Cognitive Discrepancy Models for Specific Learning Disabilities Identification: Simulations of Psychometric Limitations

W. Pat Taylor, **Jeremy Miciak**, **Jack M. Fletcher**, and **David J. Francis**
University of Houston

Abstract

Few studies have investigated specific learning disabilities (SLD) identification methods based on the identification of patterns of processing strengths and weaknesses (PSW). We investigated the reliability of SLD identification decisions emanating from different achievement test batteries for one method to operationalize the PSW approach: the concordance/discordance model (C/DM; Hale & Fiorello, 2004). Two studies examined the level of agreement for SLD identification decisions between two different simulated, highly correlated achievement test batteries. Study 1 simulated achievement and cognitive data across a wide range of potential latent correlations between an achievement deficit, a cognitive strength and a cognitive weakness. Latent correlations permitted simulation of case-level data at specified reliabilities for cognitive abilities and two achievement observations. C/DM criteria were applied and resulting SLD classifications from the two achievement test batteries were compared for agreement. Overall agreement and negative agreement were high, but positive agreement was low (0.33 – 0.59) across all conditions. Study 2 isolated the effects of reduced test reliability on agreement for SLD identification decisions resulting from different test batteries. Reductions in reliability of the two achievement tests resulted in average decreases in positive agreement of 0.13. Conversely, reductions in reliability of cognitive measures resulted in small average increases in positive agreement (0.0 – 0.06). Findings from both studies are consistent with prior research demonstrating the inherent instability of classifications based on C/DM criteria. Within complex ipsative SLD identification models like the C/DM, small variations in test selection can have deleterious effects on classification reliability.

Keywords

decision reliability; specific learning disabilities; Monte Carlo; response to intervention; cognitive assessment; measurement error

Fletcher (2012) identified three classification frameworks for defining and identifying specific learning disabilities (SLD): (a) neurological, (b) cognitive discrepancy, and (c) instructional. Each of these frameworks represents a hypothetical framework that reflects a conception of the nature of SLD and an operationalized definition that leads directly to methods identifying individual students with SLD. In recent years, neurological frameworks

Address all correspondence to: Pat Taylor, Ph.D., Texas Institute for Measurement, Evaluation, and Statistics, University of Houston, Pat.Taylor@times.uh.edu.

have generally fallen from favor (Fletcher et al., 2007), leaving two competing frameworks for consideration: cognitive discrepancy and instructional frameworks.

There is active debate about the most reliable and valid framework for defining and identifying SLD (see for example Hale et al., 2010; Consortium for Evidence Based Early Intervention Practices [CEBEIP], 2010). However, questions about the most reliable and valid frameworks for defining and identifying SLD are ultimately empirical (Morris & Fletcher, 1998). Proposed frameworks can be operationalized and evaluated to determine their validity through comparisons of resulting subgroups that represent the hypothetical classification against external variables not used to define the classification as well as to determine the reliability of identification decisions. In the present study, we investigated the reliability of one proposed operationalization of the cognitive discrepancy framework: the concordance/discordance model (C/DM; Hale & Fiorello, 2004). The C/DM is one example of a subset of methods that propose that SLD identification should be based on an intraindividual pattern of cognitive processing strengths and weaknesses (PSW methods). The C/DM is often described as an evidence-based method for the identification of SLD (Hale et al., 2010; Hale & Fiorello, 2004). It has been presented at national conferences and is frequently presented in practitioner guidance documents as a specific operationalization of more general PSW methods and cognitive discrepancy frameworks for the identification of SLD (Hanson, Sharman, & Esparza-Brown, 2008).

Cognitive Discrepancy Frameworks

Cognitive discrepancy frameworks for the definition and identification of SLD hypothesize that SLD is marked by intraindividual discrepancies in cognitive and academic performance. Such discrepancies operationalize the cardinal characteristic of SLD: unexpected academic underachievement. In 1977, the Office of Special Education Programs (OSEP) codified the cognitive discrepancy framework as the only method for the identification of SLD, mandating the identification of a discrepancy between ability (typically measured through IQ tests or language comprehension tests) and academic achievement. However, in subsequent decades serious questions emerged about the reliability and validity of IQ-achievement discrepancy methods for the identification of SLD (for an historical review see Hallahan & Mercer, 2002). Identification decisions utilizing IQ-achievement discrepancies are unreliable at the individual level because the group membership of students scoring near the discrepancy cut point fluctuates in response to differences in testing occasion, test forms, or formulae used for calculating the IQ-achievement discrepancy (Francis et al., 2005; Macmann, Barnett, Lombard, Belton-Kocher, & Sharpe, 1989). Further, IQ-achievement discrepancy methods lack validity because groups of low achieving students with and without IQ achievement discrepancies do not differ on external cognitive and achievement measures, and assessments of brain function used to validate the classification (Fletcher et al., 1994; Stuebing et al., 2002; Tanaka et al., 2011; Ysseldyke, Algozzine, Shinn, & McGue, 1982). Perhaps most importantly, treatment response is not predicted by IQ or IQ-achievement discrepancy (Hoskyn & Swanson, 2000; Stuebing, Barth, Molfese, Weiss, & Fletcher, 2009; Vellutino, Scanlon, & Lyon, 2000).

In recent years, a revised cognitive discrepancy model for the identification of SLD has received considerable attention in the school psychology literature (Hale et al., 2010). These methods are often referred to as PSW methods because they hypothesize that an intraindividual pattern of discrepancies in cognitive processing abilities and achievement are a definitional attribute of SLD. Proponents argue that these discrepancy patterns explain low achievement and can inform future treatment (Flanagan, Fiorello, & Ortiz, 2010; Johnson, 2014). Three methods have been proposed to operationalize the identification of an intraindividual PSW pattern and subsequent SLD identification: the C/DM (Hale & Fiorello, 2004), the Cross Battery Assessment approach (XBA approach; Flanagan, Ortiz, & Alfonso, 2007), and the Discrepancy/Consistency Method (Naglieri, 1999). Although occasionally presented as equivalent methods, these proposed methods to operationalize PSW methods differ in important ways, including: (a) the specific hypothesized relations between cognitive processing and achievement deficits, (b) the role of norms and benchmarks for decision making, and (c) the specific criteria for cognitive discrepancies.

To illustrate, the C/DM is unique in its utilization of ipsative comparisons, differentiating it from other PSW models that rely on a series of normative comparisons (e.g. Flanagan et al., 2007; Naglieri, 1999). The C/DM operationalizes a PSW profile as a series of significant and non-significant differences in individual performance across cognitive and academic dimensions. The C/DM is flexible with regards to theoretical orientation and test selection, which proponents cite as a strength of the model (Hale & Fiorello, 2004). To evaluate the existence of a PSW profile, three psychometric criteria must be met: (a) a non-significant difference between a theoretically linked academic deficit and a cognitive processing deficit (concordance); (b) a significant difference between an unrelated cognitive processing strength and the cognitive processing deficit (discordance); and (c) a significant difference between an unrelated cognitive processing strength and an academic deficit. Significant and non-significant differences are evaluated by comparing difference scores against a critical value based on the standard error of the difference (SED).

Research on the Reliability of the PSW Methods for SLD Identification

Few empirical studies have investigated the reliability PSW methods for SLD identification. Proponents of the C/DM and other PSW methods cite the relation of discrete cognitive processes with specific academic skills (Evans, Floyd, McGrew, & Leforgee, 2001; Hale, Fiorello, Kavanagh, Hoepner, & Gaither, 2001; Johnson, Humphrey, Mellard, Woods, & Swanson, 2010) and recent theoretical advances in conceptualizing, organizing, and understanding cognitive abilities (e.g. Cattell-Horn-Carroll Theory [CHC]; Fiorello & Primerano, 2005; McGrew, & Wendling, 2010). However, neither of these lines of research provides support for the reliability and validity of the C/DM or other PSW methods as processes for the identification of SLD. It is well known that cognitive processes are correlated with academic skills and that students with SLD have cognitive weaknesses that reflect the deficient academic domain. For example, the importance of phonological processing to early reading is well established (Torgeson, Wagner, Rashotte, Burgess, & Hecht, 1997; Speece, Ritchey, Cooper, Roth, & Schatschneider, 2004). Similarly, vocabulary and background knowledge have demonstrated strong relationships to reading comprehension, particularly in late elementary (Catts, Adlof, & Weismer, 2006; Evans et al.,

2002). However, these relations do not necessarily mean that measurement of cognitive processes facilitates identification or treatment of SLD. Proposed identification criteria for SLD (or any disorder) represent hypotheses about the nature of the disorder and can be tested to determine the reliability and validity of the classification (Morris & Fletcher, 1998). A proposed classification accrues validity when evidence mounts that resulting groups are consistent and differ on important external dimensions (e.g. intervention response, cognitive functioning, or neuroimaging results). However, the validity of classifications in the aggregate as evidenced by group differences does not mean that the approach reliably and validly classifies individual children with SLD. The validity argument for any classification system must also address the utility of the individual decisions produced by the system (Messick, 1995).

Recent studies have raised questions about the reliability of the C/DM and other PSW methods for SLD identification. Stuebing et al. (2012) utilized simulated data to evaluate the technical adequacy of three proposed PSW methods: (a) the Cross Battery Assessment (XBA) method (Flanagan et al., 2007), (b) the Discrepancy Consistency Method (DCM; Naglieri, 1999) and (c) the C/DM (Hale & Fiorello, 2004). Latent variables were created based on relations between reading and cognitive constructs and observed scores were generated based on the psychometric properties of common academic and cognitive tests. The utilization of both latent and observed variables permitted an investigation of SLD identification rates, positive predictive value (PPV; the probability of being truly SLD if a SLD result is found), and negative predictive value (NPV; the probability of being truly not SLD if a not SLD result is found). Across methods, the base rates for SLD identification were low, indicating that a large number of students would need to be assessed in order to identify a group of students with SLD. Low base rates ensured that NPV was high (median NPV = .99, range: .94 - .99). However, PPV was low across methods (median PPV = .22; range: .01 - .56) suggesting that even if "true" PSW profiles exist, proposed methods for identification would not reliably identify students demonstrating these profiles because of the emphasis on "not SLD" decisions and the high false positive rate associated with positive decisions.

These findings were substantially replicated in a study investigating classification agreement of decisions emerging from the XBA method (Kranzler, Floyd, Benson, Zaboski, & Thibodaux, in press). The authors utilized standardization data for the Woodcock-Johnson III Normative Update Tests of Cognitive Abilities and Tests of Achievement (Woodcock, McGrew, Schrank, & Mather, 2007) to conduct analyses for each of the of the CHC broad cognitive abilities identified by McGrew and Wendling (2010) as related to basic academic skills. Student data was input in the XBA PSW Analyzer (Flanagan et al., 2010) permitting an empirical classification of PSW status. This classification was then compared to scores on the broad cognitive abilities identified by McGrew and Wendling and 2×2 contingency tables were crated. Classifications were deemed in agreement when the identified PSW profile corresponded with the theoretical relations identified by McGrew and Wendling. Disagreements represented PSW profiles not aligned with these theoretical relations. Similar to Stuebing et al., (2012), identification rates were low and the XBA method was good at identifying "not SLD" students (mean specificity and NPV were 92% and 89%). However,

PPV was generally low, with mean sensitivity of 21% and mean PPV of 34%, suggesting that many students with true SLD status would not be identified by the XBA method.

Miciak, Fletcher, Stuebing, Vaughn, and Tolar (2014) utilized a sample of inadequate responders in middle school to investigate the consistency of identification decisions emerging from XBA and C/DM methods. The study utilized cognitive and academic data to examine the base rates of SLD identification and the interchangeability of the two methods. Both methods identified a low percentage of students, dependent upon the cut point for academic low achievement (range: 17.3% - 47.5%). Comparisons of the identification decisions of the XBA method and C/DM were poor (kappa range: -.04 – .31), especially when comparing the models with higher cut points for academic deficits and higher identification rates. For three of the four operationalizations of these methods, agreement between the two models for SLD identification decisions was not statistically different from chance. The authors concluded that these results raise doubts about the efficiency and reliability of the two PSW models.

Another potential source of unreliability for SLD identification decisions relates to differences in the tests utilized to establish a PSW profile. The C/DM does not specify what tests should be utilized in the SLD identification process. As a result, practitioners who utilize different tests may arrive at different identification decisions, even when measuring the same latent constructs (Macmann et al., 1989). This possibility was directly investigated by Miciak, Taylor, Fletcher, and Denton (2015) with a sample of second grade students who demonstrated inadequate response to a reading intervention. C/DM criteria were applied with two psychoeducational batteries consisting of the same cognitive measures but different, highly-correlated, and reliable achievement measures of the same latent constructs (i.e. decoding, reading fluency, and reading comprehension). Agreement between the two batteries for SLD identification decisions was low (kappa = .29). The low level of agreement observed between the two batteries was particularly noteworthy because it resulted from a difference in achievement tests only; the testing occasion, decision rules, and cognitive measures were constant.

The Present Studies

Miciak, Taylor, et al. (2015) raised questions about the reliability of SLD identification decisions within the C/DM at the observed level. The utilization of different achievement tests resulted in poor agreement across different psychoeducational batteries. However, the results of that study were specific to the tests utilized and the sample assessed. In contrast, simulated data allow for generalization beyond a specific sample (Hallgren, 2013). The current simulations allow for the analysis of multiple large datasets that are based on observed relations among variables. These data can be manipulated and analyzed to illustrate underlying psychometric principles. Such analyses are particularly useful when there are few empirical data to evaluate and/or the cost of data collection is high, as is the case for the C/DM.

To the extent possible, methods for the identification of SLD should be reliable and valid at the level of individual classification decisions to ensure that resulting educational and

programming decisions are based on true individual differences, rather than psychometric error. In the present studies, we utilized simulated data to investigate the effects of test selection on identification decisions within the C/DM. The goal was to evaluate the potential for agreement between different tests when using the C/DM for SLD identification decisions. If the C/DM is not robust to differences in test selection, it would suggest that the method, as currently prescribed, is inappropriate for SLD identification, which confers special legal protections to students and serves as the basis for educational programming decisions. Two studies of simulated data are utilized to answer our primary research questions:

1.   How generalizable are SLD identification decisions across different selections of assessments?

2.   What are the upper limits of agreement for distinct test batteries applying C/DM criteria?

Two hypotheses guided the study:

Hypothesis 1. Based on prior work (Miciak, Taylor, et al., 2015; Stuebing et al., 2012), we hypothesized that SLD identification decisions based on different test batteries would demonstrate poor agreement, especially poor positive agreement because of low identification rates.

Hypothesis 2. Based on previous work examining fluctuation around cut points and poor agreement across test batteries (Francis et al., 2005; Macmann et al., 1989), we hypothesized that the identification decisions of different test batteries would be adversely affected by reductions in test reliability.

Study 1

Methods

All simulations and analyses were conducted using SAS 9.4 (SAS Institute, 2013). The study simulated data for two observed indicators of a latent achievement dimension and single observed indicators for a latent cognitive strength and a latent cognitive weakness in a three step process. The simulation was designed to include a range of tests that were moderately related to each other to tests that were strongly related to each other. In addition, the tests should vary in reliability from marginally acceptable to very good. In order to adequately cover the parameter space for this study, the simulation stipulated that latent correlations between achievement, cognitive strength, and cognitive weakness would range from a minimum of 0.5 to a maximum of 0.95 while test reliabilities ranged from 0.7 to 0.9. These parameters are consistent with values achieved by and between most cognitive and academic tests in educational assessments, in which tests generally demonstrate good reliability and relationships among tests are generally positive and moderate to large. Correlations were incremented by 0.05 and test reliabilities were incremented by 0.1.

Step one involved creating the set of correlation matrices that included all possible combinations of correlations between achievement, cognitive strength, and cognitive weakness within the specified minimum and maximum. With three latent variables there are

three correlations each of which could take on any one of the ten potential correlation values. This resulted in $10^3 = 1000$ potential latent correlation matrices. The first eigenvalue of each correlation matrix was evaluated. If the first eigenvalue was < 0 the matrix was deemed not positive semi-definite and was discarded. This process yielded 898 viable correlation matrices. Similarly, the levels of reliabilities for two observed measures of ability and the two observed cognitive measures were fully crossed. Full crossing the three levels of reliability with the four observed measures resulted in $3^4 = 81$ combinations of reliabilities for the observed measures. The 81 combinations of reliabilities were fully crossed with the 898 viable correlation matrices resulting in 72,738 unique combinations of latent correlations and observed reliabilities that served as the basis for the study.

Step two involved generating normally distributed latent variable scores based on the latent correlation matrices from step one. For each of the unique combinations of latent correlations and observed reliabilities in step one, sets of 10,000 z-scores were simulated for each latent variable such that the simulated data reproduced the latent correlations. Generating data to recreate the desired correlation matrices was achieved using proc simnormal in SAS. Correlations of the simulated variables were calculated for comparison with study parameters to evaluate the adequacy of the simulation.

Step three involved the generation of observed variable scores. At this point an additional 10,000 z-scores were created to represent random error for each observed variable using the random normal generator in SAS. Observed scores where then created using the following formula:

$$\text{Observed Score} = \sqrt{\text{reliability}} * \text{latent score} + \sqrt{1 - \text{reliability}} * \text{random error}$$

Reliability can be expressed as a ratio of true variability to observed variability. The above equation creates a set of observed scores with known reliability by creating a sum of true scores (latent score) and random errors with proportions controlled by the desired reliability. Observed scores were generated for each of 10,000 members of the 72,738 unique combinations of latent correlations and observed reliabilities. These observed scores were then rescaled to have a mean of 100 and SD of 15 and rounded to whole numbers. This process resulted in observed scores for two achievement (A) measures (Achievement Test 1 and Achievement Test 2), a cognitive weakness measure (Cognitive Weakness), and a cognitive strength measure (Cognitive Strength).

The observed scores and simulation parameters were used to calculate differences between each of the achievement and cognitive measures. Thresholds for significant differences were calculated utilizing the standard error of the difference at $p < .05$, following procedures specified by Hale and Fiorello (p. 102, 2004). The observed patterns of significant and non-significant differences were used to determine SLD status for each case using cut points of 85 and 90 for both measures of achievement. These cut points were based on recommendations for practice for PSW methods (e.g. Flanagan et al., 2007) and match cut points utilized in previous studies of PSW methods (Miciak, Fletcher, et al., 2014; Miciak, Taylor, et al., 2015). Inclusion was limited to cases with at least one cognitive measure   70

to remove cases with scores that may be consistent with identification criteria for intellectual disabilities. While the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013) criteria for intellectual disorders allow for the utilization of confidence intervals and clinical judgment in the identification of intellectual disabilities, the requirements of the simulation required a clear decision rule. We therefore limited inclusion to cases with one score > 70, reflecting the mid-point of this confidence interval. Additionally, inclusion was limited to case that had at least one achievement score < 92 to ensure that all cases in the inclusion sample potentially demonstrated an academic deficit. These selection samples served as the data for further analysis. If SLD identification rates were higher, then overall agreement would be an appropriate index of agreement between the two tests. Given that base rates for SLD identification are relatively low, positive agreement will be the primary result of interest. This index will not be inflated due to the very large number of negative agreements found when low base rates are present.

Results

The adequacy of the simulation was evaluated by comparing the specified latent correlations with the simulated latent correlations. Absolute values of the differences ranged from <0.0001 to 0.0298 with a mean of 0.0037 and a standard deviation of 0.0024.

The mean number of observations meeting selection criteria for evaluation of SLD status from each population of 10,000 was 3,670.6 with a standard deviation of 141.5. Table 1 shows the average cell size for SLD status designation by each test for achievement < 85 and achievement < 90. Across all conditions only 4% of the selection sample was designated as SLD by at least one test when achievement was less than 85. When achievement was less than 90 this proportion increased to 6%.

Overall agreement (PO), positive agreement (PA), and negative agreement (NA) were calculated as measures of agreement for SLD status between the two simulated achievement tests within each selection sample. Average PO for the study was 0.97 (SD = 0.02) when achievement was less than 85 and 0.96 (SD = 0.02) when achievement was less than 90, which appears high. Similarly, NA for the study was 0.98 (SD = 0.01) for both achievement conditions, which also appears high. In contrast to overall agreement and negative agreement, the average PA was only 0.42 (SD = 0.13) when the achievement criterion was less than 85. When the achievement criterion was less than 90, the average PA increased to only 0.45 (SD = 0.12). Thus, relatively few cases identified as SLD were consistently identified across two measures of the same achievement dimension even though reliability for measures ranged from .7 to .9 and agreement was only sought between two such measures of achievement. The high overall and negative agreement reflects the agreement for negative cases; a positive case is more likely a false positive than a true positive.

Values of PO and NA were consistent for different values of test reliabilities while PA values varied as a function of the test reliabilities. Table 2 shows the average PO, PA, and NA values when the achievement criterion was less than 85 or 90 at different combinations of test reliability. From Table 2 it can be seen that for PA to approach 0.5, the reliability of one test had to be 0.9 while the other had to be 0.8 (product = 0.72). Even when reliabilities of both tests were 0.9 the average PA did not exceed 0.6.

The PA results presented in Table 2 are shown graphically in Figures 1 and 2 when the achievement criterion is less than 85 and 90, respectively. These graphics show that with one exception, every condition had the potential to produce a PA value of 0.0. From Figures 1 and 2 it can be seen that the likelihood of such an event is highest when both achievement tests have low reliabilities and that the likelihood diminishes as the reliabilities increase. Similarly, the likelihood of no agreement is inversely related to the achievement criterion, with the likelihood of no agreement increasing as the cut score is lower, i.e., requiring more impaired performance to meet the disability criterion. Conversely, the graphics show that agreement between two different tests of achievement can, although rarely, approach 0.8 when the reliabilities of both tests are high.

Study 2

Methods

Results from study 1 demonstrated that positive agreement for SLD identification decisions is negatively impacted by changes in test selection across a wide range of plausible correlations between variables and different test reliabilities. However, the aggregated results are somewhat complicated to contextualize and do not allow for the parsing of effects due to differences in test reliabilities. Thus, the purpose of study 2 was to isolate the relative impact that changes in test reliabilities have on agreement by holding the correlations between latent variables constant. The study simulated data following procedures similar to study 1.

In order to specify plausible correlations between latent variables, we identified commonly utilized, psychometrically sound measures that could be utilized to identify SLD with deficits in reading comprehension. One psychoeducational battery that could be utilized to evaluate SLD status within a PSW framework is the Woodcock-Johnson- III (WJ-III; Woodcock, McGrew, & Mather, 2007). The WJ-III has been recommended for use in the SLD identification process by PSW proponents (e.g. Flanagan et al., 2007) and is explicitly mentioned in guidance documents for school practitioners implementing the PSW approach to SLD identification (e.g. Portland Public Schools, 2013). Additionally, subtests from the WJ-III have been utilized in previous empirical investigations of PSW methods (Miciak, Fletcher, et al., 2014; Miciak, Taylor, et al., 2015). Subtests that could be utilized include: WJ-III Passage Comprehension, WJ-III Verbal Comprehension, and WJ-III Concept Formation as the measures of achievement, cognitive weakness, and cognitive strength, respectively. Observed correlations between these measures are attenuated by the imperfect reliabilities of the measures. The disattenuated correlations can be derived using published values of correlations and reliabilities and can serve as reasonable approximations of the correlations between the latent constructs represented by the observed measures and are used as such throughout study 2. Figure 3 presents a graphical representation of the observed and latent relations among a potential set of measures for evaluating SLD status within a PSW framework. Given the set of latent correlations derived from the disattenuated observed correlations we followed the procedure in study 1 to simulate the data for study 2. Once again we simulated Achievement Test 1 (WJ-III Passage Comprehension), a highly related but different test of the same latent construct (e.g. the Gates MacGinitie Reading Test [MacGinitie, 2000]) which would serve as Achievement Test 2, a Cognitive Strength Test,

and a Cognitive Weakness Test for the derived latent correlation matrix while controlling levels of reliability.

We sought to manipulate the reliability of the measures to isolate the impact of imperfect reliability on agreement across different academic indicators. All measures demonstrate varying reliabilities and it is important to contextualize the effects of reliability on agreement. An assessment battery consisting of multiple cognitive and academic tests may demonstrate test reliabilities ranging from .70 to .95. For example, the choice to utilize a composite score for reading comprehension instead of a single subtest (passage comprehension) can improve reliability for an 11 year old from .83 to .88. Study 2 helps to contextualize the effects of these changes in reliability.

For the battery described above, if all four measures were perfectly reliable, then the threshold for every comparison becomes 0 and the only way to have a non-significant difference is in the case where two scores are identical. This would have to be the case for Cognitive Weakness and Achievement in order to meet the PSW model specifications. Given perfect reliability of all measures the following scores would indicate a positive SLD status.

Achievement	Cognitive Strength	Cognitive Weakness
84	85	84

This does not seem in keeping with the spirit of the model. For that reason, the upper level of reliability for this experiment was 0.95.

Thus, the first condition evaluated PA when reliability was 0.95 for all observed measures. The next four conditions evaluated PA when the reliability of one of the four measures was reduced to 0.85. The sixth condition reduced the reliability of both achievement measures to 0.85. The seventh condition reduced the reliabilities of both achievement measures and that of the cognitive strength measure to 0.85. The final condition reduced the reliability of all four observed measures to 0.85. Each condition was replicated 1000 times. It was believed that 1000 replications would provide a reasonable balance between computing demands and an adequate sampling distribution of PA from which means and standard deviations could be accurately calculated (Paxton, Curran, Bollen, Kirby, & Chen, 2001).

Results

The results from study 2 are presented in Table 3. Note that the largest standard deviation in Table 3 for the agreement measures was 0.03. This means that the largest resulting standard error of the mean was below 0.001 and suggests that the 1000 replications were sufficient for the task. The average PA in condition 1 when all reliabilities were 0.95 was 0.67 ($SD = 0.02$) when achievement was less than 85 and 0.69 ($SD = 0.02$) when achievement was less than 90. Decreasing the reliability of Cognitive Weakness resulted in an increase to PA when Cognitive Weakness was the only variable with reduced reliability ($M = 0.73$, $SD = 0.02$, when achievement < 85; $M = 0.75$, $SD = 0.03$ when achievement < 90). Decreased reliability of Cognitive Weakness also resulted in an increase in PA when all other reliabilities were reduced, condition 8, as compared to condition 7 ($M = 0.57$, SD = 0.03, when achievement < 85; $M = 0.60$, $SD = 0.02$ when achievement < 90). Although there was an increase in PA

when comparing condition 8 to condition 7 there was an overall decrease in PA when compared to condition 1.

The effect of reduced reliability of Cognitive Strength on PA was mixed. In condition 3 where only the reliability of Cognitive Strength was reduced there was no change in the average PA for either level of achievement as compared to condition 1. When the reliability of Cognitive Strength was reduced after the reliabilities of both achievement measures had been reduced there was an increase in average PA when compared to the prior condition (7 vs 6). When achievement was < 85 the average PA was 0.53 ($SD = 0.03$) and when achievement < 90 the average PA was 0.56 ($SD = 0.02$). As with the effect of Cognitive Strength, there was an increase in PA relative to the prior condition, but there was an overall decrease in PA relative to the baseline condition.

Decreasing the reliability of either of the achievement measures reduced the average PA to 0.54 ($SD = 0.02$) when achievement < 85 and 0.56 ($SD = 0.02$) when achievement < 90. When the reliabilities of both achievement measures was dropped to 0.85 there was a further reduction in average PA to 0.52 ($SD = 0.02$) when achievement < 85 and 0.55 (SD = 0.02) when achievement < 90. As noted earlier, the decrease in PA when both achievement measures have reliabilities of 0.85 is mitigated when the reliability of Cognitive Weakness and/or Cognitive Strength was decreased but the overall decrease still remained substantial.

Discussion

We simulated data to investigate the reliability of SLD identification decisions across different tests (indicators) of the same latent academic factor for the C/DM, a proposed cognitive discrepancy method for the identification of SLD. We further evaluated the effect of diminished reliability on agreement for SLD identification decisions. In study 1, over 70,000 unique combinations of latent correlations between a cognitive strength, cognitive weakness, and achievement factor were stipulated and utilized to generate case-specific observed values on four indicators (Cognitive Strength, Cognitive Weakness, Achievement Test 1, and Achievement Test 2) given a wide range of potential observed reliabilities and random error. This large number of simulated samples with excellent coverage of the parameter space permits generalization beyond the intercorrelations of a specific test battery or set of tests and allows us to address fundamental questions about the reliability of ipsative cognitive discrepancy methods such as the C/DM. For each simulated sample, we applied inclusionary criteria for SLD identification specified by the C/DM to determine if individual cases would meet SLD criteria with battery 1 (Cognitive Strength, Cognitive Weakness, and Achievement Test 1) and battery 2 (Cognitive Strength, Cognitive Weakness, and Achievement Test 2) and evaluated agreement for the classification decisions based on the two different measures of the same achievement domain using the same indicators of Cognitive Weakness and Cognitive Strength.

Agreement statistics for these simulations raise important questions about whether the C/DM, across any potential operationalization, is sufficiently robust to differences in test selection to recommend widespread adoption. Proponents of the C/DM (and other PSW methods) assert that flexibility in test selection represents a strength of the proposed

methods (Flanagan et al., 2007; Hale & Fiorello, 2004). However, this flexibility in test selection negatively impacts the consistency of resulting decisions. In the present simulations, the average positive agreement for SLD identification was .42 at a cut point of 85 for achievement deficits and .45 with a cut point of 90; positive agreement rarely approached .80. To be concrete, consider what the identification rates and levels of positive and negative agreement would mean in a sample of 1,000 students with below average reading and no evidence of intellectual disabilities. In this sample, 958 students would not be identified as SLD by either battery and 13 students would be identified by both batteries. However, 29 students would receive a different identification decision depending on which achievement measure was used. Thus, as a result of minor differences in the test battery used by the diagnostician, the child's classification would change.

In study 2, we evaluated the effects of reduced reliability on one potential operationalization of the C/DM. Within this simulation, we replicated latent correlations between achievement measures, a related cognitive weakness, and a potential cognitive strength. Mirroring the results of study 1, all conditions demonstrated high overall agreement and high negative agreement because of the low identification rate. As in study 1, however, the problem is low positive agreement. Indeed, simply dropping the reliability of a single measure from excellent ($r = .95$) to adequate ($r = .85$) had a significant deleterious effect on positive agreement. When the two achievement variables exhibited reduced reliability, positive agreement ranged from .52 - .55. Further, when additional measurement error was introduced into the model in the form of reduced reliability for cognitive variables, agreement statistics increased. This initially seems counterintuitive but makes sense after careful consideration. The model requires a non-significant difference between achievement and a cognitive weakness. If Achievement Test 1, Achievement Test 2, and Cognitive Weakness take on three different values then the differences of Cognitive Weakness – Achievement Test 1 and Cognitive Weakness – Achievement Test 2 will differ. Further, when the reliabilities of Achievement Test 1 and Achievement Test 2 are the same and the reliability of Cognitive Weakness is high then the threshold for a significant difference may fall between the values of the two observed differences creating disagreement. Decreasing the reliability of Cognitive Weakness increases the threshold for a significant difference. With a large enough decrease in reliability of Cognitive Weakness the differences between Cognitive Weakness and the achievement measures will both be non-significant resulting in agreement. These findings raise significant questions about how accurately complex measurement models such as the C/DM are capturing reality.

Importantly, the agreement estimates across psychoeducational batteries described above are overly optimistic because they estimate diminishments in agreement due to variation in achievement tests but not cognitive processing tests. Additionally, the present studies evaluated only one triad of a cognitive strength, cognitive weakness, and achievement factor. The measurement of additional cognitive processes (at the discretion of the school psychologist) and the utilization of different tests to measure those processes would further diminish agreement between different psychoeducational batteries. Further, the present study evaluated agreement when correlations between the achievement deficit, cognitive strength, and cognitive weakness were greater than 0.50. In practice, the correlations between these tests may be weaker, which may diminish agreement further. Nor is clinical judgment likely

to improve the consistency or validity of SLD identification decisions for the C/DM or other PSW methods (Macmann & Barnett, 1997; Canivez, 2013). The notion of the master detective who is able to integrate unreliable test data to make reliable decisions has been characterized as a shared professional myth (Watkins, 2000). As Watkins and Canivez conclude: "because ipsative subtest categorizations are unreliable, recommendations based on them will also be unreliable. Procedures that lack reliability cannot be valid" (2004; p. 137).

The results of the present simulations align with extensive research demonstrating that *all* SLD identification methods that apply a fixed cut point to continuous data are unreliable at the individual level (Miciak, Taylor, et al., 2015). This is true of low achievement methods (Francis et al., 2005), IQ achievement discrepancy methods (Francis et al.; Macmann et al., 1989), RTI methods that dichotomize based on final status (Barth et al., 2008; Fletcher et al., 2014), RTI methods that apply dual discrepancy criteria (Burns, Scholin, Kosciolek, & Livingston, 2010; Burns & Senesac, 2005), and PSW methods (Miciak, Taylor, et al. 2015). Inconsistent identification decisions at the individual level are an inherent result of imperfect test validity and reliability. Students who score close to the cut point tend to fluctuate in group membership across differences in testing occasion, measure, or criteria.

This instability in group membership is exacerbated for ipsative cognitive discrepancy methods like the C/DM that rely on applying a cut point to *difference scores*. Difference scores are often less reliable than the two positively correlated measures from which they are derived (Macmann & Barnett, 1997; McDermott et al., 1997). As a result, methods that rely on difference scores alone are generally less reliable than other psychometric methods. This fact was well documented for methods that relied on the identification of an IQ-achievement discrepancy (Francis et al., 2005; MacMann et al., 1989). The C/DM does not improve upon the reliability of the IQ- achievement discrepancy methods for the identification of SLD in any demonstrable way. Instead, by increasing the number of factors measured and increasing the number of comparisons, the C/DM adds complexity to the cognitive discrepancy framework that will negatively impact the consistency of identification decisions. The increased complexity of the assessment process and its interpretation is not unique to the C/DM, but will affect all proposed operationalizations of PSW methods, including the XBA and DCM (Flanagan et al., 2007, Naglieri, 1999). Previous investigations of other PSW methods have demonstrated similar limitations in agreement at the individual level (Kranzler, et al., in press; Miciak, Taylor, et al., 2015; Stuebing et al., 2012).

Implications

The results of this study demonstrate the inherent unreliability of cognitive discrepancy methods for SLD identification like the C/DM. This inherent unreliability has important theoretical and clinical implications. Cognitive discrepancy frameworks for the identification of SLD theorize that there exist qualitative differences between individuals who demonstrate cognitive profiles with specific strengths and weaknesses (individuals with SLD) and individuals who demonstrate generally flat cognitive profiles. However, if these profiles cannot be reliably identified and different measures will identify different individuals, the validity of the theory must be questioned. Methods that are unreliable cannot be valid;

theories that rely on inherently unreliable procedures cannot be substantiated. Further, there is little evidence for the utility of separating individuals with greater cognitive variability from individuals with low, flat cognitive profiles who demonstrate similar academic needs. Students with borderline intellectual abilities represent a traditionally underserved population (Shaw, 2005). However, meta-analyses indicate that neither I.Q. nor other baseline cognitive characteristics are robust predictors of intervention response (Stuebing et al., 2009; Stuebing et al., 2014). Until evidence emerges demonstrating strong aptitude by treatment interactions for these subgroups, we see little to recommend the practice of classifying according to cognitive variability.

The results of these simulations also have important implications for school practice. Proponents of the C/DM and PSW methods make strong evidentiary claims (Hanson, et al., 2008) and cite uncertainty about response to intervention methods for SLD identification as a rationale for a movement toward PSW methods (Hale et al., 2010; Kavale, Kauffman, Bachmeier, LeFever, 2008; Reynolds & Shaywitz, 2009). However, a move toward PSW methods would not resolve persistent reliability problems around the SLD identification process nor is it likely to augur consensus on procedures and methods to identify SLD across its heterogeneous manifestations.

In the absence of robust evidence demonstrating that extensive cognitive assessment improves treatment effectiveness (Pashler, McDaniel, Rohrer, & Bjork, 2008; Kearns & Fuchs, 2013), it is difficult to argue that PSW methods would represent a positive step towards the goal of ensuring that struggling students receive necessary and effective help. PSW methods potentially require several hours of individual testing, often conducted over a period of weeks. Such an undertaking is expensive and time consuming, significant drawbacks in a setting of limited time and resources. Further, it is important to note the very low SLD identification rates found in the present study. In study 1, across a wide range of potential latent correlations between academic and cognitive variables and different test reliabilities, the mean identification rate was 4% of the selection sample when the cut point for academic deficits was < 85 and 6% when the cut point for academic deficits was < 90. These identification rates are consistent with previous simulations of PSW methods (Stuebing et al., 2012) and suggest that considerable testing resources would be expended for every positive SLD identification.

Until such time as cognitively tailored interventions are well established and ready for widespread adoption in schools, we suggest that finite resources would be better utilized directly assessing important academic skills and providing targeted help for students who need it. This goal is most easily achieved within schools organized around a response to intervention service delivery framework, which includes universal screening for academic risk, frequent direct assessment of academic progress in the form of progress monitoring, and the provision of interventions of increasing intensity for students at-risk or students who are not making sufficient academic progress (Bradley, Danielson, & Doolittle, 2005; Fletcher & Vaughn, 2009). Importantly, SLD identification emanates from, but is not the goal of response to intervention service delivery frameworks. The goal of response to intervention service delivery frameworks is the prevention of academic difficulties and data generated as part of this prevention process can inform SLD decisions for students who

require legal protections and/or assessment and curricular modifications and accommodations from schools.

In addition to implications for school practice, the results of the present studies have implications for assessment in clinical settings, in which data regarding response to intervention may not be as readily available. Despite this challenge, the results of present studies suggest that moving instead to complex psychoeducational assessment models such as the C/DM would not result in improved diagnostic accuracy. Instead, these complex methods demonstrate inherent instability that may limit their diagnostic and prescriptive utility. Instead, we suggest clinicians should focus on the direct assessment of academic skills, consideration of other conditions that may be comorbid or inform the etiology of academic difficulties (e.g. ADHD), and work to improve collaboration between the clinician, parent, educators, and the individual with SLD to improve academic interventions and outcomes, as recommended by the DSM-5 (APA, 2013).

Conclusions

These results are consistent with previous studies demonstrating that the application of strict cut points to continuous data results in inherently unstable groups, particularly when the cut point is applied to a distribution of differences for two measures as the C/DM prescribes. Because of this problem, PSW methods like the C/DM are unlikely to resolve persistent reliability problems associated with the SLD identification process. Until such time as PSW methods are proven to positively affect intervention outcomes, there is little evidence to support their widespread adoption.

Acknowledgments

References

Barth AE, Stuebing KK, Anthony JL, Denton CA, Mathes PG, Fletcher JM, Francis DJ. Agreement among response to intervention criteria for identifying responder status. Learning and Individual Differences. 2008; 18:296–307. DOI: 10.1016/j.lindif.2008.04.004 [PubMed: 19081758]

Bradley R, Danielson L, Doolittle J. Response to intervention. Journal of Learning Disabilities. 2005; 38:485–486. DOI: 10.1177/00222194050380060201 [PubMed: 16392688]

Burns MK, Scholin SE, Kosciolek S, Livingston J. Reliability of decision-making frameworks for response to intervention for reading. Journal of Psychoeducational Assessment. 2010; 28(2):102–114. DOI: 10.1177/0734282909342374

Burns MK, Senesak BV. Comparison of dual discrepancy criteria to assess response to intervention. Journal of School Psychology. 2005; :393–406. DOI: 10.1016/j.jsp.2005.09.003

Canivez GL. Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. The Oxford handbook of child psychological assessments. 2013:84–112.

Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology. 1990; :43, 551–558. DOI: 10.1016/0895-4356(90)90159-M

Consortium for Evidence-Based Early Intervention Practices. A response to the Learning Disabilities Association of America (LDA) white paper on specific learning disabilities (SLD) identification.

2010. Downloaded on May 1, 2014 from: http://www.isbe.net/spec-ed/pdfs/LDA_SLD_white_paper_response.pdf

Evans JJ, Floyd RG, McGrew KS, Leforgee MH. The relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and reading achievement during childhood and adolescence. School Psychology Review. 2001; 31:246–262. Downloaded on June 21, 2015 from: http://www.iapsych.com/wj3ewok/LinkedDocuments/Evans2002.pdf.

Fiorello CA, Primerano D. Research into practice: Cattell-Horn-Carroll cognitive assessment in practice: Eligibility and program development issues. Psychology in the Schools. 2005; 42(5):525–536. DOI: 10.1002/pits.20089

Flanagan DP, Fiorello CA, Ortiz SO. Enhancing practice through application of Cattell-Horn-Carroll theory and research: A "third method" approach to specific learning disability identification. Psychology in the Schools. 2010; 47(7):739–760. DOI: 10.1002/pits.20501

Flanagan, D.Ortiz, S., Alfonso, VC., editors. Essentials of cross battery assessment. 2nd. Hoboken, NJ: John Wiley & Sons, Inc; 2007.

Fletcher, JM. Classification and identification of learning disabilities. In: Wong, B., Butler, D., editors. Learning about learning disabilities. 4th. New York, NY: Elsevier; 2012.

Fletcher, JM., Lyon, GR., Fuchs, LS., Barnes, MA. Learning disabilities: From identification to intervention. New York, NY: Guilford; 2007.

Fletcher JM, Stuebing KK, Barth AE, Denton CA, Cirino PT, et al. Vaughn S. Cognitive correlates of inadequate response to reading intervention. School Psychology Review. 2011; 40:3–22. Downloaded on June 21, 2015 from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3485697/. [PubMed: 23125475]

Fletcher JM, Shaywitz SE, Shankweiler DP, Katz L, Liberman IY, Stuebing KK, et al. Shaywitz BA. Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. Journal of Educational Psychology. 1994; 86(1):6.doi: 10.1037/0022-0663.86.1.6

Fletcher JM, Stuebing KK, Barth AE, Miciak J, Francis DJ, Denton CA. Agreement and coverage of indicators of response to intervention: A multimethod comparison and simulation. Topics in Language Disorders. 2014; 34(1):74–89. DOI: 10.1097/TLD.0000000000000004 [PubMed: 25364090]

Fletcher JM, Vaughn S. Response to intervention: Preventing and remediating academic deficits. Child Development Perspectives. 2009; 3:30–37. DOI: 10.1111/j.1750-8606.2008.00072.x [PubMed: 21765862]

Francis DJ, Fletcher JM, Stuebing KK, Lyon GR, Shaywitz BA, Shaywitz SE. Psychometric Approaches to the Identification of LD IQ and Achievement Scores Are Not Sufficient. Journal of Learning Disabilities. 2005; 38(2):98–108. DOI: 10.1177/00222194050380020101 [PubMed: 15813593]

Hale JB, Alfonso V, Berninger B, Bracken B, Christo C, Clark E, et al. Yalof J. Critical issues in response-to-intervention, comprehensive evaluation, and specific learning disabilities identification and intervention: An expert white paper consensus. Learning Disability Quarterly. 2010; 33(3):223–236.

Hale, JB., Fiorello, CA. School neuropsychology: A practitioner's handbook. New York, NY: The Guilford Press; 2004.

Hale JB, Fiorello CA, Kavanagh JA, Hoeppner JB, Gaither RA. WISC-III predictors of academic achievement for children with learning disabilities: Are global and factor scores comparable? School Psychology Quarterly. 2001; 16:31–55. DOI: 10.1521/scpq.16.1.31.19158

Hallahan, DP., Mercer, CD. Learning disabilities: historical perspectives. In: Bradley, R.Danielson, L., Hallahan, DP., editors. Identification of learning disabilities: Research to practice. Mahwah, NJ: Lawrence Erlbaum Associates; 2002. p. 1-68.

Hallgren KA. Conducting simulation studies in the R programming environment. Tutorials in Quantitative Methods for Psychology. 2013; 9(2):43–60. Retrieved from http://www.tqmp.org/. [PubMed: 25067989]

Hanson, J., Sharman, MS., Esparza-Brown, J. Technical Assistance Paper. Oregon School Psychologists Association; 2008. Pattern of strengths and weaknesses in specific learning

disabilities: What's it all about?. Retrieved May 4, 2015 at http://www.jamesbrenthanson.com/uploads/PSWCondensed121408.pdf

Hoskyn M, Swanson HL. Cognitive processing of low achievers and children with reading disabilities: A selective meta-analytic review of the published literature. School Psychology Review. 2000; 29(1):102–119. http://search.proquest.com/openview/26ab403ae1d6f5182f9a62ecdc2e4405/1?pq-origsite=gscholar.

Johnson ES. Understanding why a child is struggling to learn: The role of cognitive processing evaluation in learning disability identification. Topics in Language Disorders. 2014; 34(1):59–73. DOI: 10.1097/TLD.0000000000000007

Johnson ES, Humphrey M, Mellard DF, Woods K, Swanson HL. Cognitive processing deficits and students with specific learning disabilities: A selective meta-analysis of the literature. Learning Disability Quarterly. 2010; 33(1):3–18. DOI: 10.1177/073194871003300101

Kavale KA, Kauffman JM, Bachmeier RJ, LeFever GB. Response-to-intervention: Separating the rhetoric of self-congratulation from the reality of specific learning disability identification. Learning Disability Quarterly. 2008; 31(3):135–150. DOI: 10.2307/25474644

Kearns DM, Fuchs D. Does cognitively focused instruction improve the academic performance of low-achieving students? Exceptional Children. 2013; 79:263–290. Retrieved June 22, 2015 from http://eric.ed.gov/?id=EJ1013633.

Kranzler JH, Floyd RG, Benson N, Zaboski B, Thibodaux L. Classification agreement analysis of cross-battery assessment in the identification of specific learning disorders in children and youth. International Journal of School and Educational Psychology. in press.

MacGinitie, WH. Gates-MacGinitie reading tests. Itasca, IL: Riverside; 2000.

Macmann GM, Barnett DW. Myth of the master detective: Reliability of interpretations for Kaufman's 'intelligent testing' approach to the WISC–III. School Psychology Quarterly. 1997; 12(3):197–234. DOI: 10.1037/h0088959

Macmann GM, Barnett DW, Lombard TJ, Belton-Kocher E, Sharpe MN. On the Actuarial Classification of Children Fundamental Studies of Classification Agreement. The Journal of Special Education. 1989; 23(2):127–149. DOI: 10.1177/002246698902300202

McGrew KS, Wendling BJ. Cattell–Horn–Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. Psychology in the Schools. 2010; 47(7):651–675. DOI: 10.1002/pits.20497

Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist. 1995; 50(9):741–749. DOI: 10.1037/0003-066X.50.9.741

Miciak J, Fletcher JM, Stuebing KK, Vaughn S, Tolar TD. Patterns of cognitive strengths and weaknesses: Identification rates, agreement, and validity for learning disabilities identification. School Psychology Quarterly. 2014; 29(1):21.doi: 10.1037/spq0000037 [PubMed: 24274155]

Miciak J, Taylor WP, Denton CA, Fletcher JM. The Effect of Achievement Test Selection on Identification of Learning Disabilities within a Patterns of Strengths and Weaknesses Framework. School Psychology Quarterly. 2015; 30:321–334. DOI: 10.1037/spq0000091 [PubMed: 25243467]

Morris RD, Fletcher JM. Classification in neuropsychology: A theoretical framework and research paradigm. Journal of Clinical and Experimental Neuropsychology. 1998; 10:640–658. DOI: 10.1080/01688638808402801

Naglieri, JA. Essentials of CAS assessment. New York, NY: John Wiley & Sons, Inc; 1999.

Pashler H, McDaniel M, Rohrer D, Bjork R. Learning styles concepts and evidence. Psychological Science in the Public Interest. 2008; 9(3):105–119. DOI: 10.1111/j.1539-6053.2009.01038.x [PubMed: 26162104]

Paxton P, Curran PJ, Bollen KA, Kirby J, Feinian C. Monte Carlo Experiments: Design and Implementation. Structural Equation Modeling. 2001; 8:287–312. DOI: 10.1207/S15328007SEM0802_7

Portland Public Schools. Guidance for the identification of specific learning disabilities. 2013. Retrieved May 1, 2015 from http://www.pps.k12.or.us/files/special-education/PSW_Feb_2013_Guide.pdf

SAS Institute. SAS (9.4) [computer software]. Cary, NC: SAS Institute Inc; 2013.

Shaw SR. An educational programming framework for a subset of students with diverse learning needs: Borderline intellectual functioning. Intervention in School and Clinic. 2008; 43(5):291–299.

Stuebing KK, Barth AE, Molfese PJ, Weiss B, Fletcher JM. IQ is not strongly related to response to reading instruction: A meta-analytic interpretation. Exceptional Children. 2009; 76(1):31–51. DOI: 10.1177/001440290907600102 [PubMed: 20224749]

Stuebing KK, Barth AE, Trahan LH, Reddy RR, Miciak J, Fletcher JM. Are child cognitive characteristics strong predictors of responses to intervention? A meta-analysis. Review of educational research. 2014; doi: 10.3102/0034654314555996

Stuebing KK, Fletcher JM, Branum-Martin L, Francis DJ. Evaluation of the technical adequacy of three methods for identifying specific learning disabilities based on cognitive discrepancies. School Psychology Review. 2012; 41:3–22. Retrieved June 23, 2015 from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3466817/. [PubMed: 23060685]

Stuebing KK, Fletcher JM, LeDoux JM, Lyon GR, Shaywitz SE, Shaywitz BA. Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. American Educational Research Journal. 2002; 39(2):469–518. DOI: 10.3102/00028312039002469

Tanaka H, Black JM, Hulme C, Stanley LM, Kesler SR, Whitfield-Gabrieli S, et al. Hoeft F. The brain basis of the phonological deficit in dyslexia is independent of IQ. Psychological Science. 2011; 22(11):1442–1451. DOI: 10.1177/0956797611419521 [PubMed: 22006060]

Reynolds CR, Shaywitz SE. Response to Intervention: Ready or not? Or, from wait-to-fail to watch-them-fail. School Psychology Quarterly. 2009; 24(2):130–145. DOI: 10.1037/a0016158 [PubMed: 20169006]

Vellutino FR, Scanlon DM, Lyon GR. Differentiating between difficult-to-remediate and readily remediated poor readers - More evidence against the IQ-achievement discrepancy definition of reading disability. Journal of Learning Disabilities. 2000; 33(3):223–238. DOI: 10.1177/002221940003300302 [PubMed: 15505962]

Watkins MW. Cognitive profile analysis: A shared professional myth. School Psychology Quarterly. 2000; 15(4):465–479. DOI: 10.1037/h0088802

Watkins MW, Canivez GL. Temporal stability of the WISC-III subtest composite: Strengths and weaknesses. Psychological Assessment. 2004; 16(2):133–138. DOI: 10.1037/1040-3590.16.2.133 [PubMed: 15222809]

Woodcock, RW., McGrew, KS., Mather, N. Woodcock-Johnson III Tests of Achievement. Itasca, IL: Riverside; 2001.

Ysseldyke JE, Algozzine B, Shinn MR, McGue M. Similarities and differences between low achievers and students classified learning disabled. The Journal of Special Education. 1982; 16:73–85. DOI: 10.1177/002246698201600108
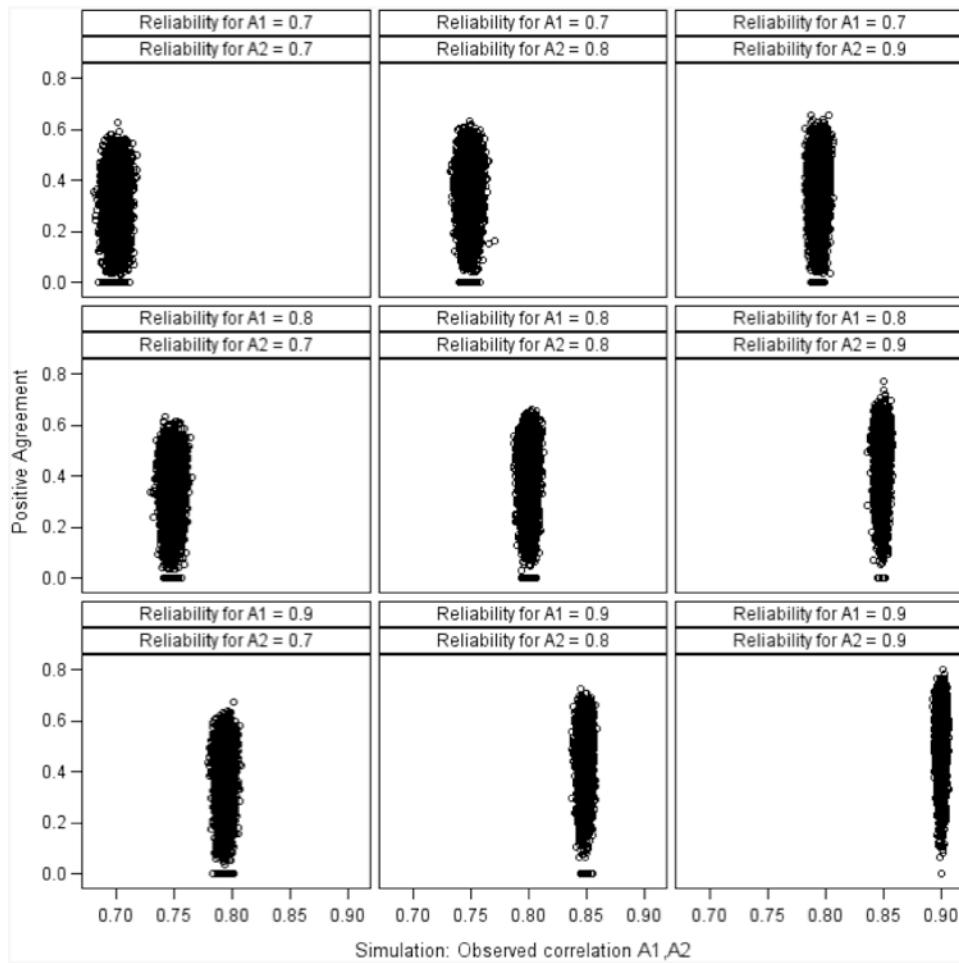
Figure 1. Positive agreement by the observed correlation for two achievement tests with a cut point < 85 across 3 levels of reliability

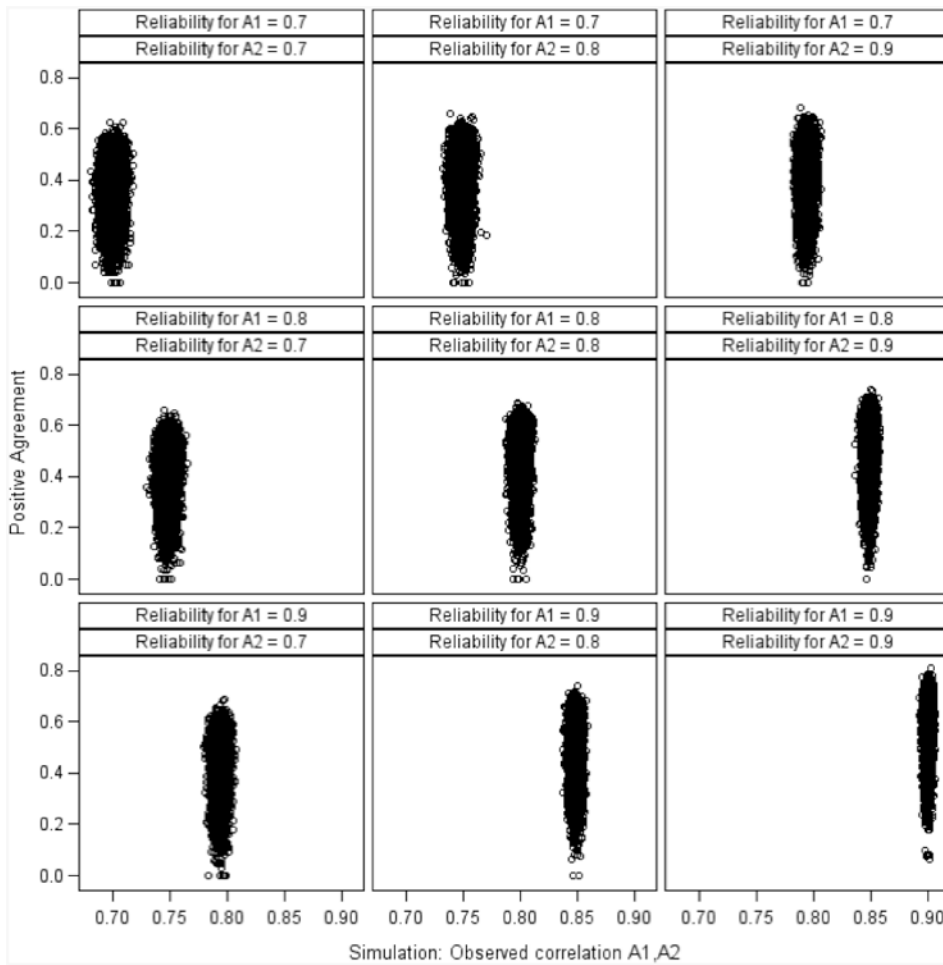Note: A1 = Achievement Test 1; A2 = Achievement Test 2.

Figure 2. Positive agreement by the observed correlation for two achievement tests with a cut point < 90 across 3 levels of reliability

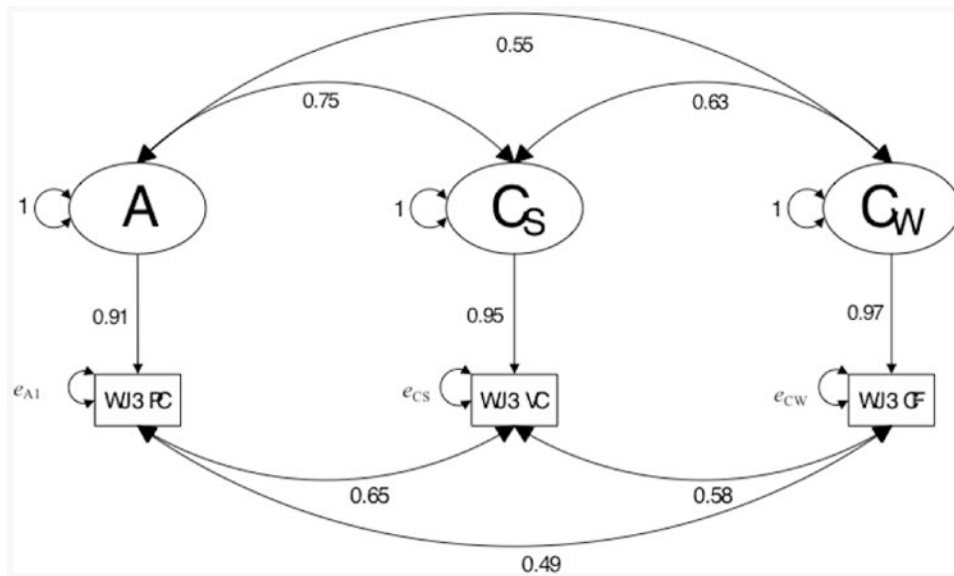Note: A1 = Achievement Test 1; A2 = Achievement Test 2.

Figure 3. Observed and latent relations for study 2

Note: A = Achievement; C_s = Cognitive Strength; C_w = Cognitive Weakness. Example Measures: WJ3 PC = Woodcock-Johnson III Passage Comprehension; WJ3 VC = Woodcock-Johnson III Verbal Comprehension; WJ3 CF = Woodcock-Johnson III Concept Formation.

Table 1

Average cell size by achievement value

Achievement Value			Test 2	
			Yes	No
< 85		Yes	46.7 (40.4)	52.4 (33.5)
	Test 1	No	52.4 (33.5)	3519.1 (170.7)
< 90		Yes	74.4 (60.7)	74.7 (45.0)
		No	74.7 (45.1)	3446.7 (196.7)

Note: Values in parentheses are standard deviations.

Table 2

Average agreement for different achievement tests at different cut points by varying levels of reliability

Product of reliabilities	Selection N	Achievement < 85			Achievement < 90		
		PO[a]	PA	NA[b]	PO[a]	PA	NA[b]
0.49	3875.8 (51.3)	0.97	0.33 (0.11)	0.99	0.96	0.37 (0.10)	0.98
0.56	3784.0 (51.3)	0.97	0.37 (0.11)	0.98	0.96	0.40 (0.11)	0.98
0.63	3688.8 (50.8)	0.97	0.40 (0.11)	0.98	0.96	0.43 (0.11)	0.98
0.64	3675.3 (51.4)	0.97	0.43 (0.11)	0.99	0.96	0.46 (0.11)	0.98
0.72	3560.3 (51.4)	0.97	0.48 (0.11)	0.99	0.96	0.50 (0.11)	0.98
0.81	3417.8 (50.5)	0.97	0.56 (0.10)	0.99	0.96	0.59 (0.10)	0.98

Note: Values in parentheses are standard deviations.

[a] Standard deviations omitted, all values = 0.02;

[b] Standard deviations omitted, all values = 0.01.

PO = Overall Agreement; PA = Positive Agreement; NA = Negative Agreement.

Table 3

Average agreement by different achievement tests at different cut points when selected measures have reduced reliability

Condition	Reduced Reliability	Selection N	Achievement < 85			Achievement < 90		
			PO[a]	PA	NA[a]	PO[b]	PA	NA[a]
1	None	3235.1 (46.4)	0.93	0.67 (0.02)	0.96	0.90	0.69 (0.02)	0.94
2	Cw	3243.5 (47.7)	0.95	0.73 (0.02)	0.97	0.93	0.75 (0.02)	0.96
3	Cs	3242.4 (46.2)	0.95	0.67 (0.02)	0.97	0.93	0.69 (0.02)	0.96
4	A2	3420.6 (47.2)	0.91	0.54 (0.02)	0.95	0.87	0.56 (0.02)	0.92
5	A1	3420.7 (47.4)	0.91	0.54 (0.02)	0.95	0.87	0.56 (0.02)	0.92
6	A1, A2	3555.3 (48.9)	0.91	0.52 (0.02)	0.95	0.87	0.55 (0.02)	0.93
7	A1, A2, Cs	3561.2 (50.3)	0.93	0.53 (0.03)	0.96	0.91	0.56 (0.02)	0.95
8	All	3562.3 (49.8)	0.95	0.57 (0.03)	0.97	0.93	0.60 (0.02)	0.96

Note: Values in parentheses are standard deviations.

[a] Standard deviations omitted, all values < 0.01;

[b] Standard deviations omitted, all values 0.01.

PO = Overall Agreement; PA = Positive Agreement; NA = Negative Agreement.