The Diagnostic Accuracy of Symptom Validity Tests when Used with Postsecondary Students with Learning Disabilities: A Preliminary Investigation

William A. Lindstrom, Jr.*, Jennifer H. Lindstrom, Chris Coleman, Jason Nelson, Noël Gregg

Regents’ Center for Learning Disorders, University of Georgia, Athens, USA

Accepted 28 August 2009

Abstract

The current exploratory investigation examined the diagnostic accuracy of the Word Memory Test (WMT), Test of Memory Malingering (TOMM), and Word Reading Test (WRT) with three groups of postsecondary students: controls, learning disability (LD) simulators, and a presumed honest LD group. Each measure achieved high overall diagnostic accuracy, yet each contributed differently to suboptimal effort detection. False-negative classifications varied by measure, yet no simulator went undetected by all three tests. The WMT and WRT identified different members of the presumed honest LD group as demonstrating poor effort, whereas the TOMM identified none. Each measure contributed unique variance in a logistic regression, with effort status best predicted by WMT Consistency. Findings provided preliminary evidence that all three measures may be useful when assessing effort during postsecondary LD evaluations. Implications for future practice and research are discussed.

Keywords: Word Memory Test; Test of Memory Malingering; Word Reading Test; Learning disability; Poor effort; Cognitive symptom exaggeration

Introduction

Whereas effort testing is commonly employed in neuropsychological settings due to the high frequency of litigation and possible financial reward (Binder, 1993; Slick, Sherman, & Iverson, 1999), such assessment has been less commonly applied during learning-related evaluations. Recent research has suggested that students in postsecondary institutions may attempt to feign learning-related disorders during testing in order to access academic accommodations, such as extended time for tests, private testing rooms, and use of assistive technology (Harrison, Edwards, & Parker, 2007; Sullivan, May, & Galbally, 2007). Academic accommodations are available to students with disabilities to provide equal access to obtain and demonstrate knowledge as mandated by the Americans with Disabilities Act of 1990 and Section 504 of the Rehabilitation Act of 1973 (LaMorte, 1999; Latham & Latham, 1996). It has been suggested that students without disabilities may perceive these accommodations as a means to an advantageous position during academic activities (Harrison et al.; Sullivan et al.). When findings of functional limitations during psychological testing may result in external gain, the possibility exists that examinees may choose to feign or exaggerate compromised abilities (Binder, 1992; Felthous, 2006; Slick et al.). Thus, the identification of instruments which are valid for accurate detection of suboptimal effort in postsecondary students seeking evaluations for learning difficulties is warranted.

* Corresponding author at: Regents’ Center for Learning Disorders, The University of Georgia, 331 Milledge Hall, Athens, GA 30602-1556, USA. Tel.: +30-706-542-4589; fax: +30-706-583-0001.
E-mail address: wlindstr@uga.edu (W.A. Lindstrom).
**The State of the Literature**

The primary purpose of the current manuscript was to generate preliminary data reflecting the utility of several symptom validity tests (SVTs) when used with postsecondary students reporting learning difficulties. Professional guidelines and ethical principles governing psychological measurement obligate clinicians to use tests whose validity and reliability have been established with members of the population tested (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; APA, 2002; Bush et al., 2005). The majority of current and frequently used instruments have been extensively validated for use with individuals with neurological disorders (e.g., Bernard & Fowler, 1990; Binder, 1993; Gervais et al., 2001; Green & Iverson, 2001; Green, Rohling, Lees-Haley, & Allen, 2001; Iverson, Green, & Gervais, 1999; Tombaugh, 1996). In contrast, only four validity studies (i.e., Frazier, Frazier, Busch, Kerwood, & Demaree, 2008; Nitch, Boone, Wen, Arnold, & Alfano, 2006; Osmon, Plambeck, Klein, & Mano, 2006; Warner-Chacon, Boone, Zvi, & Parks, 1997) addressing the use of effort measures with postsecondary students seeking assessment due to learning difficulties were identified.

This low number is surprising for several reasons. First, postsecondary accommodation and service delivery decisions are generally based solely on findings from psychoeducational and neuropsychological tests (Gregg, Heggoy, Stapleton, Jackson, & Morris, 1994; Ofiesh & McAfee, 2000). As research has demonstrated that clinical judgment is a poor detector of dissimulation (Faust, Hart, Guilmette, & Arkes, 1988; Heaton, Smith, Lehman, & Vogt, 1978), the use of effort measures within the testing battery to ensure the validity of findings is crucial; yet the research supporting the use of such measures with this population is meager at best.

Second, a fundamental tenet of effort testing is that test-takers possess the necessary ability or skill to perform at a passing level given appropriate motivation. However, research evidence indicates that postsecondary students with learning disabilities (LD) commonly possess impairments and deficits in domains that may be tapped by frequently used effort measures (e.g., word reading, reading fluency, working memory, processing speed, phonological processing; Gregg et al., 2005; Ofiesh, Mather, & Russell, 2005). As noted in the “Standards for Educational and Psychological Testing” (AERA, APA, & NCME, 1999):

> In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement (Standard 10.1; p. 106).

As such, construct validity studies are essential to ensure ability is not confounded with effort within this population.

**Clinical samples.** Two of the four extant studies included clinical samples. Nitch and colleagues (2006) found the performance of postsecondary students with LD on the Rey Word Recognition Test (RWRT; Lezak, 1995) to be equivalent to general clinical patients with no motive to feign. Both groups achieved significantly higher scores than noncredible patients. A cutoff score of <7 correct resulted in 71% sensitivity and >90% specificity. Warner-Chacon and colleagues (1997) analyzed the specificity of the RWRT, the Dot Counting Test (DCT; Lezak), the Rey Fifteen Item Test (FIT; Rey, 1964), and the Forced Choice Method (FCM; Hiscock & Hiscock, 1989) when used with postsecondary students with LD. All measures demonstrated high specificity (i.e., ≥92%). On the basis of Warner-Chacon and colleagues’ findings, Alfano and Boone (2007) concluded that deficient cognitive functions associated with LD “rarely impact effort test score performance to the extent that a task is actually failed” (p. 376), though they noted that individuals with severe academic deficits may be best served by the use of multiple effort indicators with preference for those that avoid the compromised academic skill.

Alfano and Boone (2007) also pointed out that postsecondary students with LD are often high functioning with academic achievement in the average range. Though average achievement within an intellectually bright individual might suggest an LD within an ipsative interpretation, it does not represent what most would consider a functional impairment (e.g., achievement below the average range; Dombrowski, Kamphaus, & Reynolds, 2004; Siegel, 1999). Findings reported by Merten, Bossink, and Schmand (2007) suggested credible populations with severe impairments often fail on a number of effort measures. This issue was addressed in the current study by limiting LD group members to those with standardized word identification, reading fluency, word attack, or spelling scores at or below the 16th percentile.

**Simulation studies.** Recently, Frazier and colleagues (2008) used a simulation paradigm to investigate the ability of the FIT (Rey, 1964), the Validity Indicator Profile (VIP; Frederick, 2003), and the Victoria Symptom Validity Test (VSVT; Slick, Hopp, Strauss, & Thompson, 1997) to discriminate simulated reading disorder (RD) from control performances. Findings revealed that the VSVT and VIP may be effective in detecting simulated RD, with the VSVT hard index demonstrating the strongest classification statistics (a cutoff score of 21 yielded 92.9% sensitivity and 93.7% specificity).
Osmon and colleagues (2006) developed the Word Reading Test (WRT) with the goal of discriminating between LD and suboptimal effort during psychoeducational assessments. The WRT tasks include orthographic foils the layperson would typically associate with LD (e.g., letter reversals, homophones, orthographic errors; Osmon et al.). Inclusion of such errors was thought to be a probable method of deceiving those intending to dissimulate LD. The performance of three groups of college students—controls, RD simulators, and a group instructed to simulate processing speed difficulties—was compared. For RD simulators, a cutoff of >3 errors resulted in 90% sensitivity and 100% specificity. For speed disability simulators, the same cutoff resulted in 74% sensitivity and 100% specificity. The authors concluded that the WRT possessed strong ability to distinguish reading simulators from normal controls.

The present study augments the Frazier and colleagues (2008) and Osmon and colleagues (2006) investigations by including a clinical (i.e., LD) group. Simulation studies without clinical groups examine the discrimination of good effort from simulated impairment as exhibited by unimpaired participants. The next step is to determine whether or not the measures can discriminate true LD, a diagnosis based on findings of functional impairment, from simulated impairment.

Purpose of the Study

The current exploratory investigation examined the utility of the Word Memory Test (WMT; Green, 2005), Test of Memory Malingering (TOMM; Tombaugh, 1996), and WRT (Osmon et al., 2006) for identifying incomplete effort or symptom fabrication within the postsecondary population seeking academic accommodations due to reported learning difficulties. As noted previously, research examining the construct validity of effort measures with postsecondary LD is lacking. Our study augmented the modest number of validity studies by including a clinical sample with functionally impaired achievement (<standard score 85; Dombrowski et al., 2004; Siegel, 1999) in a print-related domain (e.g., decoding, reading fluency). To our knowledge, this is the first validity investigation of the WMT, TOMM, and WRT using a clinical sample of postsecondary students with LD.

Owing to the simulation design, it should be noted that the objective of this preliminary study was not to directly inform test selection and interpretation, but to serve as an initial step toward this goal. Studies with known groups permit stronger conclusions regarding implications for clinical practice. In addition, due to the exploratory nature of the investigation and a lack of gold standard for this population, an assumption of full effort from LD group members—a group in which a large portion of members had motive to feign—was necessary.

Research Questions

Initially, we attempted to determine if the WMT, TOMM, and WRT were able to differentiate presumed honest LD and control groups from LD simulators. Second, we sought to determine the diagnostic accuracy characteristics of each measure when classifying LD and simulator groups. Third, we investigated potential differences in the diagnostic accuracy of each measure. Fourth, the incremental validity resulting from the use of multiple measures of simulation detection was assessed. Finally, we evaluated the positive and negative predictive power (PPP and NPP) of each measure as a function of base rate estimations.

Materials and Methods

Participants

The study was authorized by the Institutional Review Board and conducted in a manner consistent with its principles. Written informed consent was obtained from all participants. Seventy-five adults (47 women) who were attending or who had recently attended mid-level to large-size colleges and universities in the southeastern USA participated in the study. A significantly larger number of women than men participated (gender: $\chi^2 [1, N = 75] = 4.81, p = .03$). Ages ranged from 18.08 to 23.42 with a mean of 19.43 ($SD = .98$). The sample included 42 freshmen, 21 sophomores, 9 juniors, 1 senior, and 2 students who had previously graduated. All students were native speakers of English. Racial distribution was 91% Caucasian, 7% African American, 1% Asian American, and 1% Other. No participants were on psychotropic medication at the time of their participation.

The normally achieving participants ($N = 50$) were undergraduate students currently enrolled at a large state university seeking to meet an introductory psychology course requirement as members of a research pool. These participants had no history or suspicion of learning, neurological, or psychological disabilities. They were randomly assigned to (a) a control
group (N = 25) requested to perform their best on all tasks presented or (b) a simulator group (N = 25) instructed to feign LD on the tasks administered. All examiners were blind to non-clinical group membership.

The clinical group included students with a print-related LD (i.e., dyslexia and/or dysgraphia; N = 25). These participants sought and completed comprehensive (i.e., ~10 hr) neuropsychological evaluations at our clinic for assessment of learning disorders in adults. Of the LD group, most, though not all, had a documented history of LD dating from early childhood. Each demonstrated a diagnosis of LD based on a clinical judgment model—that is, (a) significant underachievement in aspects of reading and writing (standard scores > 1 SD below estimated ability) and (b) one or more associated cognitive processing deficits (e.g., phonemic awareness, orthographic awareness; see Gregg et al., 1994, for further description of diagnostic procedures). No diagnosis was based on a single test score; rather, diagnostic decisions resulted from patterns of score discrepancies and errors as well as historical and qualitative evidence. All members of the LD group possessed at least one basic print-related skill (e.g., word identification) falling below the 16th percentile (as assessed with the Woodcock–Johnson [WJ] III Tests of Achievement; Woodcock, McGrew, & Mather, 2001) within the context of average to above average intellectual ability (as assessed with the Reynolds Intellectual Assessment Scales; Reynolds & Kamphaus, 2003). Group means (with SD in parentheses) were as follows: RIAS Composite Index = 102.16 (7.65); WJ-III Letter-Word Identification = 78.64 (10.72); WJ-III Reading Fluency = 81.84 (8.84); WJ-III Word Attack = 73.16 (10.78); and WJ-III Spelling = 77.96 (9.63). All LD participants met the criteria for no additional learning disorder (e.g., Attention-Deficit/Hyperactivity Disorder).

The majority of the data for the LD group were collected by including the experimental battery of tests within the context of the standard 2-day evaluation conducted at our disability assessment center (N = 18). Measures were given in the same order as those administered to members of the research pool to maintain standard procedure. An additional seven participants with LD who were formerly diagnosed by our center via the same procedures were randomly contacted and agreed to voluntarily participate in the study at our center or a more convenient college site.

Whereas inclusion of a clinical sample is a strength of the study, it is important to note that 18 of 25 members of this group had motive to feign (the seven volunteers who returned to participate had no external incentive to dissimulate). In the absence of a gold standard measure for the detection of incomplete effort or symptom exaggeration in the postsecondary population with LD, it was not possible to conclude unequivocally that all members of this group exhibited optimal effort. Data were analyzed under the assumption that full effort was exhibited.

Descriptive information about the groups can be found in Table 1. Groups did not differ on age, $F(2, 72) = 1.96, p = .15,$ or education level, $F(2, 72) = 2.49, p = .09.$ $\chi^2$ tests of gender and racial distributions revealed no deviation from expectation across the three groups (gender: $\chi^2 [2, N = 75] = 1.48, p = .48$; race: $\chi^2 [6, N = 75] = 4.43, p = .62$).

**Instruments**

The measures of interest for the current study (i.e., the WMT, TOMM, and WRT) are stand-alone tests designed to identify poor effort and feigning of symptoms. All three tests employ a dichotomous, forced-choice procedure (see Gervais, Rohling, Green, & Ford, 2004). Previous research indicated that the WMT and WRT are being used to assess for suboptimal effort in

| Table 1. Group descriptives, means, standard deviations, $F$-values, and effect sizes |
|---------------------------------|---------|---------|---------|--------|--------|
|                                | Controls | Simulators | LD       | $F$    | $\eta^2$ |
| Age (M [SD])                  | 19.14 (.72) | 19.67 (.87) | 19.49 (1.24) | 1.96  |
| Education (M [SD])            | 12.48 (.71) | 13.00 (1.04) | 12.52 (.96) | 2.49  |
| Gender (% male)               | 40       | 28       | 44       | 1.48  |
| Ethnicity (% Caucasian)       | 92       | 88       | 92       | 1.43  |
| WMT Immediate Recall (M [SD])| 99.60 (.82) | 68.40 (18.10) | 95.92 (6.15) | 59.54  | .62    |
| WMT Delayed Recall (M [SD])  | 99.52 (.87) | 64.40 (16.29) | 96.52 (5.46) | 96.06  | .73    |
| WMT Consistency (M [SD])      | 99.04 (1.51) | 59.84 (15.91) | 94.48 (7.73) | 109.32 | .75    |
| TOMM Trial 1 (M [SD])         | 48.72 (1.54) | 36.20 (7.33) | 48.04 (2.01) | 61.79  | .63    |
| TOMM Trial 2 (M [SD])         | 50.00 (0.00) | 41.12 (7.56) | 49.96 (.20) | 34.34  | .49    |
| TOMM Retention (M [SD])       | 50.00 (0.00) | 38.64 (8.55) | 49.88 (.33) | 43.57  | .55    |
| WRT CRI (M [SD])              | 39.60 (.87) | 32.08 (6.31) | 38.60 (3.12) | 24.85  | .41    |
| WRT RT (M [SD])               | 770.95 (219.31) | 1 077.07 (341.13) | 1 251.58 (450.01) | 12.10  | .25    |

Notes: WMT = Word Memory Test; TOMM = Test of Memory Malingering; WRT = Word Reading Test; CRI = Correct Response Index, RT = Reaction Time in milliseconds; $\eta^2$ = partial eta squared effect size; $N = 25$ for all groups.; SD = standard deviation.

$p < .017$ (employing Bonferroni correction, $p < .05/3$).

$p < .025$ (employing Bonferroni correction, $p < .05/2$).
postsecondary populations seeking accommodations due to learning difficulties (Osmon et al., 2006; Sullivan et al., 2007). No research using the TOMM for such purposes was identified.

Supplementary measures included the grooved pegboard (Heaton, Grant, & Matthews, 1991), the Rey Complex Figure Test (Meyers & Meyers, 1995), Block Design of the Wechsler Adult Intelligence Scale-Third Edition (Wechsler, 1997), the Colorado Perceptual Speed Test (CPST; a visual matching task assessing the participant’s speed for rapidly identifying the match of a stimulus letter/number cluster [e.g., \( xp6a \)] within four alternatives; Decker, 1989; DeFries, Plomin, Vandenberg, & Kuse, 1981), and the Reading Fluency subtest of the Woodcock–Johnson Diagnostic Reading Battery (Schrank, Mather, & Woodcock, 2004).

**Design and Procedure**

All participants completed the tasks during individual testing sessions of approximately 75–90 min. Consideration was given to provide incentives to participants to reduce threats to external validity (Franzen & Martin, 1996; Rogers & Cruise, 1998); however, this was not possible given the restrictions of the research pool authorization.

For the control and simulator groups, instructions were provided to each participant in a sealed envelope such that the participant’s group assignment was unknown to the experimenter. Members of the LD and control groups were instructed to perform their best on all tasks presented. As we were particularly interested in the manifestation of simulation attempts based on layperson’s knowledge of LD, the simulation group was not coached or provided with disorder-specific information. They were given the following instructions:

You are encouraged to fake having a learning disability. Please do your best to take the tests that you are given in such a way that your results will make psychologists and experimenters like us think you have a learning disability. How you decide to do this is up to you.

**Analyses**

Initially, univariate and multivariate analyses of variance were used to determine if effort measures could discriminate the three groups. Warranted post hoc comparisons using the Tukey test followed. Next, diagnostic accuracy statistics were calculated using software developed by Watkins (2002). As we were primarily interested in the ability of effort tests to discriminate LD simulation from true LD, only these groups were included in the diagnostic accuracy evaluation. Sensitivity, specificity, PPP, and NPP statistics were calculated. Kappas were used to indicate the level of agreement beyond chance between the classification made by the measure of interest (e.g., the WMT) and actual group membership (e.g., simulation). The overall diagnostic accuracy of each measure was assessed by conducting receiver operating characteristic curve analyses (Rosen et al., 2006; Swets, Dawes, & Monahan, 2000). In addition, \( \bar{z} \)-tests comparing the area under the curve (AUC) of each measure were conducted to identify differences in overall diagnostic accuracy (see Hanley & McNeil, 1983). To assess whether combinations of indices from differing tests augmented accurate classification, logistic regression was used. Group membership (simulator vs. LD) served as the dependent variable. Independent variables included WMT Immediate Recognition (IR), Delayed Recognition (DR), and Consistency (CNS); TOMM Trials 2 and Retention; and WRT Correct Response Index (CRI). A stepwise selection method with an entry criterion of .05 and retention criterion of .10 was applied. The test index with the greatest predictive power for group membership was entered first, followed by the index with the next greatest predictive power, and so forth. Nagelkerke \( R^2 \) statistics are reported.

The initial diagnostic accuracy statistics of each measure were generated on a 50% malingering base rate; that is, 50% of participants were designated as simulators. This proportion was selected because “statistical tests are most efficient in these circumstances” (Streiner, 2003, p. 213). However, PPP and NPP of tests are not fixed characteristics, but vary with the prevalence of the condition of interest (Meehl & Rosen, 1955; Streiner). Derivations of Bayes’ theorem were used to calculate additional PPP and NPP characteristics using base rate estimates (Streiner). Currently, the prevalence of suboptimal effort in postsecondary students seeking accommodations for possible LD is unknown. With a goal of minimizing threats to external validity, we followed the lead of Rosen and colleagues (2006) and investigated the PPP and NPP with a range of base rate estimates, from a conservative 5% to the more liberal 25% estimate of Binder (1992) for symptom exaggeration during neuropsychological assessment.

The sensitivity and specificity rates identified in our sample were used in all calculations. However, as the statistics were based on an assumption of full effort by the LD group, and as it is not possible to know how extrapolation to different base rates would affect the sensitivity and specificity of a measure without extensive further research (Kraemer, 1992; Rosen et al., 2006), the results of the base rate analyses were regarded as preliminary. Replication using known groups design will be needed.
to draw firm conclusions for clinical application. The base rate analyses were included to demonstrate how the PPP and NPP of the measures are highly dependent on assumed base rates of malingering (Rosen et al.).

Results

Although scores of effort measures are not normally distributed and nonparametric statistics were more appropriate for analyses, all values were highly significant and did not change conclusions drawn; thus, parametric statistics are reported for ease of interpretation (as in Osmon et al., 2006).

Word Memory Test

Analysis of the WMT using a one-way MANOVA with group as the between-subjects variable and effort indices as the dependent variables revealed significant differences among the groups, Wilks’ $\Lambda = .22$, $F(6, 140) = 26.20$, $p = .000$. Means and standard deviations for WMT indices by group are listed in Table 1 in the form of raw scores. WMT scores are presented as percent correct.

Follow-up univariate ANOVAs revealed significant group differences among the scores of each of the three effort indices (see Table 1). Post hoc pair-wise comparisons demonstrated that, across all three indices, control scores were equivalent to LD scores, and control and LD scores were significantly higher than simulator scores. Effect sizes of significant pair-wise comparisons were consistently large (Cohen’s $d > 2.0$; Cohen, 1992). The large standard deviations exhibited by the simulator group were not surprising. Simulator groups have been shown to demonstrate greater variability in performance scores in forced-choice tests (Flowers, Bolton, & Brindle, 2008).

For the WMT, sensitivity, specificity, PPP, NPP, and the kappa were generated using Green’s (2005) recommended cutoff of IR, DR, or CNS $\leq 82.5\%$ correct. Resulting classification rates were strong; sensitivity, specificity, PPP, and NPP were all .92. The kappa of .84 is considered excellent (Fleiss, 1981). Findings revealed only 2 of the 25 simulators were not identified as such by the WMT. In addition, two individuals with LD who were presumed to give optimal effort failed to meet passing criteria on the WMT.

Test of Memory Malingering

Analysis of the TOMM using a one-way MANOVA with group as the between-subjects variable and effort indices as the dependent variables revealed significant differences among the groups, Wilks’ $\Lambda = .35$, $F(6, 140) = 16.23$, $p = .000$. Means and standard deviations for TOMM test results by group are listed in Table 1 in the form of raw scores. The TOMM uses total items correct per trial.

Follow-up univariate ANOVAs revealed significant group differences among the scores of each of the three effort indices (see Table 1). Pair-wise comparisons revealed findings consistent with those of the WMT. That is, control scores were equivalent to LD scores, and control and LD scores were significantly higher than simulator scores. Effect sizes were consistently large (Cohen’s $d > 1.65$).

The classification rates of the TOMM were determined using Tombaugh’s (1996) recommended cutoff of <45 items correct on Trial 2 or on the Retention Trial. The TOMM demonstrated perfect specificity and PPP (i.e., 1.00), rendering no false-positive classifications. Sensitivity (.68) and NPP (.76) were lower than the WMT. Eight of the 25 simulating participants were not identified as exhibiting suboptimal effort by the TOMM. The kappa of .68 is considered fair (Fleiss, 1981).

Word Reading Test

Analysis of the WRT using a one-way MANOVA with group as the between-subjects variable and CRI and Reaction Time (RT) as the dependent variables revealed significant differences among the groups, Wilks’ $\Lambda = .44$, $F(4, 142) = 17.82$, $p = .000$. Means and standard deviations for test scores by group are listed in Table 1 in the form of raw scores (i.e., total items correct and RT in milliseconds).

Univariate ANOVAs revealed significant group differences on the scores of the CRI and RT (see Table 1). For the CRI, control scores were equivalent to LD scores, and control and LD scores were significantly higher than simulator scores. In contrast, on the RT, simulator and LD scores were equivalent, and both were significantly greater than the control scores. Effect sizes were consistently large (Cohen’s $d > 1.07$). It is notable that the RT score that failed to distinguish LD and simulator scores was not used for classification purposes in Osmon and colleagues’ (2006) study and does not confound use of the error score for such purposes.
The sensitivity, specificity, PPP, NPP, and kappa of the WRT were determined using Osmon and colleagues’ (2006) recommended cutoff of ≥4 errors. The WRT demonstrated strong specificity (.92) and PPP (.89) and resulted in only two false-positive classifications. Sensitivity (.68) and NPP (.74) were generally equivalent to the TOMM and poorer than the WMT. Eight of the 25 simulating participants were not identified as exhibiting suboptimal effort by the WRT. The kappa of .60 is considered fair (Fleiss, 1981).

Examination of Overall Diagnostic Accuracy

AUC statistics were calculated for each effort test and compared via z-tests to investigate potential differences in diagnostic accuracy. All AUC indices for the effort tasks indicated high overall diagnostic accuracy (Swets, 1996). The AUCs for the WMT IR, DR, and CNS were .93, .96, and .97, respectively. The AUCs of the TOMM were .94 for the Delayed Trial and .95 for the Retention Trial. For the WRT CRI, the AUC was .89. An analysis of all possible index comparisons from the three measures using two-tailed z-tests at a 5% level of significance revealed no statistically significant differences in overall diagnostic accuracy. It is notable that small sample sizes may have resulted in comparisons with insufficient power (Keppel, 1991).

The Use of Multiple Measures in Simulation Detection

Logistic regression analysis was performed to examine the incremental validity of the WMT, TOMM, and WRT for detecting simulation. Results yielded an equation with three variables making significant ($p < .05$) contributions to predictive power. The WMT CNS was entered first ($\Delta R^2 = .81, \Delta \chi^2 [1] = 46.54$), followed by TOMM Retention, ($\Delta R^2 = .09, \Delta \chi^2 [1] = 9.17$), and WRT CRI ($\Delta R^2 = .04, \Delta \chi^2 [1] = 4.37$). No additional variables provided statistically significant incremental validity to the prediction. The $p$-value obtained via the Hosmer–Lemeshow test ($\chi^2 [8] = .81$) was large (.999), suggesting a strong fit between model and data. To rule out potential difficulties with multicollinearity, the logistic equation was run a second time after eliminating two variables (i.e., WMT DR and TOMM Trial 2) whose collinearity diagnostic statistics (i.e., variable inflation factor and tolerance) appeared problematic. Findings from the second logistic regression analysis were identical to those of the first analysis.

Base Rate Considerations

The PPP and NPP values of the WMT, TOMM, and WRT were generated for malingering base rates of 25%, 15%, and 5%. In general, NPP remained strong regardless of prevalence, with statistics ranging from .97 on the WMT to .90 on the WRT and TOMM at the 25% base rate and improving with decreasing incidence. PPP rates were more vulnerable. At the extreme base rate estimates evaluated (25% and 5%), the PPP of the WMT ranged from .80 to .38, and the PPP of the WRT ranged from .74 to .31. At a more reasonable 15% estimate (see Sullivan et al., 2007), the WMT and WRT demonstrated PPP of .67 and .60, respectively, suggesting a large portion of positive findings (i.e., 33% and 40%) might be inaccurate. In contrast, the TOMM demonstrated perfect specificity and thus flawless PPP regardless of base rate estimates. This level of performance is unrealistic to expect during clinical practice and is likely a result of the small sample size.

Discussion

The primary purpose of the current manuscript was to generate preliminary data regarding the utility of the WMT, TOMM, and WRT when used with postsecondary students seeking evaluations due to suspected LD. When interpreting the results, two caveats should be kept in mind. First, as with all simulation studies, our results should be regarded as exploratory and need to be verified using known-groups designs prior to drawing firm conclusions regarding direct implications for clinical practice (Rogers, Harrell, & Liff, 1993). Second, we presumed that the students with LD exhibited optimal effort. Some evidence suggested this to be an accurate assumption. In general, students with LD performed equivalently to controls on the classification indices of the WMT, TOMM, and WRT. Regarding specific cases, all four false positives (two for the WRT, two for the WMT) had previous diagnoses of LD. All demonstrated the appropriate pattern of average to above average intellectual functions, impaired literacy skills, associated linguistic processing deficits (e.g., phonological and/or orthographic processing), and average to above average functions in unrelated domains (e.g., visual-spatial functions). While acknowledging the limitations of clinical judgment (see Faust et al., 1988; Heaton et al., 1978), all four were deemed to exhibit adequate effort by clinicians working with them.
Overall, in the absence of a gold standard measure for the detection of incomplete effort or symptom exaggeration in the postsecondary population with LD, we proceeded under the assumption that the LD group exhibited adequate effort. The reader should note that such an assumption may have impacted the veracity of the diagnostic accuracy statistics generated. As such, all generated statistics and conclusions must be considered tentative and preliminary.

General Conclusions

Despite the limitations of exploratory studies such as this one, some general conclusions were drawn regarding the utility of the evaluated measures with this population. Findings provided preliminary evidence that all three measures may be useful when assessing effort among postsecondary students reporting learning difficulties. More specifically, students with LD performed equivalently to controls on all classification indices of the WMT, TOMM, and WRT, with simulators scoring well below both groups. In addition, all three effort measures achieved high overall diagnostic accuracy (Swets, 1996). The WMT appeared to be the most sensitive to suboptimal effort. The TOMM exhibited perfect specificity and PPP and appeared the most conservative measure. The TOMM was less sensitive than the WMT, and, in the context of its perfect specificity, the findings added to previous evidence suggesting the TOMM to be an easier task than the WMT (Gervais et al., 2004; Green, Berendt, Mandel, & Allen, 2000; Tan, Slick, Strauss, & Hultsch, 2002). A relatively new measure, the WRT demonstrated specificity equivalent to that of the WMT (.92) and sensitivity (.68) and NPP (.74) generally equivalent to the TOMM. Considering the favorable findings in Osmon and colleagues’ (2006) investigation, it shows promise for use with this population. When used in combination, findings suggested that poor effort status was best predicted by the WMT CNS score, with additional predictive validity provided by TOMM Retention and WRT CRI. It is notable that one index from each measure contributed significantly to the logistic regression equation.

Analysis of Misclassified Cases

Analysis of misclassified participants revealed that the two LD cases identified as exhibiting poor effort by the WMT were not the same cases as those identified by the WRT. As noted previously, review of scores across cognitive and linguistic processes for these four cases revealed expected deficits (e.g., phonological processing, orthographic processing), but these participants were not the most severe LD cases in the study. As such, the examined measures did not appear to be overly sensitive to LD severity in this sample.

Regarding false negatives, the three measures did not consistently fail to detect the same simulators. Of the eight simulators missed by the TOMM and the eight missed by the WRT, four were missed by both. Of the two simulators missed by the WMT, one was detected by the WRT. Consistent with Osmon and colleagues’ (2006) theoretical predictions, this individual had the poorest performance of all simulators on the orthographic measure of the battery (i.e., CPST). The other simulator classified as false negative by the WMT was detected by the TOMM. No simulator went undetected by all three tests.

A Note on Base Rate Considerations

Calculations of PPP and NPP across varying degrees of prevalence revealed consistently strong NPP across measures. PPP statistics were more vulnerable to base rates, with some falling to .67 and .60 at the 15% prevalence rate suggested by Sullivan and colleagues (2007). Given the LD group had motive to feign, sensitivity and specificity data were potentially flawed. Therefore, the use of generated PPP and NPP statistics for clinical application would be inappropriate. However, it is critical that evaluators be aware of the potentially significant impact of base rates on the predictive power of tests.

The Role of Reading Fluency

Whereas postsecondary LD performance was equivalent to control performance on untimed measures, the LD group demonstrated compromised speed on the timed word reading task of the WRT. This impairment was severe enough to be equivalent to simulator performance. The finding was not surprising given the preponderance of evidence indicating that individuals with LD struggle with reading fluency into adulthood (Gregg et al., 2005) and suggests that measures involving timed reading likely should not be used to evaluate effort in those with possible LD. However, given that the WRT CRI was sensitive to poor effort, but the RT measure was sensitive to LD and poor effort, the possibility that the differential sensitivities of the WRT CRI and RT could be useful in group discrimination via pattern analysis warrants further
study. Research has also suggested that postsecondary students with LD perform at a slower, more variable rate than peers without LD on standardized tests of processing speed (Ofiesh et al., 2005); however, some effort measures with simple processing speed requirements, including the DCT and b Test, have been shown to be insensitive to LD (e.g., Boone et al., 2000, 2002).

Limitations

As an exploratory study with a simulation design, the limitations of the current investigation are largely related to external validity (Haines & Norris, 1995; Rogers, Bagby, & Dickens, 1992). As noted previously, it was not possible to conclusively ascertain the veracity of the effort provided by the members of the LD group. Thus, this study provides general information related to the effectiveness of the WMT, TOMM, and WRT with this population, but firm conclusions regarding practical implications will require the use of known-groups designs. Second, instructions to simulators were intentionally vague in order that resulting portrayals would be based on the lay knowledge of the construct. However, investigators may have overestimated the knowledge regarding LD in the target population, resulting in decreased simulator sophistication. True malingerers would likely investigate characteristics of a disorder they were trying to replicate (Rogers, 1997), an opportunity the participants in the current study did not have. Third, an evaluation of the authenticity of simulation attempts was not conducted during debriefing, leaving the possibility that simulators failed to feign sufficiently. Fourth, the use of positive and negative incentives is recommended in investigations using simulators to reduce external validity threats (Bender & Rogers, 2004; Franzen & Martin, 1996; Rogers & Cruise, 1998). Incentives were not allowed under the restrictions of our research pool authorization. Fifth, the LD group completed a more extensive battery than controls and simulators. Battery length may affect feigning strategy selection/adherence, further limiting the generalizability of the findings. Finally, it is critical to note that the findings of the current investigation likely do not generalize to all adults with LD, as the study participants were typically high functioning and able to attain college entry despite their difficulties.

Future Directions

The primary goal of the study was to lay the groundwork from which future investigations could be generated. Such investigations should attempt to generate more clinically applicable validation evidence for SVTs used with LD populations. Simulation design is acceptable during early stages of validation, but should be followed by known group studies to confirm the generalizability of results (Rogers et al., 1993). Known group studies are critical prior to applying findings to clinical practice. Similarly, further studies involving LD samples without motive to feign will result in more externally valid findings and are encouraged to ensure that associated deficits and impairments (e.g., word reading, reading fluency, processing speed, working memory, phonological processing; Gregg et al., 2005; Ofiesh et al., 2005) do not confound the discrimination of effort from ability within this population.

To further facilitate test selection, it would be beneficial to analyze correlations between effort measure scores and cognitive, linguistic, and academic measure scores in areas commonly found to be deficits in LD. Such correlational analyses were conducted by Warner-Chacon and colleagues (1997) with the RWRT, DCT, FIT, and FCM, resulting in practical recommendations. Studies such as this are needed with the WMT, TOMM, and WRT.

Consideration of several additional variables in future studies would be useful. Coaching (e.g., Rose et al., 1998), research opportunities prior to testing, warnings of potential malingering detection devices (see Bender & Rogers, 2004), and the presence of external incentives (Franzen & Martin, 1996; Rogers & Cruise, 1998) would enhance simulator sophistication and ecological validity. During debriefing, questioning participants regarding face validity of measures and feigning strategies would be informative for clinician decision-making.

Perhaps the most critical investigations will seek to identify the base rate of suboptimal effort in postsecondary students participating in LD evaluations. In the current study, the application of Bayes’ theorem (Streiner, 2003) to the diagnostic accuracy statistics of each measure resulted in informative adjustments to predictive power statistics. Though limited in clinical utility due to weaknesses within the study, the analyses demonstrated the considerable dependence of PPP and NPP on assumed base rates of malingering (Rosen et al., 2006). Precise and reliable base rate estimates will help guide test selection, increase confidence in findings, reduce the influence of evaluator bias, and promote cautious and psychometrically sound test interpretation that will result in valid findings for those seeking academic supports.
Conflict of Interest

None declared.

References


