Table 2}).

### 2.3. Convergent and divergent validity

Given the relatively large number of variables employed to examine convergent and divergent validity, the Bonferroni correction procedure was applied to determine significance of relationships. Thus, only correlation coefficients reaching a \( P \)-value of .008 or less (i.e., .05/6) for convergent relationships and .005 or less (i.e., .05/11) for divergent relationships were considered significant. As shown in Table 3, the WCMT \( R \) score demonstrated acceptable convergent validity, with correlations ranging from \(-.37\) (i.e., DCT total errors) to \(.80\) (i.e., MDMT total correct). Results also indicated that the other measures of simulation possessed acceptable convergent validity, overall. Only DCT total errors failed to correlate significantly with all domain-specific simulation measures.

### Table 3

<table>
<thead>
<tr>
<th></th>
<th>MFIT total correct</th>
<th>DCT total errors</th>
<th>MDMT total correct</th>
<th>RMT Words</th>
<th>RMT Faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCMT ( R ) score</td>
<td>.68*</td>
<td>(-.37)</td>
<td>(.80)</td>
<td>(.73)</td>
<td>(.75)</td>
</tr>
<tr>
<td>MFIT total correct</td>
<td>(-.25)</td>
<td>(.72)</td>
<td>(.60)</td>
<td>(.59)</td>
<td></td>
</tr>
<tr>
<td>DCT total errors</td>
<td>(-.40)</td>
<td>(-.40)</td>
<td>(-.41)</td>
<td>(.92)</td>
<td>(.87)</td>
</tr>
<tr>
<td>MDMT total</td>
<td>(-.40)</td>
<td>(-.40)</td>
<td>(.92)</td>
<td>(.87)</td>
<td></td>
</tr>
<tr>
<td>RMT Words</td>
<td>(-.40)</td>
<td>(-.41)</td>
<td>(.92)</td>
<td>(.87)</td>
<td></td>
</tr>
</tbody>
</table>

Note: WCMT = Word Completion Memory Test; MFIT = Memory for 15 Items Test; DCT = Dot Counting Test; MDMT = Multi-Digit Memory Test; RMT = Recognition Memory Test.

* \( P < .008 \).
Table 4: Divergent validity of the WCMT_R score in controls and simulators (N=60)

<table>
<thead>
<tr>
<th>K-BIT IQ</th>
<th>K-BIT Voc</th>
<th>K-BIT Mat</th>
<th>FTT DH</th>
<th>FTT NDH</th>
<th>GP DH</th>
<th>GP NDH</th>
<th>GS DH</th>
<th>GS NDH</th>
<th>BNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCMT_R score</td>
<td>.40*</td>
<td>.38*</td>
<td>.30</td>
<td>.19</td>
<td>.15</td>
<td>.43*</td>
<td>-.35</td>
<td>-.04</td>
<td>-.02</td>
</tr>
<tr>
<td>K-BIT Comp IQ</td>
<td>–</td>
<td>.80*</td>
<td>.66*</td>
<td>.18</td>
<td>.15</td>
<td>.35</td>
<td>-.30</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td>K-BIT Voc</td>
<td>–</td>
<td>–</td>
<td>.37*</td>
<td>.19</td>
<td>.11</td>
<td>-.23</td>
<td>-.20</td>
<td>.10</td>
<td>.05</td>
</tr>
<tr>
<td>K-BIT Mat</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.12</td>
<td>.13</td>
<td>-.35</td>
<td>-.28</td>
<td>.00</td>
<td>.02</td>
</tr>
<tr>
<td>FTT DH</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.89*</td>
<td>-.27</td>
<td>-.24</td>
<td>-.01</td>
<td>-.02</td>
</tr>
<tr>
<td>FTT NDH</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>-.28</td>
<td>-.21</td>
<td>.05</td>
<td>.00</td>
</tr>
<tr>
<td>GP DH</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.83*</td>
<td>-.01</td>
<td>-.08</td>
</tr>
<tr>
<td>GP NDH</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.01</td>
<td>-.05</td>
</tr>
<tr>
<td>GS DH</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.91*</td>
</tr>
<tr>
<td>GS NDH</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.17</td>
</tr>
</tbody>
</table>

Note: WCMT = Word Completion Memory Test; K-BIT = Kaufman Brief Intelligence Test; Comp = Composite; Voc = Vocabulary; Mat = Matrices; FTT = Finger Tapping Test; DH = dominant hand; NDH = nondominant hand; GP = Grooved Pegboard; GS = Grip Strength; BNT = Boston Naming Test.

*P < .005.

Table 4 reveals that the WCMT_R score did not correlate significantly with the BNT, FTT, GS, GP nondominant hand performance, and K-BIT Matrices (i.e., r ranging from -.04 to -.35). However, the WCMT_R score correlated significantly, albeit moderately, with K-BIT Composite IQ and Vocabulary (i.e., r=.40 and .38, respectively) and dominant hand performance on GP (i.e., r=.43). A closer inspection of the data revealed that the significant correlations between WCMT_R score and K-BIT Composite IQ and Vocabulary resulted from significant correlations in the SIM group (r=.43 and .41, respectively), as the correlations for these variables were not significant in the CON group (r=.21 and .02, respectively). Thus, when participants were performing to the best of their abilities, there were no significant relationships between the WCMT_R score and measures of intellectual ability. In contrast, only the CON group obtained a significant correlation between the WCMT_R score and GP dominant hand performance (r=-.42), while the SIM group did not (r=-.32). Similar results were obtained with the other measures of simulation in that significant relationships, when present, were found in the simulation group only.

2.4. Test-retest reliabilities

Test-retest reliabilities for the WCMT and other simulation measures are shown in Table 5. Results revealed that all simulation measures except DCT total errors were found to possess 2-week test-retest reliability coefficients of .70 or above. Analyses of the 2-week test-retest reliabilities for the nonsimulation measures ranged from .63 on Grooved Pegboard nondominant hand to .93 for both dominant and nondominant hand Grip Strength. K-BIT Matrices was the only other nonsimulation measure with a reliability coefficient below .70 (i.e., r=.66).
Table 5
Two-week test-retest reliabilities for simulation measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pearson’s r</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCMT I score</td>
<td>.92</td>
</tr>
<tr>
<td>WCMT E score</td>
<td>.86</td>
</tr>
<tr>
<td>WCMT R score</td>
<td>.94</td>
</tr>
<tr>
<td>MFIT total correct</td>
<td>.94</td>
</tr>
<tr>
<td>DCT total errors</td>
<td>.62</td>
</tr>
<tr>
<td>MDMT total correct</td>
<td>.93</td>
</tr>
<tr>
<td>RMT Words</td>
<td>.94</td>
</tr>
<tr>
<td>RMT Faces</td>
<td>.89</td>
</tr>
</tbody>
</table>

Note: WCMT = Word Completion Memory Test; MFIT = Memory for 15 Items Test; DCT = Dot Counting Test; MDMT = Multi-Digit Memory Test; RMT = Recognition Memory Test.

2.5. Classification accuracies

Discriminant function analyses (DFA) were utilized to investigate the hypothesis that the WCMT would show superior classification accuracy over the other simulation measures and to explore the incremental validity of each measure. In the first DFA, all participants (CON, SIM, and MI groups) were classified into one of two groups, simulators or nonsimulators, after adjusting for prior probabilities due to group differences in size. The following variables, WCMT R score, MFIT total correct, and DCT total errors were entered into the DFA simultaneously. Results of the DFA were significant \(\chi^2 (6, N = 71) = 83.3, P < .001\). Total classification accuracy was 95.7%, with 97.6% of CON and MI participants correctly classified as nonsimulators and 93.3% of the SIM group correctly classified as simulators. One MI participant was misclassified as a simulator and two SIM participants were misclassified as nonsimulators.

When the independent variables were entered stepwise to determine incremental validity of each, results revealed that the WCMT R score entered into the DFA at step 1, accounting for 71% of the variance. None of the remaining indicators added incremental validity to the DFA.

Because the MI group was not administered all measures of simulation, another DFA was performed using only the CON and SIM groups. The DFA again was used to classify participants into one of two groups, simulators or nonsimulators. The following variables were entered into the DFA simultaneously: WCMT R Score, MFIT total correct, DCT total errors, MDMT total correct, RMT Words and RMT Faces. Results of this DFA were significant \(\chi^2 (6, N = 60) = 75.6, P < .001\), with 100% and 93.3% of the CON and SIM groups, respectively, correctly classified. Two of the SIM participants were incorrectly classified as nonsimulators, resulting in a total classification accuracy of 96.7%. When the six independent variables were entered stepwise into a DFA, WCMT R score entered first, accounting for 70% of the variance, and RMT Words entered at step 2, accounting for another 5% of the variance. None of the remaining four indicators added significant incremental validity.

Visual analyses were performed to identify cut-off scores for each simulation indicator that would provide the best classification accuracy in the present sample. Also, the validity and effectiveness for each measure was computed. A measure is considered valid when sensitivity divided by the false positive error rate is numerically greater than the false negative error rate.
Table 6
Cut-off scores and classification accuracies of simulation measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Cut-off</th>
<th>CON (%)</th>
<th>SIM (%)</th>
<th>MI (%)</th>
<th>Total (%)</th>
<th>VAL</th>
<th>EFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCMT I score</td>
<td>&lt;16</td>
<td>100</td>
<td>80.0</td>
<td>100</td>
<td>91.6</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>WCMT E score</td>
<td>&gt;11</td>
<td>100</td>
<td>56.7</td>
<td>100</td>
<td>81.7</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>WCMT R score</td>
<td>&lt;9</td>
<td>100</td>
<td>93.3</td>
<td>100</td>
<td>97.2</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>MFIT total correct</td>
<td>&lt;12</td>
<td>100</td>
<td>50.0</td>
<td>81.18</td>
<td>76.1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>DCT total errors</td>
<td>&gt;4</td>
<td>100</td>
<td>23.3</td>
<td>NA</td>
<td>66.2</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>MDMT total correct</td>
<td>&gt;66</td>
<td>100</td>
<td>90.0</td>
<td>95.0</td>
<td>95.0</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RMT Words</td>
<td>&lt;40</td>
<td>100</td>
<td>86.7</td>
<td>NA</td>
<td>93.3</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RMT Faces</td>
<td>&lt;33</td>
<td>100</td>
<td>66.7</td>
<td>NA</td>
<td>83.3</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Note: CON = controls; SIM = simulators; MI = memory-impaired; VAL = validity; EFF = effectiveness; WCMT = Word Completion Memory Test; MFIT = Memory for 15 Items Test; DCT = Dot Counting Test; MDMT = Multi-Digit Memory Test; RMT = Recognition Memory Test.

3. Discussion

The need for a simulation measure designed to detect more sophisticated attempts at simulating cognitive deficits has been noted by researchers (Guilmette, Hart et al., 1994; Guilmette, Sparadeo et al., 1994). The WCMT was designed specifically for this purpose, as it was developed using well-coached analogue simulators as the criterion group. The WCMT has shown promise as such in initial validity studies, correctly classifying 100% of control and memory-impaired participants and 93% of sophisticated analogue simulators in one study (Hilsabeck et al., 2001) and 100% of controls and analogue simulators in another (Merten et al., 2004). The present study examined the psychometric properties of the WCMT and compared its discriminant ability with existing measures of simulation.

Construct validity of the WCMT was demonstrated via significant group differences. Results showed that controls and memory-impaired patients obtained significantly higher WCMT R and I scores and significantly lower E scores than analogue simulators. Controls obtained significantly higher WCMT R and I scores than memory-impaired patients, but the differences were relatively small and not clinically significant (i.e., 3.23 and 2.42, respectively). The WCMT was found to possess acceptable convergent validity as indicated by significant correlations with existing simulation measures. Although all these relationships were significant, there was enough variability to suggest that the WCMT would provide non-redundant information if administered in conjunction with the other measures of simulation. The lowest significant correlation was found with DCT total errors. An examination of simulating participants’ responses on the post-experiment questionnaire revealed that many simulators did not perceive the DCT as a measure of memory, and therefore, did not simulate memory divided by specificity, and a measure is considered effective when the base rate for simulation is numerically greater than the measure’s false positive plus false negative error rates (Hayes et al., 1999). Table 6 presents the cut-off score, classification accuracy, validity, and effectiveness of each simulation measure.
impairment. This finding explains its relatively low correlation with WCMT \( R \) score, as well as with the remaining simulation measures. Additional evidence of convergent validity for the WCMT can be inferred from the significant group differences found for the other simulation measures. Convergent validity of the other simulation measures also was confirmed, although the MDMT and RMT correlated strongly, suggesting administration of both measures may provide redundant information.

Evidence for divergent validity of the WCMT was indicated by nonsignificant correlations with seven out of 10 nonsimulation measures. The WCMT \( R \) score correlated significantly with K-BIT Composite IQ and K-BIT Vocabulary, but these relationships were moderate, and examination of these relationships in control and simulating groups separately revealed significant correlations in the simulation group only. These findings may be explained by simulators attempting to show memory difficulties on Vocabulary items. The moderately significant negative correlation between WCMT \( R \) score and dominant hand GP performance may be the result of a mild memory component inherent in the test. Support for a memory component in GP is provided by Poulton and Moffitt (1995), who found a mild, yet significant, correlation between GP and a measure of nonverbal memory, the Rey-Osterrieth Complex Figure Test. Discriminant validity of the WCMT is indicated further by nonsignificant group differences on all nonsimulation measures. Acceptable divergent validity of the other simulation measures also was found.

With regard to 2-week test-retest reliability, the WCMT was found to be reliable, with reliability coefficients well above the recommended standard of .70 (Nunnally, 1978). Test-retest reliabilities were .94, .92, and .86 for WCMT \( R \), \( I \), and \( E \) scores, respectively. All other simulation indicators demonstrated test-retest reliabilities above .69, as well. MDMT total correct and RMT Words evidenced the highest test-retest reliabilities of the remaining simulation indicators. Most divergent neuropsychological measures were found to possess adequate test-retest reliabilities, indicating performances of this study’s participants were similar to those in prior studies of test-retest reliability. Only K-BIT Matrices and nondominant hand GP failed to possess 2-week test-retest reliabilities above .69.

Another primary purpose of this study was to compare the classification accuracy of the WCMT to other measures of simulation. The WCMT exhibited slightly superior classification accuracy. Using a cut-off of <9 for WCMT \( R \) score as indicative of simulating, no control or memory-impaired participant was incorrectly classified as a simulator, and only two simulating participants were incorrectly classified as nonsimulators, resulting in an overall correct classification hit rate of 97.2%. Thus, the WCMT was determined to be a valid and effective measure of simulation. These results are similar to those reported in prior studies (Hilsabeck et al., 2001; Merten et al., 2004). In addition, the WCMT \( R \) score consistently entered first into DFAs and accounted for more variance than the other measures of simulation.

In summary, the WCMT demonstrated adequate psychometric properties, including construct validity, convergent and divergent validity, and 2-week test-retest reliability. In addition, the WCMT exhibited slightly superior classification accuracy over existing simulation measures, correctly classifying 100% of controls and memory-impaired patients and 93.3% of analogue simulators for an overall classification accuracy rate of 97.2%. Further, the WCMT consistently entered into DFAs first and accounted for more variance than any other simulation measure. These results suggest the WCMT is a promising measure of simulated memory impairment. A limitation of the present study is that only three of the
six simulation measures were administered to the memory-impaired patients due to time constraints. Further, the memory-impaired group was small and heterogeneous with regard to the etiology of memory impairment.

Future research should compare the validity and effectiveness of the WCMT, as well as other simulation measures, in a variety of clinical samples, especially clinical samples suspected of simulation (i.e., known-groups design). Studies further exploring the ability of the WCMT to discriminate control participants from circumscribed neuropsychological populations would be beneficial at identifying limits to its applicability. Comparison of the utility of the WCMT and newer measures of simulation, such as the Test of Memory Malingering (Tombaugh, 1996), Word Memory Test (Green, 2003), and Amsterdam Short-term Memory Test (Schagen et al., 1997) also is needed.

Finally, it is important to remember that use of systematic multitrait-multimethod strategies as suggested by Campbell and Fiske (1959) are important in the assessment of any psychological construct. Therefore, use of the WCMT, or any simulation measure, by itself is not sufficient to draw conclusions about an examinee’s intent to mangle. Rather, a combination of domain-specific simulation measures, neuropsychological tests, interview data, physiological measures, collateral information, and self-report measures will likely provide the best diagnostic accuracy (Martin, Franzen, & Orey, 1998).

Acknowledgements

The study reported in this article was derived from the dissertation of Robin C. Hilsabeck. The authors are indebted to Sid O’Bryant for help with participant recruitment and data collection. Portions of this article were presented in 1998 at the 18th annual meeting of the National Academy of Neuropsychology in Washington, DC.

References


