The Investigation of Embedded Effort Measures Using the Children's Memory Scale (CMS)

Brianna M. Hernandez

Senior Honors Thesis

Committee: Shelley C. Heaton, Ph.D. (Chair), Lisa S. Scott, Ph.D. (Co-Chair), Erika M. Cascio, Psy.D. (Member)

University of Florida

Abstract

Effort testing is an essential component of pediatric neuropsychological assessments to validate test results. They are a necessary measure to ensure the patient's performance is reflective of their potential ability. Previous research has been largely limited to adult populations, with an absence of effort research in children. Studies in adults indicate the usefulness of multiple measures of effort throughout assessment. Embedded measures can improve effort testing. This study aimed to investigate novel, embedded effort measures using a recognition task on the Children's Memory Scale (CMS), and to compare the usefulness of the new effort measure to an established embedded effort measure, the Reliable Digit Span (RDS). The CMS Word Pairs 2 Delayed Recognition (CMS WP-DR) was analyzed for clinical use as an embedded measure of effort. The total raw score of this task was not significant in discriminating between poor and adequate effort. Five word pairs were significant in their association with RDS performance. Future examination of the CMS WP-DR should include a larger sample and other confirmatory measures.

Introduction

Validity Measures

Testing effort during neuropsychological assessments has become a necessary measure in order to validate the scores of the test battery. Effort tests, which require little effort or ability themselves, can be perfectly performed despite existing neurological or psychiatric conditions (Heilbronner et al., 2009). Failing these tests is a good indicator of invalid test results. Validity tests are designed to detect noncredible effort while remaining insensitive to ability of the examinee (Kirkwood, 2015). In a clinical setting, assessment of effort is essential to determine that the results of the performance test battery are in fact valid and reflective of the patients' abilities.

Noncredible Effort in Children

Research in the domain of effort testing is quite extensive amongst adults (Heilbronner et al., 2009; Greher & Wodushek, 2017; Rickards, Cranston, Touradji & Bechtold, 2017); however, there is very little information of that in pediatric populations. The American Academy of Clinical Neuropsychology noted in their Consensus Conference Statement that effort measures in pediatric samples should be examined in future research (Heilbronner et al., 2009). Very little is known about effort testing developed specifically for children, as most research relates to the extension of adult cutoffs and standalone measures to pediatric populations.

The paucity of research in comparison to that in adult populations may largely be due to providers not recognizing that children, like adults, are also capable of feigning or exaggerating in a clinical assessment setting (Kirkwood, 2015). Though it is unclear whether this noncredible effort is conscious, there are some clinical populations with higher rates of invalid results. The population with the highest rate of invalid effort is children with persistent problems after mild

head injury (Kirkwood, 2015). These assessments may provide opportunities to examine characteristics specific to these clinical populations that result in the lower rates of passing effort tests.

Validity Testing in Pediatric Neuropsychological Assessment

While the literature on pediatric validity measures extends across many different aims, it has become evident that multiple measures of validity, and the combination of them, may be the most effective during assessment (Boone, 2009).

Previously there has been an emphasis on clinical observations to determine noncredible effort in an examinee. Multiple observations are indicative of a participant's unwillingness to engage in the assessment. Body language such as poor eye contact, failure to acknowledge the examiner, and even sarcasm may exhibit signs of passive negativity; on the other hand, active negativity may present in openly negative comments related to the assessment (Kirkwood, 2015). These pose a few examples of examiner observations that may preface validity issues throughout testing. Using these observations, though, are more subject to bias and assume the participant is not putting forth maximum effort on tests based on their emotional state and behaviors. These therefore should not be used alone.

In terms of objective measures, there is a growing literature on tests utilized throughout examination to measure examinee effort. As is necessary, there are a number of ways to measure effort. Standalone performance validity tests (PVTs) are tests that have been developed specifically to measure patients effort, and only effort, during assessment (DeRight & Carone, 2015), while embedded measures consist of variables indicative of validity within standard neuropsychological tests that are already being administered.

Embedded Effort Measures

Notably, embedded measures are extremely time-efficient and cost effective because they are from subtests of tests that are already being administered during assessment (Brooks, Sherman, & Iverson, 2014). These effort measures can range across neurocognitive domains and be present in varying standard neuropsychological tests. This is why it is important to investigate these measures throughout various pediatric tests in order to account for the uniqueness of each clinical case. Despite the known benefits of embedded measures of effort, these are even less extensively investigated. As a result, the Reliable Digit Span (RDS) is amongst the only embedded measure that has proven to be a good indicator of suboptimal effort (DeRight & Carone, 2015; Kirkwood, Hargrave, & Kirk, 2011; Loughan, Perna, & Hertza, 2012; Miele, Gunner, Lynch, & McCaffrey, 2012; Perna, Loughan, Hertza, & Segraves, 2012; Schroeder, Twumasi-Ankrah, Baade, & Marshall, 2012; Welsh, Bender, Whitman, Vasserman, & MacAllister, 2012).

Memory Tests as Validity Measures

The established literature on validity testing includes an array of measures across memory and learning domains (Baker et al., 2014; Bigler, 2014; Brooks, 2012; Brooks et al., 2014; Kirk et al., 2011; Kirkwood et al., 2012; Kirkwood et al., 2011; Perna et al., 2012). Memory tasks are present in virtually all examined validity testing.

Previous research has demonstrated that recognition tasks, rather than recall tasks, are better indicators of good versus poor effort (Bernard, 1990; Bernard et al., 1993; Sullivan et al., 2002). Specifically, recognition tasks may require less cognitive resources, making it more extendable across clinical populations. Furthermore, this will make these tasks less sensitive to cognitive abilities, in order to measure validity of performance.

With the examination of this literature, the Children's Memory Scale (CMS) Word Pairs 2 Delayed Recognition task seems a likely indicator of suboptimal or noncredible effort versus adequate or good effort.

Study Aims and Objectives

Aim 1. To investigate a novel embedded effort measure using the Children's Memory Scale (CMS) Word Pairs 2 Delayed Recognition task total score. It is hypothesized that:

a. The CMS Delayed Recognition total raw score will be effective in discriminating between poor and adequate effort/invalid and valid performance.

Aim 2. To investigate associations between Reliable Digit Span (RDS) performance and CMS word pair item recognition. It is hypothesized that:

a. There will be some associations between individual items and RDS performance.

Aim 3. To investigate associations between RDS performance and consistency in recognition of repeated CMS word pairs. It is hypothesized that:

a. There will be some associations between repeated word pairs and RDS performance.

Aim 4. To investigate the clinical utility of significant word pair items. It is hypothesized that:

a. If there are significant word pairs, then these items will have clinical utility.

Methods

Participants

Archival data for the current study was extracted from an Institutional Review Board (IRB) approved clinical databank (IRB201500554). The clinical databank was comprised of a diagnostically heterogeneous sample that completed comprehensive neuropsychological assessments in Dr. Shelley Heaton's Pediatric Neuropsychology Clinic at the Psychology Clinic in Shands Hospital. All participants provided consent for their information to be included in the

clinical databank. Participants were eligible to be included in this study if they completed the targeted measures (CMS, RDS).

In contrast to some studies that investigated validity measures (Gunn, Batchelor, and Jones 2010; Nagle et al., 2006), the sample for this current study was not a research population instructed to feign impairment. Rather, the participants were all clinical evaluations for presenting problems or concerns. Each individual was instructed to perform to the best abilities during assessment.

Each participant received a unique battery set during their evaluation, which was based on their conditions and diagnostic history. Though the clinic setting is standardized across patient visits, each patient may see a different variety of tests during their visit. This may affect the duration of the assessment, causing some patients to fatigue while others may not. Individual attentiveness can alter test performance, producing effects similar to poor effort. Although these conditions should not affect the child's effort, participants were not excluded if they were on any medications, including those for attention deficits. Because this was a clinical sample, some participants were on medication, though this should not affect the results of this study.

Age, Gender, Race, and Ethnicity.

The data extracted for this study was comprised of individuals between the ages of 6 and 16. Participants' parents indicated their child's gender, race, and ethnicity on a Demographic Form that was given during assessment. Gender was indicated as male or female. Race categories consisted of White/Caucasian, Black/African American, Asian, Native Hawaiian/Other Pacific Islander, American Indian/Alaska Native, Mixed Race, or Other. Ethnicity was broken into Non-Hispanic/Latino or Hispanic/Latino.

There were a total of 78 participants in this study. Participants consisted of 40 females (51.3%) and 38 males (48.7%). The sample was between the ages of 6 and 16 years old. The mean age was 10.81 with a standard deviation of 3.11. The sample was 69.2% white/Caucasian, 14.1% black/African American, 1.3% American Indian/Alaska Native, 10.3% mixed race, 1.3% who identified as other, and 3.8% of the sample chose not to disclose. There were 61.5% non-Hispanic/Latino, 17.9% Hispanic/Latino, and 20.5% who did not disclose their ethnicity.

Socioeconomic Status.

Parents also filled out their education and household annual income on the Demographic Form. This information was related to the caregivers' information. Education attainment was broken into seven categories consisting of Kindergarten-6th Grade, 7th-9th Grade, 10th-11th Grade, High School Diploma, Some College, BA/BS, and Graduate Degree. Parents wrote in their occupation. The household annual income was also broken into several categories. These ranges were as follows: Less than $5,000 per year, $5,000-$10,000, $11,000-$15,000, $16,000-$20,000, $21,000-$35,000, $36,000-$50,000, and Over $50,000 per year.

The two factors collected and examined to gain a sense of the participants' socioeconomic status included average household income and education attainment of both parents. In terms of the average household income, 1.3% reported $5,000-$10,000 (N=1), 9% reported $16,000-$20,000 (N=7), 10.3% reported $21,000-$35,000 (N=8), 16.7% reported $36,000-$50,000 (N=13), 42.3% reported over $50,000 (N=33), and 20.5% did not report their average household income (N=16).

Education attainment was collected separately for the child's mother and father, where highest grade level completed was endorsed. In terms of the biological mother's education attainment characteristics, 3.8% indicated 7-9 (N=3), 5.1% indicated 10-11 (N=4), 20.5%

indicated High School Diploma (N=16), 24.6% indicated Some College (N=27), 6.4% indicated BA/BS (N=5), 21.8% indicated Graduate Degree (N=17), and 7.7% did not report (N=6). The biological father's education attainment consisted of 1.3% who endorsed K-6 (N=1), 2.6% endorsed 7-9 (N=2), 5.1% endorsed 10-11 (N=4), 32.1% endorsed High School Diploma (N=25), 21.8% endorsed Some College (N=17), 11.5% endorsed BA/BS (N=9), 7.7% endorsed Graduate Degree (N=6), and 17.9% who did not answer (N=14).

Intellectual Ability.

Though not a variable of interest, participants' Full Scale IQ (FSIQ) was collected to provide a descriptive of the population's intellectual ability. Various intellectual measures were used according to the case and the examinee's needs. The tests used to collect the FSIQ included the Wechsler Intelligence Scale for Children- Fifth Version (WISC-V), Leiter International Performance Scale, Third Edition (Leiter-3), and Reynolds Intellectual Assessment Scales (RIAS-2).

Full Scale IQ (FSIQ) was used as a measure of intellectual ability. FSIQ ranged between 55 and 118. The average FSIQ was 87.77, with a standard deviation of 14.97.

Diagnoses.

This study consisted of a diagnostically heterogeneous sample. Participants included children with various neurological or medical conditions, such as TBI, epilepsy, hematology and oncology; developmental disorders, including ADHD, learning disorders, and autism; as well as other mood and psychiatric disorders, including anxiety, depression, PTSD, and OCD. Individuals were not excluded or included in analyses based on the diagnosis.

This sample consisted of a mixed diagnostic clinical sample. Diagnosis of medical or neurological conditions (e.g. TBI, epilepsy, hematology/oncology, premature/low birth weight),

developmental disorders (e.g. ADHD, autism, learning disorders, academic problems), and other mood or psychiatric disorders (e.g. anxiety, depression, PTSD, OCD) represented 58%, 69%, and 40% of the sample, respectively. These percentages do not sum to 100% because participants can be members of multiple diagnostic groups.

Table 3 below depicts the breakdown of diagnoses within the diagnostic groups.

Table 3.

Diagnosis Characteristic*	n	% of total sample
Neurological/Medical Condition		
TBI/Concussion	13	16.7%
Epilepsy	11	14.1%
Hematology/Oncology	9	11.5%
Premature/Low Birth Weight	6	7.7%
Other Neuro/Genetic Disorder	6	7.7%
Developmental Disorder		
ADHD	8	10.3%
Autism+	5	6.4%
Autism	1	1.3%
Intellectual Disability	1	1.3%
Language Disorder	1	1.3%
Specific Learning Disorder	1	1.3%
Mixed Developmental**	37	47.4%
Mood/Other Psychiatric Disorder		
Anxiety	12	15.4%
Depression	4	5.1%
OCD	4	5.1%
PTSD	1	1.3%
Anxiety/OCD	1	1.3%
Mixed Mood**	9	11.5%

*Diagnoses percentages do not sum to 100% because participants can be members of multiple groups

**Participants with multiple diagnoses within the same diagnostic group

Study Measures

This study was focused on identifying potential validity measures in the CMS and comparing it to the gold standard, RDS.

The Children's Memory Scale.

The Children's Memory Scale (CMS) is a learning and memory assessment measure that evaluates these functions in children between the ages of 5 and 16 years old (Cohen, 1997). The CMS was normed based on a stratified random sample representative of U.S. children, accounting for age, sex, race/ethnicity, geographic region, and parent education. The internal consistency of the CMS subtests were analyzed using the split-half method for many and generalizability theory for those that are unique in item presentation formats. The corresponding reliability coefficients ranged from .61 to .94 across subtest scores and index scores. The average reliability coefficient for the Word Pairs Delayed Recognition score was .79. In terms of validity, the CMS underwent multiple revisions from 1986 to 1993 to investigate the content validity. There are moderate to high positive relationships between the immediate and the delayed recall measures of each subtest. Correlations have been analyzed between the CMS and a number of other measures across neurocognitive domains, including general cognitive ability, academic achievement, executive functioning, language processing, and memory functioning. It is a well-established and widely used learning and memory test during neuropsychological assessments. There are a total of nine subtests including Stories, Word Pairs, Word Lists, Dot Locations, Faces, Family Pictures, Numbers, Sequences, and Picture Locations. Several of the subtests have components differentiated by age, in which specific items are administered to children according to their age. These indices measure a variety of learning and memory domains ranging from auditory/verbal memory, visual/nonverbal memory, and attention/concentration.

For the Word Pairs subtest, administrators read a list of these pairs aloud. Then, they read the first word of the pair and the examinee should provide the second word from their memory of

the list. This is repeated over three trials. The examinee will score one point for each correct response and zero points for incorrect responses.

A similar task is administered later on in the CMS through the Word Pairs 2 subtest. In this case, there is a delayed recall and delayed recognition portion where the examinee must recall the word pairs learned from the previous learned lists and confirm whether or not each pair was actually on the list, respectively. The Delayed Recognition task that this study examined has 30 items for children ages 5-8 and 42 items for children ages 9-16 and contains some repeated items.

Reliable Digit Span.

The Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V) is an intelligence measure with a digit span subtest (Pearson, 2014). This subtest includes a Forward task, a Backward task, and a Sequencing task. The examinee is read a sequence of number and must recall it in the same order it was presented in. These various tasks require mental alertness (Pearson, 2014). There are two trials per each item in which the same amount of numbers are presented in the string.

The Reliable Digit Span is calculated by "summing the longest string of digits repeated without error over two trials under both forward and backward conditions" (Greiffenstein et al., 1994). This measure has been adopted to discriminate between poor and adequate effort during examination. A ≤6 cutoff score resulted in a global sensitivity score of 35% and a global specificity score of 97% (Schroeder et al., 2012). This demonstrates that a score of 6 or below indicates suboptimal effort.

Procedures

The sample underwent a full day neuropsychological evaluation that included a review of history and records, clinical interviews, intelligence tests, as well as the Reliable Digit Span (RDS) from the WISC-V and the CMS, a memory test. Other standard tests were administered, however they were not analyzed because they are beyond the scope of this research question.

The Word Pairs 2 Delayed Recognition subtest of the CMS was analyzed as the embedded measure. The potential embedded indicators included the CMS WP-DR total raw score, individual item recognition, and consistent recognition of repeated word pairs. RDS was calculated by summing the number of digits in the last trial with a perfect score on the WISC-V Digit Span Forward and Backward. The established RDS cutoff scores used were as follows: 6 and below indicated suboptimal effort, while a score of 7 or above indicated adequate effort. The most salient word pairs were identified as measuring effort. Scores of these variables were then compared to those of the Reliable Digit Span from the WISC-V, a more widely accepted embedded measure in the field.

Data Analyses

For aim 1 to investigate a novel embedded effort measure using the CMS using the recognition task total score, receiver operating characteristic (ROC) and area under the curve (AUC) analyses were used to determine the optimal classification statistics and cutoff scores for CMS WP-DR total raw scores. ROC analyses are used to capture the relationship between the sensitivity and specificity along with the range of decision thresholds of each participant, despite their level of experience or expertise (Krupinski, 2017). The ROC curve compared the test, the CMS, to the gold standard, RDS. The RDS scores were split into invalid and valid, where a score

of 6 or below characterized invalid results and a score of 7 or above characterized valid results. This was compared to the continuous measure of the total raw score from the CMS to find a similar cutoff, where there is a significant balance between true positive and false positive rates.

For aims 2 and 3, to investigate associations between RDS performance and individual word pairs as well as the consistency in recognition of repeated word pairs, chi-square tests of independence were used. This test compares the distribution of one categorical variable with the distribution of another categorical variable. The Fisher's exact test was examined, due to the smaller sample size of the current study (Hae-Young, 2017). These analyses yielded an understanding of the significance of each comparison, as well as the effect size.

For aim 4, the positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity were calculated for the significant items as identified on the chi-square tests. These analyses were used to examine the clinical utility of the significant items on the CMS that were comparable to RDS performance. The positive predictive value is the proportion of positive tests confirmed, or true positives, while the negative predictive value is the proportion of negative tests confirmed, or true negatives (Wong & Lim, 2011). The sensitivity ensures that true malingerers, those providing suboptimal effort, are detected by the test, while the specificity determines that those with good effort pass the test.

The aforementioned data analyses compare test results to the presence of the condition, as shown in Table 1, where $sensitivity = \frac{TP}{TP+FN}$, $specificity = \frac{TN}{TN+FP}$, $PPV = \frac{TP}{TP+FP}$, and $NPV = \frac{TN}{TN+FN}$. In the case of the current study, the test results from the CMS word pairs was measured against the gold standard, RDS, as presented in Table 2.

Table 1.

Test Result	Condition		
	Positive	Negative	
Positive	True Positive (TP)	False Positive (FP)	➔Positive Predictive Value (PPV)
Negative	False Negative (FN)	True Negative (TN)	➔Negative Predictive Value (NPV)
	⬇ Sensitivity	⬇ Specificity	

Table 2.

CMS Result	Gold Standard		
	RDS Fail	RDS Pass	
Incorrect	True Fail	False Fail	➔Positive Predictive Value (PPV)
Correct	False Pass	True Pass	➔Negative Predictive Value (NPV)
	⬇ Sensitivity	⬇ Specificity	

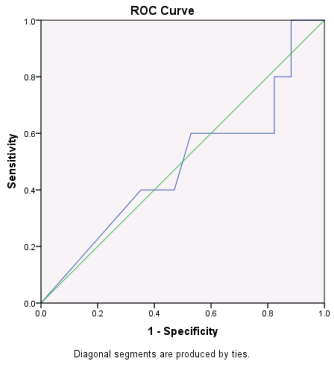
Data were analyzed with the Statistical Package for Social Sciences (SPSS, Version 24).

Results

Data

Aim 1

The first aim was to investigate the CMS recognition task as a measure to discriminate between poor and adequate effort. It was hypothesized that the total raw score of the Delayed Recognition task would discriminate between these groups. The CMS WP-DR total raw score was not useful in discriminating between poor and adequate effort for younger children ages 5-8 years ($AUC_{younger}$=0.49, p=0.94), or for the older youth (AUC_{older}=0.18, p=0.02). 12.8% of the sample had invalid RDS performance (N=10).

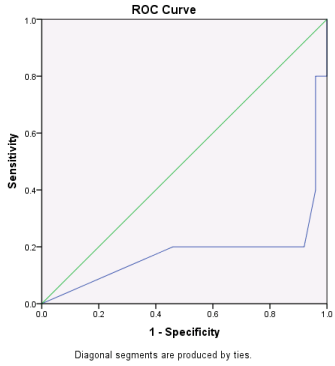


ROC Curve for Ages 5-8 years

Area Under the Curve

Test Result Variable(s): CMS_WP2DR_ysum

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.488	.157	.938	.181	.796



ROC Curve for Ages 9-16 years

Area Under the Curve

Test Result Variable(s): CMS_WP2DR_osum

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.182	.137	.020	.000	.450

Aim 2

The second aim was to investigate the associations between RDS performance and individual word pair items on the CMS. It was hypothesized that there would be some significant word pair items in comparison to RDS scores. Further examination of individual items indicated significant associations between RDS performance and the following five word pairs: item 5:

melt-rough, item 9: monkey-stove, item 19: wind-mile, item 21 (older children only): squeak-fame, and item 36 (older children only): apple-cloud (ϕ=0.31-0.61, $p<0.05$). Items 9 and 21 (older children only) appeared the most robust in discriminating between suboptimal and adequate `effort most similarly to the RDS (ϕ=0.43-0.61, p=0.006-0.007).

Aim 3

The third aim was to investigate the associations between RDS performance and the consistency in recognition of repeated word pairs on the CMS. It was hypothesized that there would be some associations between RDS scores and the recognition of repeated word pairs. Consistency in recognition of repeated word pairs was associated with RDS performance for one word pair, air-free (ϕ=0.25, $p<0.05$).

Aim 4

The fourth aim was to investigate the clinical utility of the significant word pairs. If there were significant word pairs, then it was hypothesized that these items or recognition of repeated items would have clinical utility. The sensitivity, specificity, PPV, and NPV for each significant word pair may be found in Table 4.

Table 4.

Item	p-Value	Phi	Sensitivity	Specificity	PPV	NPV
Item 5 (M-R)	0.014	0.37	97.01%	30.00%	90.28%	60.00%
Item 9 (M-S)	**0.006**	0.43	98.51%	30.00%	90.41%	75.00%
Item 19 (W-M)	0.021	0.31	91.04%	40.00%	91.04%	40.00%
Item 21*(S-F)	**0.007**	0.61	100.00%	40.00%	94.34%	100.00%
Item 36*(A-C)	0.019	0.48	98.00%	40.00%	94.23%	66.67%
Items 6&32 (A-F)	0.039	0.25	66.18%	70.00%	93.75%	23.33%

*only youth ages 9-16 years

Discussion

This study served as an exploratory examination into the potential use of a novel embedded effort measure. By comparing this potential measure to an established measure, five word pairs and one repeated word pair proved significance. Results demonstrated important information regarding the clinical application of a memory test, the Children's Memory Scale (CMS).

The results were unable to fully support the hypothesis that the total raw score would discriminate between good and poor effort. The older youth total raw score was significant for high false positives and low true positives. That is, the older youth's recognition task did not have the ability to correctly identify those examinees with good effort. Rather, this measure from the CMS had a higher proportion of valid results than there were examinees with actual valid RDS results. Due to the limited literature on the CMS as an effort measure, the exact contributors can only be speculated. Majority of previous research in recognition tasks as embedded effort measures have examined the differences between recall and recognition performance (Bernard, 1993; Johnson & Lesniak-Karpiak, 1997; Sullivan, 2002; Root et al., 2006), the results found could be due to the analysis used in this current study. The total raw score does seem to have less ability in discriminating between poor and adequate effort as compared to individual item recognition, though this is a novel approach to analyzing the ability of a recognition task as a measure of effort and was not compared to recall performance.

Previous studies regarding pediatric embedded effort measures are limited to very few investigations. The California Verbal Learning Test, Children's Version (CVLT-C) variable Recognition Discriminability proved to be a valid predictor of adequate versus suboptimal effort; a finding of recognition tasks that correlates with that found in adult populations (Baker et al.,

2014). Interestingly, this study did not show similar significance across the population of interest, and was only found to have significance in the older group of youth ages 9-16 years. Though, this significance did not indicate discrimination between valid and invalid results, rather there were high proportions of false positives in the older youth. The differences found between the recognition tasks from the older youth versus the younger youth could be attributed to the larger range of ages between 9 and 16 years rather than 5-8, therefore creating a larger sample for the older group.

Previous examination of the CMS as an embedded measure portrayed the need for a very large discrepancy between the recall and recognition scores to suggest suboptimal effort (Perna et al., 2012). Because this study investigated different measures, it is difficult to compare the two in the sense of the clinical utility. The findings of the current study do support the suggestion that the CMS may possess clinical utility for embedded effort measures, though it is most effective to analyze a combination of items, rather than an overall cutoff score.

Some adult measures have been shown to have clinical utility in pediatric samples based on the same cutoff scores (Donders, 2005; Lichtenstein et al., 2016). Examinations of pediatric measures should also establish cutoff scores from unique tests developed specifically for children. There is evidence that adult cutoffs from tests such as the Test of Memory Malingering (TOMM) are valid for use in children (Blaskewitz et al., 2008; Donders, 2005; MacAllister et al., 2009). Not all adult effort tests, however, are useful in pediatric populations, including the WMT and the Computerized Assessment of Response Bias, for example (Conder et al., 1992). Children's performance on these adult tasks may be dependent on their abilities, a characteristic that should not be present during effort testing (Green & Flaro, 2003). In the case of this study, a

test developed to examine children's memory was measured against another embedded measure, the RDS, which was extended from adult cutoff standards.

Validity measures remain insensitive to cognitive abilities, so as to detect noncredible performance across clinical populations (Kirkwood, 2015). Thus, the Full Scale IQ (FSIQ) of individuals should not affect the individual's ability to pass the proposed effort test. Post-hoc analyses revealed information regarding the intellectual functioning of the individuals who passed and failed the RDS and the CMS WP-DR significant word pairs. The FSIQ for individuals who failed the RDS ranged between 61 and 96 with an average of about 79 and a standard deviation of about 13. The mean IQ's of the individuals who passed and failed the RDS were statistically different ($p=0.03$). The FSIQ for individuals who failed the significant CMS word pair items (items 5, 9, 19, 21, 36 and consistency of 6 and 32) ranged from 55 to 116 with an average of about 86 and a standard deviation of about 15. The mean IQ's of the individuals who got these items correct and those who did not were not statistically different for any word pair ($p=0.20$-0.92). The difference between average IQ's of the performance on the RDS may be a result of the smaller sample of individuals who failed this measure. RDS performance has also shown to vary in relation to cognitive abilities and is not as effective for validity measurement in children with significant cognitive impairments (DeRight & Carone, 2015; Welsh et al., 2012; Kirkwood et al., 2011). Thus in this sample, the RDS may have not been a good measure of validity, particularly for the children with very low IQ scores. The children who answered the significant word pairs correctly and those who answered incorrectly did not show significant differences in IQ scores, and ranged similarly in intellectual functioning as the sample. This shows promising results that recognition of these word pairs does not depend on intellectual functioning.

This study was limited to a mixed pediatric clinical sample in that the groups reported to have higher rates of invalidity problems, such as those with head injury or concussions, are underrepresented in this sample. As such, only about 13% of individuals in the sample had invalid RDS performance. Therefore, the individuals with invalid results were also underrepresented for the current study.

Further research in the domain of pediatric embedded effort measures, and pediatric effort measures generally, must be conducted to determine appropriate variables for establishing credible effort. Future examination of the clinical utility of items 9 and 21 (older children only) on the CMS WP-DR should include larger pediatric samples and additional confirmatory effort measures, such as the Memory Validity Profile (MVP) or the Test of Memory Malingering (TOMM). These investigations may reveal additional significant items from this measure. Future research should also more closely examine the characteristics of those individuals who fail validity measures. This would contribute to the understanding of any patterns of factors that contribute to invalid performance.

Establishing several embedded measures of effort for pediatric neuropsychological assessments would benefit patients and mental health care providers extensively. These measures would provide valuable information to validate diagnoses in correlation to the test battery results. Measuring effort is an important component of neuropsychological assessment to ensure best practices.

References

Baker, D. A., Connery, A. K., Kirk, J. W., & Kirkwood, M. W. (2014). Embedded Performance Validity Indicators Within the California Verbal Learning Test, Children's Version. *The Clinical Neuropsychologist*, *28*(1), 116–127. https://doi.org/10.1080/13854046.2013.858184

Bernard, L. C. (1990). Prospects for faking believable memory deficits on neuropsychological tests and the use of incentives in simulation research. Journal of Clinical and Experimental Neuropsychology, 12, 715–728.

Bernard, L. C., Houston, W., & Natoli, L. (1993). Malingering on neuropsychological memory tests: Potential objective indicators. Journal of Clinical Psychology, 49, 45–53.

Bigler, E. D. (2014). Effort, symptom validity testing, performance validity testing and traumatic brain injury. *Brain Injury*, *28*(13–14), 1623–1638. https://doi.org/10.3109/02699052.2014.947627

Blaskewitz, N., Merten, T., & Kathmann, N. (2008). Performance of children on symptom validity tests: TOMM, MSVT, and FIT, *Archives of Clinical Neuropsychology*, *23*, 379-391.

Boone, K. B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. *The Clinical Neuropsychologist*, *23*(4), 729–41. https://doi.org/10.1080/13854040802427803

Brooks, B. L. (2012). Victoria Symptom Validity Test performance in children and adolescents with neurological disorders. *Archives of Clinical Neuropsychology*, *27*(8), 858–868. https://doi.org/10.1093/arclin/acs087

Brooks, B. L., Sherman, E. M. S., & Iverson, G. L. (2014). Embedded validity indicators on

CNS vital signs in youth with neurological diagnoses. *Archives of Clinical Neuropsychology*, *29*(5), 422–431. https://doi.org/10.1093/arclin/acu029

Cohen, B. (1997). Children's Memory Scale. San Antonio, TX: The Psychological Corporation.

Conder, R. J., Allen, L. L., Cox, D. R., & King, C. M. (1992). Computerized Assessment of Response Bias: Revised Edition.

DeRight, J., & Carone, D. A. (2015). Assessment of effort in children: A systematic review. *Child Neuropsychology*, *21*(1), 1–24. https://doi.org/10.1080/09297049.2013.864383

Donders, J. (2005). Performance on the Test of Memory Malingering in a Mixed Pediatric Sample. Child Neuropsychology, 11(2), 221-227. doi:10.1080/09297040490917298

Green, P., & Flaro, L. (2003). Word Memory Test Performance in Children. *Child Neuropsychology*, *9*(3). 189-207.

Greher, M. R., & Wodushek, T. R. (2017). Performance Validity Testing in Neuropsychology. *Journal of Psychiatric Practice*, *23*(2), 134–140. https://doi.org/10.1097/PRA.0000000000000218

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*, *6*(3), 218.

Gunn, D., Batchelor, J., & Jones, M. (2010). Detection of simulated memory impairment in 6-to 11-year-old children. *Child Neuropsychology*, *16*(2), 105-118.

Hae-Young, K. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. Restorative Dentistry & Endodontics, Vol 42, Iss 2, Pp 152-155 (2017), (2), 152. doi:10.5395/rde.2017.42.2.152

Heilbronner, R., Sweet, J., Morgan, J., Larrabee, G., Millis, S., & Conference Participants. (2009). American Academy of Clinical Neuropsychology Consensus Conference

Statement on the Neuropsychological Assessment of Effort, Response Bias, and Malingering. *The Clinical Neuropsychologist*, *23*(7), 1093–1129. https://doi.org/10.1080/13854040903155063

Johnson, J.L., & Lesniak-Karpiak, K. (1997). The effects of warning on malingering on memory and motor tasks in college samples. Archives of Clinical Neuropsychology, 12, 231-238.

Kirk, J. W., Harris, B., Hutaff-Lee, C. F., Koelemay, S. W., Dinkins, J. P., & Kirkwood, M. W. (2011). Performance on the test of memory malingering (TOMM) among a large clinic-referred pediatric sample. *Child Neuropsychology : A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *17*(August 2012), 242–254. https://doi.org/10.1080/09297049.2010.533166

Kirkwood, M. W. (2015). *Validity testing in child and adolescent assessment: Evaluating exaggeration, feigning, and noncredible effort.* New York, NY: The Guilford Press.

Kirkwood, M. W., Hargrave, D. D., & Kirk, J. W. (2011). The value of the WISC-IV digit span subtest in detecting noncredible erformance during pediatric neuropsychological examinations. *Archives of Clinical Neuropsychology*, *26*(5), 377–384. https://doi.org/10.1093/arclin/acr040

Krupinski, E. A. (2017). Receiver operating characteristic (ROC) analysis. *Frontline Learning Research*, *5*(2).

Lichtenstein, J. D., Erdodi, L. A., Rai, J. K., Mazur-Mosiewicz, A., & Flaro, L. (2016). Wisconsin Card Sorting Test Embedded Validity Indicators Developed for Adults can be Extended to Children. *Child Neuropsychology*, *7049*(July 2017), 1–14. https://doi.org/10.1080/09297049.2016.1259402

Loughan, A. R., Perna, R., & Hertza, J. (2012). The alue of the Wechsler intelligence scale for

children-fourth edition Digit Span as an embedded measure of effort: An investigation into children with dual diagnoses. *Archives of Clinical Neuropsychology*, *27*(7), 716–724. https://doi.org/10.1093/arclin/acs072

Macallister, W. S., Nakhutina, L., Bender, H. a, Karantzoulis, S., & Carlson, C. (2009). Assessing Effort During Neuropsychological Evaluation with the TOMM in Children and Adolescents with Epilepsy. *Child Neuropsychology : A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *15*(6), 521–531. https://doi.org/10.1080/09297040902748226

Miele, A. S., Gunner, J. H., Lynch, J. K., & McCaffrey, R. J. (2012). Are embedded validity indices equivalent to free-standing symptom validity tests? *Archives of Clinical Neuropsychology*, *27*(1), 10–22. https://doi.org/10.1093/arclin/acr084

Nagle, A., Everhart, D., Durham, T., McCammon, S., & Walker, M. (2006). Deception strategies in children: Examination of forced choice recognition and verbal learning and memory techniques. *Archives of Clinical Neuropsychology*, *21*, 777–785.

Pearson. (2014). *Weschler Intelligence Scale for Children Technical and Interpretive Manual* (5th edition). Bloomington, MN: David Weschler. .

Perna, R., Loughan, A. R., Hertza, J., & Segraves, K. (2012). The Value of Embedded Measures in Detecting Suboptimal Effort in Children: An Investigation into the WISC-IV Digit Span and CMS Verbal Memory Subtests. *Applied Neuropsychology: Child*, *2965*(April 2015), 1–7. https://doi.org/10.1080/21622965.2012.691067

Rickards, T. A., Cranston, C. C., Touradji, P., & Bechtold, K. T. (2017). Embedded performance validity testing in neuropsychological assessment: Potential clinical tools. *Applied Neuropsychology. Adult*, *5498*(July), 1–12.

https://doi.org/10.1080/23279095.2017.1278602

Root, J. C., Robbins, R. N., Chang, L., & van Gorp, W. G. (2006). Detection of inadequate effort on the California Verbal Learning Test-Second edition: Forced choice recognition and critical item analysis. *Journal of the International Neuropsychological Society*, *12*(5), 688–696. https://doi.org/10.1017/S1355617706060838

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable Digit Span: A Systematic Review and Cross-Validation Study. *Assessment*, *19*(1), 21–30. https://doi.org/10.1177/1073191111428764

Sullivan, K., Deffenti, C., & Keane, B. (2002). Malingering on the RAVLT: Part II. Detection strategies. Archives of Clinical Neuropsychology, 17, 223–233.

Welsh, A. J., Bender, H. A., Whitman, L. A., Vasserman, M., & MacAllister, W. S. (2012). Clinical utility of Reliable Digit Span in assessing effort in children and adolescents with epilepsy. *Archives of Clinical Neuropsychology*, *27*(7), 735–741. https://doi.org/10.1093/arclin/acs063

Wong, H. B., & Lim, G. H. (2011). Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV. *Proceedings of Singapore healthcare*, *20*(4), 316-318.