Published in final edited form as:

J Psychoeduc Assess. 2018 February ; 36(1): 21–33. doi:10.1177/0734282916683287.

Simulation of LD Identification Accuracy Using a Pattern of Processing Strengths and Weaknesses Method With Multiple Measures

Jeremy Miciak¹, W. Pat Taylor¹, Karla K. Stuebing¹, and Jack M. Fletcher¹

¹University of Houston, TX, USA

Abstract

We investigated the classification accuracy of learning disability (LD) identification methods premised on the identification of an intraindividual pattern of processing strengths and weaknesses (PSW) method using multiple indicators for all latent constructs. Known LD status was derived from latent scores; values at the observed level identified LD status for individual cases according to the concordance/discordance method. Agreement with latent status was evaluated using (a) a single indicator, (b) two indicators as part of a test–retest “confirmation” model, and (c) a mean score. Specificity and negative predictive value (NPV) were generally high for single indicators (median specificity = 98.8%, range = 93.4%–99.7%; median NPV = 94.2%, range = 85.6%–98.7%), but low for sensitivity (median sensitivity = 49.1%, range = 20.3%–77.1%) and positive predictive value (PPV; median PPV = 48.8%, range = 23.5%–69.6%). A test–retest procedure produced inconsistent and small improvements in classification accuracy, primarily in “not LD” decisions. Use of a mean score produced small improvements in classifications (mean improvement = 2.0%, range = 0.3%–2.8%). The modest gains in agreement do not justify the additional testing burdens associated with incorporating multiple tests of all constructs.

Keywords

learning disabilities; cognitive testing; assessment; classification accuracy

The topic for this special issue of the *Journal of Psychoeducational Assessment* is how data simulation can inform decision making from psychoeducational assessments. Data simulation techniques are particularly valuable for this purpose. Multivariate data simulations allow for the generation of thousands of samples exhibiting specified distributions and relations between variables. These samples allow researchers to manipulate a wide range of parameters and evaluate the results—manipulations which would be nearly impossible using “real” data.

Corresponding Author: Jeremy Miciak, Department of Psychology, University of Houston, 4849 Calhoun Rd., Ste 373, Houston, TX 77204-6022, USA., Jeremy.Miciak@times.uh.edu.

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 & Human Development or the National Institutes of Health.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Learning Disabilities (LDs) Identification: Current Controversies

Fletcher (2012) identified two competing conceptual frameworks for understanding the nature of “unexpected underachievement” in LDs that are relevant in current controversies about identification: (a) instructional frameworks and (b) cognitive discrepancy frameworks. Instructional frameworks hold that LDs are marked by low achievement, often adding inadequate instructional response as a marker of unexpectedness, in addition to a consideration of exclusionary factors. In contrast, cognitive discrepancy frameworks hold that LDs are marked by academic difficulties that are unexpected based on the presence of normal cognitive functioning or significant processing strengths, in addition to a consideration of exclusionary factors and the adequacy of instruction (Hale et al., 2010; Kavale & Forness, 2000). Since the passage of Public law 94–142 and the adoption of IQ-achievement discrepancy in the 1977 regulations that made LDs a category for special education services, cognitive discrepancy frameworks have been the predominant method to identify children with LDs. However, decades of research has demonstrated the poor reliability and validity of IQ-achievement discrepancy as an inclusionary criteria for LDs (Stuebing et al., 2002); few today advocate for the use of IQ-achievement discrepancies. Instead, a new cognitive discrepancy framework has been proposed, which we (and others) refer to as processing strengths and weaknesses (PSW) methods. These PSW methods assert that an intraindividual pattern of PSW is a necessary inclusionary criterion for the identification of LDs (Flanagan, Ortiz, & Alfonso, 2013; Hale & Fiorello, 2004).

Multiple methods to operationalize PSW methods have been proposed, including the concordance/ discordance method (C/DM; Hale & Fiorello, 2004), the cross-battery assessment method or dual discrepancy/consistency approach (XBA; Flanagan et al., 2013), and the discrepancy/consistency method (D/CM; Naglieri, 2011). Although presented as conceptually equivalent, these PSW methods differ in important ways, including the hypothesized relations and structure of cognitive functioning, the role of norms and benchmarks for decision making, and the specific methods utilized to identify an intraindividual pattern of PSW (Stuebing, Fletcher, Branum-Martin, & Francis, 2012).

PSW methods have gained increasing prominence in special education and school psychology in recent years. In 2015, more than half of all states permitted PSW methods for LD identification at some level; 14 states explicitly allowed for the use of PSW methods, and another 12 did not prohibit the methods (Maki, Floyd, & Roberson, 2015). In 2010, the Learning Disabilities Association of America (LDA) issued a white paper summarizing what it characterized as an expert consensus on the definition and identification of LDs. The report concluded that PSW methods for LD identification “make the most clinical and empirical sense” (Hale et al., 2010, p. 225). One rationale for a move away from RTI methods toward PSW methods is that there “is no true positive in a response to intervention (RTI) model, meaning that all children who fail to respond to quality instruction and intervention are considered specific learning disability (SLD) by default” (Hale et al., 2010, p. 226). Without a “true positive,” the paper continues, there is no way to determine the sensitivity and specificity of different methods, which may be a reason for poor reliability. It is true that RTI methods do not exhibit a “true positive” for LD; no such thing exists for any

proposed method, and identification is always based on observed measurements of the indicators of LD.

Reliability and the “Gold Standard” Problem

Across definitional and conceptual variations, LD is a latent construct (Fletcher, 2012). It is unobservable outside of attempts to measure it using imperfect psychoeducational measures. There is no blood test or brain scan available that would differentiate which students truly demonstrate LDs from those who do not. This is true not just of LD, but also of many psychological and medical conditions with heterogeneous etiologies and that are syndromically defined. Thus, definitions of LD and methods to operationalize these definitions must be understood as testable hypotheses. Criteria can be applied across different measurement occasions, samples, and test batteries, and resulting groups can be compared for reliability and validity, the latter evaluation only possible when data on functioning in external dimensions not utilized in group formation are available (Morris & Fletcher, 1988).

Empirical Research on the Reliability of PSW Methods

Previous empirical and simulated research has raised questions about the reliability of LD identification decisions emerging from PSW methods. Miciak, Fletcher, Stuebing, Vaughn, and Tolar (2014) evaluated agreement across two different methods to operationalize PSW methods: the C/DM (Hale & Fiorello, 2004) and the XBA approach (Flanagan, Ortiz, & Alfonso, 2013). Psychoeducational data from 139 students in middle school who demonstrated inadequate response to an intensive intervention were utilized to empirically classify students as either meeting or not meeting LD identification criteria according to the proposed methods, and decisions were compared. Across the two methods, agreement for individual decisions did not exceed that which would be expected via chance, with κ ranging from $-.04$ to $.31$, dependent upon the cutoff point for indexing low achievement.

These results were substantially replicated in a nonoverlapping study conducted with 203 students enrolled in fourth grade (Miciak, Taylor, Denton, & Fletcher, 2015). Psychoeducational data permitted an empirical classification of students as either meeting or not meeting LD identification criteria according to the C/DM and XBA methods. Paralleling results of the previous study, agreement for LD identification decisions did not exceed that which would be expected via chance ($\kappa = -.10$).

Miciak, Williams, Taylor, Cirino, and Fletcher (2016) utilized a third, nonoverlapping sample to investigate whether test selection reduced the reliability of LD identification decisions within the C/DM. Psychoeducational data from 139 second-grade students were utilized to empirically classify students using two batteries that were equivalent at the latent level, but differed in the selection of reading measures. The results of identification decisions emerging from these two batteries demonstrated modest agreement for which students met and which students did not meet LD criteria ($\kappa = .28$). These results suggest that LD identification decisions within the C/ DM are not robust to changes in test selection.

Kranzler et al. (2016a) utilized Woodcock–Johnson norming data to evaluate the assumption that PSW profiles, as indicated by XBA criteria (Flanagan et al., 2013), are causally related to academic deficits. Participants who demonstrated average cognitive functioning and a specific cognitive deficit in areas thought to be meaningfully related to academic achievement were identified as demonstrating a PSW profile, whereas students who did not meet these criteria were identified as not demonstrating a PSW profile. The authors then evaluated classification agreement between PSW profile (based on cognitive variables only) and low academic achievement. Consistent with the Stuebing et al. (2012) simulation, the XBA method identified a low number of children as LD; the method was very reliable in detecting true negatives and demonstrated high negative predictive validity. However, positive predictive value (PPV) and sensitivity were low. This suggests that many students who demonstrate academic deficits infrequently demonstrate a PSW profile. It also raises questions about the extent to which low academic achievement is caused by intraindividual PSW patterns.

Simulation Studies of PSW Methods

Stuebing et al. (2012) simulated data to investigate the C/DM, XBA, and D/CM methods for LD identification. The authors first identified latent profiles of cognitive processing and academic achievement and classified students as having met or not met LD criteria according to premises of PSW methods. Case-level observed data were then generated using published reliabilities for common psychoeducational assessments. LD status at the latent level (the “gold standard”) permitted an evaluation of reliability at the observed level—a significant advantage of data simulation. Results indicated that only a small percentage of all students (1%–2%) met LD identification criteria under PSW methods. Furthermore, all three approaches demonstrated high specificity and high negative predictive value (NPV) due to the low base rate of the model. However, all three approaches demonstrated moderate to low sensitivity and low PPV, suggesting that even if “true” patterns of strengths and weaknesses exist and are meaningful, classifications at the observed level will not reliably identify those students.

Taylor, Miciak, Fletcher, and Francis (2016) investigated the reliability of identification decisions emerging from the C/DM in a simulated replication of Miciak et al. (2015). The simulation compared the identification decisions resulting from an application of C/DM criteria using two theoretically equivalent assessment batteries that differed only in the selection of academic measures. In Study 1, the authors simulated cognitive and academic data across a broad range of potential relations between an academic weakness, a concordant cognitive weakness, and a discordant cognitive strength. Across the broad parameterization, overall agreement was high, reflecting low identification rates. However, percent positive agreement was low to moderate, ranging from 0.33 to 0.59 across all scenarios. These results suggest that within complex PSW methods that rely upon specific patterns of difference scores between subtests (like the C/DM), changes in test selection may have negative effects on agreement.

The Present Study

Previous investigations of the reliability of LD identification decisions emerging from PSW methods have evaluated the effect of test selection (Miciak et al., 2014; Taylor et al., 2016), method of operationalization (Miciak et al., 2014), as well as unreliability from single observed measures tapping a latent ability (Stuebing et al., 2012). However, no previous simulation or empirical study has investigated a key assumption embedded within directions for implementation of PSW methods that the collection of data from multiple subtests or measures will improve classification accuracy when implemented as part of a recursive assessment process. Proponents of PSW methods have asserted that previous studies investigating the reliability of PSW methods are flawed because “single measures are generally not sufficiently reliable ..., nor do they have adequate content validity. That is, they do not cover the broad range of abilities implied by the broad constructs in CHC theory” (Flanagan & Schneider, 2016, p. 140).

In the present study, we directly investigated this assertion to determine the extent to which the incorporation of multiple measures of a single latent construct improves classification accuracy. Classification accuracy was evaluated by comparing agreement between latent status and status emerging from observed data in three scenarios: (a) identification decisions from a single indicator for each latent construct; (b) identification decisions emerging from a recursive process in which a positive classification at Time 1 is confirmed at Time 2 with a second battery of measures of the same latent construct, often referred to as cognitive hypothesis testing (Hale & Fiorello, 2004); and (c) the use of a mean or composite score to classify individual cases, representing the use of two indicators of a latent construct recommended by other PSW proponents (Flanagan & Schneider, 2016).

Method

The C/DM and Intraindividual Discrepancies

The C/DM represents a prominent method to operationalize PSW methods for LD identification that is often characterized as evidence-based (Hale et al., 2010; Hanson, Sharman, & Esparza-Brown, 2009). The C/DM includes three inclusionary criteria, in addition to a consideration of exclusionary clauses and instructional adequacy. The C/DM is agnostic with regard to theoretical orientation and can be applied across different measures (Hale & Fiorello, 2004). Inclusionary criteria depend upon the magnitude of difference scores between three variables: (a) an academic weakness, (b) a specific cognitive weakness, and (c) a cognitive strength. The three resulting difference scores are compared against a threshold for significant differences, which is calculated based on the standard error of the difference, a standardized estimate of the distribution of differences between two measures. To be identified with LD, the assessed student must demonstrate a three-part pattern of academic and PSW, including (a) a nonsignificant difference between an academic weakness and a theoretically related cognitive weakness, (b) a significant difference between an academic weakness and a cognitive strength, and (c) a significant difference between a cognitive weakness and a cognitive strength. Within the simulation, students who met these three inclusionary criteria based on their observed scores were identified as LD, and those who did not were identified as not LD.

Simulation Procedures

We used SAS 9.4 (SAS Institute, 2013) for all simulations and analyses. The simulations for the current study were designed to evaluate the assumption that multiple testing occasions will improve the accuracy of PSW decisions. Following Stuebing et al. (2012), we assumed that the observed test scores could be represented by a true score model where each observed score was the sum of true scores and random errors, that the true scores would follow a multivariate normal distribution, and that the published correlations of observed measures would adequately serve as the population correlations.

Unlike Stuebing et al. (2012), we manipulated the reliabilities of the observed scores. The values we chose for reliabilities ranged from .75 to .95, incrementing by .05. PSW methods are premised on the assumption that meaningful differences in intraindividual cognitive processing are indicative of LD. However, none of the methods provides guidance on the exact magnitude of a meaningful difference at the latent level. The values we chose for a meaningful difference at the latent level were differences greater than 0.5 *SD*, 1.0 *SD*, and 1.25 *SD*. We fully crossed these two factors to produce $5 \times 3 = 15$ scenarios.

As in Stuebing et al. (2012), we simulated scores for the Wechsler Intelligence Scale for Children - Fourth Edition (WISC-IV; Wechsler, 2003) Verbal Comprehension Index (VCI) and Perceptual Reasoning Index (PRI), Full Scale IQ, and the Wechsler Individual Achievement Test reading composite (WIAT-II R; Wechsler, 2001). The population correlations for the true scores were calculated by disattenuating the correlations of observed measures using the published reliabilities for those measures as shown in Equation 1.

$$\rho_{xy} = \frac{r_{xy}}{(\sqrt{\text{reliability}_x} \times \sqrt{\text{reliability}_y})} \quad (1)$$

With this correlation matrix for the true scores, we then used SAS Proc Simnormal to simulate 1,000,000 observations in *z*-score metric for each of the latent variables. We then simulated another 1,000,000 unrelated observations in *z*-score metric for eight variables to serve as random errors for each of the four observed variables at two time points. All tests within a scenario had the same reliability. Observed values were created by adding the product of the simulated latent score and the square root of the reliability to the product of the simulated random error and the square root of one minus the reliability as shown in Equation 2. As the simulated observed scores were in a *z*-score metric, we applied a linear transformation to achieve means of 100, and standard deviations of 15. Values were then rounded to whole numbers, and extreme values were truncated to match published ranges.

$$Y_{ijk} = t_{ij} \times \sqrt{\text{reliability}_k} + e_{ij} \times \sqrt{\text{reliability}_k} \quad (2)$$

where Y_{ijk} is the observed score for observation i on test j with reliability k , t_{ij} is the true score for observation i on test j , and e_{ij} is random error for observation i on test j .

Classification Procedures

Classification of observations as either LD or not LD was performed for the latent scores following the classification procedures of the C/DM (Hale & Fiorello, 2004) using the three levels of latent difference criteria. The following rules were used to establish LD status: (a) VCI – PRI was negative and exceeded the difference criteria, (b) the absolute value of VCI – WIAT-II R was smaller in absolute magnitude than the difference criteria, and (c) PRI – WIAT-II R was positive and exceeded the difference criteria. This process resulted in a “true” LD status for each of the three latent difference criteria. The observed scores were then used to determine observed LD status in a similar fashion. However, the criteria for the meaningful differences between observed scores were calculated using the reliability that was used to generate the observed score. The formula to calculate the criteria for observed scores is shown in Equation 3, as recommended by Hale and Fiorello (2004).

$$\text{Threshold} = 15 \times \sqrt{1 - \text{reliability}_{\text{Test1}} - \text{reliability}_{\text{Test2}}} \times 1.96. \quad 3$$

This process was applied to observed scores to determine LD status at Time Point 1 and Time Point 2 and to the average of scores from both time points. A fourth LD status based on observed scores was created using the LD status from Time Points 1 and 2. This joint LD status was set to LD if the observation had a positive LD status at both time points and a negative LD status otherwise.

Results

Preliminary Classification Accuracy

To evaluate the quality of the data simulation, we compared the correlations of the simulated latent scores with the desired correlations based on the published correlations and reliabilities. Differences between specified and obtained correlations ranged from $-.00037$ to $.00029$ with a mean of $-.00004$ and a standard deviation of $.00027$. These results indicated that the simulated data provided an adequate representation of the population of interest. To determine the referral sample, we limited the data to be analyzed to those observations where WIAT-II R < 90 and FSIQ > 69.

Classification accuracy with a single time point was calculated using Proc Freq in SAS. Crosstabs were constructed using the referral sample for observed LD status by latent LD status for each level of reliability and latent criteria. Results from these crosstabs are presented in Table 1. Specificity ($M = 93.65$, $SD = 2.76$) and NPV ($M = 93.85$, $SD = 4.72$) were both high but sensitivity ($M = 52.02$, $SD = 18.81$) and PPV ($M = 46.04$, $SD = 13.14$) were both considerably lower. This means that a negative outcome for LD on a single observed measure is highly accurate whereas a positive outcome is not as accurate.

Recursive Assessment Classification Accuracy

To provide an assessment that more closely reflects recommended practice, we determined LD status using results from two separate testing sessions. For this assessment, LD status was positive if LD status for each of the testing sessions was positive; otherwise, LD status

was negative. Results from this assessment are shown in Table 2. NPV was generally lower in the recursive testing scenarios ($M = 92.06$, $SD = 5.28$), but specificity was higher ($M = 98.29$, $SD = 1.71$). Conversely, sensitivity ($M = 31.32$, $SD = 19.72$) was lower relative to the single testing scenarios whereas PPV ($M = 66.94$, $SD = 12.79$) showed improvements under the recursive assessment scenarios.

Average Score Classification Accuracy

As a final evaluation, we determined LD designation based on the average of the scores from both testing occasions. For this assessment, LD status was positive if the differences between the average scores satisfied the patterns specified above. Results from this assessment are shown in Table 3. Generally, results indicate improved classification accuracy when utilizing a mean score. Similarly, specificity ($M = 95.41$, $SD = 3.45$), NPV ($M = 94.14$, $SD = 4.99$), sensitivity ($M = 54.35$, $SD = 23.48$), and PPV ($M = 57.23$, $SD = 14.53$) all showed increases when based on the averages of test scores. The improvement in the number of correctly classified cases ranged from 569 to 6,067 ($M = 3,849$, $SD = 1,728$).

Discussion

The present study investigated whether the incorporation of multiple measures of a single latent construct improved the accuracy of binary decisions emerging from one prominent proposal to operationalize PSW methods: the C/DM (Hale & Fiorello, 2004). We simulated a latent data set exhibiting a normal distribution and specified relations between variables and classified individual cases as meeting or not meeting LD criteria according to the premises of the C/DM. We then generated two observed scores for each latent case at five levels of unreliability. Data at the observed level were utilized to classify individual cases as meeting or not meeting C/DM LD identification criteria in three scenarios, based on (a) a single observation of each variable; (b) two observations of each variable, implemented in a recursive test–retest in which a potential LD identification decision is confirmed with a second assessment battery; and (c) a mean score from two observations of each latent variable.

Classification accuracy for positive LD identification decisions based on a single indicator of each latent construct was generally low. Consistent with previous simulations of PSW methods and the C/DM specifically (Stuebing et al., 2012; Taylor et al., 2016), the method did better at identifying cases as “not LD” (specificity range = 86.5%–96.2%; NPV range = 86.9%–98.7%), but accuracy for positive LD identification decisions was lower (sensitivity range = 20.3%–77.1%; PPV range = 30.8%–69.6%). In even the most favorable scenarios, approximately one in four students who truly has LD would not be identified by these methods, and three in 10 students identified as LD would not truly demonstrate LD. For most scenarios, classification accuracy was worse; median PPV across the 15 scenarios was 53.52%, indicating that half of all students who are truly LD would not be identified by the proposed methods at the observed level. These results are consistent with previous research (Miciak et al., 2015; Stuebing et al., 2012; Taylor et al., 2016) suggesting that identification decisions emerging from the C/DM with a single indicator are unreliable.

The inclusion of additional measures as part of a recursive test–retest procedure did not uniformly improve classification accuracy. In 12 of the 15 scenarios, the total number of correct classifications increased. However, in each of these 12 scenarios, the total number of true positive classifications decreased. Thus, observed improvements in total classification accuracy and true negative classifications must be balanced against declines in sensitivity and PPV. Across all scenarios, median PPV was reduced to 28.2%. In scenarios in which the largest increase in correct classifications was observed, PPV was 34.0%, indicating that only one in three cases identified as LD at the observed level was truly LD. Such trade-offs are observed across all scenarios in which there was an improvement in the number of cases classified correctly.

The utilization of a mean score of two measures increased the number of correct classifications in all scenarios. The difference in cases classified correctly in comparison with a single indicator ranged from 569 to 6,067 (out of approximately 220,000 referred cases). Agreement statistics generally improved, and the methods generally demonstrated adequate specificity and NPV. Agreement for positive identification decisions was generally lower (sensitivity range = 20.1–86.12). Median sensitivity improved, but still indicated that a large portion of truly SLD cases would not be identified by the C/DM using mean scores (Median sensitivity = 57.34%).

Improvement at What Cost?

Proponents of PSW methods have asserted that previous investigations of the reliability of PSW methods are flawed because they have relied on single indicators of a construct, which cannot reliably and fully measure the latent construct (Flanagan & Schneider, 2016) and assert that reliability would be improved through the incorporation of additional measures of each construct. The results of this simulation indicate that these assertions result in modest improvements in accuracy when using a mean score, but not a test–retest procedure.

However, these modest improvements in classification accuracy must be interpreted in a context of finite resources. The use of mean scores for identification decisions resulted in a 3% increase in correct classifications in the most favorable scenario. In the very best scenario, using the mean score improved the classification of 6,067 of 223,590 cases in the “referred sample.” To improve the classification of 6,067 kids, the equivalent of another 223,590 needed to be tested. Stated differently, to increase the number of correct classifications by 3%, the total amount of testing would double (an increase of 100%). The length and complexity of assessment batteries recommended by PSW proponents (Flanagan et al., 2013) require significant resource expenditures, which is why the recommended dual test procedure deviates from common practice in schools (Kranzler et al., 2016a). Furthermore, among the 6,067 corrections observed in this most favorable scenario, only 299 decisions (4.9%) were true positives. This pattern holds across most scenarios; however, in only three scenarios, more than half of the increase in correct classifications was accounted for by a change in true positives. This finding highlights a fundamental issue for PSW methods: As latent LD rates are generally low, increased classification accuracy will be driven primarily by additional true negatives. In most scenarios, the use of twice as many measures to derive a mean score increased the number of true positives by less than 1% to

2% over a single indicator and still exhibited a nontrivial number of classification errors. For many schools, the testing burden associated with PSW methods would represent significant resource allocation and shift resources from intervention.

Limitations

Data simulation does not permit a comprehensive application of exclusionary criteria, such as other neurosensory disorders, second language acquisition, inadequate instructional opportunity, or other factors that might explain low academic achievement. We excluded students with an $IQ < 70$, as these scores might be indicative of an intellectual deficit. The goal of this data simulation was to evaluate the technical adequacy of the proposed methods and answer questions related to the incorporation of multiple measures. Data simulation cannot evaluate a fully implemented, recursive assessment process that includes instructional activities, progress monitoring, and other adaptive elements. Nor can this data simulation account for permitting the application of clinical judgment. Only one PSW method was evaluated, although the results are consistent with evaluations of other methods and represent general limitations of methods in which the measured attributes are correlated, have small degrees of unreliability, and apply bright thresholds to dimensional constructs. Methods based on other approaches emanating from instructional methods demonstrate similar problems of low agreement in empirical (Barth et al., 2008; Brown Waesche, Schatschneider, Maner, Ahmed, & Wagner, 2011; Burns, Scholin, Kosciulek, & Livingston, 2010) and simulated data (Fletcher et al., 2014). These problems are universal for psychometric approaches to LD identification. At the very least, a range of scores that incorporate measurement error (e.g., confidence intervals) should be used in decision making.

Conclusion

The burden for evidence for overcoming the inherent psychometric limitations of PSW methods should fall upon those advocating their use (Kranzler et al., 2016b). However, little empirical, peer-reviewed research from advocates is available, which is why simulations are so important.

As expected, applying the C/DM with single indicators demonstrated generally low agreement with latent status. The incorporation of an additional indicator to calculate a mean score, but not as a test–retest procedure, did modestly improve classification accuracy. However, these modest improvements in number of cases correctly classified (0.3%–2.9%) must be balanced against a 100% increase in testing time. Finite resources would be better spent directly assessing key academic skills and providing targeted interventions for students in need of assistance rather than administering more cognitive tests. Inherited wisdom instructs us there is no education in the second kick of a mule. In psychoeducational assessment, there is little value in repeating flawed assessment practices.

Acknowledgments

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Award Number P50 HD052117, Texas Center for Learning Disabilities, from the Eunice Kennedy Shriver National Institute of Child Health & Human Development to the University of Houston.

References

- Barth AE, Stuebing KK, Anthony JL, Denton CA, Mathes PG, Fletcher JM, & Francis DJ (2008). Agreement among response to intervention criteria for identifying responder status. *Learning and Individual Differences*, 18, 296–307. [PubMed: 19081758]
- Brown Waesche JS, Schatschneider C, Maner J, Ahmed Y, & Wagner R (2011). Examining agreement and longitudinal stability among traditional and RTI-based definitions of reading disability using the affected-status agreement statistic. *Journal of Learning Disabilities*, 44, 296–307. [PubMed: 21252372]
- Burns MK, Scholin SE, Kosciulek S, & Livingston J (2010). Reliability of decision-making frameworks for response to intervention for reading. *Journal of Psychoeducational Assessment*, 28, 102–114.
- Flanagan DP, Ortiz SO, & Alfonso VC (2013). *Essentials of cross-battery assessment* Hoboken, NJ: John Wiley.
- Flanagan DP, & Schneider WJ (2016). Cross-Battery Assessment? XBA PSW? A case of mistaken identity: A commentary on Kranzler and colleagues' "Classification agreement analysis of Cross-Battery Assessment in the identification of specific learning disorders in children and youth." *International Journal of School & Educational Psychology*, 4, 137–145.
- Fletcher JM (2012). Classification and identification of learning disabilities. In Wong B & Butler D (Eds.), *Learning about learning disabilities* (4th ed., pp. 1–25). New York, NY: Elsevier.
- Fletcher JM, Stuebing KK, Barth AE, Miciak J, Francis DJ, & Denton CA (2014). Agreement and coverage of indicators of response to intervention: A multi-method comparison and simulation. *Topics in Language Disorders*, 34, 74–89. [PubMed: 25364090]
- Hale JB, Alfonso V, Berninger V, Bracken B, Christo C, Clark E, ... Dumont R (2010). Critical issues in response-to-intervention, comprehensive evaluation, and specific learning disabilities identification and intervention: An expert white paper consensus. *Learning Disability Quarterly*, 33, 223–236.
- Hale JB, & Fiorello CA (2004). *School neuropsychology: A practitioner's handbook* New York, NY: Guilford Press.
- Hanson J, Sharman MS, & Esparza-Brown J (2008). Pattern of strengths and weaknesses in specific learning disabilities: What's it all about? Technical Assistance Paper. Oregon School Psychologists Association Retrieved October 5, 2012 from <http://www.jamesbrenthanson.com/uploads/PSWCondensed121408.pdf>.
- Kavale KA, & Forness SR (2000). What definitions of learning disability say and don't say a critical analysis. *Journal of Learning Disabilities*, 33, 239–256. [PubMed: 15505963]
- Kranzler JH, Floyd RG, Benson N, Zaboski B, & Thibodaux L (2016a). Classification agreement analysis of Cross-Battery Assessment in the identification of specific learning disorders in children and youth. *International Journal of School & Educational Psychology*, 43, 124–136.
- Kranzler JH, Floyd RG, Benson N, Zaboski B, & Thibodaux L (2016b). Cross-Battery Assessment pattern of strengths and weaknesses approach to the identification of specific learning disorders: Evidence-based practice or pseudoscience? *International Journal of School & Educational Psychology*, 43, 146–157.
- Maki KE, Floyd RG, & Roberson T (2015). State learning disability eligibility criteria: A comprehensive review. *School Psychology Quarterly*, 30, 457–469. [PubMed: 25581001]
- Miciak J, Fletcher JM, Stuebing KK, Vaughn S, & Tolar TD (2014). Patterns of cognitive strengths and weaknesses: Identification rates, agreement, and validity for learning disabilities identification. *School Psychology Quarterly*, 29, 21–37. [PubMed: 24274155]
- Miciak J, Taylor WP, Denton CA, & Fletcher JM (2015). The effect of achievement test selection on identification of learning disabilities within a patterns of strengths and weaknesses framework. *School Psychology Quarterly*, 30, 321–334. [PubMed: 25243467]

- Miciak J, Williams JL, Taylor WP, Cirino PT, & Fletcher JM (2016). Do processing patterns of strengths and weaknesses predict differential treatment response? *Journal of Educational Psychology*, 108, 898–909. [PubMed: 27616784]
- Morris RD, & Fletcher JM (1988). Classification in neuropsychology: A theoretical framework and research paradigm. *Journal of Clinical and Experimental Neuropsychology*, 10, 640–658. [PubMed: 3066798]
- Naglieri JA (2011). The discrepancy/consistency approach to SLD identification using the PASS theory. In Flanagan DP & Alfonso VC (Eds.), *Essentials of specific learning disability identification* (pp. 145–172). Hoboken, NJ: John Wiley.
- SAS Institute (2013). SAS (9.4) [computer software] Cary, NC: SAS Institute.
- Stuebing KK, Fletcher JM, Branum-Martin L, & Francis DJ (2012). Evaluation of the technical adequacy of three methods for identifying specific learning disabilities based on cognitive discrepancies. *School Psychology Review*, 41, 3–22. [PubMed: 23060685]
- Stuebing KK, Fletcher JM, LeDoux JM, Lyon GR, Shaywitz SE, & Shaywitz BA (2002). Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal*, 39, 469–518.
- Taylor WP, Miciak J, Fletcher JM, & Francis DJ (2016). Cognitive discrepancy models for specific learning disabilities identification: Simulations of psychometric limitations. *Psychological Assessment* 10.1037/pas0000356
- Wechsler D (2001). *Wechsler individual achievement test –Second Edition (WIAT-II)* San Antonio, TX: The Psychological Corporation.
- Wechsler D (2003). *Wechsler intelligence scale for children–Fourth Edition (WISC-IV)* San Antonio, TX: The Psychological Corporation.

Table 1.
Classification Agreement for Identification Decisions Based on Single Indicators With Latent SLD Status ($N = 1,000,000$).

Parameters		Latent classifications		Observed classifications							
Reliability	Meaningful difference	Referred sample ^a	Latent LD	LD	True positive	False positive	False negative	True negative	Sen.	Spec.	PPV
<i>r</i>	<i>SD</i>	<i>n</i>	<i>n</i>	%	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>	%	%
.95	0.5	221,721	39,594	17.86	25,795	11,246	13,799	170,881	65.15	93.83	69.64
.95	1	221,721	20,874	9.41	15,654	21,387	5,220	179,460	74.99	89.35	42.26
.95	1.25	221,721	11,329	5.11	8,695	28,346	2,634	182,046	76.75	86.53	23.47
.9	0.5	222,363	39,240	17.65	17,506	10,541	21,734	172,582	44.61	94.24	62.42
.9	1	222,363	20,634	9.28	14,019	14,028	6,615	187,701	67.94	93.05	49.98
.9	1.25	222,363	11,198	5.04	8,630	19,417	2,568	191,748	77.07	90.8	30.77
.85	0.5	223,154	38,783	17.38	12,372	8,846	26,411	175,525	31.9	95.2	58.31
.85	1	223,154	20,324	9.11	10,877	10,341	9,447	192,489	53.52	94.9	51.26
.85	1.25	223,154	10,994	4.93	7,138	14,080	3,856	198,080	64.93	93.36	33.64
.8	0.5	223,590	38,270	17.12	9,752	7,989	28,518	177,331	25.48	95.69	54.97
.8	1	223,590	20,022	8.95	8,742	8,999	11,280	194,569	43.66	95.58	49.28
.8	1.25	223,590	10,821	4.84	5,906	11,835	4,915	200,934	54.58	94.44	33.29
.75	0.5	224,247	37,769	16.84	7,684	7,001	30,085	179,477	20.34	96.25	52.33
.75	1	224,247	19,649	8.76	6,884	7,801	12,765	196,797	35.03	96.19	46.88
.75	1.25	224,247	10,618	4.73	4,712	9,973	5,906	203,656	44.38	95.33	32.09

Note. LD = learning disability; Sen. = sensitivity; Spec. = specificity; PPV = positive predictive value; NPV = negative predictive value.

^a Referred sample = all students with standard score < 90 on academic measure and IQ > 70.

Classification Agreement for Identification Decisions Based on Recursive Test–Retest Procedure With Latent SLD Status and Comparison With Single Measure Indicators ($N = 1,000,000$).

Parameters		Latent classifications				Observed classifications				Agreement				Increase in correct classifications ^b		Increase in true positive ^c	
Reliability	Meaningful difference	Referred sample ^a	Latent LD	LD	True positive	False positive	False negative	True negative	Sen.	Spec.	PPV	NPV	%	%	n(%)	%	%
<i>r</i>	<i>SD</i>	<i>n</i>	<i>n</i>	%	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>	%	%	%	%	%	%			
.95	0.5	221,721	39,594	17.86	17,986	3,165%	21,608	178,962	45.4	98.3	85.0	89.2		272 (0.1)		-2,871.0	
.95	1	221,721	20,874	9.41	12,497	8,654%	8,377	192,193	59.9	95.7	59.1	95.8		9,576 (4.9)		-33.0	
.95	1.25	221,721	11,329	5.11	7,197	13,954	4,132	196,438	63.5	93.4	34.0	97.9		12,894 (6.8)		-11.6	
.9	0.5	222,363	39,240	17.65	9,450	3,126%	29,790	179,997	24.1	98.3	75.1	85.8		-641 (-0.3)		1,256.8	
.9	1	222,363	20,634	9.28	9,497	3,079%	11,137	198,650	46.0	98.5	75.5	94.7		6,427 (3.2)		-70.4	
.9	1.25	222,363	11,198	5.04	6,588	5,988%	4,610	205,177	58.8	97.2	52.4	97.8		11,387 (5.7)		-17.9	
.85	0.5	223,154	38,783	17.38	5,072	2,172%	33,711	182,199	13.1	98.8	70.0	84.4		-626 (-0.3)		1,166.1	
.85	1	223,154	20,324	9.11	5,757	1,487%	14,567	201,343	28.3	99.3	79.5	93.3		3,734 (1.8)		-137.1	
.85	1.25	223,154	10,994	4.93	4,443	2,801%	6,551	209,359	40.4	98.7	61.3	97.0		8,584 (4.2)		-31.4	
.8	0.5	223,590	38,270	17.12	3,152	1,625%	35,118	183,695	8.2	99.1	66.0	84.0		-236 (-1)		2,796.6	
.8	1	223,590	20,022	8.95	3,759	1,018%	16,263	202,550	18.8	99.5	78.7	92.6		2,998 (1.5)		-166.2	
.8	1.25	223,590	10,821	4.84	3,046	1,731%	7,775	211,038	28.2	99.2	63.8	96.5		7,244 (3.5)		-39.5	
.75	0.5	224,247	37,769	16.84	1,925	1,103%	35,844	185,375	5.1	99.4	63.6	83.8		139 (0.1)		-4,143.2	
.75	1	224,247	19,649	8.76	2,321	707%	17,328	203,891	11.8	99.7	76.7	92.2		2,531 (1.2)		-180.3	
.75	1.25	224,247	10,618	4.73	1,921	1,107%	8,697	212,522	18.1	99.5	63.4	96.1		6,075 (2.9)		-45.9	

Note. LD = learning disability; Sen. = sensitivity; Spec. = specificity; PPV = positive predictive value; NPV = negative predictive value.

^aReferred sample = all students with standard score < 90 on academic measure and IQ > 70.

^bImprovement over classification accuracy of one indicator (see Table 1).

^cTrue positive percentage represents proportion of change in correct classifications accounted for by increase in true positive classifications.

Classification Agreement for Mean Score of Two Indicators With Latent SLD Status and Comparison With Single Measure Indicators ($N = 1,000,000$).

Parameters		Latent classifications				Observed classifications				Agreement				Increase in correct classifications ^b		Increase in true positive ^c	
		Meaningful Difference	Referred sample ^a	Latent LD	LD	True positive	False positive	False negative	True negative	Sen.	Spec.	PPV	NPV				
<i>r</i>	<i>SD</i>	<i>N</i>	<i>N</i>	<i>n</i>	%	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>	%	%	%	%	<i>n</i> (%)	%	%	%
.95	0.5	221,721	39,594	17.86	28,668	8,588	10,926	173,539	72.4	95.3	77.0	94.1	5,531 (2.8)	51.9			
.95	1	221,721	20,874	9.41	16,697	20,559	4,177	180,288	80.0	89.8	44.8	97.7	1,871 (1.0)	55.8			
.95	1.25	221,721	11,329	5.11	9,087	28,169	2,242	182,223	80.2	86.6	24.4	98.8	569 (0.3)	68.9			
.9	0.5	222,363	39,240	17.65	18,483	7,964	20,757	175,159	47.1	95.7	69.9	89.4	3,554 (1.9)	27.5			
.9	1	222,363	20,634	9.28	15,904	10,543	4,730	191,186	77.1	94.8	60.1	97.6	5,370 (2.7)	35.1			
.9	1.25	222,363	11,198	5.04	9,644	16,803	1,554	194,362	86.1	92.0	36.5	99.2	3,628 (1.8)	28.0			
.85	0.5	223,154	38,783	17.38	11,943	5,980	26,840	178,391	30.8	96.8	66.6	86.9	2,437 (1.3)	-17.6			
.85	1	223,154	20,324	9.11	11,905	6,018	8,419	196,812	58.6	97.0	66.4	95.9	5,351 (2.6)	19.2			
.85	1.25	223,154	10,994	4.93	8,123	9,800	2,871	202,360	73.9	95.4	45.3	98.6	5,265 (2.6)	18.7			
.8	0.5	223,590	38,270	17.12	7,907	4,365	30,363	180,955	20.7	97.6	64.4	85.6	1,779 (1.0)	-103.7			
.8	1	223,590	20,022	8.95	8,449	3,823	11,573	199,745	42.2	98.1	68.9	94.5	4,883 (2.4)	-6.0			
.8	1.25	223,590	10,821	4.84	6,205	6,067	4,616	206,702	57.3	97.2	50.6	97.8	6,067 (2.9)	4.9			
.75	0.5	224,247	37,769	16.84	5,564	3,279	32,205	183,199	14.7	98.2	62.9	85.1	1,602 (0.9)	-132.3			
.75	1	224,247	19,649	8.76	6,065	2,778	13,584	201,820	30.9	98.6	68.6	93.7	4,204 (2.0)	-19.5			
.75	1.25	224,247	10,618	4.73	4,604	4,239	6,014	209,390	43.4	98.0	52.1	97.2	5,626 (2.7)	-1.9			

Note. LD = learning disability; Sen. = sensitivity; Spec. = specificity; PPV = positive predictive value; NPV = negative predictive value.

^a Referred sample = all students with SS < 90 on academic measure and IQ > 70.

^b Improvement over classification accuracy of one indicator (see Table 1).

^c True positive percentage represents proportion of change in correct classifications accounted for by increase in true positive classifications.