Published in final edited form as:

Sch Psychol Q. 2015 September ; 30(3): 321–334. doi:10.1037/spq0000091.

The Effect of Achievement Test Selection on Identification of Learning Disabilities within a Patterns of Strengths and Weaknesses Framework

Jeremy Miciak,

University of Houston

Pat Taylor,

University of Houston

Carolyn A. Denton, and

The University of Texas Health Science Center Houston

Jack M. Fletcher

University of Houston

Abstract

Purpose—Few empirical investigations have evaluated learning disabilities (LD) identification methods based on a pattern of cognitive strengths and weaknesses (PSW). This study investigated the reliability of LD classification decisions of the concordance/discordance method (C/DM) across different psychoeducational assessment batteries.

Methods—C/DM criteria were applied to assessment data from 177 second grade students based on two psychoeducational assessment batteries. The achievement tests were different, but were highly correlated and measured the same latent construct. Resulting LD identifications were then evaluated for agreement across batteries on LD status and the academic domain of eligibility.

Results—The two batteries identified a similar number of participants as having LD (80 and 74). However, indices of agreement for classification decisions were low ($\kappa = .29$), especially for percent positive agreement (62%). The two batteries demonstrated agreement on the academic domain of eligibility for only 25 participants.

Conclusions—Cognitive discrepancy frameworks for LD identification are inherently unstable because of imperfect reliability and validity at the observed level. Methods premised on identifying a PSW profile may never achieve high reliability because of these underlying psychometric factors. An alternative is to directly assess academic skills to identify students in need of intervention.

Keywords

decision reliability; diagnostic agreement; learning disabilities; response to intervention; cognitive testing

In 2010, the Learning Disabilities Association of America (LDA) published a white paper documenting what it described as an expert consensus on learning disability (LD) identification and treatment (Hale et al., 2010). The white paper recommended that LD identification should be based on the identification of an intra-individual pattern of strengths and weaknesses (PSW) in cognitive processing. This method reflects an underlying classification hypothesis that discrepancies in cognitive skills are a marker of the cardinal characteristic of LD: unexpected underachievement. It is consistent with conceptualizations of LD as a cognitive weakness in a sea of strengths (Shaywitz, 2003). However, the assertion that this view of LD is valid (Johnson, 2014) and that there exists an expert consensus to utilize cognitive processing strengths and weaknesses for LD identification has not been supported by empirical research (Fletcher, Lyon, Fuchs, & Barnes, 2007). It is not the view of all putative experts on LD identification (Consortium for Evidence Based Early Intervention Practices [CEBEIP], 2010). Nonetheless, several states (e.g. Georgia, Idaho, and Maine) require an evaluation of cognitive processes as part of the comprehensive evaluation required for special education placement (Hauerwas, Brown, & Scott, 2013).

Few empirical studies have directly investigated proposed methods to operationalize the PSW model. To our knowledge, no empirical study has investigated the reliability of identification decisions across measures within a PSW model. This is an important consideration because neither federal statutes nor state directives provide explicit guidance on the selection of measures or specific assessment procedures. Thus, different school psychologists may use different tests to measure achievement and cognitive processing, due to different district directives or idiosyncratic preferences. Proponents of PSW methods often assert that the approaches are robust across measures (Flanagan, Ortiz, & Alfonso, 2007; Hale & Fiorello, 2004). However, the utilization of different measures may impact the reliability of decision making within the PSW framework. Thus, questions emerge: Will two school psychologists arrive at the same identification decision if they follow the same procedures but administer different tests? Because of differences in the reliability of the tests, their intercorrelations, and different normative samples, it seems possible that identification decisions could be different even if the tests measure the same latent construct. In the present study, we directly investigate this possibility.

The Concordance/Discordance Model

Overview

Hale and Fiorello (2004) proposed a specific operationalization of the PSW model: the Concordance/Discordance Model of LD (C/DM). Within the C/DM, LD is marked by an intraindividual pattern of cognitive processing strengths and weaknesses that differentiates it from garden variety low achievement. According to this model, the academic underachievement of students with LD is explained by a specific cognitive weakness yet is unexpected because of the presence of cognitive strengths. To identify LD, observed scores must demonstrate a specified pattern of concordance between academic and cognitive processing weaknesses and discordance between cognitive processing strengths and those weaknesses.

Research on the C/DM

Proponents find support for PSW models like the C/DM in studies investigating cognitive processing correlates of academic achievement (Evans, Floyd, McGrew, & Leforgee, 2002; McGrew & Wendling, 2010), specific cognitive profiles (Fiorello, Hale, & Snyder 2006) or the implications of cognitive profile variability (Fiorello et al., 2006; Hale et al., 2010). Proponents argue that the identification of a profile of cognitive processing strengths and weaknesses aligns PSW models for LD identification with long-standing theory and definitions of LD and can improve treatment outcomes. They conclude:

Although there have been countless arguments both for and against the utility of this type of information (e.g. knowledge of specific CHC [Cattell-Horn-Carroll] cognitive strengths and weaknesses) for identification and treatment of SLD [Specific Learning Disability], there is currently no evidence to demonstrate that current cognitive tests are ineffective for either purpose. (Flanagan, Fiorello, & Ortiz, 2010; p. 741)

However, there is little evidence indicating that PSW methods like the C/DM are effective for identification and treatment. Further, the assertion that there is no evidence against the utilization of cognitive processing strengths and weaknesses for LD identification is potentially contradicted by research documenting: (a) the limitations of approaches based on identifying specific cognitive profiles (Macmann & Barnett, 1997; Watkins, 2000; Watkins & Canivez, 2004); (b) the psychometric limitations of ability-achievement discrepancy methods (Francis et al., 2005; Macmann, Barnett, Lombard, Belton-Kocher, & Sharpe, 1989); and (c) recent research raising methodological concerns specific to the reliability and validity of proposed PSW methods, including the C/DM (Miciak et al., 2014; Stuebing, Fletcher, Branum-Martin, & Francis, 2012).

Psychometric Issues

Psychometric approaches to the identification of LD demonstrate inherent instability in identification decisions for individual students. Psychometric approaches such as low achievement and discrepancy models for LD identification apply fixed cut points to a continuous distribution, which may represent performance on a single measure or the distribution of differences for two measures. However, imperfect reliability ensures that there will be instability around the cut point regardless of the threshold (Francis et al., 2005). This is exacerbated for ipsative methods utilizing a discrepancy score as the primary index for comparison because the discrepancy score is always less reliable than the positively correlated subtests from which it is derived (Macmann & Barnett, 1997; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992). Further, no test perfectly measures the latent construct of interest. Thus, variability in identification decisions may occur when alternate measures are utilized to measure the same latent construct (Macmann et al., 1989). These limitations apply to *all* psychometric approaches to the identification of LD, including PSW methods like the C/DM.

Research Studies of the Reliability of PSW Methods

Beyond these general limitations, recent studies have found evidence that proposed operationalizations of the PSW model demonstrate poor technical adequacy and reliability, and there is little evidence for the validity of these methods as classification systems. Stuebing et al. (2012) used simulated data to evaluate the technical adequacy of three operationalizations of the PSW model: (a) the C/DM (Hale & Fiorello, 2004), (b) the Cross Battery Assessment approach (XBA; Flanagan, et al., 2007), and (c) the Discrepancy/Consistency Method (D/CM; Naglieri, 1999). All three methods demonstrated very low identification rates (range: .1% – 17.9%), raising important questions about efficiency. Perhaps more troubling, results across the three methods demonstrated consistently poor positive predictive value (range: .01 – .71), suggesting that even if the identification of a PSW profile were educationally relevant, the methods would be unable to reliably identify true positives at the observed level. Within a model premised on matching instruction to individual cognitive profiles, this limited reliability for positive classifications may result in a large number of students receiving interventions not matched to their true cognitive and academic profile because of psychometric error.

Miciak et al., (2014) evaluated the identification rates, convergent validity, and external validity of the C/DM and XBA method with a sample of middle school students who had responded inadequately to intensive reading intervention. Results indicated that both methods identified a low percentage of students (range: 17% – 48%), dependent upon the cut point used to indicate low achievement. Agreement statistics comparing the classification decisions of the two methods were poor (kappa range: $-.04$ – $.31$) and did not represent a statistically significant improvement over chance except in cases comparing models utilizing the same low cut points for academic achievement deficits (< 1 SD below population mean). Models with lower cut points resulted in lower identification rates and high negative agreement, potentially inflating kappa which weights both positive and negative agreement equally. Additionally, comparisons of the academic performance of resulting groups on measures not utilized for classification did not reveal qualitative differences between groups, raising questions about the external validity of the methods. The authors concluded that the methods may not be interchangeable and that both methods would need to be validated independently if they are to be considered for adoption.

The Present Study

The C/DM has been cited as a “research-based” method for identifying LD (Hanson, Sharman, & Esparza-Brown, 2008) and is frequently recommended as a method for implementing the PSW model (Flanagan et al. 2010; Hale, Kaufman, Naglieri, & Kavale, 2006). Using extant data, we empirically classified a sample of at-risk second grade students to compare the LD identification decisions resulting from applying C/DM procedures to data from two comparable psychoeducational batteries, both of which had been administered to all participants. Cognitive processing measures were the same for both batteries. However, the two batteries utilized different achievement measures that were equivalent at the latent level but differed at the measurement level. Recent studies of reading skill development suggest that component reading skills (i.e. decoding, fluency, and comprehension) are

separable at the latent level (Cirino et al., 2013; Vellutino, Tunmer, Jaccard, & Chen, 2007). Well-developed indicators of the same latent reading skill tend to be highly correlated with each other and with the latent factor despite potential differences at the observed level on those indicators (Cirino et al., 2013). The present study directly investigated the effects of these observed score differences on the classification decisions of the C/DM. We asked:

1. What is the level of agreement achieved by two comparable, but different assessment batteries utilized for LD identification within the C/DM?
2. What is the level of agreement achieved by the two assessment batteries on the academic domain of eligibility for LD?

Method

Overview

The study took place with the approval of the institutional review boards of all participating universities. All data were collected as part of a larger study evaluating the efficacy of an explicit reading intervention and a guided reading intervention for improving reading outcomes for second grade students at-risk for reading difficulties (Denton, Fletcher, Taylor, Barth, & Vaughn, 2014). Pretest data from 177 second graders who had failed to respond to school-based reading interventions were utilized to empirically identify students as meeting or not meeting LD criteria using the C/DM method for implementing the PSW framework. C/DM criteria were applied with two assessment batteries that were equivalent at the latent level, but utilized different reading tests to measure academic deficits. The resulting LD identification decisions of the C/DM utilizing the two assessment batteries were then compared to evaluate agreement.

Study Context

Schools—This study took place in two school districts in the southwestern United States. Five schools were drawn from one district in a large urban area and four schools were drawn from a smaller district on the outskirts of a mid-sized city. The mean enrollment at participating schools was 660 (range: 378–886). All participating schools included a large number of students who qualified for free or reduced lunch ($M = 88.8\%$, range: 79.6% – 97.9%).

Criteria for participation—All first grade students in participating schools were screened for reading difficulties in spring of first grade ($n = 1,942$). All students who received a composite standard score less than or equal to 92 on the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999) or the Woodcock-Johnson Tests of Achievement Basic Reading Skills Cluster (WJ-III; Woodcock et al., 2001) were selected for participation. A total of 218 students met criteria. Thus, all students had received first grade instruction in school districts that implemented response to intervention frameworks and all students scored at or below a standard score of 92 on one of two nationally normed standardized measures of reading, indicating continued reading risk.

Student participants—Among the 218 students who met eligibility criteria, a total of 29 students did not complete the testing battery in second grade because they left the participating school. Five students left participating schools during summer break. Two students were removed from the study by their parents and two others were withdrawn at the request of their school due to scheduling conflicts. One additional student was removed due to special education placement and the decision of the individualized education program (IEP) team. Two students did not complete the full pretest battery and were excluded because they could not be classified. The final sample ($N = 177$) included a large number of students who qualified for free or reduced lunch (90.7%). The mean age was 7.8 years ($SD = .51$; range: 7.0–9.8). The sample was 47.5% female. The racial/ethnic composition of the sample was 61.8% African-American, 31.6% Hispanic, and 6.6% Caucasian (not Hispanic). Racial/ethnic data were unavailable for 25 students. Only 11 (8.4%) students were previously identified for special education, primarily with a speech/language impairment ($n = 7$).

Measures

The test battery was selected to incorporate measures that assess student performance across multiple domains of reading, as well as performance across cognitive domains implicated as correlates of reading performance in early elementary and as constructs associated with reading disabilities (Evans et al., 2001; Nelson, Benner, & Gonzalez, 2003). All testing was completed prior to participation in the researcher-provided intervention. Trained examiners completed all testing in a quiet area at participating students' schools over 2–3 sessions.

Academic measures

Woodcock-Johnson III Test of Achievement (WJ-III; Woodcock et al., 2001): The WJ-III is a nationally-normed, individually administered test of academic achievement. We administered two subtests assessing basic decoding skills (Letter-Word Identification and Word Attack) and one subtest assessing reading comprehension (Passage Comprehension). The three subtests demonstrate strong psychometric properties. Reliability coefficients for students ages 6–10 for Letter-Word Identification range from .93–.98, for Word Attack from .88–.94, and for Passage Comprehension from .89–.96. We utilized age-based standard scores for all analyses. The published correlation between the Letter-Word Identification and Word Attack subtests for students aged 6–8 is .80. Within our sample, the correlation for the two basic reading measures was .72. The WJ-III reading subtests correlate positively with other standardized measures of reading such as the Wechsler Individual Achievement Test (Wechsler, 2003), suggesting good validity. Basic reading tasks for the two batteries correlate at .82, while reading comprehension tasks correlate at .79.

Test of Word Reading Efficiency (TOWRE; Torgesen et al., 1999): The TOWRE is a nationally-normed, individually administered test of single-word reading fluency. We administered two subtests: Phonemic Decoding Efficiency and Sight Word Efficiency. The test-retest reliability for students ages 6–9 is .90 for Phonemic Decoding Efficiency and .97 for Sight Word Efficiency. We utilized age-based standard scores for all analyses. The published correlation between Phonemic Decoding Efficiency and Sight Word Efficiency for the entire normative sample is .78. Within our sample, the correlation between the two

fluency subtests was .71. The two TOWRE subtests are indicators of a single fluency factor and correlate highly with measures of passage fluency and word list fluency tests (Cirino et al., 2013).

Gates-MacGinitie Reading Test (MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2000): The Gates-MacGinitie is a nationally-normed, group administered test. We administered the Reading Comprehension subtest. The Kuder-Richardson Formula 20 reliability coefficient for Reading Comprehension is .92. We converted national curve equivalent scores to a standard score metric ($M = 100$, $SD = 15$) to match the metric of all other tests and used this standard score for all analyses. Within our sample, the correlation of Gates-MacGinitie Reading Comprehension with WJ-III Passage Comprehension subtest was .44.

Cognitive measures

Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen, & Rashotte, 1999): The CTOPP is a nationally-normed, individually administered test of phonological processing. We administered two subtests: Elision and Rapid Letter Naming. Test-retest reliability coefficients are .79 for Elision and .72 for Rapid Letter Naming. The Elision subtest is an indicator of phonological awareness and Rapid Letter Naming is an indicator of rapid automatized naming. Both constructs have been identified as significant correlates of reading difficulty (Barth, Catts, & Anthony, 2009; Catts, Adlof, & Weismar, 2006)

The Kaufman Brief Intelligence Test (KBIT-2; Kaufman & Kaufman, 2004): The KBIT-2 is nationally-normed, individually administered test of cognitive functioning. We administered two subtests: Verbal Knowledge and Matrix Reasoning. Test-retest reliability coefficients (corrected for restriction of range) are .88 for Verbal Knowledge and .76 for Matrix Reasoning. We utilized age-based standard scores for all analyses. Both the Matrix Reasoning and Verbal Knowledge subtests correlate with other measures of intelligence, with adjusted correlations ranging from .79–.84 for the verbal composite to other verbal measures and .47–.81 for the nonverbal composite to other perceptual reasoning measures. Additionally, deficits in verbal knowledge and vocabulary are associated with deficits in reading comprehension (Catts et al., 2006).

Operationalizing the Concordance/Discordance Model

Overview—LD identification within the C/DM requires that the student's test results demonstrate: (a) a discordance (i.e. significant difference) between an observed cognitive processing strength and an observed cognitive processing weakness, (b) a discordance between an observed cognitive processing strength and a specific observed achievement deficit, and (c) a concordance (i.e. non-significant difference) between an observed cognitive processing weakness and an observed academic deficit that are theoretically related. Thresholds for determining significant and non-significant differences between observed test scores are based on the standard error of the difference, a standardized estimate of the distribution of differences between two measures.

Specification of the model—The algorithm to establish a C/DM profile is well-defined and easily applied. However, Hale and Fiorello (2004) do not provide specific guidelines for: (a) pre-referral procedures, (b) a specific threshold for establishing academic deficits, (c) specific measures to assess academic and cognitive processing skills, (d) procedures for establishing a theoretical relation between concordant cognitive processing skills and academic deficits, and (e) the application of exclusionary criteria. In the sections that follow, we specify the procedures and criteria we followed to classify participants as meeting or not meeting C/DM LD criteria. Because of the flexibility of the procedures outlined by Hale and Fiorello, the C/DM as applied in this paper is but one instantiation. Others may operationalize it differently. However, all procedural choices are based on sound theory and empirical evidence and permit an investigation of psychometric issues that may impact classification accuracy within the C/DM, an important consideration not addressed by Hale and Fiorello.

Pre-referral procedures: All participants in the present study were enrolled in a participating school during first grade and thus participated in a tiered intervention framework. Screening for the present study took place in spring of first grade, after students had participated in a school-based response to intervention framework for nearly one year at which time all students fell below a final status test indicating reading risk (i.e. the screening test for the larger intervention study; Denton et al., 2014). This is consistent with general recommendations for pre-referral practices for the PSW framework (Feifer, 2008; Hale et al., 2010). It is important to note that pre-referral procedures were utilized to identify a potential research sample only and did not impact subsequent LD identification decisions.

Academic deficit thresholds: We chose a cut point for low achievement of a standard score less than 90 on a nationally normed achievement measure in any of the three reading related domains of eligibility: basic reading, reading fluency, and reading comprehension. A cut point of 90 is frequently utilized in studies of learning disabilities and inadequate response to intervention (e.g. Fletcher et al., 2011; Vellutino et al., 2006).

Measures selection: The C/DM does not require specific academic or cognitive measures. Instead, Hale and Fiorello (2004) emphasize the flexibility of the model, arguing that it is compatible with assessments from different batteries and even different psychometric paradigms. In the present study, we utilized nationally-normed, standardized assessments of reading and cognitive processes implicated as correlates of reading proficiency. Because the goal of the present study was to investigate the effect of measure selection on LD identification decisions, we utilized distinct academic measures to operationalize the C/DM; we selected measures that were equivalent at the latent level but differed at the level of measurement. The measures utilized in batteries 1 and 2 are provided in Table 1. Intercorrelations of the different indicators within each eligibility domain are provided in the measures description.

Theoretically linking academic and cognitive weaknesses: We sought to implement the C/DM model in a limited manner based on cognitive processes with strong theoretical and empirical ties to specific reading deficits, similar to a previous study we conducted

investigating the reliability and validity of PSW models for LD identification in middle school students (Miciak et al., 2014).

The specific measures utilized to assess each cognitive process are provided in Table 1. We assessed reading performance in the three reading related domains of eligibility for LD: (a) basic reading, (b) reading fluency, and (c) reading comprehension. Within each domain, we specified a cognitive processing skill that corresponds with the domain of reading based on empirical evidence documenting a close relationship. For example, considerable empirical evidence has linked phonological awareness and performance in basic decoding (Catts et al., 2006; Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997). Thus, to meet the C/DM requirement of theoretically linked academic and cognitive processing deficits, students with specific decoding deficits had to demonstrate corresponding deficits in phonological awareness. Similarly, students with specific reading fluency deficits had to demonstrate corresponding deficits in rapid naming because of the strong association between proficiency in fluency tasks and rapid naming tasks (Torgesen et al., 1997). Finally, students with specific deficits in reading comprehension had to demonstrate corresponding deficits in verbal knowledge because of the strong relationship between comprehension, vocabulary, and verbal knowledge (Catts et al., 2006).

Data Analysis

Agreement for LD identification—To empirically classify participants as meeting or not meeting C/DM LD criteria according to assessment batteries 1 and 2, we calculated observed score differences and compared those scores to a calculated threshold for *significant* differences. Thresholds were calculated following procedures from Hale and Fiorello (2004; p. 102) based on the standard error of the difference. LD identification was then based on the pattern of concordance and discordance, as measured by assessment batteries 1 and 2.

We evaluated agreement between the identification decisions resulting from batteries 1 and 2 with four indices of classification agreement, based on recommendations by the Food and Drug Administration (FDA) for reporting results from studies evaluating diagnostic tests (U.S. Department of Health and Human Services Food and Drug Administration, 2007).

Percent agreement: Percent agreement represents the total number of agreements (both positive and negative classifications) divided by the total number of observations. It represents general agreement and is easily interpretable. However, percent agreement does not correct for chance agreement, so it may represent an inflated estimate of agreement when comparing dichotomous classifications.

Kappa: Kappa represents the improvement over chance achieved by the two classification models for both positive and negative agreements. While kappa is preferred over percent agreement, both measures may be misleading in situations in which the portion of positive or negative classifications is low. For this reason, we also report agreement on both positive and negative decisions.

Percent positive agreement: Percent positive agreement represents agreement on positive classifications between one classification model and a non-reference standard. This index is useful in situations where there is no “true” status with which to calculate positive predictive value. In the present study, we compared assessment battery 1 to assessment battery 2, with battery 2 representing the non-reference standard. Thus, the percent positive agreement is calculated by dividing the total number of positive classification agreements by the total identified by battery 2.

Percent negative agreement: Percent negative agreement represents agreement on negative classifications between one classification model and a non-reference standard. Similar to percent positive agreement, percent negative agreement is useful when there is no “true” negative status with which to calculate negative predictive value. Percent negative agreement is calculated by dividing the total number of negative classification agreements by the total identified by assessment battery 2.

Agreement for academic domain of eligibility—To evaluate research question two, which investigates the eligibility domain in which students qualified for LD, we report a full matrix of identification decisions by single domains and combinations. Additionally, we report the percentage of overall (exact) agreement and the exact agreement for positive identification decisions.

Results

Descriptive statistics for the entire sample on all measures are provided in Table 2. Mean performance on academic measures ranged from 79.9–90.2, indicating that, on average, participants demonstrated significantly impaired reading performance across a range of reading measures. On cognitive processing measures, mean performance ranged from 83.2 to 91.2, which was roughly commensurate with performance on achievement measures.

Classification Results

For assessment battery 1 with a cut point for academic deficits of $SS < 90$, 80 (45.2%) students met LD criteria and 97 (54.8%) did not meet LD criteria. The most common reading domain for qualification was basic reading ($n = 49$; 27.6%). Fewer students met LD criteria for reading comprehension ($n = 32$; 18.1%) or reading fluency ($n = 36$; 20.3%). For battery 2 with a cut point for academic deficits of $SS < 90$, a similar number of students met LD criteria ($n = 74$; 41.8%). The most common reading area in which students met LD criteria for battery 2 was reading comprehension ($n = 39$; 22%). Fewer students met LD criteria in basic reading ($n = 28$; 15.8%) or reading fluency ($n = 21$; 11.9%).

Agreement between Batteries 1 and 2 for LD Identification

Agreement across the two batteries was low. Crosstabs for the LD classification decisions of batteries 1 and 2 are reported in Table 3. The two models agreed on 69 negative classifications and 46 positive classifications, resulting in 65% agreement. The two models achieved a kappa of .29, which is generally considered poor agreement (Cicchetti & Sparrow, 1981). Further, the percent positive agreement was low (62%), representing a

marginal improvement over chance agreement. The percent negative agreement was somewhat higher at 67%, likely reflecting the preponderance of negative classification decisions across both batteries.

Agreement between Batteries 1 and 2 for Eligibility Domain

Table 4 presents the specific academic domains of eligibility in which students were identified as having LD utilizing batteries 1 and 2. As reported above, the batteries agreed on 69 negative identification decisions. However, among the 108 students identified as having LD by at least one battery, the two batteries agreed on the exact domain of eligibility for only 25 students and disagreed for 83 students.

Discussion

Results of the present study suggest that the use of different measures within the C/DM for LD identification results in low agreement for LD identification decisions. Moreover, there was low agreement on the specific academic domain of eligibility for LD. For example, a student might be identified as primarily impaired in decoding by one assessment battery and as impaired in reading comprehension when a different battery measuring the same latent achievement constructs was utilized. This variability may have significant implications for intervention design and as a result many students may receive interventions sub-optimally matched to their academic needs.

All identification methods that rely on applying fixed cut points to a continuous distribution are inherently unstable (Francis et al., 2005; Macmann et al., 1989) because of imperfect reliability and validity at the observed level. However, this instability is exacerbated within the C/DM because it relies on the identification of a specific pattern of cognitive processing and academic achievement *difference scores* with fixed cut points for significant differences. Difference scores of positively correlated tests are always less reliable than their constituent subtests (Macmann & Barnett, 1997; McDermott et al., 1992), which introduces even greater measurement error and instability in classification decisions.

In the present study, the effect of measurement error was clearly evident. Results indicated that the selection of different academic measures had a striking effect on resulting LD identification decisions. Overall percent agreement for the two batteries was 65%, with a kappa of .29, which is generally considered poor agreement. However, a holistic index of agreement such as kappa or percent agreement likely understates the extent of the problem, because both indices weight positive and negative agreement equally. A consideration of positive agreement is more revealing because it illustrates the level of agreement on students who potentially have LD, an important educational designation. Among the 108 students identified by one of the two theoretically equivalent batteries, only 46 children were identified by both batteries. In practical terms, this means that for 62 (35%) of the participants in this study, relying on data from batteries 1 or 2 may have resulted in a different special education placement decision.

The deleterious effects of measurement error on LD identification decisions are not likely to be reduced simply by using better or repeated tests of the same construct(s). First, all of the

measures in the present study are standardized measures with good psychometric properties that are frequently used in practice and research. Second, an iterative assessment process in which specific deficits and hypotheses are tested then retested for confirmation does not eliminate the effects of psychometric error. Rather, a test-retest process shifts the effects of measurement error towards Type II errors (false negatives [not LD identifications]) to minimize Type I errors (false positives [LD identifications]). We argue that this shift would represent an undesirable tradeoff because the most critical consideration is that necessary interventions are delivered in a timely fashion, even at the expense of some Type I errors.

Finally, there was low agreement on the academic domain of eligibility between the two operationalizations of the C/DM method. The two batteries agreed on the domain of eligibility for LD for only 25 students (14% of the sample; 23% of students who were identified as LD by at least one battery). Given that these data are to be used not only to determine special education eligibility but also to plan specific treatments, such low levels of agreement are particularly troublesome. In response, the multi-disciplinary team charged with identifying LD and forming an intervention plan may decide to ignore the specific domain of eligibility and create an intervention plan independent of the eligibility process. Such actions may reduce confidence in the eligibility process. Alternately, the team could limit intervention to the specific academic domains of eligibility, which may result in students not receiving instruction in academic areas of need. Neither scenario is ideal.

Implication for practice

Messick (1995) proposed a unified theory of validity in psychological assessment that included both the interpretation and use of a specific assessment. Thus, validity represents a holistic evaluation of the meaning and consequences of a particular instrument or process. Barnett and colleagues (2006) summarized Messick's framework as addressing the *confidence* allowed by assessment data and its *helpfulness*. We apply this heuristic to structure our considerations of this study and the PSW model for school psychologists, special educators, and more generally those concerned with helping children learn to read and write.

How much confidence should we have in the classification decisions of PSW methods for LD identification? While we recognize that no single study settles a scientific controversy, it would be a disservice to interpret this study as a singular example. Instead, we suggest situating its findings as additional evidence aligned with considerable empirical research raising significant questions about the technical adequacy of cognitive discrepancy frameworks for LD identification (Fletcher et al., 2007). The findings of the present study are largely consistent with previous studies investigating the technical adequacy of methods based on an ability-achievement discrepancy (Francis et al., 2005; Macmann et al., 1989), which find instability in identification decisions related to differences in measures, testing occasion, and formulae for calculating intraindividual discrepancies. Studies of PSW methods (and specifically the C/DM) also find problems of technical adequacy (Miciak et al., 2014; Stuebing et al., 2012). Indeed, these issues extend beyond cognitive discrepancy frameworks to any classification framework based on applying strict cut points to psychometric data, including low achievement and response to intervention.

Are PSW methods for LD identification helpful? If all methods for LD identification based on strict interpretation of psychometric data fail to achieve high reliability for individual classification decisions, one might conclude that PSW methods are no better or worse than any other method. However, this conclusion fails to account for problems specific to PSW methods: the low identification rate and poor agreement for positive LD identifications. Previous studies suggest that PSW methods identify few students (Miciak et al., 2014; Stuebing et al., 2012). In the present study, less than half the sample was identified as LD in any of the eligibility domains for each battery, which is surprising because the base rate in a sample of inadequate responders should be relatively high. However, this may be related to the bias of PSW methods towards avoidance of Type 1 errors. Decisions concerning who does “not have LD” are more accurate than decisions of who actually has LD (Stuebing et al., 2012). These patterns were apparent in the present study, which also showed low positive agreement, even more pronounced at the level of the academic domain of eligibility. This disagreement on the domain of eligibility occurs because the C/DM method does not simply use a low achievement cut point. Rather, it searches for a pattern of achievement and cognitive processing. Because the intercorrelations and reliabilities of the academic and cognitive measures vary, different achievement (eligibility) domains may be identified as discrepant across different assessment batteries. As a result, entirely different intervention plans could emerge simply because of measurement error.

PSW proponents often assert that the etiology of specific academic deficits should be discovered to help design interventions tailored to the student’s individual cognitive strengths and weaknesses (Hale et al., 2010; Johnson, 2014). However, if the design, implementation, and assessment of academic interventions tailored to individual cognitive attributes were sufficiently straightforward to recommend national and international adoption of PSW methods, we should think that the many and varied attempts to validate these practices would have long ago succeeded in creating a defensible evidence base. They have not. Recent reviews fail to find compelling evidence for treatment interactions with learning style or cognitive characteristics (Pashler, McDaniel, Rohrer, Bjork, 2009; Kearns & Fuchs, 2013).

Limitations

The findings of the present study are limited to the sample and methods described. Specific attention should be paid to the manner in which we operationalized the C/DM. Other operationalizations may differ in their methodological choices, utilizing different: (a) cut points for academic deficits, (b) cognitive processing measures to assess different processing skills, and (c) theoretical links between cognitive processes and specific academic deficits. These different choices would result in different specific findings, but it is unlikely the general pattern of findings would change. Our operationalization of the C/DM is conservative and replicable. We would not expect that greater reliability for classification decisions would be achieved by introducing additional layers of complexity to the operationalization of the model.

Finally, the pre-referral and screening procedures utilized in the present study should be considered. The pre-referral procedures may differ from the recommendations for PSW

models, and the screening cut point of 92 is likely more inclusive than would be utilized in school practice. However, it is important to note that these procedures impact the selection of the participant sample only. Although different pre-referral and screening procedures may change the nature and number of participants, it is unlikely that these changes would affect the pattern of results, which relate to the psychometric limitations of cognitive discrepancy models.

Conclusions

In the present study, we found that the utilization of psychoeducational batteries that differed in their academic achievement tests resulted in high levels of instability in the identification of individual students, particularly for positive identifications of LD. Such instability in individual identification decisions is inherent to any identification method based on strict interpretations of psychometric data but is exacerbated when multiple measures are utilized to calculate difference scores and identify a specific pattern of test scores. The use of multiple measures could reduce this instability, but would likely increase Type II errors to minimize Type I errors, a tradeoff that may deprive students of necessary academic instruction. In the absence of compelling research documenting improved academic outcomes through cognitive assessment, the findings of the present study suggest that PSW methods are at best premature for widespread adoption. Limited resources may be better utilized directly assessing academic skills of interest to identify students in need of additional help.

Acknowledgments

This research was supported by grant P50 HD052117, Texas Center for Learning Disabilities, from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development or the National Institutes of Health.

References

- Denton CA, Fletcher JM, Taylor WP, Barth AE, Vaughn S. An experimental evaluation of guided reading and explicit intervention for primary-grade students at-risk for reading difficulties. *Journal of Research on Educational Effectiveness*. 2014; 7:268–293.
- 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; 29:21–37. [PubMed: 24274155]
- Barnett DW, Elliott N, Graden J, Ihlo T, Macmann G, Nantais M, Prasse D. Technical adequacy for response to intervention practices. *Assessment for Effective Intervention*. 2006; 32(1):20–31.
- Barth AE, Catts HW, Anthony JL. The component skills underlying reading fluency in adolescent readers: A latent variable analysis. *Reading and Writing*. 2009; 22:567–590.
- Catts HW, Adlof SM, Weismar SE. Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, & Hearing Research*. 2006; 49:278–293.
- Cicchetti DV, Sparrow SS. Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American journal of Mental Deficiency*. 1981; 86:127–137. [PubMed: 7315877]
- Cirino PT, Romain MA, Barth AE, Tolar TD, Fletcher JM, Vaughn S. Reading skill components and impairments in middle school struggling readers. *Reading and writing*. 2013; 26(7):1059–1086. [PubMed: 24000271]

- 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*. 2002; 31>:246–262.
- Feifer SG. Integrating response to intervention (RTI) with neuropsychology: A scientific approach to reading. *Psychology in the Schools*. 2008; 45(9):812–825.
- 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.
- Flanagan, D.; Ortiz, S.; Alfonso, VC., editors. *Essentials of cross battery assessment*. 2nd ed.. Hoboken, NJ: John Wiley & Sons, Inc.; 2007.
- 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, Vaughn S. Cognitive correlates of inadequate response to reading intervention. *School Psychology Review*. 2011; 40:3–22. [PubMed: 23125475]
- Fiorello CA, Hale JB, Snyder LE. Cognitive hypothesis testing and response to intervention for children with reading problems. *Psychology in the Schools*. 2006; 43:835–853.
- 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. [PubMed: 15813593]
- Hale JB, Alfonso V, Berninger B, Bracken B, Christo C, Clark E, 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, Kaufman A, Naglieri JA, Kavale KA. Implementation of IDEA: Integrating response to intervention and cognitive assessment models. *Psychology in the Schools*. 2006; 43(7):753–770.
- Hanson J, Sharman MS, Esparza-Brown J. Pattern of strengths and weaknesses in specific learning disabilities: What’s it all about? Technical Assistance Paper. Oregon School Psychologists Association. 2008 <http://www.jamesbrenthanson.com/uploads/PSWCondensed121408.pdf>.
- Hauerwas LB, Brown R, Scott AN. Specific learning disability and response to intervention: State-level guidance. *Exceptional Children*. 2013; 80:101–120.
- 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.
- Kaufman, AS.; Kaufman, NL. *Kaufman Brief Intelligence Test*. 2nd ed.. Minneapolis, MN: Pearson Assessment; 2004.
- Kearns DM, Fuchs D. Does cognitively focused instruction improve the academic performance of low-achieving students? *Exceptional Children*. 2013; 79:263–290.
- MacGinitie, W.; MacGinitie, R.; Maria, K.; Dreyer, L.; Hughes, K. *Gates-MacGinitie Reading Tests-4*. 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.
- Macmann GM, Barnett DW, Lombard TJ, Belton-Kocher E, Sharpe MN. On the actuarial classification of children: Fundamental studies of classification agreement. *Journal of Special Education*. 1989; 23:127–149.
- McDermott PA, Fantuzzo JW, Glutting JJ, Watkins MW, Baggaley AR. Illusions of meaning in the ipsative assessment of children’s ability. *The Journal of Special Education*. 1992; 25:504–526.

- 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.
- 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.
- Naglieri, JA. Essentials of CAS assessment. New York, NY: John Wiley & Sons, Inc.; 1999.
- Nelson RJ, Benner GJ, Gonzalez J. Learner characteristics that influence the treatment effectiveness of early literacy interventions: A meta-analytic review. *Learning Disabilities Research & Practice*. 2003; 18(4):255–267.
- Pashler H, McDaniel M, Rohrer D, Bjork R. Learning styles: Concepts and evidence. *Psychological Science in the Public Interest*. 2009; 9(3):105–119. [PubMed: 26162104]
- Shaywitz, S. Overcoming dyslexia: A new and complete science-based program for reading problems at any level. New York: Alfred A. Knopf; 2003.
- 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(1):3–22. [PubMed: 23060685]
- Torgesen, J.; Wagner, R.; Rashotte, C. Test of Word Reading Efficiency. Austin, TX: Pro-Ed.; 1999.
- Torgesen JK, Wagner RK, Rashotte CA, Burgess S, Hecht S. Contributions of phonological awareness and rapid automatic naming ability to the growth of word-reading skills in second- to fifth-grade children. *Scientific Studies of Reading*. 1997; 1:161–185.
- U. S. Department of Health and Human Services: Food and Drug Administration. [Retrieved on May 1, 2014] Statistical guidance on reporting results from studies evaluation diagnostic tests. 2007. from <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm>
- Vellutino FR, Scanlon DM, Small S, Fanuele DP. Response to intervention as a vehicle for distinguishing between children with and without reading disabilities: Evidence for the role of kindergarten and first-grade interventions. *Journal of Learning Disabilities*. 2006; 39:157–169. [PubMed: 16583795]
- Vellutino FR, Tunmer WE, Jaccard JJ, Chen R. Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading*. 2007; 11:3–32.
- Wagner, RK.; Torgesen, JK.; Rashotte, CA. Comprehensive Test of Phonological Processing. Austin, TX: PRO-ED Inc.; 1999.
- Watkins MW. Cognitive profile analysis: A shared professional myth. *School Psychology Quarterly*. 2000; 15(4):465–479.
- Watkins MW, Canivez GL. Temporal stability of WISC-III subtest composite: strengths and weaknesses. *Psychological Assessment*. 2004; 16(2):133–138. [PubMed: 15222809]
- Wechsler, D. Wechsler Intelligence Scale for Children-WISC-IV. San Antonio, TX: Psychological Corporation; 2003.
- Woodcock, RW.; McGrew, KS.; Mather, N. Woodcock-Johnson III Tests of Achievement. Itasca, IL: Riverside; 2001.

Table 1

Academic and cognitive processing measures for C/DM assessment batteries 1 and 2

Reading Domain	Academic Measures		Related Cognitive Processing Measures ¹
	Assessment Battery 1	Assessment Battery 2	
Basic Reading	WJ-III Letter/Word ID	WJ-III Word Attack	CTOPP Phonological Awareness
Reading Fluency	TOWRE Phonemic Decoding	TOWRE Sight Words	CTOPP Rapid Letter Naming
Reading Comprehension	WJ-III Passage Comprehension	Gates MacGinitie Reading Comprehension	KBIT-2 Verbal Knowledge

¹ Related cognitive processing measures do not vary across batteries 1 and 2. WJ-III = Woodcock Johnson Tests of Achievement-Third Edition; TOWRE = Test of Word Reading Efficiency; Phonemic Decoding = Phonemic Decoding Efficiency; Sight Words = Sight Word Efficiency; CTOPP = Comprehensive Test of Phonological Processing; KBIT-2 = Kaufman Brief Intelligence Test Second Edition.

Table 2

Descriptive statistics for all academic and cognitive processing measures

Variable	<i>M</i>	<i>SD</i>
<i>Academic measures</i>		
WJ-III Letter-Word Identification	85.94	10.16
WJ-III Word Attack	90.21	11.67
TOWRE Sight Word Efficiency	79.94	10.14
TOWRE Phonemic Decoding Efficiency	82.98	8.21
WJ-III Passage Comprehension	81.77	9.82
Gates MacGinitie Reading Comprehension	84.51	11.10
<i>Cognitive processing Measures</i>		
KBIT-2 Verbal Knowledge	83.15	16.44
KBIT-2 Nonverbal Reasoning	85.44	13.86
CTOPP Rapid Letter Naming	91.24	8.82
CTOPP Elision	83.28	12.57

N = 177; WJ-III = Woodcock Johnson Tests of Achievement-Third Edition; TOWRE = Test of Word Reading Efficiency; KBIT-2 = Kaufman Brief Intelligence Test-Second Edition; CTOPP = Comprehensive Test of Phonological Processing.

Table 3

Agreement of learning disability classifications based on two assessment batteries

Battery 2	Battery 1		Total
	Non LD	LD	
Non LD	69	34	103
LD	28	46	74
Total	97	80	177

Kappa = .28; percent agreement = 65%; percent positive agreement = 62%; percent negative agreement = 67%; LD = Learning Disability.

Note: Academic battery 1 includes Woodcock-Johnson Tests-III (WJ-III) Letter-Word Identification, TOWRE Phonemic Decoding Efficiency, and WJ-III Passage Comprehension; Battery 2 includes WJ-III Word Attack, TOWRE Sight Word Efficiency, and Gates MacGinitie Reading Comprehension

Table 4.
Comparison of academic area of learning disability eligibility identified with batteries 1 and 2

Academic Area of Learning Disability Eligibility for Battery 2												
	NO	LD	BR	RF	RC	BR & RF	BR & RC	RF & RC	ALL AREAS	Total		
Academic Area of Learning Disability Eligibility for Battery 1												
NOLD	N	69	3	4	19	.	1	1	.	97		
BR	N	8	9	.	2	2	1	.	.	22		
RF	N	7	.	8	.	1	.	.	1	17		
RC	N	4	.	.	5	.	.	.	.	9		
BR & RF	N	7	.	.	.	.	1	1	.	9		
BR & RC	N	3	8	.	1	.	1	.	.	13		
RF & RC	N	3	.	.	.	.	.	2	.	5		
ALL AREAS	N	2	.	1	1	.	1	.	.	5		
Total	N	103	20	13	28	3	5	4	1	177		

Note: Bold cells indicate exact agreement between battery 1 and battery 2.

BR = basic reading; RF = reading fluency; RC = reading comprehension; NO LD = student not identified with LD. Academic battery 1: BR = Woodcock-Johnson-III (WJ-III) Letter-Word Identification, RF = TOWRE Phonemic Decoding Efficiency, RC = WJ-III Passage Comprehension. Academic battery 2: BR = WJ-III Word Attack, RF = TOWRE Sight Word Efficiency, RC = Gates MacGinitie Reading Comprehension.

N = 177; 94/177 (53%) = exact agreement. 25/108 (23%) = exact agreement for positive classifications only.