**Reexamining the Diagnostic Accuracy of PSW Methods:**   
**Dispelling the Illusion of Inaccuracy with Buffer Zones**

W. Joel Schneider1, Dawn P. Flanagan2, Christopher R. Niileksela3, and Joseph R. Engler4

1Policy, Organizational and Leadership Studies, Temple University

2Department of Psychology, St. John’s University

3Department of Educational Psychology, University of Kansas

4Department of School Psychology, Gonzaga University

**Author Note**

Correspondence concerning this article should be addressed to W. Joel Schneider, Temple University, 1301 Cecil B. Moore Ave. Philadelphia, PA 19122. Email: [schneider@temple.edu](mailto:schneider@temple.edu)

**Abstract**

One of the most hotly debated and controversial topics in school psychology involves the identification of students with specific learning disabilities (SLDs). Within the past decade, there has been an increase in the use of patterns of strengths and weaknesses (PSW) as a viable method for identifying SLDs. Several researchers, however, have found that the diagnostic accuracy of PSW methods to be unacceptably low by using strict thresholds to identify students with and without SLDs. This study reexamined the diagnostic accuracy of PSW methods using standard error of measurement (i.e., buffer zones) around thresholds to mirror school psychological practice more closely. Results of the study demonstrated that most diagnostic errors are a result of scores falling within 5 points of the threshold. Thus, by using buffer zones, diagnostic accuracy of PSW methods improves significantly. The practical application of buffer zones to improve the identification of SLDs in practice is also discussed.

*Keywords:* specific learning disabilities, patterns of strengths and weaknesses, diagnostic accuracy, buffer zones

Reexamining the Diagnostic Accuracy of PSW Methods Through the Use of Buffer Zones

Although our understanding of what constitutes a specific learning disability (SLD) continues to evolve and scholars propose alternative definitions (Kavale & Forness, 1985), the SLD definition that is to be used within school systems within the United States has been specified by the Individuals with Disabilities Education Improvement Act (IDEIA, 2004):

a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, that may manifest itself in an imperfect ability to listen, think, speak, read, write, spell, or to do mathematical calculations, including conditions such as perceptual disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia (C.F.R. § 300.8[c][10].

Specific assessment procedures or models are not included in this definition. IDEIA (2004) expanded the methods states can use to identify a SLD beyond the Ability-Achievement Discrepancy (AAD) method to include Response to Intervention (RtI) and alternative research-based procedures, namely Patterns of Strengths and Weaknesses (PSW). While a number of models have been developed to identify SLDs (Alfonso & Flanagan, 2018) there has been disagreement on the most appropriate methodology for identifying SLDs (Grapin, 2018).

PSW methods use assessment data (e.g., standardized tests of cognitive abilities and academic skills) to document a pattern of cognitive and academic strengths and weaknesses that is consistent with the SLD construct as defined in IDEA. PSW approaches of SLD identification have grown over the past decade. States and practicing school psychologists have responded accordingly and adopted these methods. Maki et al. (2015) reported that approximately 25% of states allow the use of PSW for identification of a SLD, while another 25% of states do not specify whether PSW can be used for SLD identification. Additionally, school psychologists reported that approximately 35% of school districts use PSW methods, 34% use RtI methods, and 30% use ability-achievement discrepancy (AAD) methods (Maki & Adams, 2019). Of the PSW methods, the Dual Discrepancy/Consistency (DD/C; Flanagan et al., 2013) method was the most common PSW method used within the school districts. Recently, Benson et al. (2020) found that the DD/C method was used by approximately 67% of school psychologists who use the PSW method. PSW methods have also become a major part of training programs in school psychology (Lockwood & Farmer, 2019).

Although there has been a substantial increase in the use of PSW methods in the field, they have not been without critique (e.g., McGill et al., 2016). A number of studies have suggested these methods have poor diagnostic utility and do not perform well when identifying SLD (Kranzler et al., 2016, 2019; Miciak et al., 2018; Stuebing et al., 2012). These studies have been important in promoting dialogue and research into improving SLD identification practices, although there are some important psychometric issues that are important to address and affect the interpretation of these studies. In this paper, we aim to address one of the main issues that is a critique of PSW models, namely the use of strict cut scores, or thresholds, in PSW models. First, we summarize the previous studies that have examined the diagnostic utility of PSW methods. Second, we review the psychometric issues related to using strict thresholds in research and practice. Finally, we present a simulation to provide a more nuanced description about how diagnostic errors are distributed around strict thresholds, how these impact diagnostic accuracy statistics, and propose some possible solutions to this issue. We hope to show that when practitioners consider a reasonable amount of measurement error around thresholds when making diagnostic decisions, PSW models are not as fatally flawed as previously thought.

## Research on the Technical Adequacy and Diagnostic Utility of PSW

PSW methods for SLD identification have been in use for well over a decade (Flanagan, 2006; Hale & Fiorello, 2004; Naglieri & Das, 1997) although only recently they been the subject of critical review and empirical analysis. These studies on the reliability and validity of PSW methods are a welcome addition to the field. They have laid the foundation for future studies on the utility of PSW methods and paved the way for continued scientific discussions, investigations, and debate.

In the first of these studies by Steubing et al (2012), simulated data were used to examine the technical adequacy of three PSW methods, including Naglieri and Das’s Discrepancy/Consistency method (D/C; 1997), Flanagan and colleagues Dual Discrepancy/Consistency (DD/C; 2011)[[1]](#footnote-1) method, and Hale and Fiorello’s Concordance/Discordance method (C/DM; 2004). First, simulated true scores were estimated using the correlations among cognitive and academic scores. Then, the reliabilities for the tests were used to derive observed scores for the cognitive and academic tests for each case in the simulation. For each case, SLD identification procedures were performed on both the true scores and the observed scores, where the true scores were used to identify those who “truly” had SLD using each of the models, and the observed scores were used to determine how well the different PSW models correctly identified those with and without SLD.

Diagnostic accuracy statistics (i.e., overall accuracy, sensitivity, specificity, positive predictive value [PPV], and negative predictive value [NPV]) were calculated to examine the technical adequacy of each models. Steubing et al. (2012) showed that all PSW methods had high specificities and NPVs, indicating that they were excellent at identifying cases that were not SLD. Sensitivities varied from poor to excellent across conditions. Poor sensitivity suggests it is unlikely that all cases whose true scores meet the PSW method definition of SLD will also have observed scores that meet the PSW definition of SLD. All methods had medium to small PPVs, meaning that most cases identified as SLD using observed scores did not have true scores that meet SLD criteria. It was determined that measurement error in cognitive and academic scores causes PSW methods to be too inaccurate for practical use. Further simulations (Miciak et al., 2018; Taylor et al., 2017) and the application of PSW methods to real cases (Miciak et al., 2014) have shown similar results.

The simulation studies evaluating PSW methods have been notable and instructive, opening an important dialogue regarding the current methods used to identify children with and without SLDs. These findings rightly call into question whether PSW methods are appropriate for use in practice due to their poor ability to accurately identify those with and without SLDs based on the findings. However, there are some psychometric limitations in this research that are worth examining more closely, namely the use of thresholds on continuous variables to make dichotomous diagnostic decisions. This issue has important implications for how the results of these studies are interpreted.

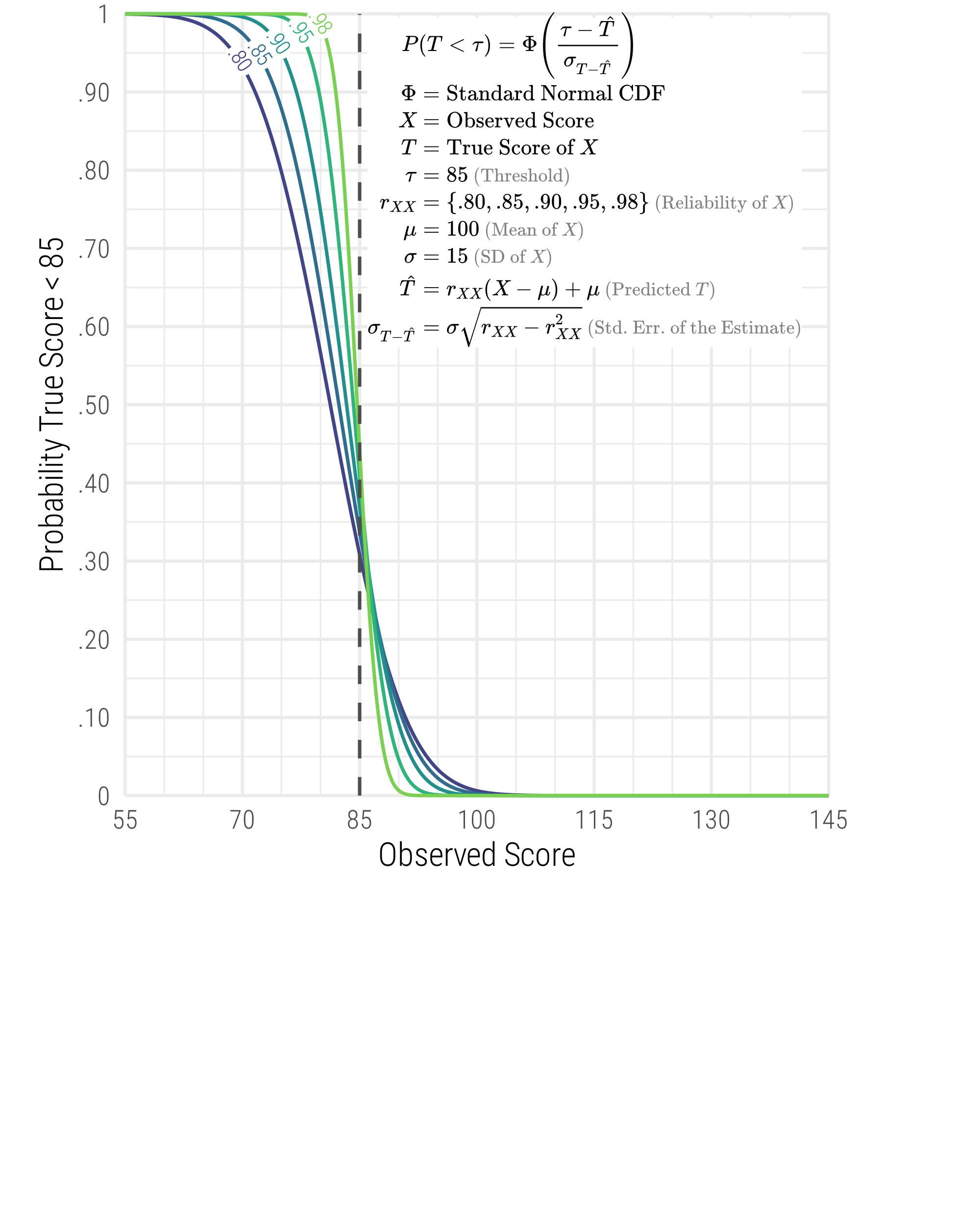
## Diagnostic Thresholds and Diagnostic Decision Making

Diagnostic decision making is an important part of school psychological practice. One of the major reasons diagnostic errors may be made is that diagnostic systems are based on categories that require dichotomous decisions (i.e., a person either does or does not meet the criteria for a disability), but the measures used to make those decisions are continuous and reflect the dimensional nature of psychological constructs. It is unlikely that there are specific thresholds on psychological measures that signify a person does or does not have a disability because there are not clear categorical boundaries on them (Haslam et al., 2020; Meehl, 1995). Differences between people on psychological measures likely represent quantitative *differences in degree* rather than qualitative *differences in kind*, making classification difficult (Hengartner & Lehmann, 2017; Psyridou et al., 2020; Zachar & Kendler, 2017). SLD classification decisions are supposed to represent differences in kind (SLD/not SLD), yet those decisions are made with continuous measures. All methods of SLD identification allowable by law (i.e., AAD, RtI, PSW) use thresholds of some type, thus they all share the same psychometric limitations (Miciak et al., 2016).

Continuous variables with meaningful thresholds that separate differences in kind exist in the natural world, but they are rare. For example, the freezing point of a liquid is a threshold that matters because that is the temperature at which the liquid undergoes a phase transition to a solid state. For SLD, there are no phase transitions from “normal” to “disordered” states. Instead, as relevant cognitive abilities decrease, the risk of academic difficulties increases along smooth gradients (e.g., Fitzpatrick et al., 2015). To the degree that thresholds are used in SLD diagnoses, they are convenient boundaries established by tradition and consensus. In the same way that *α* = .05 has no inherent theoretical meaning for hypothesis testing, an ability score of 85 is merely a convenient threshold at which average scores are distinguished from below average scores. In this way, diagnostic thresholds are like speed limits. As driving speed increases, the risk of injury increases along a smooth gradient. A 55mph speed limit is imposed not because driving at 56mph is abruptly more dangerous than driving at 55mph but because the risk of injury becomes increasingly unacceptable at high speeds. We accept the arbitrary specificity of speed limits because they roughly correspond to the highest driving speeds still associated with low risk.

When strict thresholds are imposed on continuous variables, the meaning of those thresholds may shift over time. At first, they may simply be convenient placeholders, but they may become imbued with excess meaning, where they are interpreted as representing meaningful differences. Treating a threshold of convenience as inherently meaningful cascades into multiple substantive interpretive errors when evaluating the accuracy of PSW methods. Imagine that we consider a diagnosis that depends on having a below-average score on a continuous construct measured with a standard score (mean = 100, *SD* =15), and we choose a score of 85 (at the 16th percentile) as a convenient threshold to distinguish average from below average scores. The reliability of our measures greatly impacts the number of diagnostic errors that occur with this threshold, and this is shown in Figure 1. As the reliability of a measure changes, the probability that an observed score will be on the opposite side of a threshold as the true score also changes. The probability that a true score is less than 85 is high if the observed score is well below 85. Alternatively, the probability that a true score is less than 85 is quite low if the observed score is well above 85. The reliability of the observed score determines how quickly the probability curve drops from near 1 to near 0. If the reliability coefficient is high, the downward slope of the probability curve is steep near the threshold but nearly flat elsewhere. For lower reliability coefficients, the curve begins to fall well before the threshold, falls more gently, and ends well after the threshold. When an observed score is near the threshold, there is a substantial risk that its true score is on the opposite side of the threshold, even when the reliability coefficient is quite high.

**Figure 1.** The probability that a true score is less than 85 depends on the observed score’s position and reliability coefficient.



Most diagnostic errors will occur near the threshold unless the distribution is clearly bimodal and derived from a mixture of qualitatively different groups. This is not specific to SLD identification; it will be true for any test score that is continuous (e.g., curriculum-based measures, norm-referenced test scores, rating scales) and used for differentiating people into diagnostic categories (e.g., SLD, giftedness, intellectual disability). The placement of cut scores on the distribution, reliability of measures, and base rates of the diagnostic categories can greatly influence the diagnostic accuracy statistics.

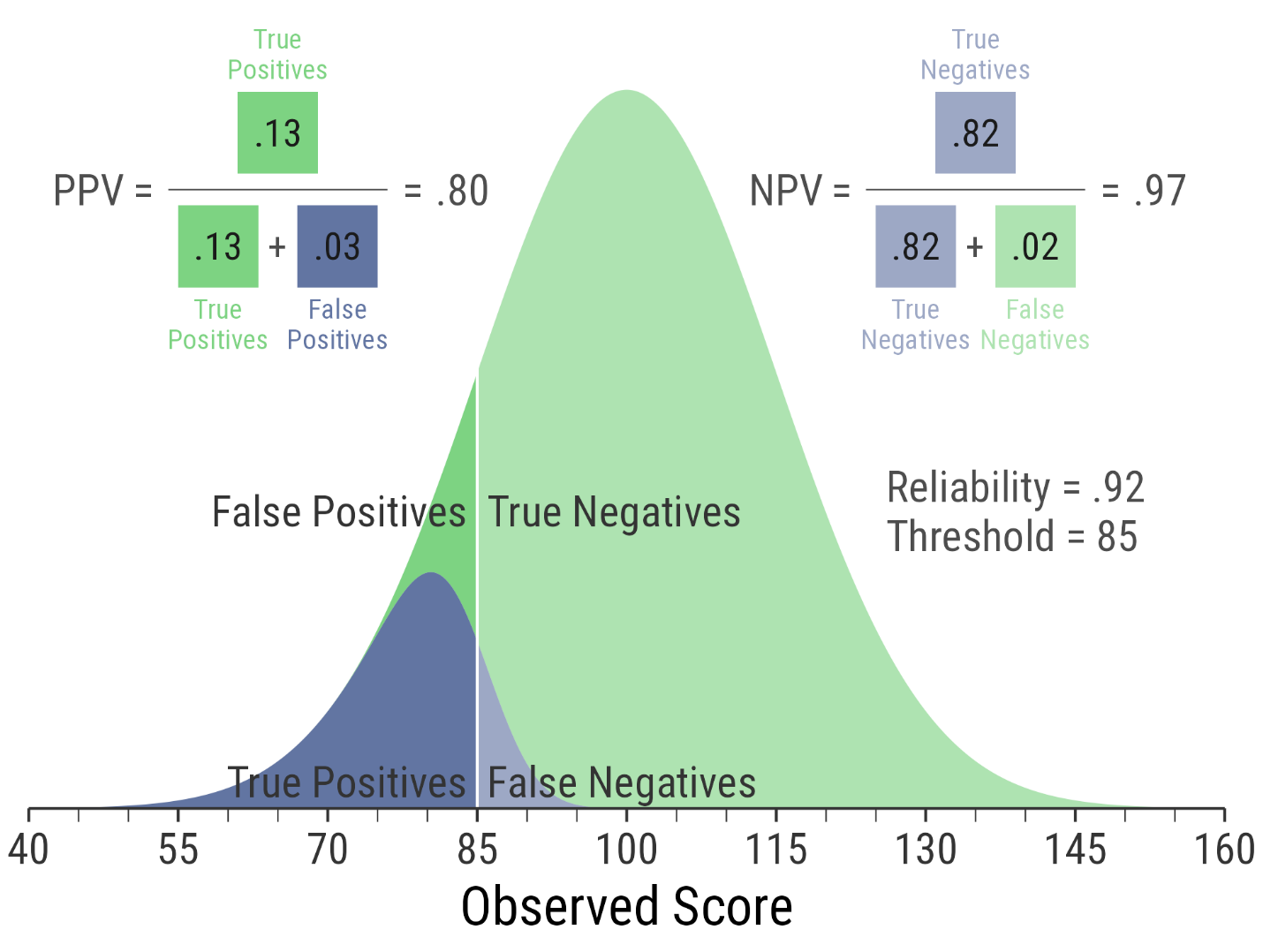
Diagnostic accuracy statistics are used to help demonstrate how well specific measures or models of diagnosis differentiate people into categories. The thresholds placed on continuous variables are integral to calculating diagnostic accuracy statistics, and any issues related to using those thresholds will also transfer to the calculation of these statistics. There are several considerations that should be made when using these statistics to evaluate diagnostic accuracy. First, diagnostic accuracy statistics are primarily designed to identify categorical differences, and SLD is not categorical (Fletcher et al., 2019). Second, although it is relatively straightforward to use them with dichotomous or polytomous test scores, using them with continuous predictors requires finding an arbitrary place in the continuum to cut the variable in in half (i.e., the cut score). While it is possible to evaluate the optimal place for a cut score to maximize both sensitivity and specificity, there is a substantial loss of information with this practice, and diagnostic errors will be common near the threshold.

When diagnostic accuracy statistics are calculated using strict thresholds on continuous variables, diagnostic error rates can appear misleadingly large. A demonstration of this is included in Figure 2. Suppose a standard score with a classical reliability coefficient of .93 is normally distributed, and we somehow know each observed score’s associated true score. We dichotomize the observed and true scores at a threshold of 85 and count how often the observed and true scores fall on opposite sides of the threshold. As seen in Figure 2, the PPV of .80 is unimpressive, but the NPV .97 is high. Almost all the false positive and false negatives are within 10 points of the threshold. For observed scores below 75 or above 95, there is a low probability that their corresponding true scores are on the wrong side of 85. Viewed simply, the reason that NPV is higher than PPV is that there are a lot more scores above 95 than there are scores below 75. If the diagnostic threshold were above mean, the PPV would be higher than the NPV because more people would meet the diagnostic criterion than not.

To say that a PPV of .80 means that we have a 20% chance of making an error every time we identify a weakness is to gloss over the fact that almost all of those “errors” are near the threshold and are unlikely to result in gross errors of interpretation. A weakness might be a little milder or more severe than the observed score leads us to believe, but for reliable variables, it is rare for an observed score to be far enough from its true score that we mistake a weakness for a strength or vice versa. The further an observed score is from a threshold, the less likely its true score is on the other side of the threshold. For this reason, low population-based (unconditional) PPVs and NPVs are of little concern when the PPVs and NPVs conditioned on the individuals’ scores are high. We will show that for individual decision making the conditional PPV and NPV (i.e., PPV and NPV that are based on an individual’s scores) are more important than the unconditional PPV and NPV (i.e., based on population statistics).

Figure 2

False positives and false negatives (i.e., when observed and true scores are on opposite sides of a threshold) tend to occur near the threshold.



The simulation studies previously mentioned (e.g., Steubing et al. 2012) used strict thresholds to categorize the true and observed scores in order to calculate diagnostic accuracy statistics as a way to evaluate PSW models. However, in these studies all diagnostic errors were treated equally, whether they were very minor (threshold is missed by 1 point) or egregiously incorrect (threshold is missed by 30 points). When continuous space is dissected into categories using thresholds and no consideration is given to the distance from the threshold a diagnostic error is, conclusions regarding the diagnostic accuracy of any model, including PSW or other SLD identification models, may be misleading. Importantly, the magnitude of diagnostic errors seems relevant when using thresholds because a “near miss” (e.g., one or two points) is unlikely to change a case conceptualization for a practicing school psychologist. However, an “egregious miss” would likely change the case conceptualization and may result in an incorrect decision about a student. What is unknown is how many diagnostic errors might be considered near misses versus how many diagnostic errors might be considered egregious in these studies. This knowledge is potentially valuable for diagnostic decision making, both in research and in practice.

One proposed solution to address the issue of making dichotomous decisions using continuous variables is by using a “buffer zone” that takes inherent measurement error into account. Psyridou and colleagues (2020) examined this issue with students who do and do not show difficulties in reading fluency and reading comprehension. In this longitudinal study, they showed that when strict thresholds are used there is significant instability in identifying children with reading difficulties over time. However, when using “buffer zones” near those thresholds to account for measurement error, classification accuracy improves. This finding is not unexpected, but it has an important implication for diagnostic decision making. If most diagnostic errors occur close to the threshold, then diagnostic errors may be mitigated by considering a buffer zone near that threshold. Scores near a diagnostic threshold should flag practitioners that they need to look at their data carefully and reduce the uncertainty in their decision with additional data.

Practitioners have long known that clinical judgement is required when measurement error makes interpretation ambiguous (Kaufman, 1979). The inclusion of the buffer zone may better mirror school psychological practice because practitioners would likely not apply thresholds as strictly as they are in research.

## Purpose of the Current Study

The current study was designed to examine how specific thresholds affect diagnostic decisions. The purpose of our study was to answer the following research questions.

1. When diagnostic errors occur, how far are they from the threshold?
2. How do diagnostic accuracy statistics change when a buffer zone (i.e., 5 points) is included to account for a reasonable amount of measurement error?

# Method

## Simulation of Test Scores

To illustrate how diagnostic accuracy statistics can give a distorted view of the inaccuracy of SLD classification decisions, we simulate data using a simplified PSW model. For our purposes, we need only simulate a single general ability, a single specific cognitive ability and a single academic ability. If we were to simulate the all the scores practitioners typically obtain, the illustration would become cluttered with unnecessary detail.

Figure 1 includes our simulation model in which there are true scores for general (*g*), specific (*s*), and academic abilities (*a*), along with their respective observed scores, *G*, *S*, and *A*. The observed general ability score *G* has a reliability coefficient of .97, and the other observed scores *S* and *A* have a reliability coefficient of .92. In Table 1 and Figure 2, we have specified our simplified criteria for SLD. The strict criteria in Table 1 are similar to the criteria by which previous simulation studies were conducted in that the scores are dichotomized and if a single criterion is not met, the decision is “Not SLD” (e.g., Stuebing et al., 2012). In our simulation, we specify three zones in which scores can fall: SLD-Likely, SLD-Possible (i.e., the buffer zone), and SLD-Unlikely. The three zones were created by specifying thresholds for the SLD-Likely criteria and then relaxing the criteria by 5 points for the SLD-Possible criteria. In addition, we specify criteria for general ability > specific cognitive ability and general ability > academic ability in the same way, relaxing criteria by 5 points for SLD-Possible. Cases were classified by the lowest criterion set for which all criteria were satisfied. Thus, if all 5 criteria in the SLD-Strict criterion set were met, the decision was “SLD-Strict.” If even one score was in a lower zone, the case was classified as belonging in the lower group. For example, if 4 criteria were met for SLD-Strict but one score was in the SLD-Adjacent zone, the case was classified as “SLD-Adjacent.” The decision rules were applied to the true and observed scores and then the diagnostic accuracy statistics were calculated from the number of cases in which the decisions based on the true and observed scores agree and disagree.

**Figure 3**

Model for simulation of general, specific, and academic ability true scores (g, s, and a) and observed scores (G, S, and A).

Diagram, schematic

Description automatically generated**Table 1**

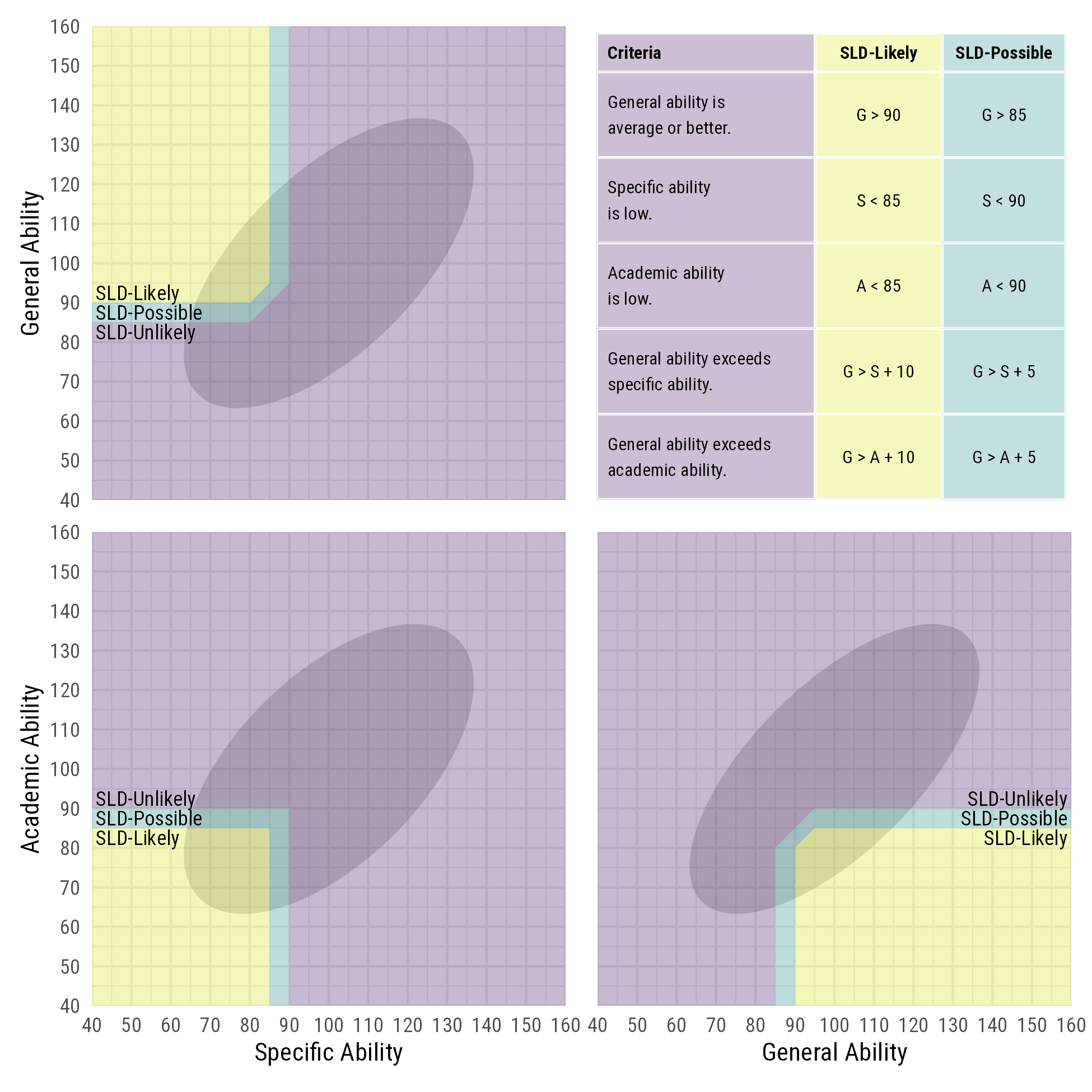
Simplified PSW Criteria for SLD Identification

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SLD-Likely | |  | SLD-Possible | |
| Criteria | True Scores | Observed Scores |  | True Scores | Observed Scores |
| Average General | *g > 90* | *G* > 90 |  | *g > 85* | *G* > 85 |
| Low Specific | *s* < 85 | *S* < 85 |  | *s* < 90 | *S* < 90 |
| Low Academic | *a* < 85 | *A* < 85 |  | *a* < 90 | *A* < 90 |
| General > Specific | *g* > *s* + 10 | *G* > *S* + 10 |  | *g* > *s* + 5 | *G* > *S* + 5 |
| General > Academic | *g* > *a* + 10 | *G* > *A* + 10 |  | *g* > *a* + 5 | *G* > *A* + 5 |

*Note*: Cases that meet all SLD-Likely criteria are classified as SLD-Likely. The remaining cases that meet all SLD-Possible or a combination of SLD-Likely and SLD-Possible criteria are classified as SLD-Possible. Cases that do not meet *all* SLD-Possible criteria are classified as SLD-Unlikely.

**Figure 3**

*Visual Representation of Simplified PSW Criteria for SLD Identification*



**Data Analysis**

Data were simulated in the R programming environment using the lavaan package (Rosseel, 2012). Analyses were conducted with tidyverse packages (Wickham et al., 2021). All code for this project can been downloaded from ??.

# Results

## Descriptive Statistics

Descriptive statistics for 10,000,000 simulated cases can be seen in Table 2. The contingency table for decisions based on observed and true scores is provided in Table 3.

Table ??. Correlations, means, and standard deviations for simulated cases n = 10,000,000

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | True Scores | | |  | Observed Scores | | |
|  | *g* | *s* | *a* |  | *G* | *S* | *A* |
| *g* | — |  |  |  |  |  |  |
| *s* | .67 | — |  |  |  |  |  |
| *a* | .71 | .64 | — |  |  |  |  |
| *G* | .98 | .66 | .70 |  | — |  |  |
| *S* | .64 | .96 | .62 |  | .63 | — |  |
| *A* | .68 | .62 | .96 |  | .67 | .59 | — |
| Mean | 100.00 | 100.01 | 100.00 |  | 100.00 | 100.01 | 100.00 |
| SD | 14.78 | 14.39 | 14.39 |  | 15.00 | 15.00 | 15.00 |

## Distribution of Errors Around Thresholds

In the top panel of Figure 6, it can be seen that 24% of false positives according to SLD-Strict criteria have no true scores on the wrong side of the threshold. They are false positives because their general true score does not exceed their specific true score or their academic true score by at least 10 points. The remaining 76% of false positives have at least one true score that crosses the relevant threshold. For about 90% of false positive cases, the largest distance a true score crosses outside the SLD-Strict threshold is 5 points. Less than 1% of false positives have true scores that are 10 points or more beyond the SLD-Strict threshold.

In the bottom panel of Figure 6, it can be seen that 26% of false negatives according to SLD-Strict criteria have no true scores on the wrong side of the threshold. They are false negatives because their general true score does not exceed their specific or academic true score by at least 10 points. The remaining 74% of false negatives have at least one true score that crosses the relevant threshold. For about 91% of false negative cases, the largest distance by which the true score crosses outside the SLD-Strict threshold is 5 points. Less than 1% of false negatives have true scores that are 10 points or more outside the SLD-Strict zone.

In this illustration, we see that most false positives and false negative cases are not true diagnostic errors. Rather, they represent cases in which the severity of the problem was slightly overestimated (false positives) or slightly underestimated (false negatives). Misestimation of the severity of SLD is certainly a problem, but misestimations that are within five points of a strict threshold are not true diagnostic errors and therefore, should not be classified as diagnostic errors. Rather, slight over- and underestimates represent quantitative differences in degree (e.g., mild to severe); they do not represent differences in kind (i.e., SLD/Not SLD).

**How does including a buffer zone effect diagnostic decision making?**

First, we calculate diagnostic accuracy statistics on the strict criterion set only. These statistics are similar to results presented in previous simulation studies. The diagnostic accuracy statistics in Table 4 are not intended to be estimates of diagnostic accuracy in everyday practice. Our purpose, rather, is to illustrate how accuracy statistics can be misleadingly low. It is enough for this purpose that, like all previous simulation studies, the PPV and sensitivity values are low and the NPV and specificity values are high. The prevalence of SLD is low in this simulation study because it reflects the prevalence of SLD due to a single specific deficit manifesting in a single academic ability. Had we simulated scores for multiple specific cognitive and academic abilities the prevalence of SLD would have been many times higher. The other diagnostic accuracy statistics in Table 4 are similarly influenced by the simplicity of our model in Figure 1.

## Misleadingly Low PPV

What drives down the PPV statistic is the presence of false positives. A false positive in this context is a case with observed scores that meet criteria for SLD and at least one corresponding true score that does not. If a large number of false positives have true scores barely on the wrong side of a threshold, then we should be less concerned about ostensibly low PPV values. In this illustration, 76.5% of the false positives have true scores that meet the SLD-Relaxed criteria. That is, more than three quarters of the false positives have true scores that meet the less stringent SLD criteria of g > 85, s < 90, a < 90, g > s +5, and g > a +5. These are not exactly diagnostic mistakes.

To get a sense of what the false positive true scores look like, see Figure 3. Because of the limits of two-dimensional visualization, Figure 3 displays only specific and academic observed scores. The arrows represent false positive cases that start on their observed scores and end on their true scores. Most arrowheads end in the 5-point-wide SLD-Relaxed zone, and only a few arrowheads end in the SLD-Adjacent zone. Observed scores that start in the SLD-Strict zone and their associated true scores end more than 10 points beyond the threshold are quite rare.

As seen in Figure 4, if the observed scores are consistent with Not SLD, then there is a .98 probability that corresponding true scores are also consistent with Not SLD. However, if the observed scores meet the SLD-Strict criteria, the probability that the corresponding true scores also meet the SLD-Strict criteria is .49. The probability that these same observed scores have corresponding true scores that meet SLD-Relaxed criteria is .88. This finding indicates that 88% of the time an observed score meeting SLD-Strict criteria will have a corresponding true score that meets SLD-Relaxed criteria. If we took the PPV of .48 (see Table 4) at face value, we would believe that over half of cases identified as SLD are not in fact SLD. In reality, most of those false positives are cases in which the true score is within five points of the observed score and therefore are consistent with mild SLD (i.e., SLD-Relaxed).

## Misleadingly Low Sensitivity

Low sensitivity is worrisome because it means that many people with SLD will not be identified as having SLD. What drives down the sensitivity statistic is the presence of false negatives. A false negative in this context is a case with observed scores that do not meet SLD-Strict criteria and corresponding true scores that do. . Figure 5 includes a two-dimensional visualization of specific cognitive and academic observed scores. The arrows represent false negative cases that start on their observed scores and end on their true scores. Most false negative cases have observed scores that meet SLD-Relaxed criteria with arrowheads representing true scores that end in the SLD-Strict zone. As a result, the low sensitivity statistic is less concerning.

Under the conditions specified for these simulated scores, 83.8% of false negatives have observed scores that meet SLD-Relaxed criteria (i.e., all observed scores are within 5 points of SLD-Strict criteria). Thus, if we took the .61 sensitivity statistic at face value, we would believe that 39% of people with observed scores that are within five points of the SLD-Strict criteria are not SLD. In reality, only 6.3% of cases whose true scores meet SLD-Strict criteria would have observed scores that meet neither the SLD-Strict nor the SLD-Relaxed criteria. Thus, most cases (93.7%) with true scores that meet the SLD-Strict criteria would be identified as having SLD as defined by strict and relaxed criteria.

Previous studies of the accuracy of PSW methods have examined the unconditional probability of an observed score and its associated true score (or latent construct score) falling on the same or opposite sides of a specified diagnostic threshold score. From such probabilities, the sensitivity, specificity, PPV, NPV, and overall accuracy of a diagnostic decision can be calculated (citations). Although these analyses give correct answers to a narrow kind of question, it is not the question we believe is most relevant to the evaluation of PSW methods’ accuracy. Such analyses lend themselves to misinterpretation because they invest the threshold score with surplus meaning.

**Some errors are more important than others**

If we invest the diagnostic threshold with too much meaning, all cases in which observed and true scores fall on opposite sides of the threshold are considered diagnostic errors. Such a decision rule means that trivially small and egregiously large differences between observed and true scores are treated alike. For example, if the threshold is 85, an observed score of 84 is treated as the same kind of error whether its true score is 87 or 99. In the first case, the observed score of 84 is off by a small amount—3 points (i.e., about 0.2 *SD*). Assuming a test reliability of .95 and multivariate normality, a difference of 3 points or more will occur 37% of the time, indicating it is common. This so-called diagnostic error would not meaningfully alter the interpretation of the score because both scores (84 and 87) indicate mild difficulties. In the latter case, 84 and 99 differ by 15 points (1 *SD*). Under the same assumptions as before, rarely do observed scores and true scores differ by 15 points or more—less than 1 in 100,000 cases. This degree of mismeasurement will likely result in a true interpretive error because the observed score leads us to label the ability as a mild difficulty when it is not.

Importantly, so-called diagnostic errors that are trivially small far outnumber diagnostic errors that result in substantive misinterpretations. We ought to regard with suspicion any analysis that fails to distinguish between trivial and substantive errors. If most of the so-called errors result from small differences near the thresholds, then the charge that DD/C is unreliable or inaccurate is misleading.

## The misapplication of PPV and NPV give exaggerated error rates.

Suppose a standard score with a classical reliability coefficient of .93 is normally distributed, and we know each observed score’s associated true score. We dichotomize the observed and true scores at a threshold of 85 and count how often the observed and true scores fall on opposite sides of the threshold. As seen in Figure 9, the PPV of .80 is unimpressive, but the NPV of .97 is high. Almost all the false positive and false negatives are within 10 points of the threshold. For observed scores below 75 or above 95, there is a low probability that their corresponding true scores are on the wrong side of 85. Viewed simply, the reason that NPV is higher than PPV is that there are a lot more scores in the diagnostic safe zone above 95 than there are scores in the safe zone below 75. If the diagnostic threshold were above the mean, the PPV would be higher than the NPV because more people would meet the diagnostic criterion than not.

To say that a PPV of .80 means that we have a 20% chance of making an error every time we identify a weakness is to gloss over the fact that almost all of those “errors” are near the threshold and are unlikely to result in gross errors of interpretation. A weakness might be a little milder or slightly more severe than the observed score leads us to believe, but for reliable variables, it is rare for an observed score to be far enough from its true score that we would mistake a weakness for a strength or vice versa.

## For individual decisions, the conditional PPV matters more than the overall PPV

Imagine that a well-trained and conscientious practitioner assesses a student whose scores unambiguously meet DD/C criteria. This student’s specific cognitive and academic weaknesses are below 70, and the student’s general ability is above 110. The practitioner need not worry about studies suggesting that DD/C has a low PPV in general. What matters more is the probability of making a false positive identification in this particular case. That is, we would like to know the conditional PPV, the probability of making a correct identification of SLD, given the observed scores in this case.

It is possible to calculate a conditional PPV (and NPV) based on the individual’s observed scores with well known, relatively simple equations (See Appendix A and the web app at <https://w-joel-schneider.shinyapps.io/conditionalppv/>). For example, one can estimate that an observed cognitive weakness of 65 is quite unlikely to be associated with a true score above 85 when the score’s reliability coefficient is above .90. If a child’s cognitive weaknesses are below 70, it is certainly possible that some of them are not quite as low as their observed scores suggest, but it is unlikely that they are substantially above 85. Likewise, if a child’s strengths are above 100, perhaps some of them are overestimated, but it is unlikely that they are, in fact, weaknesses.

The further an observed score is from a threshold, the less likely its true score is on the other side of the threshold. For this reason, low population-based (unconditional) PPVs and NPVs are of little concern when the PPVs and NPVs conditioned on the individuals’ scores are high.

## Diagnostic Buffer Zones

Conditional and unconditional PPVs based on strict diagnostic thresholds are often misleading and usually unnecessary. We should worry more about making gross interpretative errors than about whether our scores are barely on the wrong side of an arbitrary threshold.

An alternative to a strict threshold is to set a convenience threshold along with a buffer zone in which the diagnostic implications of the score are provisional, indefinite, and/or contingent on other factors. Suppose a cognitive score is below a threshold of 85. If the true score is also below 85, we can say that the score is consistent with SLD. We can ask what is the probability that the true score is 85 or higher but distinguish between scores that are within a 5-point buffer (85–89) and those that are inconsistent with SLD (i.e., 90 or higher). In the left panel of Figure 9, we can see that for observed scores less than 85, the true score is also likely to be consistent with SLD. Except when the reliability is unrealistically high, there is a substantial risk that the true score is in the buffer zone. However, provided that the reliability coefficient is above .90, the risk that the true score is inconsistent with SLD is low.

On the right panel of Figure 9, we see that the overall risk of false negatives is low. Because most scores above 85 are well above 85, this finding can be misleading. It does not mean that we need not worry about false negatives ever. When observed scores are barely above 85, the probability that the true score is on the other side of the threshold is high. However, we should worry more about whether the true score is below the buffer of 81–85. Furthermore, we should be less concerned about unconditional PPVs and NPVs depicted in Figure ?? and more concerned about conditional PPVs and NPVs.

## Implications for Practice

In Figure 10, we display a hypothetical case with observed scores consistent with mild SLD. For each score, the expected distribution of true scores is shown along with the probabilities that the true scores cross the diagnostic threshold and the associated buffer zone. The reliability coefficients were given plausible values for general cognitive, specific cognitive, and academic ability composites.

For the general ability composite to be consistent with SLD, it must be 90 or higher. If its true score is in the buffer zone of 85–89, it is neither consistent nor inconsistent with SLD. Its meaning remains indeterminate until additional data are considered. If the general ability observed score is above 90 but its true score is below 85, the true score is inconsistent with SLD and represents a diagnostic error. For the specific cognitive and academic weaknesses, true scores below 85 are consistent with SLD and true scores above 90 are inconsistent with SLD.

For this case, the multivariate conditional PPV (the probability that all three true scores are on the SLD-consistent side of the threshold) is only .32. There is a .52 probability that one or more true scores is in the 5-point buffer zone and a .16 probability that at least one true score is on the SLD-inconsistent side of the buffer. Thus, although the probability that all scores are SLD-consistent is low, there is an .84 probability that no true score is on the SLD-inconsistent side of the buffer zone, indicating that a serious interpretive error has been avoided.

For cases with observed scores that indicate mild SLD such as those in Figure 10, most likely there really is a mild problem. However, we have to accept that there is a substantial risk that one or more true scores are on the wrong side of our threshold and to a much lesser extent our buffer zone. In cases with scores near the threshold, we need not consider them diagnostic errors but rather near misses that are unlikely to change one’s case conceptualization. . We need to remind ourselves about the likely magnitudes of the errors that we are talking about. For example, the true scores might be General Ability = 97, Specific Ability = 91 (not 83), and Academic Ability = 83. So, in this case, we thought that the specific ability might be a partial explanation for the academic difficulty, but in truth, its contributions to the problem are minor at most. There is a .16 chance of making an interpretive error of that magnitude.

In Figure 11, we see a more severe case of SLD in which the observed scores are further away from the diagnostic thresholds. The conditional PPV is .97, which is high enough for most of us to feel comfortable with the SLD classification decision. The probability that one or more scores is on the wrong side of the buffer zone is only 0.01.

**Discussion**

Paul Meehl, one of the most important figures in psychology in the 20th century, stated that “*No statistical method should ever be applied blindly, unthinkingly, paying no attention to anything else about the situation than the way the numbers behave.* A statistical procedure is not an automatic, mechanical truth-generating machine for producing or verifying substantive causal theories.” (Meehl, 1992, p. 143, emphasis in original). Studies on the diagnostic accuracy of SLD identification methods have been important and essential to the field. However, we feel that some of the results have been interpreted very strictly. The presence of measurement error in test scores is well known, but this has not been adequately considered in some of the previous research on PSW methods. Rather, diagnostic accuracy statistics have been applied without a close eye as to where those diagnostic errors fall.

**Limitations**

The misapplication of PPV and NPV give exaggerated error rates.

Suppose a standard score with a classical reliability coefficient of .93 is normally distributed, and we somehow know each observed score’s associated true score. We dichotomize the observed and true scores at a threshold of 85 and count how often the observed and true scores fall on opposite sides of the threshold. As seen in Figure ??, the PPV of .80 is unimpressive, but the NPV .97 is high. Almost all the false positive and false negatives are within 10 points of the threshold. For observed scores below 75 or above 95, there is a low probability that their corresponding true scores are on the wrong side of 85. Viewed simply, the reason that NPV is higher than PPV is that there are a lot more scores in the diagnostic safe zone above 95 than there are scores in the safe zone below 75. If the diagnostic threshold were above mean, the PPV would be higher than the NPV because more people would meet the diagnostic criterion than not (e.g., vision-related difficulties in older adults).

To say that a PPV of .80 means that we have a 20% chance of making an error every time we identify a weakness is to gloss over the fact that almost all of those “errors” are near the threshold and are unlikely to result in gross errors of interpretation. When reliable scores are interpreted properly, they can be misleading, but usually only by small amounts. A weakness might be a little milder or more severe than the observed score leads us to believe, but for reliable variables, it is rare for an observed score to be far enough from its true score that we mistake a weakness for a strength or vice versa.

Figure ??. False positives and false negatives (i.e., when observed and true scores are on opposite sides of a threshold) tend to occur near the threshold. In this illustration, the threshold is 85 and the reliability coefficient is .93.

For individual decisions, the conditional PPV matters more than the overall PPV

Imagine that a well-trained and conscientious practitioner assesses a student whose scores unambiguously meet DD/C criteria. This student’s cognitive and academic weaknesses are all below 70, and the student’s cognitive strengths are well above 110. The practitioner need not worry about studies suggesting that DD/C has a low PPV in general. What matters more is the probability of making a false positive identification in this particular case. That is, we would like to know the conditional PPV, the probability of making a correct identification of SLD, given the observed scores in this case.

It is possible to calculate a conditional PPV (and NPV) based on the individual’s observed scores with well known, relatively simple equations (See Appendix A and the web app at <https://w-joel-schneider.shinyapps.io/conditionalppv/>). For example, one can estimate that an observed cognitive weakness of 65 is quite unlikely to be associated with a true score above 85 when the score’s reliability coefficient is above .90. If all a child’s cognitive weaknesses are below 70, it is certainly possible that some of the child’s weaknesses are not quite as low as their observed scores suggest, but it is unlikely that they are substantially above 85. Likewise, if all the child’s strengths are above 100, perhaps some of the child’s strengths are overestimated, but it is unlikely that they are, in fact, weaknesses.

The further an observed score is from a threshold, the less likely its true score is on the other side of the threshold. For this reason, low population-based (unconditional) PPVs and NPVs are of little concern when the PPVs and NPVs conditioned on the individuals’ scores are high.

**References**

Fitzpatrick, C., Archambault, I., Janosz, M., & Pagani, L. S. (2015). Early childhood working memory forecasts high school dropout risk. Intelligence, 53, 160–165. <https://doi.org/10.1016/j.intell.2015.10.002>

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. <https://www.jstatsoft.org/v48/i02/>.

Alfonso, V. C., & Flanagan, D. P. (2018). *Essentials of Specific Learning Disability Identification*. John Wiley & Sons.

Benson, N. F., Maki, K. E., Floyd, R. G., Eckert, T. L., Kranzler, J. H., & Fefer, S. A. (2020). A national survey of school psychologists’ practices in identifying specific learning disabilities. *School Psychology*, *35*(2), 146–157. https://doi.org/10.1037/spq0000344

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. H.R.1350 (2004). https://www.congress.gov/bill/108th-congress/house-bill/1350

Fitzpatrick, C., Archambault, I., Janosz, M., & Pagani, L. S. (2015). Early childhood working memory forecasts high school dropout risk. *Intelligence*, *53*, 160–165. https://doi.org/10.1016/j.intell.2015.10.002

Flanagan, D. P. (Ed.). (2006). *The achievement test desk reference: A guide to learning disability identification* (2nd ed). John Wiley & Sons.

Flanagan, D. P., & Alfonso, V. C. (2011). *Essentials of specific learning disability identification*. Wiley.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment*. John Wiley & Sons.

Grapin, S. L. (2018). Contemporary perspectives on the identification of specific learning disabilities (SLD): Introduction to the special issue. *School Psychology Forum, Research in Practice*, *12*(1), 2–5.

Hale, J. B., & Fiorello, C. A. (2004). *School neuropsychology: A practitioner’s handbook*. Guilford Press.

Haslam, N., McGrath, M. J., Viechtbauer, W., & Kuppens, P. (2020). Dimensions over categories: A meta-analysis of taxometric research. *Psychological Medicine*, *50*(9), 1418–1432. https://doi.org/10.1017/S003329172000183X

Hengartner, M. P., & Lehmann, S. N. (2017). Why Psychiatric Research Must Abandon Traditional Diagnostic Classification and Adopt a Fully Dimensional Scope: Two Solutions to a Persistent Problem. *Frontiers in Psychiatry*, *8*, 101. https://doi.org/10.3389/fpsyt.2017.00101

Kaufman, A. S. (1979). *Intelligent Testing with the WISC-R*. Wiley.

Kavale, K. A., & Forness, S. R. (1985). Learning disability and the history of science: Paradigm or oaradox? *Remedial and Special Education*, *6*(4), 12–24. https://doi.org/10.1177/074193258500600404

Kranzler, J. H., Floyd, R. G., Benson, N., Zaboski, B., & Thibodaux, L. (2016). 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*(3), 124–136. https://doi.org/10.1080/21683603.2016.1155515

Kranzler, J. H., Gilbert, K., Robert, C. R., Floyd, R. G., & Benson, N. F. (2019). Further Examination of a Critical Assumption Underlying the Dual-Discrepancy/Consistency Approach to Specific Learning Disability Identification. *School Psychology Review*, *48*(3), 207–221. https://doi.org/10.17105/SPR-2018-0008.V48-3

Lockwood, A. B., & Farmer, R. L. (2019). The cognitive assessment course: Two decades later. *Psychology in the Schools*, *57*(2), 265–283. https://doi.org/10.1002/pits.22298

Maki, K. E., & Adams, S. R. (2019). A current landscape of specific learning disability identification: Training, practices, and implications. *Psychology in the Schools*, *56*(1), 18–31. https://doi.org/10.1002/pits.22179

Maki, K. E., Floyd, R. G., & T., R. (2015). State learning disability eligibility criteria: A comprehensive review. *School Psychology Quarterly*, *30*, 457. https://doi.org/10.1037/spq0000109

McGill, R. J., Styck, K. M., Palomares, R. S., & Hass, M. R. (2016). Critical Issues in Specific Learning Disability Identification: What We Need to Know About the PSW Model. *Learning Disability Quarterly*, *39*(3), 159–170. https://doi.org/10.1177/0731948715618504

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist*, *50*(4), 266–275. https://doi.org/10.1037/0003-066X.50.4.266

Miciak, J., Fletcher, J. M., & Stuebing, K. K. (2016). Accuracy and validity of methods for identifying learning disabilities in a response-to-intervention service delivery framework. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention* (pp. 421–440). Springer.

Miciak, J., Fletcher, J. M., Stuebing, K. K., Vaughn, S., & Tolar, T. D. (2014). Patterns of cognitive strengths and weaknesses: Identification rates, agreement, and validity for learning disabilities identification. *School Psychology Quarterly*, *29*(1), 21–37.

Miciak, J., Taylor, W. P., Stuebing, K. K., & Fletcher, J. M. (2018). Simulation of LD identification accuracy using a pattern of Processing Strengths and Weaknesses method with multiple measures. *Journal of Psychoeducational Assessment*, *36*(1), 21–33. https://doi.org/10.1177/0734282916683287

Naglieri, J. A., & Das, J. P. (1997). *Cognitive Assessment System: Interpretive Handbook*. Riverside.

Psyridou, M., Tolvanen, A., Lerkkanen, M.-K., Poikkeus, A.-M., & Torppa, M. (2020). Longitudinal Stability of Reading Difficulties: Examining the Effects of Measurement Error, Cut-Offs, and Buffer Zones in Identification. *Frontiers in Psychology*, *10*, 2841. https://doi.org/10.3389/fpsyg.2019.02841

Stuebing, K. K., Fletcher, J. M., Branum-Martin, L., & Francis, D. J. (2012). Evaluation of the technical adequacy of three methods for identifying specific learning disabilities based on cognitive discrepancies. *School Psychology Review*, *41*(1), 3–22. https://doi.org/10.1080/02796015.2012.12087373

Taylor, W. P., Miciak, J., Fletcher, J. M., & Francis, D. J. (2017). Cognitive discrepancy models for specific learning disabilities identification: Simulations of psychometric limitations. *Psychological Assessment*, *29*(4), 446–457. https://doi.org/10.1037/pas0000356

Zachar, P., & Kendler, K. S. (2017). The Philosophy of Nosology. *Annual Review of Clinical Psychology*, *13*(1), 49–71. https://doi.org/10.1146/annurev-clinpsy-032816-045020

1. At that time, the DD/C method was termed, “CHC-based Operational Definition of SLD” (Flanagan & Alfonso, 2011). Stuebing et al. (2012) referred to this model as “XBA” though this is a misnomer. XBA is an assessment procedure to ensure all appropriate data necessary to measure the cognitive abilities and processes related to the referral were adequately represented in an SLD evaluation. [↑](#footnote-ref-1)