Failure Base Rates for the Word Reading Test

Virginia Kleman

*Pacific University*
Failure Base Rates for the Word Reading Test

Abstract
Research has demonstrated that effort accounts for over 50% of score variability on neuropsychological tests, yet effort can be difficult to evaluate (Green, Rohling, LeesHaley, & Allen, 2001). Because of the integral role effort plays in neuropsychological testing, specific measures called symptom validity tests (SVTs) have been developed to determine if a client’s level of effort is sufficient during testing. While most SVTs are face-valid as memory tests (Lezak, Howieson, & Loring, 2004), some research suggests that domain-specific (or symptom specific) SVTs are necessary when evaluating clients for conditions or disorders not normally associated with “memory problems” (Osmon, Plambeck, Klein, & Mano, 2006). The Word Reading Test (WRT) was developed to address this domain-specific issue and is designed to be used in Learning Disability (LD) evaluations (Osmon et al., 2006). The purpose of the present study was to contribute to the normative base of the WRT by determining the base rate of failure in a clinical population. Subjects were 30 outpatients referred to a university doctoral clinical psychology training and research clinic for neuropsychological evaluation for academic purposes. Using the recommended cut-off score of ~4 errors, three participants (10%) failed the test; using a cut-off of ~3 errors, four participants (13%) failed the test. There were no significant differences in WRT scores for age, years of education, presence of secondary gain, WAIS-III indices, or diagnostic categories.

Degree Type
Thesis

Rights
This work is licensed under a Creative Commons Attribution-No Derivative Works 3.0 License.
Copyright and terms of use

If you have downloaded this document directly from the web or from CommonKnowledge, see the “Rights” section on the previous page for the terms of use.

If you have received this document through an interlibrary loan/document delivery service, the following terms of use apply:

Copyright in this work is held by the author(s). You may download or print any portion of this document for personal use only, or for any use that is allowed by fair use (Title 17, §107 U.S.C.). Except for personal or fair use, you or your borrowing library may not reproduce, remix, republish, post, transmit, or distribute this document, or any portion thereof, without the permission of the copyright owner. [Note: If this document is licensed under a Creative Commons license (see “Rights” on the previous page) which allows broader usage rights, your use is governed by the terms of that license.]

Inquiries regarding further use of these materials should be addressed to: CommonKnowledge Rights, Pacific University Library, 2043 College Way, Forest Grove, OR 97116, (503) 352-7209. Email inquiries may be directed to: copyright@pacificu.edu

This thesis is available at CommonKnowledge: http://commons.pacificu.edu/spp/20
FAILURE BASE RATES FOR THE WORD READING TEST

A THESIS

SUBMITTED TO THE FACULTY

OF

SCHOOL OF PROFESSIONAL PSYCHOLOGY

PACIFIC UNIVERSITY

HILLSBORO, OREGON

BY

VIRGINIA M.M. KLEMAN

IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

OF

MASTER OF SCIENCE IN CLINICAL PSYCHOLOGY

APRIL 17, 2009

APPROVED.

Michael Daniel, Ph.D.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>REVIEW OF LITERATURE</td>
<td>6</td>
</tr>
<tr>
<td>Effort in Neuropsychological Testing</td>
<td>6</td>
</tr>
<tr>
<td>The Measurement of Effort</td>
<td>7</td>
</tr>
<tr>
<td>Identifying Effort in Standard Neuropsychological Tests</td>
<td>8</td>
</tr>
<tr>
<td>The Development of Symptom Validity Testing</td>
<td>10</td>
</tr>
<tr>
<td>The Next Generation of Symptom Validity Test</td>
<td>11</td>
</tr>
<tr>
<td>General-Global vs. Domain Specific Tests</td>
<td>13</td>
</tr>
<tr>
<td>Symptom Validity Testing with Learning Disability and ADHD Evaluations</td>
<td>14</td>
</tr>
<tr>
<td>The Creation of the Word Reading Test</td>
<td>15</td>
</tr>
<tr>
<td>The Purpose of the Study</td>
<td>17</td>
</tr>
<tr>
<td>METHODS</td>
<td>18</td>
</tr>
<tr>
<td>RESULTS</td>
<td>21</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>29</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>32</td>
</tr>
</tbody>
</table>
ABSTRACT

Research has demonstrated that effort accounts for over 50% of score variability on neuropsychological tests, yet effort can be difficult to evaluate (Green, Rohling, Lees-Haley, & Allen, 2001). Because of the integral role effort plays in neuropsychological testing, specific measures called symptom validity tests (SVTs) have been developed to determine if a client's level of effort is sufficient during testing. While most SVTs are face-valid as memory tests (Lezak, Howieson, & Loring, 2004), some research suggests that domain-specific (or symptom specific) SVTs are necessary when evaluating clients for conditions or disorders not normally associated with "memory problems" (Osmon, Plambeck, Klein, & Mano, 2006). The Word Reading Test (WRT) was developed to address this domain-specific issue and is designed to be used in Learning Disability (LD) evaluations (Osmon et al., 2006). The purpose of the present study was to contribute to the normative base of the WRT by determining the base rate of failure in a clinical population. Subjects were 30 outpatients referred to a university doctoral clinical psychology training and research clinic for neuropsychological evaluation for academic purposes. Using the recommended cut-off score of ≥ 4 errors, three participants (10%) failed the test; using a cut-off of ≥ 3 errors, four participants (13%) failed the test. There were no significant differences in WRT scores for age, years of education, presence of secondary gain, WAIS-III indices, or diagnostic categories.
LIST OF TABLES

Table 1 Means and Standard Deviations for Participant Variables ........................................... 21
Table 2 Independent T-Tests for WRT Pass/Fail Using a Cut-off Score of $\leq 4$ ......................... 22
Table 3 Chi-Square Results for WRT Pass/Fail Using a Cut-off Score of $\leq 4$ and Presence of Secondary Gain ............................................................................................................. 23
Table 4 Chi-Square Results for WRT Pass/Fail Using a Cut-off Score of $\leq 4$ and Final Diagnosis .................................................................................................................................................. 24
Table 5 Independent T-Tests for WRT Pass/Fail Using a Cut-off Score of $\leq 3$ ......................... 25
Table 6 Chi-Square Results for WRT Pass/Fail Using a Cut-off Score of $\leq 3$ and Presence of Secondary Gain ............................................................................................................. 26
Table 7 Chi-Square Results for WRT Pass/Fail Using a Cut-off Score of $\leq 3$ and Final Diagnosis .................................................................................................................................................. 26
Table 8 One-way ANOVA Comparisons of WRT Total Errors Means Between Final Diagnoses .................................................................................................................................................. 27
Table 9 One-way ANOVA Comparisons of Reaction Time Means Between Final Diagnoses .......... 28
INTRODUCTION

Neuropsychological assessment is often utilized during the diagnostic process for a variety of psychological and medical disorders. Recently, there has been increasing research on the influence of effort on the validity of neuropsychological test results. Effort accounts for about 50% of score variability among neuropsychological test findings when all other factors are held constant (Green, Rohling, Lees-Haley, & Allen, 2001). These findings have led to an increased emphasis on the importance of assessing effort during testing as well as the development of more sophisticated tests designed to measure effort, called symptom validity tests (SVTs).

Some of the more obvious factors that influence effort and ability to concentrate are ambient stimuli not related to the task at hand. Room temperature, lighting, and noise can all affect a client’s ability to put forth full effort. Likewise, physical and medical conditions of the client, such as hunger, illness, injury, or pain can also affect his ability to concentrate. Psychological factors also play a role in determining amount of effort put forth. Within the context of a neuropsychological examination, the client may feel anxious about his performance and the outcome of the examination, or he may be experiencing any number of psychological symptoms, including feelings of depression, excitement, or anxiety, all of which could affect his effort (Lezak, Howieson, & Loring, 2004).
Effort can also be influenced by the presence of external motives or incentives. These external motives, referred to as secondary gains, are often present in legal contexts and other circumstances where the client stands to benefit from a diagnosis of impairment. Secondary gains may include workers compensation payments, disability benefits, academic tutoring and accommodations, or less punitive treatment from the judicial system (Bianchini, Mathias & Greve, 2001). When clients consciously choose to modify their performance in an attempt to obtain external incentives, their behavior is termed malingering (Lezak et al., 2004).

For these reasons, it is critical that clinicians be able to objectively describe clients' level of effort. If a clinician cannot accurately account for a client's effort, it is not possible to know if the test scores produced by the client are valid. Traditionally, client effort has been measured through behavioral observations and clinical judgment (Lezak et al., 2004). But research on the reliability and validity of clinical judgment has shown that clinicians have only been able to detect deceit approximately half of the time (Cripe, 2002). The classic research by Paul Meehl, in which clinical judgment is compared to statistical prediction demonstrates that the objective, formal method of statistical prediction is at least equal, if not superior to, subjective clinical judgment in describing or predicting behavior (Meehl, 1954).

Because of the limited reliability of clinical judgment, researchers have suggested ways of analyzing the internal consistency of test scores to evaluate effort. Both formulas and cut-off scores on the Wechsler Adult Intelligence Scale – III (WAIS-III; The Psychological Corporation, 1997), the Wechsler Memory Scale – III (WMS-III; The Psychological Corporation, 1997), the California Verbal Learning Test (CVLT; Delis,
Kramer, Kaplan, & Ober, 2000), and the Wisconsin Card Sorting Test (WCST; Heaton, Chelune, Talley, Kay, & Curtiss, 2000) have been developed to help identify inadequate effort. However, given the potential importance of many neuropsychological evaluations, especially when secondary gains are relevant, it is clear that specific effort tests must be integrated into neuropsychological test protocols. These tests can offer a concrete, objective measure of client effort, which in turn strengthens confidence that test scores are valid.

Early symptom validity tests included the Digit Memory Test (Hiscock & Hiscock, 1989), the Dot Counting Test (Rey, 1941), and the Portland Digit Recognition Test (Binder & Willis, 1991), which paved the way for the next generation of SVTs, including the Test of Memory Malingering (TOMM; Tombaugh, 1997), the Word Memory Test (WMT; Green, Allen, & Astner, 1996), and the Medical Symptom Validity Test (MSVT; Richman et al., 2006). All of these tests, save for the Dot Counting Test (Rey, 1941), employ a memory test format to detect suboptimal effort. But recent research has raised questions about using a memory test format to assess effort in some types of neuropsychological evaluations. Osmon, Plambeck, Klein, and Mano, (2006) distinguish between a “general-global” orientation and a “domain-specific” orientation in SVTs. In the “general-global” model, suboptimal effort would be evidenced in memory-based SVTs regardless of the purpose of the neuropsychological evaluation or the particular disorder being assessed. The “domain-specific” model assumes that suboptimal effort will only be evident on tests that involve the type of cognitive ability specific to the disorder being evaluated. Osmon et al. (2006) argued that SVTs should be based on a “domain-specific” perspective and employ test formats that are specific to a symptom of
the disorder for which the client is being evaluated. Therefore, memory-based SVTs would be used for evaluation of disorders for which memory impairment is a salient symptom, such as dementia, while language-based SVTs would be used for language-related disorders, such as Learning Disability of Written Expression or Reading Disorder. According to Osmon et al. (2006) and the domain-specific hypothesis, someone who intentionally gives inadequate effort may be more likely to do so on tests that appear to be related to the symptoms that they assume are part of the disorder for which they are being evaluated. Conversely, clients will be less likely to give inadequate effort on tests that do not appear to be related to the disorder for which they are being evaluated.

To test this hypothesis, Osmon et al. (2006) devised a new SVT called the Word Reading Test (WRT), designed to detect suboptimal effort unique to clients referred for Learning Disability (LD). Using this test, Osmon et al. (2006) compared three groups of student participants, two of which were asked to simulate different symptoms. One group presented as though they had reading difficulties, one group completed the test with slowed processing speed, and the third group was asked to give their best effort.

In comparing these three groups, the researchers found that the reading simulators produced the lowest accuracy scores, while the processing speed simulators produced the slowest reaction times. These results demonstrated that the simulators behaved differently, depending on the type of symptomatology they were attempting to effect. These findings have direct implications for the debate between general-global and domain-specific SVTs, indicating that a client with suboptimal effort will behave differently depending on the type of evaluation.
The current study will extend Osmon et al.’s (2006) original study with simulators by using clients referred for neuropsychological testing for academic purposes.
Effort in Neuropsychological Testing

Effort significantly influences the results of a neuropsychological evaluation. For this reason, clinicians must be able to account for clients' effort in order to make valid conclusions and recommendations based on their test performance. Particularly in those cases where no previous diagnosis of cognitive disorder or brain injury exists, the results of a neuropsychological evaluation may be the only evidence of any cognitive inefficiency and therefore must be held to the most stringent measures of validity (Lezak et al., 2004).

As previously discussed, effort may be affected by any number of variables, including physical health, psychological adjustment, environmental distracters, or the presence of an external incentive. Clients with somatic and factitious disorders may also put forth inadequate effort, which is directly related to their psychological diagnosis. Regardless of the cause of poor effort, it is clear that a valid and reliable method of assessing effort is needed.

Lezak et al. (2004) suggest that four factors be considered when evaluating effort: the consistency in performance among tests and reported history, how well the neuropsychological profile and reported symptomatology fit the description of a known disease or disorder, the client's personal and social history, as well as his current
stressors, and the manner in which the client describes his symptoms, including his emotional reactions.

The Measurement of Effort

In order to address these factors, Slick et al. (1999) offer a formalized set of criteria to consider when evaluating possible neurocognitive deficits. The first criterion is the presence of an external incentive or secondary gain. The second factor is evidence of either fabrication or exaggeration of symptomatology as demonstrated by at least one of the following: negative response bias on at least one forced-choice test of cognitive function, performance on a symptom validity test that is consistent with substandard effort, discrepancy between test results and known brain functioning, discrepancy between test results and behavioral observations of the client, discrepancy between test results and the client’s daily activities, or discrepancy between test results and background information (Slick et al., 1999).

The third factor to consider is any discrepancy between the client self-report and other objective sources. These sources include but are not limited to: self-reported history and documented history, self-reported symptoms and known patterns of brain functioning, self-reported symptoms and behavioral observations, self-reported symptoms and reports of other objective observers, or evidence of either fabrication or exaggeration of psychological dysfunction (Slick et al., 1999).

The final consideration according to Slick et al. (1999) is that the factors described are not better accounted for by some other psychiatric, neurological or developmental disorder. Based on the specific combination of factors, Slick et al. (1999) propose three diagnostic categories for malingering cognitive dysfunction (MND):
definite, probable, and possible. According to these criteria the most compelling evidence of suboptimal effort is negative response bias on a forced-choice test. Although performance on SVT are of paramount importance in assessing effort, performance on some standard neuropsychological tests can offer information regarding client effort as well.

Identifying Effort in Standard Neuropsychological Tests

Larrabee (2003) found that questionable effort and malingering can be identified through specific patterns of performance on a range of neuropsychological tests, including those for problem-solving, motor functioning, visual perception, memory, and attention. Methods have been developed for analyzing scores or responses for inadequate effort on each of the following tests: the WAIS-III (The Psychological Corporation, 1997), the WMS-III (The Psychological Corporation, 1997), the Rey Auditory Verbal Learning Test (AVLT; Schmidt, 1996), the CVLT (Delis et al., 2000), and the WCST (Heaton et al., 2000). When the overall test profile for a neuropsychological examination is considered with behavioral information and SVT results, the clinician’s ability to account for client effort is significantly strengthened (Lezak et al., 2004).

On the WAIS-III, specific subtests such as Digit Span and Vocabulary can offer insight into a client’s effort. With clients who have no history of brain damage or attention problems, a scaled score of less than 7 on Digit Span is highly unusual and can provide evidence to suggest poor effort (Trueblood, 1994). Reliable digit span, or the sum of the highest number of digits correctly recalled in both forward and back conditions, can also serve as an indicator of substandard effort when equal to 7 or less (Greiffenstein, Baker, & Gola, 1994). Additionally, if a client’s Vocabulary scaled score is considerably
higher than the Digit Span scaled score, effort is called into question (Mittenburg, Theroux-Fichera, Zielinski, & Heilbronner, 1995).

Memory tests like the WMS-III are particularly useful for examining client effort because these tests usually contain both recognition and recall items. In good effort clients, the number of free recall responses will be lower than the number of cued recognition response. If a client consistently recognizes fewer words than he or she can recall without cues, effort should be evaluated further. Additionally, Killgore and DellaPietra (2000a,b) identified six items from the delayed recognition portion of Logical Memory that 70-80% of clients consistently answered correctly. The researchers devised the Rarely Missed Index from weighting the six items, which has successfully discriminated between good and poor effort volunteers with 98% accuracy.

The AVLT and CVLT require clients to recall a list of words that is read to them five times. It is expected that recall will increase as the word list is repeated, and that recognition and cued recall scores will be higher than free recall scores. In both of these tests, if the client recognizes the same number or fewer words than they immediately recalled following the distraction trial, effort is definitely a concern. However, none of these guidelines hold true with clients who have severe brain injury (Lezak et al., 2004).

The WCST requires clients to modify their response strategy to stimuli six different times over the course of the test. Most good-effort clients who have no cognitive deficits will recognize the change in demanded response and modify their response strategy. Suhr and Boyer (1999) found that poor effort volunteers were more likely to fail to maintain the response set and spontaneously alter their response strategy despite receiving positive feedback. Bernard, McGrath and Houston (1996) found that poor-
effort volunteers were more likely to perseverate on the WCST. These findings suggest both failure to maintain set and perseverative responses can be indicators of inadequate effort.

*The Development of Symptom Validity Testing*

While many standard neuropsychological tests can provide some information regarding client effort, specific SVTs allow a more complete assessment of effort. The pioneer of effort testing, Loren Pankratz, coined the term ‘symptom validity testing’ and described measuring client effort by “test[ing] exactly what the patient says he can’t do. You devise a test for each person, an individual strategy for each patient. In this way, you motivate the patient to demonstrate the deficiency in what he says or believes he can’t do” (Lezak et al., 2004, p. 770).

Pankratz’s model is adaptable and works well because it is straightforward, face valid, and because any deviation from the norm is obvious. The model is logical and largely accurate when replicated, but it lacks practicality and has very limited reliability (Lezak et al., 2004). Therefore, specific, standardized, and psychometrically-sound SVTs are necessary.

The first SVT to build upon Pankratz’s research was the Digit Memory Test, devised by Hiscock & Hiscock (1989). The Digit Memory Test consists of three blocks of 24 trials in which five-digit strings of numbers are presented. The client is required to correctly choose the previous number from two options after a time delay. The delay period between the presentation of the numbers and client recall increases throughout the test, providing the semblance of increased difficulty. The Digit Memory Test has proven to be quite easy for all clients except for those with severe cognitive dysfunction, so even
a few errors in recall are sufficient to question client effort (Prigatano & Amin, 1993; Guilmette, Hart & Giuliano, 1993).

The successor to the Digit Memory Test was the Portland Digit Recognition Test (PDRT), created by Binder and Willis (1991). The PDRT follows the same format as the Digit Memory Test except during the delay periods a counting distracter is used. Like the Digit Memory Test, three item sets of increasing delay periods are used, which in combination with the counting distracter produce the illusion of a memory test that increases in difficulty over time. The PDRT is unique in that it has demonstrated sensitivity to suboptimal effort in various clinical populations, in addition to simulator groups. In one study comparing clients with suspected brain injury and the possibility of secondary gains against clients with confirmed brain injury and no secondary gains, over 25% of the compensation-seeking clients scored below the worst performance of all the confirmed brain injury no secondary gain clients (Binder & Willis, 1991).

In contrast to the Digit Memory Test and the PDRT’s forced-choice format, the Dot Counting Test (Rey, 1941) requires clients to count dots presented on index cards as quickly as they can. Cards with ungrouped, randomly scattered dots should take good effort clients longer to count than cards with systematically grouped dots. Any deviation from this pattern signals poor effort and incorrect counts could be a sign of malingering (Binks, Gouver, Waters, 1997).

The Next Generation of SVTs

The Digit Memory Test, PDRT, and the Dot Counting Test rely on below-chance cut-off scores to identify poor effort. However, subsequent research demonstrated that even a few errors on these tests, which is a much higher level of performance than
chance, are suggestive of suboptimal effort. Pankratz's version of symptom validity testing would also be considered a below-chance cutoff, since a failure is attributed to any performance that is worse than 50% correct. These SVTs do not require the use of norms and score interpretation is very simple (Bianchini et al., 2001). However, the items on these SVTs are exceedingly easy and the correct answers are obvious. Therefore, many malingering clients will not perform worse than chance and pass the test, which may in turn give the rest of their test results credence with the impression of adequate effort.

The next generation of SVTs were developed in response to the limitations of below-chance cut-off tests. Recently developed SVT's rely on normed scores produced by good-effort volunteers and malingering simulators to develop expected score ranges for adequate effort. Three such tests are the Test of Memory Malingering (TOMM; Tombaugh, 1997), the Word Memory Test (WMT; Green, Allen, & Astner, 1996), and the Medical Symptom Validity Test (MSVT; Richman et al., 2006).

The TOMM is a recognition test utilizing line drawings of common objects. Fifty drawings are presented to clients one at a time. Clients must correctly discriminate between repeated and novel pictures both immediately and after a delay. Even in client populations that were diagnosed with neurological impairments, clients were still able to achieve over a 92% correct response rate. Additionally, age and level of education did not appear to influence performance on the TOMM (Tombaugh, 1997).

The WMT (Green et al., 1996) requires the client to learn twenty semantically related word pairs and then recall and recognize the words after a delay. Similar to the TOMM, performance on the WMT is largely unaffected by all but the most extreme
forms of cognitive impairment (Green et al., 1996). Additionally, age and years of education do not have a significant impact on WMT performance (Green et al., 1996). The WMT is unique compared to other SVTs in that it was originally normed on clinical forensic patients, as opposed to simulators. It has also proven to be more sensitive to suboptimal effort than the TOMM; in a study of 519 cases in which secondary gains were present, 11% of clients failed the TOMM, while 32% failed the WMT (Gervais, Rohling, Green, & Ford, 2004).

The MSVT (Richman et al., 2006) was designed to be used by physicians. Like the WMT, the MSVT requires the client to learn ten semantically related word pairs. After a delay, the client must correctly recognize and recall the word pairs. The MSVT is extraordinarily easy; the authors found that clients who were tested in a foreign language that they did not speak were able to perform almost perfectly on the test. In contrast, 42% of compensation-seeking clients tested in their own language failed the test (Richman et al., 2006).

**General-Global vs. Domain-Specific Tests**

As with the TOMM, WMT, and MSVT, the majority of SVTs have been developed using the layperson’s knowledge or conception of brain damage as a guide (Lezak et al., 2004). Laypeople often equate brain damage or cognitive dysfunction with memory problems. Accordingly, most SVTs have been devised to have face validity as tests of memory. Most often, these “memory tests” appear to be much more difficult than they actually are and only the most cognitively impaired clients are unable to attain near perfect scores.
New research has suggested that memory-based SVTs may not be as successful in discriminating between good and poor effort when evaluating clients for conditions or disorders not normally associated with memory problems (Osmon, Plambeck, Klein, & Mano, 2006). For example, memory impairment is not typically associated with learning disorders or ADHD. Therefore, memory-based SVTs may not be sensitive to suboptimal effort in these types of evaluations.

Osmon et al. (2006) identify this issue as an argument for domain-specific, as opposed to general-global SVTs. The general-global hypothesis assumes that variability in effort will be evident across all neuropsychological tests, including memory-based SVTs, regardless of the disorder being evaluated. Conversely, the domain-specific hypothesis holds that variability in effort is rooted in the individual person’s conception of what symptoms are characteristic of a particular disorder. Therefore, under the domain-specific hypothesis, a client who does not have a learning disorder and wants to be diagnosed with one, probably does not believe memory deficits are a consequence of a learning disorder and will not display poor effort on a memory-based SVT. Instead, suboptimal effort will be evident on domain-specific SVTs, which correspond to the symptoms associated with the disorder in question.

Symptom Validity Testing with Learning Disability and ADHD Evaluations

Neuropsychological evaluations for LD and ADHD are unique because of the wide range of external incentives available to clients with these diagnoses. Academic accommodations, disability benefits, and the possibility of stimulant medications are all potential secondary gains for having a diagnosis of LD or ADHD. Additionally, there has been an increase in the number of LD and ADHD evaluations in recent years (Harrison,
And because information about the symptoms of the disorders is widely available, it is fairly simple for clients to misrepresent their abilities. In one sample of 127 college students referred for ADHD evaluations, an estimated 20% significantly exaggerated their symptoms, based on criteria developed by Mittenburg (Harrison, 2006). Sullivan et al. (2007) described a sample of 66 college students who were self-referred for either ADHD, LD, or combined ADHD and LD evaluations. Of the 21 students that were evaluated for ADHD only, 47.6% failed the WMT, showing evidence of decreased effort. Two of the 13 cases (15.4%) evaluated for LD only failed the WMT. Fifty-three of the 66 students were evaluated for ADHD, either individually or in combination with LD, and 24.5% of those students failed the WMT. These results imply that there is a fairly high base rate of SVT failure among clients being evaluated for LD and ADHD, which in turn suggests that scrutinizing effort in these evaluations is particularly important.

**The Creation of the Word Reading Test**

Osmon et al. (2006) devised the Word Reading Test (WRT) to evaluate effort for clients completing LD assessments. Clients are presented with a target word on a computer screen for a brief duration, then two words are presented on the screen without delay. The client must correctly choose between the target word and a foil; the foil is similar to the target word but modified slightly. The foil words contain "errors" which individuals who put forth poor effort may think are typical of mistakes made by people with a learning disorder, such as homophones, mirrored letters, and additions or deletions of letters. Clients are instructed to choose the target word as quickly as possible while maintaining accuracy.
Osmon et al. (2006) compared WRT performance of three groups of students, two of which were simulators. The students were also administered the WMT. One group of simulators was asked to portray clients with reading difficulties, one group was asked to portray clients with slowed thinking associated with learning disabilities, and the third group was asked to give their best effort. With regard to overall accuracy, the good-effort students made the fewest mistakes, while the reading difficulty simulators made the most errors. The slowed thinking simulators scored in between the other two groups on number of errors. When reaction time was considered, the slowed thinking simulators performed the worst and no significant difference was found between the reading difficulty simulators and the good effort students.

Additionally, none of the good-effort students made more than three errors over the course of the test, so Osmon et al. (2006) suggested a cut-off score of 4 or more incorrect responses to be indicative of suboptimal effort. There was little overlap between simulator and non-simulator performance on the WRT, but there were many overlapping scores between simulators and non-simulators on the WMT. However, the failure rate for both simulator groups was similar for the WMT and WRT, which argues against the domain-specific hypothesis. Osmon et al. (2006) concluded that the WMT is effective for distinguishing between malingering and good effort when evaluating reading-related disorders, but it appears that the WRT has better specificity because there was no overlap in scores for simulator and good effort groups. This research demonstrates support for both the domain-specific and general-global hypotheses, suggesting that while decreased effort and malingering can, and often is, demonstrated in a range of cognitive abilities,
poor effort may be most pronounced in tests that are associated with the specific cognitive ability in question.

Osmon et al.'s (2006) use of simulators limits the WRT's application in clinical contexts. The next step in determining the clinical utility of the WRT is to establish the baseline performance of actual patients undergoing evaluation for LD.

The Purpose of the Study

The purpose of this study is extend Osmon et al.'s (2006) original study with simulators by using clients referred for neuropsychological testing for academic purposes. Using Osmon et al.'s (2006) cut-off score of 4 or more incorrect answers, descriptive statistics of clients will be analyzed, including age, years of education, WAIS-III indices, final diagnoses, and reaction time. Additionally, other cut-off scores will be evaluated.
METHOD

Participants

Participants were thirty adults (16 women and 14 men) referred to a university doctoral clinical psychology training and research clinic for neuropsychological evaluation for academic purposes. The mean age for the group was 31.6 (SD = 10.4); mean years of education was 14.1 years (SD = 2.2). Three ethnic groups were identified from the sample: Twenty-six were Caucasian (86.7%), three were Latino (10%), and one was classified as Other (3.3%). Twenty-one (70%) clients were referred for assessment from a college or university health center. Three referrals came from private psychologists (10%), one came from vocational rehabilitation (3.3%), one came from a medical doctor (3.3%), and four came from other sources (13.3%). The reasons for referral were split between four categories: 16 for Learning Disorder (53.3%), 11 for cognitive disorders (36.7%), two for Asperger’s Disorder (6.7%), and one for Attention-Deficit/Hyperactivity Disorder (3.3%). Participants were divided into three groups reflecting the diagnosis they were given after evaluation: 15 clients were diagnosed with a cognitive disorder (50%), eight were diagnosed with ADHD (27%), and seven were diagnosed with other disorders such as Major Depressive Disorder and Asperger’s Disorder (23%).

Medical diagnoses were represented in the group as follows: one endorsed a pain disorder (3.3%), four endorsed migraine headaches (13.3%), one endorsed endocrine
problems (3.3%), and nine described other medical diagnoses (30%). The remaining 50% of the sample group did not endorse any medical diagnoses. Based on interview with clients, family members and review of records, 16 (53%) participants were judged to have potential external incentives for poor test performance, including academic accommodations, the possibility of stimulant medications, and grounds for disability benefits.

Procedure

Prior to testing, participants were informed that their data may be used in future research. The clients who agreed to such use of their information signed an informed consent acknowledging this information and the Institutional Review Board overseeing the clinic approved this use of client data (IRB #046-08). Prior to data analysis, all client data was deidentified and entered into a password-protected computer database kept independent from client records.

Clients were administered a comprehensive neuropsychological battery by clinical psychology graduate students and interns under the supervision of a neuropsychologist. The battery included the following tests: the WAIS-III (The Psychological Corporation, 1997), the Finger Tapping Test from the Halstead-Reitan Neuropsychological Battery (Reitan, 1993), the Grooved Pegboard Test (Trites, 1989), the Woodcock-Johnson-3 (WJ3; Woodcock, McGrew, Mather, & Schrank, 2001) subtests: Letter-Word Identification, Calculation, Math Fluency, Writing Fluency, Writing Samples, Word Attack, and Spatial Relations; the Judgment of Line Orientation (JLO; Benton, Varney, & Hamsher, 1978), the Rey-Osterrieth Complex Figure (Rey-O; Osterrieth, 1944): Copy, Immediate Recall, & Delayed Recall; the NEPSY (Korkman, 1998): Phonological...
Processing, the Boston Diagnostic Aphasia Examination (BDAE; Goodglass, Kaplan, Barresi, & Weintraub, 2001): Boston Naming Test and Complex Ideation Test; the Delis-Kaplan Executive Function System (DKEFS; Delis, Kaplan, & Kramer, 2001) subtests: Verbal Fluency and Trail Making; the Nelson-Denny Reading Test (NDRT; Brown, Fishco, & Hanna, 1993), the WMS-III (The Psychological Corporation, 1997) subtests: Logical Memory I, II, and Recognition, Visual Reproduction I, II, and Recognition, Spatial Span Forward and Backward; the WCST (Heaton et al., 2000), the CVLT2 (Delis et al., 2000), the IVA Continuous Performance Test (IVA; Sanford & Turner, 2007), the MMPI-2 (Hathaway et al., 1989), the MSVT (Richman et al., 2006), and the WRT (Osmon et al., 2004). As recommended by Osmon, ≥ 4 errors was considered WRT failure.

Participants were tested individually over a time period ranging from two days to three weeks. Participants were evaluated from September 2007 through August 2008. The total testing time for each client ranged from 10-12 hours.
RESULTS

Table 1 displays the means and standard deviations for age, education and WAIS-III indices for all participants.

Table 1.

Means and Standard Deviations for Participant Variables

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Years of Education</th>
<th>WAIS-III</th>
<th>WAIS-III</th>
<th>WAIS-III</th>
<th>WAIS-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>31.6</td>
<td>14.1</td>
<td>112.1</td>
<td>108.1</td>
<td>96.3</td>
<td>97.8</td>
</tr>
<tr>
<td>SD</td>
<td>10.4</td>
<td>2.2</td>
<td>14.8</td>
<td>17.5</td>
<td>18.1</td>
<td>12.4</td>
</tr>
</tbody>
</table>

Using a cut-off score of four or more incorrect responses, three clients (10%) failed the WRT. Independent-sample $t$ tests were conducted to examine differences in age, years of education, WAIS-III indices, and reaction time for participants who passed vs. failed the WRT. Table 2 displays the means, standard deviation, $t$ and $p$ values for each independent-sample $t$ test. No significant differences were found for any of the variables, equal variances not assumed.
Table 2.

*Independent T-Test Results for WRT Pass/Fail Using a Cut-off of \( \geq 4 \)*

<table>
<thead>
<tr>
<th></th>
<th>Pass</th>
<th></th>
<th>Fail</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>t</td>
<td>p</td>
</tr>
<tr>
<td>Age</td>
<td>31.6</td>
<td>10.5</td>
<td>31.7</td>
<td>11.6</td>
<td>2.4</td>
<td>1.0</td>
</tr>
<tr>
<td>Years of education</td>
<td>14.2</td>
<td>2.2</td>
<td>14.0</td>
<td>1.0</td>
<td>-0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>WAIS-III VCI</td>
<td>112.4</td>
<td>15.0</td>
<td>109.7</td>
<td>14.6</td>
<td>-0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>WAIS-III POI</td>
<td>108.4</td>
<td>18.4</td>
<td>105.0</td>
<td>4.0</td>
<td>-0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>WAIS-III WMI</td>
<td>95.7</td>
<td>17.7</td>
<td>102.0</td>
<td>24.9</td>
<td>0.4</td>
<td>0.7</td>
</tr>
<tr>
<td>WAIS-III PSI</td>
<td>98.0</td>
<td>12.9</td>
<td>95.7</td>
<td>6.4</td>
<td>-0.53</td>
<td>0.2</td>
</tr>
<tr>
<td>Reaction time</td>
<td>1.0</td>
<td>0.3</td>
<td>1.5</td>
<td>0.34</td>
<td>2.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

(seconds)
A chi-square test of independence was used to analyze differences between pass/fail groups for presence of secondary gain and final diagnosis given in the evaluation. Tables 3 and 4 display these results. There were no significant differences in the proportions of participants in the pass/fail groups that had potential secondary gain or between the three different categories of final diagnoses.

Table 3.

<table>
<thead>
<tr>
<th></th>
<th>Pass WRT ( (N) )</th>
<th>Fail WRT ( (N) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary gain</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>No secondary gain</td>
<td>12</td>
<td>3</td>
</tr>
</tbody>
</table>

\( \chi^2 = 2.9, \ df = 1, \ p = 0.1. \)
Table 4.

*Chi-Square Results* for WRT Pass/Fail Using a Cut-off of ≥ 4 and Final Diagnosis

<table>
<thead>
<tr>
<th></th>
<th>Pass WRT (N)</th>
<th>Fail WRT (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive D/O diagnosis</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>ADHD D/O diagnosis</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Other diagnoses*</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

*χ² = 3.0, df = 2, p = 0.2.

*Other diagnoses included Major Depressive Disorder and Asperger’s Disorder.*
When a more stringent cut-off score of $\geq 3$ errors was used, four participants (13%) failed the WRT. Independent-sample $t$ tests were conducted to examine differences in age, years of education, WAIS-III indices, and reaction time for participants who passed vs. failed the WRT using this more stringent standard. Table 5 displays these results. No significant difference was found between pass/fail groups based on age, years of education, WAIS-III indices, or reaction time.

Table 5.

*Independent T-Test Results WRT Pass/Fail Using a Cut-off of $\geq 3$*

<table>
<thead>
<tr>
<th></th>
<th>Pass</th>
<th>Fail</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>31.9</td>
<td>29.3</td>
<td>-0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>Years of education</td>
<td>14.3</td>
<td>13.3</td>
<td>-1.1</td>
<td>0.3</td>
</tr>
<tr>
<td>WAIS-III VCI</td>
<td>113.3</td>
<td>104.5</td>
<td>-1.1</td>
<td>0.4</td>
</tr>
<tr>
<td>WAIS-III POI</td>
<td>109.0</td>
<td>102.0</td>
<td>-1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>WAIS-III WMI</td>
<td>96.7</td>
<td>93.8</td>
<td>-0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>WAIS-III PSI</td>
<td>98.9</td>
<td>90.8</td>
<td>-1.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Reaction time</td>
<td>1.0</td>
<td>1.6</td>
<td>2.9</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Using a chi-square test of independence, differences in pass/fail groups for presence of secondary gain and diagnosis given at the end of the end of evaluation were evaluated. Tables 6 and 7 display these results. No significant difference was found between pass/fail groups based on presence of secondary gain or diagnosis given at the end of the evaluation.

Table 6.

*Chi-Square Results* for WRT Pass/Fail Using a Cut-off of ≥ 3 and Presence of Secondary Gain

<table>
<thead>
<tr>
<th>Presence of Secondary Gain</th>
<th>Pass WRT (N)</th>
<th>Fail WRT (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary gain</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>No secondary gain</td>
<td>12</td>
<td>3</td>
</tr>
</tbody>
</table>

*χ² = 0.87, df = 1, p = 0.35

Table 7.

*Chi-Square Results* for WRT Pass/Fail Using a Cut-off of ≥ 3 and Final Diagnosis

<table>
<thead>
<tr>
<th>Final Diagnosis</th>
<th>Pass WRT (N)</th>
<th>Fail WRT (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive disorder</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>ADHD disorder</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Other disorder*</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

*χ² = 5.7, df = 2, p = 0.58.

*Other diagnoses included Major Depressive Disorder and Asperger's Disorder.
A one-way analysis of variance was used to compare the total errors and reaction time across the three final diagnosis groups. Tables 8 and 9 display these results. A significant difference was found in comparing the mean reaction times for final diagnosis. The ADHD group had significantly slower reaction time than the Other diagnoses group. All other comparisons of errors and reaction time were not significant.

Table 8.

*One-Way ANOVA Comparisons* of WRT Total Errors for Final Diagnoses

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive D/O</td>
<td>0.8</td>
<td>1.15</td>
<td>0.26</td>
</tr>
<tr>
<td>ADHD D/O</td>
<td>2.0</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>Cognitive D/O</td>
<td>0.8</td>
<td>1.15</td>
<td>0.99</td>
</tr>
<tr>
<td>Other D/O*</td>
<td>0.29</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>ADHD D/O</td>
<td>2.0</td>
<td>2.5</td>
<td>0.12</td>
</tr>
<tr>
<td>Other D/O*</td>
<td>0.29</td>
<td>.049</td>
<td></td>
</tr>
</tbody>
</table>

*ANOVA was nonsignificant, f(2) = 2.6, p = .095.

*Other diagnoses included Major Depressive Disorder and Asperger's Disorder.*
Table 9.

*One-Way ANOVA* Comparisons of WRT Reaction Time for Final Diagnoses

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive D/O</td>
<td>1.01</td>
<td>0.22</td>
<td>0.09</td>
</tr>
<tr>
<td>ADHD D/O</td>
<td>1.3</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>Cognitive D/O</td>
<td>1.01</td>
<td>0.22</td>
<td>0.99</td>
</tr>
<tr>
<td>Other D/O*</td>
<td>0.88</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>ADHD D/O</td>
<td>1.3</td>
<td>0.45</td>
<td>0.025</td>
</tr>
<tr>
<td>Other D/O*</td>
<td>0.88</td>
<td>0.19</td>
<td></td>
</tr>
</tbody>
</table>

*The ANOVA test is significant, $f(2) = 4.44$, $p < 0.05$.*

*Other diagnoses included Major Depressive Disorder and Asperger’s Disorder.*
DISCUSSION

The purpose of this study was to establish baseline performance on the Word Reading Test for patients referred for neuropsychological evaluation for academic purposes. Using a cut-off score of $\geq 4$ as recommended by Osmon et al. (2004), three clients (10%) failed the WRT. None of the participants that failed the WRT had potential secondary gains. The proportion of participants who had secondary gain was not different between the pass and fail groups. And no differences were found in the final diagnosis based on passing or failing the test. No differences were found between pass and fail groups for age, years of education, WAIS-III indices, or reaction time. This information suggests that these variables do not predict performance on the WRT when a cut-off score of $\geq 4$ is used.

Since secondary gains were unrelated to WRT total errors, it is unlikely that the participants that did fail the test were malingering. As previously noted, other factors can cause suboptimal effort. However, neither were there differences for participants who passed and failed on other factors such as diagnosis, age or education.

This raises the question of what a “failure” on the WRT actually signifies. Despite the apparent easy content of the WRT, it requires a certain amount of working memory. Participants must discriminate between the target word and a foil after seeing the target word previously. Failure to correctly complete this task could signify poor working memory, not suboptimal effort. However, there was not significant relationship between
ADHD and WRT failure in this study. Small sample size likely did not provide an adequate test of this relationship. Examining WRT performance of larger groups of ADHD patients would shed light on this possibility.

When a more stringent cut-off score of $\geq 3$ was used, similar results were found when compared to the cut-off score of $\geq 4$. Four participants (13%) failed the WRT, in contrast to the three failures (10%) when a cut-off of $\geq 4$ was used. There still were no significant differences between the pass/fail groups based on age, years of education, WAIS-III indices, or reaction time. The proportion of participants who had secondary gain was not different between the pass and fail groups. And no differences were found in the final diagnosis based on passing or failing the test.

Analysis of WRT total errors and reaction time for the three diagnoses yielded interesting results. The comparison of number of errors made by diagnosis group trended towards significance ($p = .095$), with the largest difference appearing between the ADHD group and the other two groups. Additionally, the diagnosis groups’ reaction times were significantly different ($p < .05$). The ADHD group was significantly slower in reaction time compared to the Other diagnoses group. The ADHD group was also slower in reaction time than the Cognitive disorders group, but the difference was only marginally significant ($p = .09$).

While it could be concluded that participants diagnosed with ADHD are more susceptible to suboptimal effort than participants with other diagnoses, a more likely explanation is that the significant difference in reaction time is due to the nature of ADHD. One of the hallmark symptoms of ADHD is variable attention, which could easily explain the longer reaction times for those participants. However, more research is
necessary to identify the relationship between diagnosis and performance on the WRT.

One limitation of this study is the variability in the participant group. Participants were referred for a variety of different types of cognitive evaluations from a multitude of referral sources. Additionally, participants were diagnosed with a variety of different disorders which may have clouded the results. The WRT is intended for use in LD evaluations, but this study included clients who were evaluated for other diagnoses and some were ultimately diagnosed with diagnoses other than LD. In order to more precisely evaluate the WRT, future research should specifically analyze the WRT performance of participants being evaluated for LD.

Another limitation was the small sample size of the participants. The small sample size reduces the power of the analysis, which makes it difficult to detect significant results. Future research should be conducted with larger sample sizes to address this issue. Nonetheless, the significant difference found in reaction time and nearly significant difference in diagnostic category using a cut-off score of three may be a factor of importance for future research.

While this study has a relative strength in that the participants were from a clinical sample rather than simulators, a weakness is that there was no method to independently verify the adequacy of effort. If it were possible to determine which participants gave their adequate and inadequate effort in this study, it would be possible to analyze the sensitivity of the WRT as well as specific characteristics of participants who were more likely to give suboptimal effort. Future research is needed to analyze whether the WRT can actually discriminate between good effort and poor effort in LD evaluations, as well as whether domain-specific SVTs are necessary.
REFERENCES


delayed subtests of the Wechsler Memory Scale-III. *Psychological Reports*, 86, 851-857.


Sullivan, B.K., May, K., & Galbally, L. (2007). Symptom exaggeration by college adults
in Attention-Deficit Hyperactivity Disorder and learning disorder assessments.

*Applied Neuropsychology, 14*(3), 189-207.


