Use of the Wechsler Adult Intelligence Scale Digit Span subtest for malingering detection: A meta-analytic review

Lindsey J. Jasinski, David T. R. Berry, Anni L. Shandera, and Jessica A. Clark

University of Kentucky, Lexington, KY, USA

Twenty-four studies utilizing the Wechsler Adult Intelligence Scale (WAIS) Digit Span subtest—either the Reliable Digit Span (RDS) or Age-Corrected Scaled Score (DS-ACSS) variant—for malingering detection were meta-analytically reviewed to evaluate their effectiveness in detecting malingered neurocognitive dysfunction. RDS and DS-ACSS effectively discriminated between honest responders and dissimulators, with average weighted effect sizes of 1.34 and 1.08, respectively. No significant differences were found between RDS and DS-ACSS. Similarly, no differences were found between the Digit Span subtest from the WAIS or Wechsler Memory Scale (WMS). Strong specificity and moderate sensitivity were observed, and optimal cutting scores are recommended.

Keywords: Digit span; Malingering; Feigning; Wechsler Adult Intelligence Scale–Third Edition; Wechsler Memory Scale–Third Edition; Wechsler; Meta-analysis.

Neuropsychological assessments provide information vital to diagnosis, prognosis, and treatment following an injury or illness; however, test results may be rendered useless when feigning or suboptimal effort occurs. Thus, it has been strongly recommended that an examination of malingering should be included in every neuropsychological test battery (Bush, et al., 2005). According to the Diagnostic and Statistical Manual of Mental Disorders–Fourth Edition, Text Revision (DSM–IV–TR; American Psychiatric Association, 2000), malingering is the “intentional production of false or exaggerated physical or psychological symptoms, motivated by external incentives” (p. 739). Malingering is thought to be a multidimensional construct, with distinctions made between psychiatric, neurocognitive, and somatic feigning (Larrabee, 2007). In general, estimates of the base rate of neurocognitive feigning have demonstrated substantial rates of malingering, particularly in forensic settings. For instance, Frederick, Crosby, and Wynkoop (2000) estimated 44% of criminal defendants had an invalid test-taking approach, while Mittenberg, Patton, Canyock, and Condit (2002) showed rates of malingered around 41% in litigants after a traumatic brain injury (TBI). Thus, it seems that 40% is a reasonable approximation of the base rates of neurocognitive feigning in forensic assessments. In nonforensic contexts, estimates of feigning are lower, around 7.3%, but are significant enough to warrant attention (Berry & Schipper, 2007). These estimates point to the need for well-validated measures of malingering and their routine utilization.

Current malingering assessment practices

Dedicated malingering tests

Tests developed specifically to assess feigning, also known as dedicated malingering tests, have received significant attention in recent years. Several such tests have been well supported in the literature (Grote & Hook, 2007). The most common type of dedicated malingering test utilizes a forced-choice memory paradigm, where patients are shown a stimulus to remember and then are asked to choose the correct response from several options after a brief delay. In this model, feigning is suspected in the presence of either worse-than-chance performance or scores falling below an a priori cutting score. Examples of well-validated forced-choice measures for detecting malingered neurocognitive dysfunction include the Test of Memory Malingering.
(Tombaugh, 1996), Word Memory Test (Green, Allen, & Astner, 1997), Letter Memory Test (Inman et al., 1998), Validity Indicator Profile (Frederick & Crosby, 2000), Portland Digit Recognition Test (Binder, 2002), and Victoria Symptom Validity Test (Slick, Hopp, Strauss, & Thompson, 1997). See Vickery, Berry, Inman, Harris, and Orey (2001), and Grote and Hook (2007) for more comprehensive reviews.

While dedicated malingering tests are by far the most widely used in clinical practice and boast the most empirical support, they suffer from some limitations. One of the strongest criticisms of dedicated malingering tests is their vulnerability to coaching, or the provision of information to examinees prior to an evaluation regarding the presence and/or nature of malingering indices. For example, attorneys might discuss specific tests with clients or warn them not to fail tests where they have to choose answers from two options. The impact of attorney coaching has been well documented in the malingering literature to date (Gutheil, 2003; Lees-Haley, 1997; Suhr & Gunstad, 2007; Wetter & Corrigan, 1995). Additionally, increasing amounts of information available via the Internet make self-coaching an easy task when preparing for a neuropsychological evaluation (Bauer & McCaffrey, 2006; Ruiz, Drake, Glass, Marcotte, & van Gorp, 2002). For example, Allen and Green (2001) documented a steady decline in failure rates of the Computerized Assessment of Response Bias (CARB) over a six-year period corresponding closely to increasing Internet availability, and Berry and Schipper (2007) describe a recent search of the Psychological Assessment Resources (PAR, Inc.) website that provided extensive information on the names, acronyms, and purpose of several malingering measures. This, coupled with the finding that nearly 20% of Americans surveyed felt that intentional distortion or misrepresentation within the context of the legal compensation system was acceptable (Public Attitude Monitor, 1992), points to the importance of developing accurate tests for detecting malingering.

In addition to vulnerability to coaching, another concern with dedicated malingering tests is the administration time required. For many well-validated procedures, upwards of 20 minutes additional time is needed for a single dedicated malingering test. Given Rogers’s (1997) recommendation that multiple indicators of feigning be employed, the added testing time could be significant. Further, dedicated tests of malingering provide data relevant only to effort, but no information on other cognitive functions, making them time consuming with very little relevant diagnostic information obtained (Mathias, Greve, Bianchini, Houston, & Crouch, 2002).

**Embedded indices**

In response to the notable limitations of dedicated malingering tests, interest in indices embedded within standard neuropsychological tests has grown in recent years. Malingering indices developed within large-scale batteries have included subtests from the Wechsler Adult Intelligence Scale Revised (WAIS–R; Wechsler, 1981) and Third Edition (WAIS–III; Wechsler, 1997a; Digit Span: Trueblood & Schmidt, 1993; Vocabulary–Digit Span, WAIS Discriminant Function: Mittenberg, Theroux-Fichera, Zielinski, & Heilbrunner, 1995; Reliable Digit Span, RDS: Greiffenstein, Baker, & Gola, 1994), from the Wechsler Memory Scale–Revised (WMS–R; Wechsler, 1987) and Third Edition (WMS–III; Wechsler, 1997b; Logical Memory recognition: Killgore & DellaPietra, 2000; Discriminant Function: Mittenberg, Patton, & Legler, 2003), and from the Halstead–Reitan Battery (Discriminant Function: Mittenberg, Rothole, Russell, & Heilbrunner, 1996). Additionally, several individual tests have been examined for their ability to detect malingering, including the Wisconsin Card Sorting Test (Bernard, McGrath, & Houston, 1996), Auditory Verbal Learning Test (Binder, Villaneuva, Howieson, & Moore, 1993), California Verbal Learning Test (Millis, Putnam, Adams & Ricker, 1995), Finger Tapping Test (Arnold et al., 2005), and Rey Complex Figure Test (Meyers & Volbrecht, 1999), among others. See Larrabee (2007) for a more comprehensive review of the malingering research on several of these instruments.

**Methodological considerations**

**Research methodology**

Methodological rigor has been a topic of considerable attention with regard to malingering research. While case studies and differential prevalence designed studies have been utilized in the past, they are not currently accepted as standard practice for malingering research. In contrast, the most frequently utilized methodologies include simulation and known-groups design studies. Simulation studies involve a group of participants asked to feign symptoms of a disorder while completing a series of tests, who are compared to a group of “normal” participants asked to respond honestly or to perform to the best of their ability on the same test battery. The honest group often has no prior psychiatric or neurological diagnosis. However, Rogers (2008) points out that such simulation designs have only “marginal relevance to malingering because they do not differentiate feigned from genuine disorders” (p. 413). Rather, the stronger version of the simulation design study, enhanced simulation design, employs a clinically relevant honest comparison group, thus providing direct information on the differences between the feigned and authentic disorder. For example, in traumatic brain injury research, the clinically relevant group would be participants who had sustained a brain injury but were not seeking compensation and had verified honesty (Rogers, 2008).

Simulation studies are often utilized due to their strong internal validity, control over group assignment, and convenience. However, concerns are often raised regarding the generalizability of simulation studies. Several methodological considerations have been recommended to strengthen the external validity of a study. For instance, in addition to the malingering group being clinically relevant or having experienced the condition...
of interest (e.g., brain injury), honest and malingering groups should be matched demographically. Further, Rogers (2008) recommends that malingering instructions be clear and comprehensible to the participants, as well as specific. A warning for participants to feign believably and relevant proportional incentives, both positive for following instructions, and negative for failing to do so, should be included to increase motivation in the malingering group. Instructions should include information on relevant symptoms of interest, as well as detection strategies and avoidance techniques. Additionally, allowing preparation time for participants after they are given symptom information greatly enhances similarity to real-world evaluations, as does the inclusion of both malingering and nonmalingering tests within the research battery. Finally, a postexperimental check for understanding and cooperation with instructions is necessary to exclude data from subjects who were not compliant. If utilized, each of these methodological considerations can help maintain strong internal validity, while improving external validity (Rogers, 2008).

Known-groups methodology, in contrast, is often utilized for its strong external validity, at the expense of some internal validity and study control. Known-groups designs employ a fully clinically relevant malingering group thought to be feigning based on failure of one or more well-validated malingering tests. This “known” malingering group is then compared to a clinically relevant honest group of similar background and injury severity. The generalizability of such results is very high, given that both groups have experienced the necessary antecedent events and motivation to malingering similar to those seen in clinical practice. However, known-groups methodology is limited based upon the quality of the malingering indicators used to determine group status, and imperfect measures result in imperfect groups and subsequently lower internal validity (Rogers, 2008).

Research methodology is an important factor to consider when conducting meta-analytic studies, because differing methodologies greatly impact study outcomes, resultant operating characteristics, and diagnostic classification parameters of any malingering detection tool. Because each has strengths and weaknesses, well-validated malingering measures should have supportive evidence from both types of methodologies, per Rogers’s (1997) recommendations. Further, specific methodological characteristics should be coded and explored as confounds and contributing factors to the findings, as is the case in the current meta-analytic review.

**Statistical methods**

Diagnostic validity refers to a “test’s ability to differentiate persons with and without a specified disorder” (Larrabee & Berry, 2007, p. 14), in this case malingering. A cutting score is a score at or below which, based on prior research, an individual is thought to be malingering. Sensitivity and specificity are the most commonly reported statistics in malingering detection studies and provide information on the test’s ability to classify malingering. Sensitivity (SN) refers to the probability of a positive test sign (i.e., failure of malingering test) in individuals who are malingering. Specificity (SP), in contrast, refers to the probability of a negative test sign (i.e., passing malingering test) in individuals who are responding honestly. The hit rate is the overall efficiency of the test, taking into account both sensitivity and specificity rates. It is important to note that distributions from which these scores are derived usually overlap, and thus there is no score that perfectly separates malingering and honest groups, resulting in the need to choose the most effective cutting score based on minimization of relevant errors (Greve & Bianchini, 2004). While sensitivity and specificity rates are the most frequently cited statistics in malingering detection, they do not represent the entire picture. When derived from malingering studies, these statistics are typically based on equal sample sizes with known base rates of malingering approximating 50%. Positive and negative predictive powers, in contrast, take into account the relevant base rate and therefore provide a more pertinent statistic for clinicians. Additionally, they are more relevant for clinicians in that they predict the probability that an individual is either honest or malingering based on test scores. Specifically, positive predictive power (PPP) refers to the probability of malingering given a positive test score, while negative predictive power (NPP) is the probability of the absence of malingering, or honest responding, given a negative test score. However, given that these values vary with the base rate of each malingering study, they are not explored in the present meta-analysis, but can be easily calculated using the relevant sensitivity and specificity values provided subsequently. Finally, an effect size describes a test’s ability to differentiate between two groups, with larger effect sizes indicative of larger average differences between the groups. Greater average differences, then, translate into more accurate diagnostic classification statistics. Each of these statistics are examined in further detail in the present study.

**Goals of present study**

The goal of the present study is to provide a comprehensive, meta-analytic evaluation of the ability of the Digit Span subtest in detecting malingered neurocognitive dysfunction. The Digit Span subtest is embedded within the revised, third, and fourth editions of the Wechsler Adult Intelligence Scale (WAIS–R, WAIS–III, WAIS–IV), as well as the revised, third, and fourth editions of the Wechsler Memory Scale (WMS–R, WMS–III, WMS–IV). It requires an examinee to repeat increasing strings of digits in forward and reverse order, until two consecutive strings of the same length are missed. A variation, Reliable Digit Span (RDS; Greiffenstein et al., 1994), has been developed using the sum of the longest strings of digits forward and backward where both trials were passed. RDS was developed specifically as a malingering detection tool and has been extensively studied.
The Digit Span subtest was chosen as the focus for the present review for several reasons. Primarily, it is embedded within the Wechsler Adult Intelligence Scale (WAIS–R, WAIS–III, WAIS–IV) and Wechsler Memory Scale (WMS–R, WMS–III, WMS–IV), which are among the most frequently utilized neuropsychological instruments in clinical practice today (Camera, Nathan, & Puente, 2000; Rabin, Barr, & Burton, 2005). As such, retrospective malingering analyses can be conducted if a clinician administered either of these instruments and is in need of information regarding malingering status. Such retrospective analyses can also be particularly useful in attaining research goals. Further, it is virtually identical across several revisions of the WAIS and WMS, thereby increasing the number of studies available for the present meta-analysis, subsequently improving the reliability of the findings. Finally, the Digit Span subtest is easily and quickly administered either individually or as part of a battery, so is particularly user friendly in a clinical setting. Thus, the focus of the present study is on a meta-analytic review of the Digit Span subtest and its ability to detect malingering.

METHOD

A priori inclusion criteria

Malingering studies utilizing the Digit Span subtest of the revised or third editions of the WAIS and WMS were surveyed through March 2009. Studies investigating the usefulness of either the total age-corrected scaled score (DS-ACSS) or Reliable Digit Span (RDS) were included. To be considered for inclusion, manuscripts were required to come from a peer-reviewed journal, peer-reviewed article in press, or English-language dissertation. Conference presentations and papers were not included in the present study, as they do not typically undergo as thorough a review process. In cases where dissertation results were later published, only the published paper was included for evaluation.

Publications meeting these criteria were examined and were included if they utilized known-groups or enhanced simulation methodology (e.g., with a clinically relevant honest comparison group). Thus, studies using student or community volunteers instructed to malingering were included provided they were compared to a group with a history of traumatic brain injury or other neurological disorder. Studies were excluded if comparisons were made only with normal controls, as they do not provide information on a test’s ability to discriminate feigned from true injury (Rogers, 1997). In the present study, clinical samples primarily included traumatic brain injury and chronic pain patients, although several mixed neurological groups were also included. Chronic pain patients were included because they often feign neurocognitive symptoms in addition to somatic complaints (Iverson & McCracken, 1997). Groups feigning psychiatric symptoms or mental retardation were not included in the present review to reduce the heterogeneity among groups and the possibility of comparing “apples to oranges,” a common criticism of meta-analysis (Lipsey & Wilson, 2001). Thus, the primary focus of this paper was on feigned neurocognitive symptoms. The use of the Digit Span test in detecting other types of feigning should be explored in future studies.

Data collection

Keyword-guided searches of the PsycINFO and MedLine databases were undertaken to identify studies of any version of the WAIS or WMS used to detect malingering. Combinations of the keywords “malinger,” “feign,” “dissimulate,” “fak,” and “pattern analysis” were crossed with “WAIS,” “WAIS-R,” “WAIS-III,” “WMS,” “WMS-R,” and “WMS-III,” both in acronym and written form, as well as “Wechsler.” The asterisks are used in search databases to allow for variable suffixes (such as -ing, -ed, -s, -ion). Additionally, several of the prominent pattern analysis techniques used in malingering literature were searched for with either the WAIS or the WMS, including “performance curve,” “magnitude of error,” “violation of learning principle,” and “atypical pattern analysis.” Reference sections of review papers, book chapters, test manuals, and empirical studies were scrutinized to identify additional manuscripts. Lastly, email contact with highly published individuals using the WAIS or WMS for malingering detection was attempted in order to solicit any papers in press or under review.

As a result of these keyword-guided searches, 74 papers were identified, 43 of them utilizing a version of the WAIS, 24 reviewing a version of the WMS, and 4 studies using both the WAIS and the WMS. It was necessary to exclude studies on the following basis: lack of a clinically relevant control group (n = 10), lack of any feigning group (n = 5), the use of a correlational or differential prevalence design (n = 6), or because the Digit Span subtest was used to determine malingering status for a known-groups design (n = 2). Further, 21 studies were excluded because they examined subtests of the WAIS or WMS other than Digit Span for malingering detection. Finally, 6 papers were also excluded on the basis of being a conference presentation. A total of 24 papers (WAIS: n = 23; WMS: n = 1) were considered appropriate based on the a priori inclusion criteria and were coded. These studies are listed in the Appendix.

Selection of data for review

In some cases, multiple contrasts were available from a single study. Here, only the independent contrasts were selected because using effect sizes from overlapping groups increases statistical dependence and inflates the mean effect sizes for the study (Lipsey & Wilson, 2001). One source of nonindependence involved comparing a single feigning group with multiple honest groups. For example, a few studies included multiple honest groups of mild and severe TBI patients, all compared to a single malingering group. In this case, the contrast using a group of mild TBI patients was chosen because it was
thought to be most clinically relevant, given that mild TBI patients are most likely to present for neuropsychological testing and have the greatest motivation to exaggerate symptoms, compared to those with a severe TBI where compensation could theoretically be based upon physical findings and notable functional decline. Another source of nonindependence of contrasts was the use of several malingering groups compared to a single honest group. In this situation, the more stringent classification criteria for the malingering group were used (e.g., choosing definite over possible malingered neurocognitive dysfunction, MND, based on the Slick, Sherman, & Iverson, 1999, criteria).

A final source of nonindependent contrasts arose from the overlapping nature of the Digit Span age-corrected scaled score (DS-ACSS) and Reliable Digit Span (RDS) variants, where the same trial scores are used to compute each scale. Four of the studies included provided enough information to calculate effect sizes for both variants. In order to create independent contrasts, two studies were randomly assigned to contribute to the DS-ACSS analyses, while the other two were assigned to the RDS analyses. While this method slightly reduced the number of studies available for each of the analyses, the benefits of independent effects outweighed the desire for additional studies.

Data extraction

Demographic variables such as age, gender, and mean education level were extracted from each of the studies and are presented in the Appendix. Additionally, the type of group was coded according to the following system: traumatic brain injury, mixed neurological, and chronic pain. The type of participants in the malingering group was also coded, including patients, students, or community volunteers. Means and standard deviations for RDS and DS-ACSS were extracted when available, as were sensitivity and specificity, hit rate, and methodological characteristics of each study, specific to the research design type. Methodological characteristics deemed relevant in this study were used to rate the methodological strength of each study and are based on recommendations from prior literature reviews (Berry, Baer, Rinaldo, & Wetter, 2002; Berry & Schipper, 2007). A total methodological strength score for each contrast was calculated and was divided by the number of possible points for each design type (e.g., Sim, enhanced simulation; KG, known groups). This was done because 16 variables were coded for simulation design studies, while known-groups studies had only 7. Thus, a proportion score was used to directly compare the designs. Higher scores represent methodologically stronger studies.

In order to ascertain coding accuracy, 42% (10) of publications were cross-coded by a second rater. Reliabilities were found to be adequate for demographic variables ($r = .97$), methodological characteristics ($r = .92$), means and standard deviations ($r = 1.0$), and diagnostic accuracy statistics (sensitivity $r = 1.0$, specificity $r = 1.0$). Discrepancies were discussed until 100% consensus was reached.

RESULTS

A total of 24 studies were selected and were coded according to the inclusion criteria. Of these, two studies did not provide enough information (i.e., standard deviations) to calculate an effect size, while an additional two studies provided enough information to compute two independent effect sizes. In other words, each of these two studies included at least two feigning and two honest groups for independent comparison. Thus, the analysis of effect sizes included 24 independent contrasts. Further, of the 24 studies examined, only 19 studies provided information on the relevant diagnostic accuracy statistics. See Appendix for a summary of each study, including effect sizes and diagnostic accuracy statistics.

Calculation of effect sizes was completed utilizing the Hedges’s $d$ statistic from the meta-analytic software program DSTAT (Johnson, 1989). Further analyses were conducted using the Statistical Package for the Social Sciences (SPSS) software. The Hedges’s $d$ value represents the distance between two sample means in terms of pooled standard deviation units. Although the alternative effect size Cohen’s $d$ is often used, it has been shown to be biased with small sample sizes, such as those often utilized in neuropsychology research (Hedges & Olkin, 1985). Hedges’s $d$ is a standardized metric thought to be an unbiased estimate of effect size by correcting for sample size with the inverse variance, thus giving more weight to studies with larger sample sizes (Lipsey & Wilson, 2001). Thus, in the text below, all $d$ values are Hedges’s $d$.

The next step in data analysis was screening the effect sizes for outliers, or unusually high or low values that might skew subsequent analyses. An important consideration when conducting outliers analysis is the balance between retaining all possible comparisons in order to investigate any methodologically interesting determinants of variability in the dataset, and excluding extreme values unlikely to be representative of observations within the population of interest. In the present study, adequate variability was seen in the effect sizes, so the desire for a representative set of findings from the population was a priority, and outliers were excluded. Further, such extreme effect sizes have disproportionate influence on meta-analytical statistics (e.g., means, variances) and may distort them (Lipsey & Wilson, 2001). Thus, a step-and-fences procedure (Tukey, 1977) was completed for the set of 24 independent contrasts. This procedure uses the median $d$ value to determine the likelihood of a value belonging to the same distribution. Values outside of the inner fence, located approximately 2.67 standard deviations away from the median $d$ value, are considered outliers. Application of this process resulted in two outliers being identified from the set of independent contrasts ($d$ values: 2.97 and 3.46), both of which contributed to a positively skewed distribution. It is noteworthy that one outlier was from a study with a very small sample size ($n = 7$ for the feigning group) while the other
came from the only study utilizing inmates as part of the feigning group. Removal of these outliers brought greater normality to the distribution.

Following outlier removal, 22 independent contrasts using either the DS-ACSS ($n = 12$) or RDS ($n = 10$) variant remained for analysis. The distribution of the sample was roughly normal. The $d$ values for the independent contrasts ranged from 0.92 to 2.46, with a weighted mean effect size of $1.25$ (95% confidence interval, CI = 1.13–1.37) and a median of $1.23$. Overall, these mean values indicate that the two variants of the Digit Span test separate malingering and honest groups by just over one pooled standard deviation, a significant effect size that is considered large (Cohen, 1988). Thus, honest participants had higher scores on the DS-ACSS and RDS subtests than did those thought to be malingering, indicating that malingerers suppressed their scores beyond what would be observed in a clinical population responding honestly.

Representativeness of study findings

While these effect sizes are substantial, several factors should be considered before having confidence in the results of a meta-analysis, including the possibility that the findings are not characteristic of all studies conducted in an area. Specifically, meta-analysis has often been challenged on the basis of the “file drawer problem,” whereby only statistically significant results are accepted for publication in peer-reviewed journals, leaving open the possibility of small effects or negative findings going unpublished. The result would be effect sizes that overestimate the actual effect in the population. In response to such criticism, Rosenthal (1991) developed the fail-safe $N$ statistic, using it to estimate the number of studies with a null effect that would have to exist in researcher’s “file drawers” in order to reduce the observed mean effect size from the sample studies to a nonsignificant level. Using this procedure, the fail-safe $N$ based on the set of 22 independent contrasts in this study resulted in an estimated 244 unpublished null findings necessary to render the overall effect nonsignificant ($p > .05$). Based on this analysis, it appears highly improbable that this number of studies exist, thereby increasing confidence in the findings of this study as representative of the population at large.

An additional consideration is the extent to which methodological rigor influenced the overall effects derived from the studies. Intuitively, one would assume that studies with increased methodological rigor would result in less variable, and thus more reliable, effect sizes. In order to ascertain the relationship between methodological quality and the effect sizes, a correlation coefficient was calculated using the methodological strength score assigned to each study and resulted in $r = -.24$ ($p = .28$). This indicates that, overall, methodological quality was not significantly associated with the magnitude of effect size.

Thus, based on the sizable mean effect sizes, large value of the fail-safe $N$, and lack of a significant relationship between methodological rigor and effect size, the results of this study can be interpreted as representative of the population of studies in the area of malingering and the Digit Span test, including both DS-ACSS and RDS. This subtest is able to differentiate honest and malingering groups by more than one pooled standard deviation on average. However, what is not yet known is extent of variability in the effect sizes and the diagnostic accuracy of DS-ACSS and RDS. These issues will be explored subsequently.

Variability of effect sizes

When summarizing data across studies, a further consideration is the homogeneity of the effect sizes derived, with significant heterogeneity suggesting that a single mean effect size should not be used to represent an overall effect. In other words, the effect sizes may differ from the population mean by more than just sampling error, rather by systematic differences that can be further examined. The $Q$ statistic is commonly used in meta-analysis, with a significant result indicating heterogeneity. It represents an asymptotic chi-square distribution with $k - 1$ degrees of freedom (Hedges & Olkin, 1985). Results from the 22 independent contrasts indicate that a significant amount of variability in the effect sizes was not observed, $Q(21) = 30.58, p < .08$, although the results approached significance. Hence, the overall mean effect size does appear to adequately represent the overall ability of the Digit Span test to detect malingering.

Within- and between-measure differences

Although the overall effect size showed no statistically significant heterogeneity, theoretically derived moderators were explored to examine their potential impact on the data, as the idea of predetermined focused contrasts is potentially unrelated to overall variability and is therefore useful to examine (Rosenthal, 1991). Further, since the $Q$ statistic noted above approached significance, moderators may be useful in explaining the observed variability, despite lack of formal statistical significance. This exploration included the variant utilized (RDS or DS-ACCSS) and the test battery from which Digit Span scores were derived (versions of WAIS or WMS), in addition to other methodological moderators discussed below.

A major source of hypothesized variability is between-variant differences, given that the Digit Span test can be used in the form of DS-ACSS or RDS. To examine this, separate effect sizes were calculated for each variant after outlier extraction, and the results are presented in Table 1. Overall, both RDS and DS-ACSS showed robust effect sizes, with RDS having a slightly larger effect size ($d = 1.34, 95\% CI = 1.18 - 1.50$) than DS-ACSS ($d = 1.08, 95\% CI = 1.01 - 1.50$). Further analysis of these effect sizes showed within-measure homogeneity tests to be significant for RDS ($Q = 15.89, p < .05$), indicating that the overall effect size for RDS is not able to describe the dataset accurately, and further moderator analyses could help explain the variability seen within the
RDS data. Homogeneity tests were not significant, however, for the DS-ACSS data ($Q = 12.25, p > .05$).

In addition to examining the homogeneity of each variant, the $Q_B$ statistic was used to test for significant differences between the ability of the RDS and DS-ACSS to separate honest and feigning groups. The $Q_B$ statistic is analogous to an $F$-test and uses the Scheffé method to protect against Type I error. In essence, this analog to analysis of variance (ANOVA) partitions the overall $Q$ statistic for the study into that part that can be explained by sampling error within the two groups ($Q_a$; for RDS and DS-ACSS) and the portion that can be explained by the independent variable ($Q_b$; Lipsey & Wilson, 2001). Essentially, this is the within- and between-groups variance used in ANOVA and here applied to the differences between the RDS and DS-ACSS. Using this method, no significant differences were found in the ability of the RDS or DS-ACSS to separate honest and feigning groups ($Q_B = 2.44, p > .05$). In general, it appears that both RDS and DS-ACSS show strength in malingering detection, and either could be used with confidence.

Another potential moderator was the test battery in which Digit Span was administered. Table 2 shows the results from this analysis, and as can be seen, nearly all versions of the WAIS and WMS resulted in strong effect sizes, with the lowest seen in the WMS–R. However, it is important to note that only one study was available for both the WMS–III and WMS–R; thus, these data should be interpreted with caution. When examining tests of homogeneity across each test version, results were found to be homogenous for the WAIS–III ($Q = 6.11, p > .05$) and mixed WAIS/WMS studies ($Q = 4.95, p > .05$). In contrast, significant heterogeneity was found for the WAIS–R ($Q = 18.48, p < .05$), again suggesting that the overall effect size is not an accurate descriptor, and further moderator analyses would be helpful. It is important to note that the other scales may not have shown significant variability in part because of small sample size. Homogeneity testing was not completed for the WMS–III or WMS–R given that only one effect size was available from these studies. Again, the $Q_B$ statistic was used, and no significant differences were found in the ability of the WAIS test versions or WMS test versions in differentiating malingering and honest groups ($Q_B = 1.03, p > .05$). This finding of similar effect sizes across test versions is crucial given the new revisions of both tests that have been recently developed. The results of this study suggest that the Digit Span subtest could be used effectively as a malingering detection tool regardless of the test version administered and likely can be used in newer or future versions of the test. Further, retrospective data analysis that relies on older test versions can be conducted with confidence.

### Methodological moderators

Overall, very little variability was found in the effect sizes across both the type of variant and test version. However, as mentioned previously, additional theoretically derived methodological moderators may be an interesting and important area of consideration. Further, the heterogeneity found in the RDS effect sizes as well as those derived from the WAIS–R warrant further analyses to explain variability that extends beyond sampling error. A major hypothesized source of additional variability may be the methodological design with which the study was conducted, given differences in reliability and validity found between these methods (Rogers, 1997). Further, the exploration of simulation versus known-groups design has been a topic of interest in much literature, and examination of these variables may shed light onto the quality and applicability of each design.

As such, a focused contrast of the effect of study design on the effect sizes was conducted to see whether the real-world setting, and hence increased external validity, of those in known-groups studies ($n = 15$) moderated the effect size outcome relative to a simulation or laboratory design ($n = 7$). Results of these and other methodological moderator analyses appear in Table 3. Again using the $Q_B$ statistic, results showed no significant differences between study design type ($Q_B = 2.17, p > .05$). Interestingly, this suggests that results from both simulation ($d = 1.27, 95\% CI = 1.05–1.23$) and known-groups studies ($d = 1.29, 95\% CI = 1.22–1.36$) provided similar effect sizes, and that both can be considered valuable sources of information on malingering and the Digit Span test.

Turning to other methodological moderators, those hypothesized in advance from enhanced simulation studies included the presence of incentive offered for effort.
preparation time, warning participants to feign believably, and the use of coaching, while those from known-groups studies included having malingering status based on objective criteria. Although each of these moderators contributed to the methodological strength score discussed previously, each was hypothesized as particularly important and warrants individual exploration as a possible contributing moderator. Interestingly, the effect of preparation time and coaching on malingering indices or symptoms could not be examined due to the lack of any enhanced simulation design studies utilizing those methodological characteristics.

The next methodological moderator that was examined was the presence of financial incentive in enhanced simulation designed studies. Because the incentive was not consistently awarded to all participants (e.g., use of a drawing) and due to the very small number of studies utilizing simulation design with financial incentive \((n = 3)\), this variable was coded as either being present or not. Results of this analysis are presented in Table 3 and indicate that the presence of financial incentive resulted in larger effect sizes than when no such incentive was used \((Q_B = 19.58, p < .001)\). Given that the RDS variant and WAIS–R battery showed significant heterogeneity, a specific contrast examining the moderating effect of incentive was undertaken for those groups. For the WAIS–R, the presence of financial incentive resulted in larger effect sizes when using an incentive \((Q_B = 9.00, p < .01)\). For RDS, the analysis could not be completed due to having only one study using the RDS and financial incentive and one study without incentive. Intuitively, one would assume that increased effort via the use of an incentive would result in less blatant malingering and thus smaller effect sizes. However, in this case incentives resulted in more obvious feigning.

Turning to the moderating effect of warning participants to feign believably, results in Table 3 showed that the presence of a warning significantly moderated effect size outcome \((Q_B = 19.71, p < .001)\). The use of a warning significantly reduced the effect size. The results should be interpreted with caution given the very small sample of those not provided a warning \((n = 2)\); however, this may indicate that the use of a warning results in less blatant malingering. Again turning to the RDS and WAIS–R, results for the WAIS–R were similar, with smaller effect sizes and less blatant malingering observed in studies utilizing a warning \((Q_B = 8.07, p < .01)\). Analysis for the RDS could not be completed because all RDS studies used a warning.

Finally, turning to the known-groups design studies, a hypothesized moderator was whether the malingering classification was based on objective criteria (e.g., well-validated dedicated malingering tests). Results in Table 3 show that the presence of objective criteria significantly moderated the effect size outcome \((Q_B = 15.91, p < .001)\). As might be expected, those studies employing objective criteria for establishing the known malingering groups had smaller effect sizes. However, this finding may be an artifact of having very few studies that did not employ objective criteria and thus should be interpreted with caution. When examining the RDS and WAIS–R, the use of objective criteria significantly moderated the effect for both groups \((RDS, Q_B = 4.49, p < .05; WAIS–R, Q_B = 16.82, p < .001)\), with smaller effect sizes seen in studies utilizing this methodological quality.

### Diagnostic accuracy

In addition to examining the ability of the Digit Span subtest in separating groups based on the effect size, another important consideration for clinical practice is the diagnostic accuracy of a test. As such, the sensitivity (proportion of feigners correctly identified by a test) and specificity (the proportion of honest individuals correctly identified as such by a test) were examined across studies. As noted previously, a few studies provided information relevant for effect size coding, but sensitivity and specificity values were not available, and vice versa. Table 4 provides the mean diagnostic accuracy statistics for the 19 independent contrasts from 18 studies (1 study used separate malingering and honest groups for cross-validation and thus allowed for 2 independent contrasts). For the overall sample of contrasts, mean sensitivity values were moderate \((M = 61.4\%, SD = 20.5)\) while mean specificity

<table>
<thead>
<tr>
<th>Manipulation</th>
<th>Condition</th>
<th>k</th>
<th>Mean d-value</th>
<th>(Q_B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study design</td>
<td>KG</td>
<td>14</td>
<td>1.29</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>Sim</td>
<td>8</td>
<td>1.17</td>
<td></td>
</tr>
<tr>
<td>Financial incentive</td>
<td>Yes</td>
<td>3</td>
<td>1.34</td>
<td>19.58***</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>5</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>Warning believable</td>
<td>Yes</td>
<td>6</td>
<td>1.16</td>
<td>19.71***</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2</td>
<td>1.41</td>
<td></td>
</tr>
<tr>
<td>KG obj. criteria</td>
<td>Yes</td>
<td>11</td>
<td>1.22</td>
<td>15.91***</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>3</td>
<td>1.51</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Analyses conducted with outlier-parsed dataset. \(k\) = number of studies. KG = known-groups design, Sim = simulation design. ***\(p\) significant at <.001.
values were moderately high (M = 86.3%, SD = 10.9). This resulted in an overall hit rate of 76.3% (SD = 7.0).

A closer examination of the diagnostic accuracy separately for RDS and DS-ACSS showed a mean sensitivity of 63.3% (SD = 17.6) for RDS (N = 9), while specificity was 86.1% (SD = 11.0). The overall hit rate for RDS was high average (M = 75.7%, SD = 6.6) at a mean cutting score of 7.1 (SD = 0.3). Similarly, the mean sensitivity for the DS-ACSS (N = 10) was adequate (M = 59.7%, SD = 23.7), with high average specificity (M = 86.5%, SD = 11.5) and hit rate (M = 76.8%, SD = 7.6) at a mean cutting score of 6.8 (SD = 1.9). Overall, there was no significant difference between the measures on sensitivity (t = 0.376, p = .712), specificity (t = −0.075, p = .941), or hit rate (t = −0.345, p = .734). These findings again suggest that either variant can be used effectively in classifying malingering. However, it is noteworthy that the cutting score for RDS was much less variable as indicated by a small standard deviation, suggesting more consistency across studies in recommended cutting scores.

**DISCUSSION**

This purpose of the present study was to provide a comprehensive meta-analytic review of the ability of the Digit Span subtest of the Wechsler Adult Intelligence Scale (WAIS) and Wechsler Memory Scale (WMS) in detecting malingering, particularly in forensic settings where simply assuming honesty is not acceptable. More specifically, the ability of the Reliable Digit Span (RDS) variant (Greiffenstein et al., 1994) and the Digit Span age-corrected scaled score variant (DS-ACSS) in separating and accurately classifying malingering and honest groups was examined across all studies in the area through March 2009. The WAIS and WMS are two of the most frequently utilized neuropsychological tests, and the examination of embedded malingering indices within these measures has the potential to allow clinicians to verify honesty in patients without adding unnecessary testing time to a battery. Further, retrospective analyses of effort can be conducted when no malingering tests were given, and embedded indices on standard neuropsychological tests have been hypothesized to be less susceptible to coaching (Mathias et al., 2002).

**Summary of results: Effect size analysis**

Overall, results of the present study indicate that the Digit Span test is sensitive to malingering and on average differentiates malingering and honest groups by more than one pooled standard deviation (d = 1.25, 95% CI = 1.13–1.37). Similarly, both the RDS and DS-ACSS variants produced large effect sizes (RDS: d = 1.34, 95% CI = 1.18–1.50; DS-ACSS: d = 1.08, 95% CI = 1.01–1.50). These effect sizes are similar to the means reported in a recent meta-analysis of several dedicated malingering tests, suggesting that the Digit Span may be a viable alternative to those more time-consuming tests (Vickery et al., 2001). A file drawer analysis suggested that 244 null studies would need to exist to render the effect insignificant, an unlikely prospect. Homogeneity testing revealed consistency among the effect sizes and variability consistent with what is expected given sampling error. Results such as these can be considered positive, with the observed consistency in the effect sizes resulting in increased confidence in the use of the scale, particularly in forensic settings. These results aid in the Digit Span test meeting Daubert standards of admissibility in court (Daubert v. Merrell Dow Pharmaceuticals, 1993).

Further, the effect sizes derived from studies on the Digit Span subtest were found to be similar in magnitude across the variants (RDS vs. DS-ACSS) or test battery (WAIS vs. WMS) utilized. Overall, although the effect sizes for the RDS and WAIS–R showed significant variability, there were no significant differences between the two variants or between any of the test versions in their ability to separate malingering and honest groups. Effect sizes from both variants and all test versions were large. These results are promising for clinicians who may choose to use either the WAIS or the WMS, depending on the referral question, indicating that the Digit Span subtest from either battery could be used interchangeably to adequately detect malingering. Further, should a legal issue arise after an evaluation has been completed, a statement regarding the presence or absence of malingering can be made using the Digit Span results in conjunction with other malingering indices, given the high likelihood that either a WAIS or a WMS was administered during the original testing. For researchers, retrospective data analyses are possible, and results from various test versions can be combined to produce larger sample sizes and hence

---

**TABLE 4**

Operating characteristics for overall sample and by subscale

<table>
<thead>
<tr>
<th>Operating Characteristics</th>
<th>Cut Score M</th>
<th>k</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Hit Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sample</td>
<td>6.9</td>
<td>19</td>
<td>61.4 (20.5)</td>
<td>86.3 (10.9)</td>
<td>76.3 (7.0)</td>
</tr>
<tr>
<td>RDS</td>
<td>7.1</td>
<td>9</td>
<td>63.3 (17.6)</td>
<td>86.1 (11.0)</td>
<td>75.7 (6.6)</td>
</tr>
<tr>
<td>DS-ACSS</td>
<td>6.8</td>
<td>10</td>
<td>59.7 (23.7)</td>
<td>86.5 (11.5)</td>
<td>76.8 (7.6)</td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td>.712</td>
<td>.941</td>
</tr>
</tbody>
</table>

*Note.* Analyses conducted on independent contrasts, outlier-parsed dataset. k = number of studies. RDS = reliable digit span. DS-ACSS = digit span age-corrected scaled score. Standard deviations in parentheses.
more reliable outcomes. Further, with the recent release of the newest versions of the WAIS–IV (Wechsler, 2008) and WMS–IV (Wechsler, 2009), Digit Span research and application to malingering detection is supported and based on these results should generalize to the nearly identical forms of the subtest in these new test versions.

Moreover, the impact of study methodology on the effect sizes was explored, again with promising results. For instance, homogeneity was found among effect sizes derived from both enhanced simulation and known-groups studies, implying that both methodologies are justifiable approaches to determining a test’s ability to classify malingering. This is consistent with Rogers’s (1997) recommendation that both designs are vital to test validation, although it is in contrast with the theoretical notion that known-groups designs are typically the stronger methodology. Moreover, it is in contrast with Rogers and Cavanaugh’s (1983) criticism of simulation studies, specifically that they involve generalizing results from individuals asked to comply with instructions to fake to patients who fake when asked to comply. The findings here are promising for research using enhanced simulation studies conducted with the recommended methodological controls (Rogers, 2008), indicating they are ecologically valid and may provide a reliable estimate of a test’s ability to detect malingering in clinical settings.

Additional moderators were hypothesized and explored, as they can provide some insight into methodological characteristics that may be important despite overall consistency in the derived effect sizes. Further, significant heterogeneity was found for the subset of RDS effect sizes as well as those derived from the WAIS–R, suggesting that moderator variables could aid in explaining such variability. Hypothesized moderators included the use of incentives for effort, the use of a warning to feign believably in enhanced simulation studies, and the use of objective criteria in classifying feigning groups within the known-groups methodology. Each resulted in significant moderation of the overall effect sizes and for RDS and WAIS–R effect sizes where they could be analyzed, indicating that such methodological controls should be routinely applied in malingering studies. The use of coaching on symptoms or feigning indices, as well as the effect of preparation time, could not be analyzed due to lack of relevant studies.

With regard to the use of incentives, the present results are limited in that the use of incentives was coded as a dichotomy, due to inconsistency in how the incentives were awarded and the small number of studies included in these analyses (n = 3). In looking at these studies, the incentive amount ranged from $20 per study participant to a $100 drawing for a single participant. The results of the moderator analyses showed that effect size increased for those offered an incentive, and this was true when eye-balling the amount of incentive. The study providing the largest incentive (Iverson & Franzen, 1996) also showed the largest effect size (d = 2.46). However, intuition suggests that a larger incentive is associated with a more ecologically valid study and less blatant malingering, thus reducing the differences between the feigning and honest groups. Prior research in this area has been mixed, with Orey, Cragar, and Berry (2000) finding no effect of incentive, while Rogers and Cruise (1998) reported a main effect of incentive on an inventory for psychiatric malingering. It may be the case that incentive has a differential effect based on the type of feigning (neurocognitive vs. psychiatric) or based on the type of test administered (e.g., self-report vs. oral administration); however, this has yet to be examined in the literature. Given the very small sample size, the current results should be interpreted with extreme caution, but raise an interesting point for further exploration.

Another methodological characteristic thought to have an impact on the ecological validity of a study is the use of a warning to feign believably. While the practice of providing a warning regarding malingering instruments is not consistent in clinical practice (Slick, Tan, Strauss, & Hultsch, 2004), Rogers (1997) strongly recommends using a warning in simulation research. Again, little research has focused on this specific aspect of simulation methodology, but intuition suggests that inclusion of a warning might result in smaller group differences, or more credible feigning. This was the case in the present study, although again the results should be interpreted with extreme caution given the majority of studies utilized such a warning.

Finally, the use of objective, valid measures for classifying the malingerers in a known-groups study significantly moderated the effect sizes in the present meta-analysis. Specifically, those studies employing well-validated dedicated tests of malingering as well as using the Slick et al. (1999) criteria to aid in the determination of group status showed smaller effect sizes, while those not utilizing such criteria had larger and more variable differences. The use of objective criteria is now considered standard practice in known-groups studies, as it increases the internal validity of the study. These results are not surprising, but are in need of further exploration with a larger number of studies.

Summary of results: Diagnostic accuracy statistics

Overall, the Digit Span test showed strong specificity (M = 86.3%), with moderate sensitivity (M = 61.4%) and overall hit rate (M = 76.3%). More specifically, the RDS and DS-ACSS variants also showed strong specificity (M = 86.1%; M = 86.5%, respectively) with good sensitivity (M = 63.3%; M = 59.7%, respectively). There were no significant differences in the diagnostic accuracy of the RDS and DS-ACSS, indicating that clinicians may choose the one best suited to their needs. Although RDS showed slightly better consistency across recommended cutting scores, the DS-ACSS is calculated automatically when scoring the WAIS and is conceptually easier to understand, making it the likely choice by clinicians. Further, DS-ACSS utilizes a standardized score and is not calculated as a simple sum of digits forward and backward as RDS is, so may be more difficult for the sophisticated malingerer to gauge the amount of feigning required to produce a given standard score.
Limitations

The primary shortcoming of the present review is the small number of studies available for analysis. More specifically, the limited number of studies across versions of the WAIS and WMS batteries precluded in-depth exploration among these tests. Thus, the results of those analyses should be considered tentative and an area in need of further research, particularly in light of the new test versions being released. Further, the effects of coaching and preparation time are potentially important moderators that could not be explored here.

Additionally, although the Digit Span subtest may hold promise as a screening instrument, the present study was not able to explore this issue thoroughly based on a limited range of cutting scores presented to maximize sensitivity, a requirement for using either RDS or DS-ACSS as a screening tool for malingering. Future research should focus on publishing the full range of cutting scores and associated sensitivity and specificity values, to allow clinicians and researchers to further examine this potentially time-saving diagnostic approach.

In summary, the present review provided evidence to support the use of the Digit Span subtest, including RDS and DS-ACSS, in adequately detecting malingering. The Digit Span subtest performs similarly to many well-validated dedicated malingering tests. The advantages of embedded indices, in addition to the large effect sizes and strong diagnostic accuracy statistics, make the Digit Span test an essential tool for clinicians and researchers alike.

Original manuscript received 11 May 2010
Revised manuscript accepted 28 July 2010

REFERENCES

Note: References marked with an asterisk designate studies included in the meta-analysis.


## APPENDIX

Demographic and methodological characteristics, effect sizes, and classification accuracy of studies included in independent contrasts analyses

<table>
<thead>
<tr>
<th>Study</th>
<th>Comp type</th>
<th>Grp</th>
<th>N</th>
<th>Subj.</th>
<th>Age (years)</th>
<th>Ed. (years)</th>
<th>Method</th>
<th>Coa</th>
<th>Prep</th>
<th>Inc</th>
<th>Bel.</th>
<th>MS</th>
<th>Subtest</th>
<th>Ed</th>
<th>CS</th>
<th>SN</th>
<th>SP</th>
<th>HR</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axelrod (2006)</td>
<td>KG</td>
<td>M</td>
<td>36</td>
<td>TBI</td>
<td>43.1</td>
<td>11.9</td>
<td>1.6</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.86</td>
<td>DS</td>
<td>6.3</td>
<td>2.1</td>
<td>≤ 7</td>
<td>.75</td>
<td>.72</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>29</td>
<td>TBI</td>
<td></td>
<td>38.2</td>
<td>17.8</td>
<td>2.1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.86</td>
<td>RDS</td>
<td>6.7</td>
<td>2.4</td>
<td>≤ 7</td>
<td>.62</td>
<td>.69</td>
<td>0.69</td>
</tr>
<tr>
<td>Babikian (2006)</td>
<td>KG</td>
<td>M</td>
<td>66</td>
<td>MixedN</td>
<td>42.5</td>
<td>12.3</td>
<td>2.4</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.86</td>
<td>RDS</td>
<td>6.7</td>
<td>2.4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>2.18</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>56</td>
<td>Mix</td>
<td></td>
<td>48.4</td>
<td>16.4</td>
<td>2.4</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.43</td>
<td>RDS</td>
<td>4.2</td>
<td>2.4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1.0</td>
</tr>
<tr>
<td>Chafetz (2008)</td>
<td>KG</td>
<td>M</td>
<td>16</td>
<td>Dis.</td>
<td>24.1</td>
<td>8.5</td>
<td>1.5</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.43</td>
<td>RDS</td>
<td>4.2</td>
<td>2.4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>2.18</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>21</td>
<td>Dis</td>
<td></td>
<td>30.2</td>
<td>8.9</td>
<td>1.6</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.43</td>
<td>RDS</td>
<td>8.0</td>
<td>2.0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1.0</td>
</tr>
<tr>
<td>Duncan (2002)</td>
<td>KG</td>
<td>M</td>
<td>54</td>
<td>Crim.</td>
<td>34.7</td>
<td>10.1</td>
<td>2.2</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.57</td>
<td>RDS</td>
<td>5.8</td>
<td>3.4</td>
<td>≤ 7</td>
<td>.68</td>
<td>.71</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>124</td>
<td>Crim</td>
<td></td>
<td>38.1</td>
<td>9.5</td>
<td>10.5</td>
<td>2.7</td>
<td>8.9</td>
<td>2.1</td>
<td>.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Etherton (2000)</td>
<td>KG</td>
<td>M</td>
<td>32</td>
<td>Pain</td>
<td>43.0</td>
<td>8.5</td>
<td>3.0</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.86</td>
<td>DS</td>
<td>6.4</td>
<td>3.0</td>
<td>≤ 7</td>
<td>.75</td>
<td>.75</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>36</td>
<td>TBI</td>
<td></td>
<td>32.2</td>
<td>14.1</td>
<td>NA</td>
<td>NA</td>
<td>9.8</td>
<td>3.1</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Etherton (2006)</td>
<td>KG</td>
<td>M</td>
<td>35</td>
<td>Pain</td>
<td>43.3</td>
<td>11.6</td>
<td>NA</td>
<td>NA</td>
<td>10.5</td>
<td>2.3</td>
<td>.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>35</td>
<td>Pain</td>
<td></td>
<td>43.0</td>
<td>8.4</td>
<td>NA</td>
<td>NA</td>
<td>10.5</td>
<td>2.3</td>
<td>.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greiffenstein (1994)</td>
<td>KG</td>
<td>M</td>
<td>43</td>
<td>TBI</td>
<td>39.3</td>
<td>10.0</td>
<td>2.2</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.57</td>
<td>RDS</td>
<td>6.7</td>
<td>1.2</td>
<td>≤ 7</td>
<td>.68</td>
<td>.77</td>
<td>1.90</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>30</td>
<td>PPCS</td>
<td></td>
<td>37.6</td>
<td>13.9</td>
<td>11.7</td>
<td>0.8</td>
<td>8.9</td>
<td>1.1</td>
<td>.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greiffenstein (1995)</td>
<td>KG</td>
<td>M</td>
<td>68</td>
<td>TBI</td>
<td>38.3</td>
<td>11.7</td>
<td>2.9</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.57</td>
<td>RDS</td>
<td>6.6</td>
<td>1.8</td>
<td>≤ 7</td>
<td>.89</td>
<td>.80</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>53</td>
<td>TBI</td>
<td></td>
<td>24.6</td>
<td>12.5</td>
<td>2.6</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.57</td>
<td>RDS</td>
<td>9.2</td>
<td>1.7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.68</td>
</tr>
<tr>
<td>Heaton (1978)</td>
<td>Sim</td>
<td>M</td>
<td>16</td>
<td>Stud/</td>
<td>CV</td>
<td>24.4</td>
<td>7.5</td>
<td>2.4</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>.50</td>
<td>DS</td>
<td>7.0</td>
<td>2.9</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>16</td>
<td>TBI</td>
<td></td>
<td>26.7</td>
<td>6.5</td>
<td>1.6</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>.50</td>
<td>RDS</td>
<td>6.1</td>
<td>2.8</td>
<td>≤ 5</td>
<td>.36</td>
<td>.75</td>
<td>1.24</td>
</tr>
<tr>
<td>Heinly (2005)</td>
<td>KG</td>
<td>M</td>
<td>18</td>
<td>TBI</td>
<td>18.7</td>
<td>0.9</td>
<td>1.0</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.86</td>
<td>RDS</td>
<td>9.2</td>
<td>2.3</td>
<td>&lt; 8</td>
<td>.27</td>
<td>.65</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>45</td>
<td>TBI</td>
<td></td>
<td>18.7</td>
<td>1.7</td>
<td>1.3</td>
<td>0.6</td>
<td>11.6</td>
<td>2.4</td>
<td>.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inman (2002)</td>
<td>Sim</td>
<td>M</td>
<td>21</td>
<td>TBI</td>
<td>18.7</td>
<td>0.9</td>
<td>1.0</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>.75</td>
<td>RDS</td>
<td>9.2</td>
<td>2.3</td>
<td>&lt; 8</td>
<td>.27</td>
<td>.65</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>24</td>
<td>TBI</td>
<td></td>
<td>18.7</td>
<td>1.7</td>
<td>1.3</td>
<td>0.6</td>
<td>11.6</td>
<td>2.4</td>
<td>.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iverson (1994)</td>
<td>Sim</td>
<td>M</td>
<td>40</td>
<td>Stud/Inm</td>
<td>28</td>
<td>6.9</td>
<td>14.2</td>
<td>2.0</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>.50</td>
<td>DS</td>
<td>2.7</td>
<td>1.7</td>
<td>&lt; 5</td>
<td>.90</td>
<td>.90</td>
<td>2.97</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>20</td>
<td>TBI</td>
<td></td>
<td>28</td>
<td>7.5</td>
<td>13.9</td>
<td>1.9</td>
<td>8.3</td>
<td>2.3</td>
<td>.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study (Year)</td>
<td>Comp Type</td>
<td>Age</td>
<td>Ed.</td>
<td>Test Scale</td>
<td>MS</td>
<td>Note</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----------</td>
<td>-----</td>
<td>-----</td>
<td>------------</td>
<td>----</td>
<td>------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iverson (1996)</td>
<td>Sim</td>
<td>M</td>
<td>20</td>
<td>Psych.</td>
<td>20.0</td>
<td>1.7</td>
<td>13.7</td>
<td>0.7</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>.50</td>
<td>DS</td>
<td>2.6</td>
<td>1.4</td>
<td>&lt; 4</td>
<td>0.78</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>20</td>
<td>MemImp</td>
<td>34.0</td>
<td>13.2</td>
<td>13.6</td>
<td>2.3</td>
<td>8.2</td>
<td>2.8</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Langeluddecke (2003)</td>
<td>KG</td>
<td>M</td>
<td>25</td>
<td>TBI</td>
<td>33.9</td>
<td>12.0</td>
<td>10.8</td>
<td>2.1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>1.0</td>
<td>DS</td>
<td>11.6</td>
<td>3.6</td>
<td>≤ 8</td>
<td>0.16</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>50</td>
<td>TBI</td>
<td>38.4</td>
<td>11.3</td>
<td>11.1</td>
<td>3.3</td>
<td>16.3</td>
<td>3.9</td>
<td>.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Larrabee (2003)</td>
<td>KG</td>
<td>M</td>
<td>24</td>
<td>TBI</td>
<td>39.3</td>
<td>11.8</td>
<td>12.5</td>
<td>2.3</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.71</td>
<td>RDS</td>
<td>7.4</td>
<td>1.7</td>
<td>≤ 7</td>
<td>.50</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>12</td>
<td>TBI</td>
<td>34.8</td>
<td>16.8</td>
<td>12.6</td>
<td>2.6</td>
<td>10.0</td>
<td>2.2</td>
<td>.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Martin (1996)</td>
<td>Sim</td>
<td>M</td>
<td>51</td>
<td>Stud./TBI</td>
<td>22.1</td>
<td>5.5</td>
<td>14.0</td>
<td>1.4</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>.44</td>
<td>DS</td>
<td>11.3</td>
<td>6.3</td>
<td>≤ 11</td>
<td>.35</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>70</td>
<td>TBI</td>
<td>21.6</td>
<td>5.5</td>
<td>14.6</td>
<td>1.3</td>
<td>16.4</td>
<td>3.4</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathias (2002)</td>
<td>KG</td>
<td>M</td>
<td>24</td>
<td>TBI</td>
<td>38.9</td>
<td>12.3</td>
<td>12.5</td>
<td>2.8</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>1.0</td>
<td>RDS</td>
<td>6.8</td>
<td>1.5</td>
<td>≤ 7</td>
<td>.67</td>
<td>.82</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>30</td>
<td>TBI</td>
<td>33.6</td>
<td>13.5</td>
<td>12.5</td>
<td>2.4</td>
<td>10.1</td>
<td>2.0</td>
<td>.93</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mittenberg (1993)</td>
<td>Sim</td>
<td>M</td>
<td>39</td>
<td>CV</td>
<td>32.8</td>
<td>12.4</td>
<td>15.5</td>
<td>2.3</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>.56</td>
<td>DS</td>
<td>11.2</td>
<td>3.9</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>KG</td>
<td>M</td>
<td>36</td>
<td>TBI</td>
<td>42.9</td>
<td>16.0</td>
<td>11.0</td>
<td>3.4</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.43</td>
<td>DS</td>
<td>6.1</td>
<td>2.2</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>36</td>
<td>TBI</td>
<td>30.3</td>
<td>12.9</td>
<td>11.9</td>
<td>1.5</td>
<td>9.5</td>
<td>3.3</td>
<td>1.53</td>
<td>4.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strauss (2002)</td>
<td>Sim</td>
<td>M</td>
<td>11</td>
<td>TBI</td>
<td>37.6</td>
<td>8.7</td>
<td>12.7</td>
<td>2.2</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>1.0</td>
<td>DS</td>
<td>6.3</td>
<td>2.4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>16</td>
<td>TBI</td>
<td>35.4</td>
<td>9.5</td>
<td>13.1</td>
<td>2.9</td>
<td>9.0</td>
<td>3.0</td>
<td>1.03</td>
<td>2.9</td>
<td>.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suhr (1997)</td>
<td>KG</td>
<td>M</td>
<td>31</td>
<td>TBI</td>
<td>37.6</td>
<td>8.7</td>
<td>12.7</td>
<td>2.2</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>1.0</td>
<td>DS</td>
<td>6.3</td>
<td>2.4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>20</td>
<td>TBI</td>
<td>35.4</td>
<td>9.5</td>
<td>13.1</td>
<td>2.9</td>
<td>9.0</td>
<td>3.0</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trueblood (1994)</td>
<td>KG</td>
<td>M</td>
<td>12</td>
<td>TBI</td>
<td>34.9</td>
<td>11.6</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.86</td>
<td>DS</td>
<td>4.8</td>
<td>NA</td>
<td>&lt; 7</td>
<td>.75</td>
<td>.75</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>12</td>
<td>TBI</td>
<td>35.0</td>
<td>11.3</td>
<td>NA</td>
<td>NA</td>
<td>8.3</td>
<td>NA</td>
<td>NA</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trueblood (1993)</td>
<td>KG</td>
<td>M</td>
<td>8</td>
<td>TBI</td>
<td>37.4</td>
<td>11.6</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.86</td>
<td>DS</td>
<td>5.9</td>
<td>NA</td>
<td>≤ 7</td>
<td>.62</td>
<td>.75</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>8</td>
<td>TBI</td>
<td>37.4</td>
<td>12.0</td>
<td>NA</td>
<td>NA</td>
<td>8.9</td>
<td>NA</td>
<td>NA</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. For studies, only the first author is listed. n/a = not applicable; NA = information not available. Comp type = comparison type (KG = known groups; Sim = enhanced simulation). Grp = group (M = malingering; H = honest). N = number of subjects. Subj. = subject sample (Crim. = criminal pretrial defendants, CV = community volunteers, Dis. = disability claimants, Inm = Inmates, MemImp = memory impaired, MixedN = mixed neurological, Pain = chronic pain, PPCS = persistent postconcussive syndrome, Psych. = psychiatric patients, Stud. = students, TBI = traumatic brain injury). Age (M = mean, SD = standard deviation). Ed. = education level. Method (Coa = coaching provided on either symptoms or avoiding detection, Prep = preparation time given, Inc = incentive given, Bel. = warning to be believable; N = not present, Y = included in methodology). MS = relative method score. Test scale (RDS = reliable digit span, DS = digit span age-corrected scaled score; M = mean test score, SD = standard deviation, CS = cutting score, SN = sensitivity, SP = specificity, HR = hit rate, d = corrected Hedges's g effect size).

aStudy using a version of the Wechsler Adult Intelligence Scale (WAIS).
bStudy using a version of the Wechsler Memory Scale (WMS).
cStudy excluded as an outlier.