Confronting the Base Rate Problem: More Ups and Downs for Cognitive Scatter Analysis

Ryan J. McGill

College of William & Mary

Author Note

A preliminary version of this paper was presented at the 2017 annual meeting of the Southeastern Psychological Association, Atlanta, GA.

Standardization data from the *Kaufman Assessment Battery for Children, Second Edition (KABC-II)*. Copyright © 2004 NCS Pearson, Inc. Used with permission. All rights reserved.

Correspondence concerning this paper should be addressed to Ryan J. McGill, School of Education, College of William & Mary, P. O. Box 8795 Williamsburg, VA. 23187. E-Mail: rmcgill@wm.edu

Abstract

Within the professional literature, it is frequently suggested that interpretation of cognitive profile scatter may be useful for generating a host of clinical inferences. To wit, Hale and colleagues (2008) posit that cognitive scatter is a defining characteristic of specific learning disability and that individuals with learning disabilities may have higher levels of scatter compared to normal controls. To investigate the tenability of this claim, the present study employed diagnostic efficiency statistics and other recommended psychometric methods (e.g., receiver operative characteristic curve, Bayesian nomogram) to test whether cognitive scatter could accurately distinguish between individuals with and without a known learning disability (LD) diagnosis in the Kaufman Assessment Battery for Children-Second Edition (KABC-II; Kaufman & Kaufman, 2004a) normative sample. Results indicated that increasing levels of cognitive profile scatter identified individuals with LD at no better than chance levels. The current negative results add to a growing corpus of research questioning the utility of many of the interpretive procedures that are utilized by school psychologists for commercial ability measures. In particular, it is suggested that clinicians who interpret cognitive profile scatter may risk diagnostic overconfidence and in the case of LD identification, unacceptable levels of false positive decisions attributable to error. Implications for evidence-based assessment in school psychology are discussed.

Keywords: Cognitive scatter, KABC-II, Evidence-based practice

Confronting the Base Rate Problem: More Ups and Downs for Cognitive Scatter Analysis

For decades, school psychologists have been encouraged within the professional literature to carefully evaluate variability across, between, and within composite and subtest scores on intelligence tests and the procedures for engaging in some variant of scatter analysis are described in virtually every technical manual and interpretive guidebook available to practitioners. The intuitive nature and perceived utility of such analyses also contribute to their ubiquity. To wit, Courville and colleagues (2016) suggest that “index-level and subtest-level variability often have clinically meaningful implications and provide the very foundation for assessment of an individual’s intellectual strengths and weaknesses” (p. 225).

The use of scatter analysis presage debate about its relationship to the nature of learning disorders (Kavale, 2002). Over 70 years ago, Rapaport, Gil, and Schafer (1945) proposed an interpretive framework that provided clinicians with a step-by-step process for analyzing intra-individual cognitive strengths and weaknesses based upon the belief that variability in cognitive test performance served as a marker for the presence of behavioral pathology and a multitude of related approaches have been developed within the psychological assessment literature (e.g., Kaufman, 1994; Naglieri, 2000; Priftera & Dersh, 1993). More recently, Kaufman’s (1994) *intelligent testing* approach (which bears many similarities to the approach articulated by Rapaport et al. (1945), in which test users are encouraged to use the analytical powers of a “master detective” to interpret the meaning of significant scatter and its potential implications for the clinical utility of obtained scores, has become a bedrock of modern IQ test interpretation in clinical practice (McGill, 2016). As Jastak (1949) presciently noted long ago, “The main reason for the psychologist’s persistent preoccupation with test scatter is the amount of valuable information it yields as a supplement to quantitative indices” (p. 177). Thus, it is not surprising that in a survey of 354 national certified school psychologists 70% of respondents reported that they found cognitive scatter/profile analysis to be clinically useful (Pfeiffer et al., 2000). Though an update on this survey is much needed[1], the rise of cross-battery assessment and profile of strengths and weakness-based approaches to LD identification (PSW) over the last 15 years suggest that interest in these and other related profile analytic methods has likely not diminished.

Despite the popularity and perceived clinical utility of these methods among practitioners, scientific support has been less consistent. Previous investigations of several profile analytic methods (i.e., ipsative analysis, subtest-level profiles, individual cognitive weakness models) have consistently found that unique cognitive profiles are psychometrically weak, lacking adequate reliability and validity for confidant clinical interpretation. (e.g., Kranzler, Floyd, Benson, Zaboski, & Thibodaux, 2016; McDermott, Fantuzzo, & Glutting, 1990; Macmann & Barnett, 1997; Watkins, 2000, 2003). More germane to the present discussion, in a quantitative meta-analysis, Kavale and Forness (1984) found that the average PIQ-VIQ difference for children with LD was only 3.5 points, a difference found in approximately 79% of the normal population. As a result, it was concluded that cognitive scatter was of "little value in LD diagnosis" (p. 139).

In spite of these findings, proponents of profile analysis continue to advance claims that the interpretation of intra-individual scatter is clinically useful and that the mere presence of scatter renders some composite scores such as FSIQ invalid (Fiorello et al., 2002, 2007). As an example, Hale and colleagues (2008) suggest that cognitive scatter is a defining characteristic of individuals with academic disorders, noting that children with learning disabilities typically have

[1] This author is aware of several attempts to update Pfeiffer et al. (2000) and although data collection is still underway, it is anticipated these results are likely to be replicated.

higher levels of scatter when compared to normal controls. Following this logic, it can be hypothesized that higher and/or statistically significant levels of cognitive scatter may be useful as a marker for the presence of an academic disorder (see Flanagan and Alfonso, 2010 for a review of several diagnostic models that align with this premise). This is an empirical question that can be addressed using a diagnostic utility approach (Wiggins, 1988).

According to Hunsley and Mash (2007), evidence-based assessment (EBA) emphasizes the use of research and theory to inform the selection of assessment targets in clinical practice. EBA guidelines stress that the diagnostic techniques employed by mental health professionals must be supported with appropriate psychometric evidence examining their potential accuracy. Determining the accuracy of cognitive scatter as a potential marker for LD requires the computation of sensitivity and specificity statistics from a 2 x 2 contingency table that cross-tabulates decisions made from test scatter with those from a gold standard diagnosis (a sample diagnostic contingency array is provided in Figure 1). In the context of the present discussion, sensitivity is the proportion of individuals with LD who present with meaningful test scatter (i.e., true positive decisions). Specificity represents the proportion of individuals without significant test scatter who are not LD (i.e., true negative decisions). Diagnostic accuracy is often represented as the rate of “hits” (i.e., true positive and true negative decisions) in a selected population. It is important to note that while a highly sensitive and highly specific test is desired, there is a tradeoff between sensitivity and specificity: as one increases the other decreases (McFall & Treat, 1999). However, for diagnostic tests, it has been argued that protection against making a Type I error or a false positive decision is preferred (Meehl & Rosen, 1955).

While there is an active debate in the field over the utility and validity of cognitive scatter analysis, there does exist a compelling body of literature that heavily warns against these practices (Daley & Nagle, 1996; Watkins, 1999, 2000, 2003, 2005). As an example, Watkins (2005) examined the diagnostic utility of four different operationalizations of WISC-III subtest scatter and found that use of any of these indices to diagnose LD resulted in chance levels of accuracy. However, previous investigations have focused more narrowly on the diagnostic utility of subtest-level scatter or the impact of scatter on the integrity of global composite scores (Freberg et al., 2008; Watkins & Glutting, 2000) from older versions of the Wechsler Scales. Since, the turn of the century, the landscape of cognitive testing has changed dramatically as a result of the rise of the Cattell-Horn-Carroll theory of cognitive abilities (CHC; Schneider & McGrew, 2012). Contemporary tests now provide users with a multitude of factor-based scores that are thought to have more clinical utility than subtest-level indices (Flanagan, Ortiz, & Alfonso, 2013). Unfortunately, the potential diagnostic utility of test scatter, as produced from factor-level scores from a contemporary measure of cognitive ability, has thus far evaded empirical scrutiny.

Purpose of the Current Investigation

To address this gap in literature, the present investigation employed diagnostic utility statistics (e.g., Kessel & Zimmerman, 1993; McFall & Treat, 1999; Swets, 1998) to determine whether significant factor-level test scatter could accurately distinguish between participants ages 6:0 to 18:11 in the Kaufman Assessment Battery for Children-Second Edition (KABC-II; Kaufman & Kaufman, 2004a) normative sample with (*n* = 107) and without (*n* = 2,116) a known LD diagnosis. Given the fact that cognitive variability is endemic in the population, it is important to determine the degree to which the distributions of cognitive attributes for clinical and non-clinical subgroups may overlap (McGill, Styck, Palomares., & Hass, 2016). As recommended by Youngstrom et al. (2015), Bayesian methods were also used to determine the degree to which test scatter improved posterior probabilities and diagnostic likelihood ratios from *a priori* base rates of the prevalence of LD in the population. It is believed that the information furnished from this investigation will be instructive for evaluating the tenability of the claims made by Hale et al. (2008) and for advancing use of the EBA model in school psychology practice.

Method

Participants

Participants were children and adolescents ages 6:0 to 18:11 (*N* = 2,223) drawn from the KABC-II standardization sample. The standardization sample was obtained using stratified proportional sampling across demographic variables of age, sex, race/ethnicity, parent educational level, and geographic region. Examination of the tables in the manual (Kaufman & Kaufman, 2004b) revealed a close correspondence to the 2001 U. S. census estimates across the stratification variables. The present sample was selected on the basis that it corresponded to the age ranges at which the CHC interpretive model could be fully specified and that it permitted analyses during the age span in which LD is most frequently diagnosed.

Measurement Instrument

The KABC-II is a multidimensional test of cognitive abilities for children and adolescents aged 3 to 18 years. The measure is comprised of 10 core subtests which contribute to the measurement of five CHC-based scale scores: Crystallized Ability (Gc), Fluid Reasoning (Gf), Visual Processing (Gv), Long-Term Storage and Retrieval (Glr), and Short-Term memory (Gsm). The core subtests are linearly combined to form a full scale Fluid-Crystallized (FCI) composite score. All scale and composite variables on the KABC-II are expressed as standard scores with a mean of 100 and a standard deviation of 15. The total norming sample (*N* = 3,025) is nationally representative based upon 2001 U.S. census estimates. Extensive normative and psychometric data can be found in the KABC-II manual (Kaufman & Kaufman, 2004b).

Data Analyses

Data analyses involved several steps. First, pairwise comparisons for all KABC-II factor scores were created in the normative dataset to determine the level of profile scatter (i.e., difference between the highest and lowest scale score as per Kaufman, Lichtenberger, Fletcher-Janzen, & Kaufman, 2005). This produced a continuous scatter index without the imposition of an arbitrary cut-point. As noted by MacCallum, Zhang, Preacher, and Rucker (2002), the problems associated with artificially dichotomizing continuous data in psychological research have long been known. Next, a one-way analysis of variance (ANOVA) was used to determine the degree to which KABC-II scores differed between normative participants with and without a known LD diagnosis. The Levine's test of homogeneity of variances was examined to test the degree to which score variances were statistically different and the Welch approximate *F* test was used to evaluate the omnibus test.

Diagnostic utility statistics (e.g., Kessel & Zimmerman, 1993) were then computed using the *Diagnostic Utility* program by Watkins (2002). A number of attributes of a test, collectively known as *diagnostic efficiency statistics*, can be derived from these numbers. Sensitivity and specificity can be combined into a single number called the likelihood ratio (LR). According to Streiner (2003), LR+ and LR- that are ≥ 1 indicate the test is not useful as a rule-in/out indictor of an attribute. Alternatively, positive predictive power and negative predictive power refer to the ratio of individuals who are correctly classified based upon the presence or absence of scatter. Because diagnostic statistics are influenced by prevalence rates, these statistics were also plotted on a receiver operating characteristic curve (ROC) and the area under the curve (AUC) was used to quantify the ROC (McFall & Treat, 1999; Swets, 1988). A nonparametric approach was used to fit the curve in SPSS version 23.0. Youngstrom (2014) recommended the following guidelines for interpreting AUC: values between 0.50 and 0.70 characterize low accuracy, values between 0.70 and 0.90 represent medium accuracy, and values between 0.90 and 1.00 characterize high accuracy.

Finally, Bayesian methods were employed to construct a probability nomogram using the EBA framework advocated by Youngstrom and colleagues (2015). This approach to diagnosis focuses on determining the probability of a child having a diagnosis. In such circumstances, Meehl (1954) encourages clinicians to “bet against the base rate.” Put another way, if the base rate of LD in the population is approximately 15%, prior to engaging in any information gathering activities, there is a 15% chance that a student who is referred for clinical assessment may have LD. What clinician’s need to know is the degree to which their assessment procedures help to improve the odds of a correct diagnosis. Probability nomograms combine base rates and the aforementioned likelihood ratios to quantify the utility of the information provided by assessment data (Markon, 2013). If the post-assessment (posterior) probability of diagnosis is meaningfully improved from the base rate (prior), then it can be concluded that procedures such as scatter analysis may be a useful in the context of the process of LD identification.

Results

KABC-II composite and index scores from the LD sample were significantly lower than the scores from participants without a known diagnosis. Table 1 contains the means and standard deviations of all of the KABC-II index and composite scores for normative participants disaggregated by LD status. One-way ANOVA indicated that these score differences were all statistically significant (*p* < .001)[2]. Standardized mean differences ranged from -0.69 to -1.14 representing moderate to large effect sizes. Interestingly, clinically significant levels of cognitive scatter were observed for both the LD (24.4) and Non-LD (25.2) groups challenging the widely held belief that significant variability is rare.

To remove the effects of prevalence, sensitivity and 1-specificity were plotted on a ROC graph to investigate accurate LD classification based upon the presence of meaningful scatter (see Figure 2). The diagonal line in the graph represents chance agreement or the diagnostic equivalent of flipping a coin. An AUC value of 0.49 resulted when all possible cut scores were used. Thus, the probability that a randomly selected participant with LD in the normative sample would have a higher level of scatter compared to a participant without LD failed to exceed chance levels (Swets, 1988; Youngstrom, 2014).

Diagnostic utility statistics for increasingly higher levels of cognitive profile scatter (10 to 30 points) are presented in Table 2. It should be noted that although it is suggested in various KABC-II interpretive resources (i.e., Kaufman, Lichtenberger, Fletcher-Janzen, & Kaufman, 2005; Lichtenberger, Sotelo-Dynega, & Kaufman, 2009) that a 23-point difference between an individual’s highest and lowest scale scores may be clinically significant, lower thresholds have been proposed in the assessment literature for other instruments. At all levels, diagnostic accuracy failed to exceed chance levels (AUCs from .47 to .51). However, negative predictive values were consistently strong suggesting that a relatively flat cognitive profile may function well as a potential *rule-out* test for the presence of LD. However, the positive predictive values at all levels of scatter (.04 to .05) were hopelessly weak indicating that scatter did not function as a

[2] Levene’s test for equality of variance were statistically significant for the Glr and FCI scores. Thus, robust tests were conducted using the Brown-Forsyth correction (Brown & Forsyth, 1974).

useful *rule-in* test for LD which is of primary concern to a clinician. As a result, the nomogram plots in Figure 3 illustrate well that scatter analysis did little to improve the probability of a correct diagnosis from the base rate prior.

Discussion

In an era that stresses evidence-based practice, there is a need to emphasize the importance of using science to guide clinical assessment activities (Hunsley & Mash, 2007). Typical assessment training and practice have not kept pace with advances in evidence-based practices in school psychology due to shortcomings in clinical judgement and literature gaps about empirically supported practices (Lilienfeld, Ammirati, & David, 2012). As a result, much of what clinicians do remains impressionistic and prone to biases that complicate decision making in the presence of diagnostic uncertainty (Watkins, 2000; Youngstrom & Van Meter, 2016). As a potential remedy, EBA has many benefits including concrete guidance on essential psychometric criteria for the appropriate use of assessment instruments and the relative value afforded by popular interpretive schemes and heuristics (i.e., profile analysis). As a consequence, the present study sought to apply the EBA approach suggested by Youngstrom et al. (2015) to better examine the degree to which scatter analysis helped to discriminate between individuals with and without a known LD diagnosis. Despite previous research suggesting that scatter analysis may have limited clinical utility (e.g., Watkins, 1999, 2005), it continues to be encouraged in many omnibus texts devoted to clinical assessment (Kaufman, Raiford, & Coalson, 2016; Sattler, 2008). As a result, the current study was conducted to investigate and determine the utility of the information yielded by these procedures. To this author’s knowledge, this is the first application of this framework to examine an applied assessment practice in school psychology.

Although ANOVA results indicated that there were significant group differences in the KABC-II scores, this outcome has little relevance at the level of the individual as cognitive profile variability is endemic in the population (McGill, Styck, Palomares, & Hass, 2016). Despite the suggestion that significant test scatter is rare and thus worthy of additional clinical consideration (e.g., Hale & Fiorello, 2004), over half of the present sample (57%) had a difference of 23 points or more between their highest and lowest scale scores on the KABC-II. Inexplicably, the level of scatter for the non-LD group was found to be slightly higher than the LD group. As noted by Watkins (2003), those asserting that scatter is clinically meaningful often fail to take into consideration that such variation is common in the population. According to Wiggins (1988), when psychologists generate hypotheses from markers that have high rates of occurrence in the population, this can result in impressions that are true of virtually *all* people of the type that is under consideration.

As a result, diagnostic utility statistics indicated that test scatter did not function as a useful indicator of LD regardless of the threshold (10 to 30 points) at which it was considered to be clinically meaningful. AUC values ranged from .46 to .51, values that reflect chance agreement[3]. Whereas, sensitivity estimates were optimal at the 10- and 15-point thresholds, these values were degraded as the level of test scatter increased. However, at the lower thresholds, the false alarm rates (1-Specificity) were significantly elevated resulting in an unacceptable level of Type I error. It is important to point out that this is the tradeoff that occurs when adopting a liberal diagnostic threshold and thus casting a wide net. Whereas more true positive cases will be

[3] AUC values below .50 indicate that a sign is less accurate than a coin flip as a classification method.

accurately identified, a high number of false positive tests will also occur. As a consequence, the PPV values and overall classification accuracy at all levels of test scatter were quite low.

According to Stuebing et al. (2012), low PPV suggest that we are unlikely to find all of the observations that truly meet the definition of LD. Furthermore, the probability nomograms (see Figure 3) revealed that the presence of scatter did not improve the posterior odds of correct LD identification, or put another way, clinicians employing these methods "will spend a great deal of time conducting assessments that have a very low probability of accurately identifying true SLD" (Kranzler et al., 2016, p. 11).

Attempts to analyze scatter on IQ tests date back to the very inception of standardized intelligence testing and are well ensconced in school and clinical psychology practice. However, despite their intuitive appeal and the degree to which they are referenced in popular interpretive handbooks and technical resources, researchers have consistently found that the claims of support for scatter analysis and profile analysis more generally greatly outstrip available research evidence (e.g., Canivez, 2013; Freberg et al., 2008; McGill, 2016; Kranzler et al., 2016; Watkins, 2000; Watkins & Glutting, 2000; Watkins, Glutting, & Youngstrom, 2005). The present results add to the growing body of scientific evidence questioning the use of these and other related techniques in clinical practice.

Limitations

As with any study, the current investigation is not without limitations that should be considered when evaluating the results. First, it should be noted that LD research is plagued by the fact that there is no acceptable diagnostic gold standard for this condition. As this is a precursor for diagnostic validity analyses, estimating the correct "hit rate" for various assessment methods will at some level always represent a best guess for LD identification. Although it is not disclosed in the KABC-II manual how the participants in the normative sample with a known LD diagnosis were identified, it can be assumed, given the time at which the sample was obtained, that it is likely that many were identified using the discrepancy model. Thus, it can be argued that the current results may be an artifact of previously preferred identification practices that may no longer be applicable. However, Peterson and Shinn (2002) demonstrated that similar children are identified as LD regardless of the method of identification that is endorsed locally due to the social pressures imposed upon examiners by the presence of low achievement. Thus, it is unlikely that utility of scatter analysis would be enhanced by the imposition of a different identification method.

Also, it should be noted that the prevalence rate in the current sample (approximately 5%) is somewhat lower than the estimated prevalence rate for LD in the population (currently estimated to fluctuate between 10% and 15%). Thus, it is possible that the current results may be an artifact of under-sampling. Finally, whereas the present results indicate that test scatter in the aggregate lacked diagnostic utility, consideration of more focal patterns of strengths and weaknesses may be more useful. However, it is worth noting that a recent empirical investigation by Kranzler et al. (2016) produced similar sensitivity and PPV estimates for individual scores using a popular version of PSW assessment in which it has been proposed that specific patterns of strengths and weaknesses in cognitive-achievement scores may be indicative of a learning disorder (Flanagan & Alfonso, 2010).

Future research applying the EBA framework to examine the utility of scatter/profile analytic techniques on other measurement instruments and clinical samples would be instructive. Given the sample used in this study, substantive inferences generated from the results are in a strict sense constrained to the KABC-II thus, it is possible that more positive results may emerge for different instruments, especially in target samples in which test scatter may be less prevalent. Nevertheless, until these results are produced in the professional literature, professionals are left with no other choice but to speculate about such outcomes.

Implications for Practice

In spite of the time and cost involved, practitioners often go to great lengths (e.g., cross-battery assessment, engaging in numerous sophisticated procedures of profile/scatter analysis) to extract information from commercial ability measures to better understand the etiology of LD. Yet, no evidence to date has conclusively demonstrated that these high inference procedures are reliable and valid. Thus, critics may argue that the present results do nothing but confirm what is already known. Unfortunately, Bray and colleagues (1998) noted previously that many practitioners appear to be unaware of this countering body of evidence.

Surveys continue to reveal that in spite of these negative findings, “practitioners—operating in the complexity of the clinical world—find considerable value in using clinical interpretive techniques such as profile analysis” (Pfeiffer et al., 2000, p. 384) and these interpretive techniques continue to be emphasized in many training programs (Decker, Hale, & Flanagan, 2013). Reconciling this discrepancy is not terribly difficult. As noted long ago by Macmann and Barnett (1997), defensible or not, the inferences generated from these procedures “reduces the perception of uncertainty, providing a means through which clinicians can ‘understand,’ that is, construct meaning or make sense of an otherwise ambiguous situation” (p. 228). To be clear, these inferences are not always wrong however practitioners should not conflate their unique capacity to speculate with the ability to accurately predict behavioral phenomena (Dawes, Faust, & Meehl, 1989; McGill & Busse, 2017).

According to Watkins, Glutting, and Youngstrom (2005), “scientific psychological practice cannot be sustained by clinical conjectures, personal anecdotes, and unverifiable personal beliefs that have consistently failed empirical validation” (p. 265). Given the primacy of assessment in our business, studies employing the EBA framework outlined here may be a particularly useful source of information for clinicians seeking to distinguish between practices that are empirically supported and those that subsist on the basis of clinical tradition.

In sum, many conventional beliefs about cognitive scatter appear to be based on mythology. Significant test scatter does not have any meaningful impact on the validity of composite scores (Freberg et al., 2008; McGill, 2016; Watkins & Glutting, 2000) and its relationship to functional outcome criteria (i.e., treatment utility) is ambiguous. The results of the current study coupled with previous research provide clear and compelling evidence that little can be said with confidence when an examinee presents with significant test scatter in their cognitive profile. Despite the genuine sense of satisfaction felt by professionals who engage in scatter analysis, practitioners are encouraged to eschew these procedures until scientific evidence is more supportive of their use. No amount of psychological detective work can overcome the psychometric difficulties associated with these idiographic assessment procedures that are now well-known (Meehl, 1978; Watkins, 2003, 2009).

Compliance with Ethical Standards

Conflict of Interest: The author declares no conflict of interest.

This article was not funded.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

Bray, M. A., Kehle, T. J., & Hintze, J. M. (1998). Profile analysis with the Wechsler Scales: Why does it persist? *School Psychology International, 19,* 209-220. doi: 10.1177/0143034398193002

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16,* 129-132. doi: 10.1080/00401706.1974

Canivez, G. L. (2013). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.). *The Oxford handbook of child psychological assessment* (pp. 84-112). New York: Oxford University Press.

Courville, T., Coalson, D. L., Kaufman, A. S., & Raiford, S. E. (2016). Does WISC-V scatter matter? In A. S. Kaufman, S. E. Raiford, & D. L. Coalson (Eds.). *Intelligent testing with the WISC-V* (pp. 209-226). Hoboken, NJ: Wiley.

Daley, C. E., & Nagle, R. J. (1996). Relevance of WISC-III indicators for the assessment of learning disabilities. *Journal of Psychoeducational Assessment, 14,* 320-333. doi: 10.1177/073428299601400401

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgement. *Science, 243,* 1668-1674. doi: 10.1126/science.2648573

Decker, S. L., Hale, J. B., & Flanagan, D. P. (2013). Professional practice issues in the assessment of cognitive functioning for educational applications. *Psychology in the Schools, 50,* 300-313. doi: 10.1002/pits.21675

Fiorello, C. A., Hale, J. B., Holdnack, J. A., Kavanaugh, J. A., Terrell, J., & Long, L. (2007). Interpreting intelligence test results for children with disabilities: Is global intelligence relevant? *Applied Neuropsychology, 14,* 2-12, doi: 10.1080/09084280701280338

Fiorello, C. A., Hale, J. B., McGrath, M., Ryan, K., & Quinn, S. (2002). IQ interpretation for children with flat and variable test profiles. *Learning and Individual Differences, 13,* 115-125. doi: 10.1016/S1041-6080(02)00075-4

Flanagan, D. P., & Alfonso, V. C. (Eds.). (2010). *Essentials of specific learning disability identification*. Hoboken NJ: Wiley.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd. ed.). Hoboken, NJ: John Wiley.

Freberg, M. E., Vandiver, B. J., Watkins, M. W., & Canivez, G. L. (2008). Significant factor score variability and the validity of the WISC-III full scale IQ in predicting later academic achievement. *Applied Neuropsychology, 15,* 131-139. doi: 10.1080/09084280802084010

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

Hale, J. B., Fiorello, C. A., Dumont, R., Willis, J. O., Rackley, C, & Elliot, C. (2008). Differential Ability Scales-Second Edition (neuro)psychological predictors of math performance for typical children and children with math disabilities. *Psychology in the Schools, 45,* 838-858. doi: 10.1002/pits.20330

Hsu, L. M. (2002). Diagnostic validity statistics and the MCMI-III. *Psychological Assessment, 14,* 410-422. doi: 10.1037/1040-3590.14.4.410

Hunsely, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3,* 29-51. doi: 10.1146/annurev.clinpsy.3.022806.091419

Jastak, J. (1949). Problems of psychometric scatter analysis. *Psychological Bulletin, 46,* 177-197.

doi: 10.1037/h0054912

Katz, D. L. (2001). *Clinical epidemiology & evidence-based medicine*. Thousand Oaks, CA: Sage.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York: John Wiley.

Kaufman, A. S., & Kaufman, N. L. (2004a). *Kaufman Assessment Battery for Children* (2nd ed.). Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (2004b). *Kaufman Assessment Battery for Children manual* (2nd ed.). Circle Pines, MN: American Guidance Service.

Kaufman, A. S., Lichtenberger, E. O., Fletcher-Janzen, E., & Kaufman, N. L. (2005). *Essentials of KABC-II assessment*. Hoboken, NJ: John Wiley.

Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (Eds.). (2016). *Intelligent testing with the WISC-V*. Hoboken, NJ: John Wiley.

Kavale, K. A. (2002). Discrepancy models in the identification of learning disability. In R. Bradley, L. Donaldson, & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 369-426). Mahwah, NJ: Erlbaum.

Kavale, K. A., & Forness, S. R. (1984). A meta-analysis of the validity of Wechsler scale profiles and recategorizations: Patterns or parodies? *Learning Disability Quarterly, 7,* 136-156. doi: 10.2307/1510314

Kessel, J. B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self-administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. *Psychological Assessment, 5,* 395-399. doi: 10.1037/1040-3590.5.4.395

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,* 124-136. doi: 10.1080/21683603.2016.1155515

Lichtenberger, E. O., Sotelo-Dynega, M., & Kaufman, A. S. (2009). The Kaufman Assessment Battery for Children-Second Edition. In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 61-94). Hoboken, NJ: John Wiley.

Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing between science pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology, 50,* 7-36. doi: 10.1016/j.jsp.2011.09.006

MacCallum, R. C., Zhang, S., Preacher, K. J., Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19-40. doi: 10.1037/1082-989X.7.1.19

Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "intelligent testing" approach to the WISC-III. *School Psychology Quarterly, 12,* 197-234. doi: 10.1037/h0088959

Markon, K. E. (2013). Information utility: Quantifying the total psychometric information provided by a measure. *Psychological Methods, 18,* 15-35. doi: 10.1037/a0030638

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8,* 290-302. doi: 10.1177/073428299000800307

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215-241. doi: 10.1146/annurev.psych.50.1.215

McGill, R. J. (2016). Invalidating the full scale IQ score in the presence of significant factor score variability: Clinical acumen or clinical illusion? *Archives of Assessment Psychology, 6 (1),* 49-79.

McGill, R. J., & Busse, R. T. (2017). When theory trumps science: A critique of the PSW model for SLD identification. *Contemporary School Psychology, 21,* 10-18. doi: 10.1007/s40688-016-0094-x

McGill, R. J., Styck, K. S., 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*, 159-170. doi: 10.1177/0731948715618504

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806-834. doi: 10.1037/0022-006X.46.4.806

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194-216. doi:: 10.1037/h0048070

Naglieri, J. A. (2000). Can profile analysis of ability test scores work? An illustration using the PASS theory and CAS with an unselected cohort. *School Psychology Quarterly, 15,* 419-433. doi: 10.1037/h0088798

Peterson, K. M., & Shinn, M. R. (2002). Severe discrepancy models: Which best explains school identification practices for learning disabilities? *School Psychology Review, 31,* 459-476. Retrieved from http://www.nasponline.org

Pfeiffer, S. I., Reddy, L. A., Kletzel, J. E., Schmelzer, E. R., & Boyer, L. M. (2000). The practitioner's view of IQ testing and profile analysis. *School Psychology Quarterly, 15,* 376-385. doi: 10.1037/h0088795

Rapaport, D., Gil, M., & Schafer, R. (1945). *Diagnostic psychological testing: The theory, statistical evaluation, and diagnostic application of a battery of tests* (Vol. 1). Chicago: Yearbook Medical.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). La Mesa, CA: Sattler Publishing.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99-144). New York: Guilford Press.

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening test. *Journal of Personality Assessment, 81,* 209-219. doi: 10.1207/S15327752JPA8103_03

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 disability. *School Psychology Review, 41,* 3-22.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293. doi: 10.1126/science.3287615

Watkins, M. W. (1999). Diagnostic utility of WISC-III subtest variability among students with learning disabilities. *Canadian Journal of School Psychology, 15,* 11-20. doi: 10.1177/082957359901500102

Watkins, M. W. (2000). Cognitive profile analysis: A shared professional myth. *School Psychology Quarterly, 15,* 465-479. doi: 10.1037/h0088802

Watkins, M. W. (2002). *Diagnostic Utility Statistics* [Computer software]. Phoenix, AZ: Ed & Psych Associates.

Watkins, M. W. (2003). IQ subtest analysis: Clinical acumen or clinical illusion? *The Scientific Review of Mental Health Practice, 2,* 118-141.

Watkins, M. W. (2005). Diagnostic validity of Wechsler subtest scatter. *Learning Disabilities: A Contemporary Journal, 3,* 20-29.

Watkins, M. W. (2009). Errors in diagnostic decision making and clinical judgment. In T. B. Gutkin & C. R. Reynolds (Eds.), *Handbook of school psychology* (4th ed.; pp. 210-229). New York, NY: Wiley.

Watkins, M. W., & Glutting, J. J. (2000). Incremental validity of WISC-III profile elevation, scatter, and shape information for predicting reading and math achievement. *Psychological Assessment, 12,* 402-408. doi: 10.1037//1040-3590.12.4.402

Watkins, M. W., Glutting, J. J., & Youngstrom, E. A. (2005). Issues in subtest profile analysis. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed.; pp. 251-268). New York: Guilford Press.

Wiggins, J. S. (1988). *Personality and prediction: Principles of personality assessment*. Malabar, FL: Krieger Publishing.

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology, 39,* 204-221. doi: 0.1093/jpepsy/jst062

Youngstrom, E. A., Choukas-Bradley, S., Calhoun, C. D., & Jensen-Doss, A. (2015). Clinical guide to the evidence-based assessment approach to diagnosis and treatment. *Cognitive and Behavioral Practice, 22*, 20-35. doi: 10.1016/j.cbpra.2013.12.005

Youngstrom, E. A., & Van Meter, A. (2016). Empirically supported assessment of children and adolescents. *Clinical Psychology: Science and Practice, 23,* 327-347. doi: 10.1111/cpsp.12172

Table 1

Means and Standard Deviations of the KABC-II Index and FCI Scores for Participants Ages 6-18 in the Normative Sample Based Upon Known LD Condition (N = 2,223)

	Non-LD		LD		
KABC-II Score	*M*	*SD*	*M*	*SD*	*d*
Short-Term Memory (Gsm)	100.7	14.8	88.9	14.2	-0.81
Visual Processing (Gv)	100.5	14.8	90.7	13.5	-0.69
Long-Term Memory (Glr)	101.0	14.9	87.4	11.8	-1.01
Fluid Reasoning (Gf)	100.6	14.7	88.8	13.9	-0.82
Crystallized Ability (Gc)	100.6	14.7	88.6	13.8	-0.84
Fluid-Crystallized Index (*g*)	100.8	14.6	85.7	11.5	-1.14

Note. KABC-II = Kaufman Assessment Battery for Children-Second Edition (Kaufman & Kaufman, 2004b); *g* = general intelligence; Differences significant at $p < .001$ for all 6 score comparisons across groups based upon known LD condition. Nevertheless, given the unequal variances for the Glr and FCI scores, the potential threat of inflated Type I error (false positive) must also be considered. Average cognitive profile scatter for the Non-LD group (25.2) and LD group (24.4) were both clinically significant.

Table 2

Diagnostic Efficiency Statistics for Different Levels of Cognitive Scatter on the KABC-II Predicting the Presence of LD

Scatter Level	AUC	Sensitivity	Specificity	LR+	LR-	PPV	NPV	DOR	IPPP	Accuracy	K
10 Points	.501	.964	.039	1.00	0.91	.050	.954	1.11	.000	.086	.000
15 Points	.511	.882	.140	1.02	0.84	.051	.957	1.22	.001	.177	.002
23 Points	.498	.576	.420	0.99	1.01	.049	.949	0.98	.000	.428	.000
30 Points	.469	.243	.696	0.80	1.08	.040	.945	0.73	-.009	.673	-.018

Note. AUC = area under curve, LR+ = likelihood ratio for positive test results LR- = likelihood ratio for negative test results, PPV = positive predictive value, NPV= negative predictive value, DOR = diagnostic odds ratio, IPPP = incremental validity of positive test diagnosis (Hsu, 2002), Accuracy = agreement/hit rate, K = kappa coefficient for chance agreement (Streiner, 2003).

Diagnostic Condition

		LD	Not LD	
Test Outcome	Significant Scatter	True Positive	False Positive *Type I Error*	**Positive Predictive Value (PPV):** Probability that scatter will be observed when LD is present.
	No Scatter	False Negative *Type II Error*	True Negative	**Negative Predictive Value (NPV):** Probability that scatter is not present when LD is not observed.
		Sensitivity: Probability that there will be scatter when LD is identified.	**Specificity:** Probability that there will not be scatter when LD is not identified.	

Prevalence Rate: Base rate of LD in the sample

Accuracy: Rate at which true positives and true negatives are correctly identified

False Alarm Rate (1-Specificity): Rate at which individuals present with scatter but do not have LD

Figure 1. Model for evaluating the degree to which scatter accurately predict learning disability (LD) within a diagnostic decision framework.

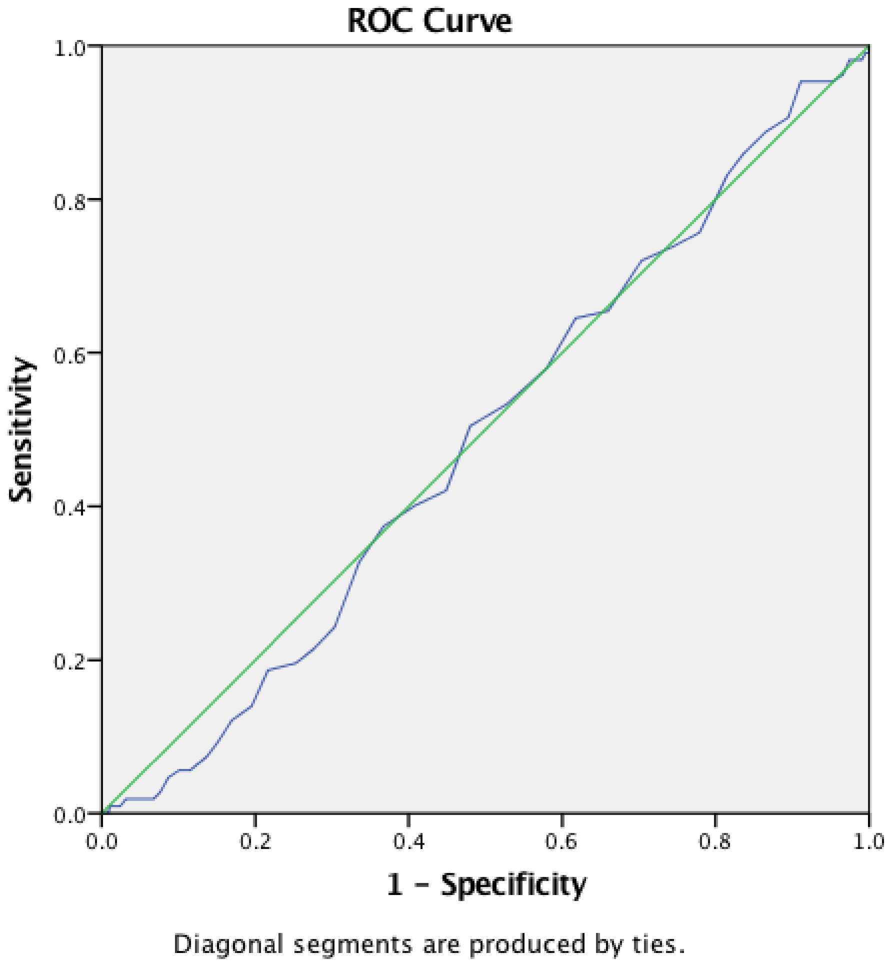
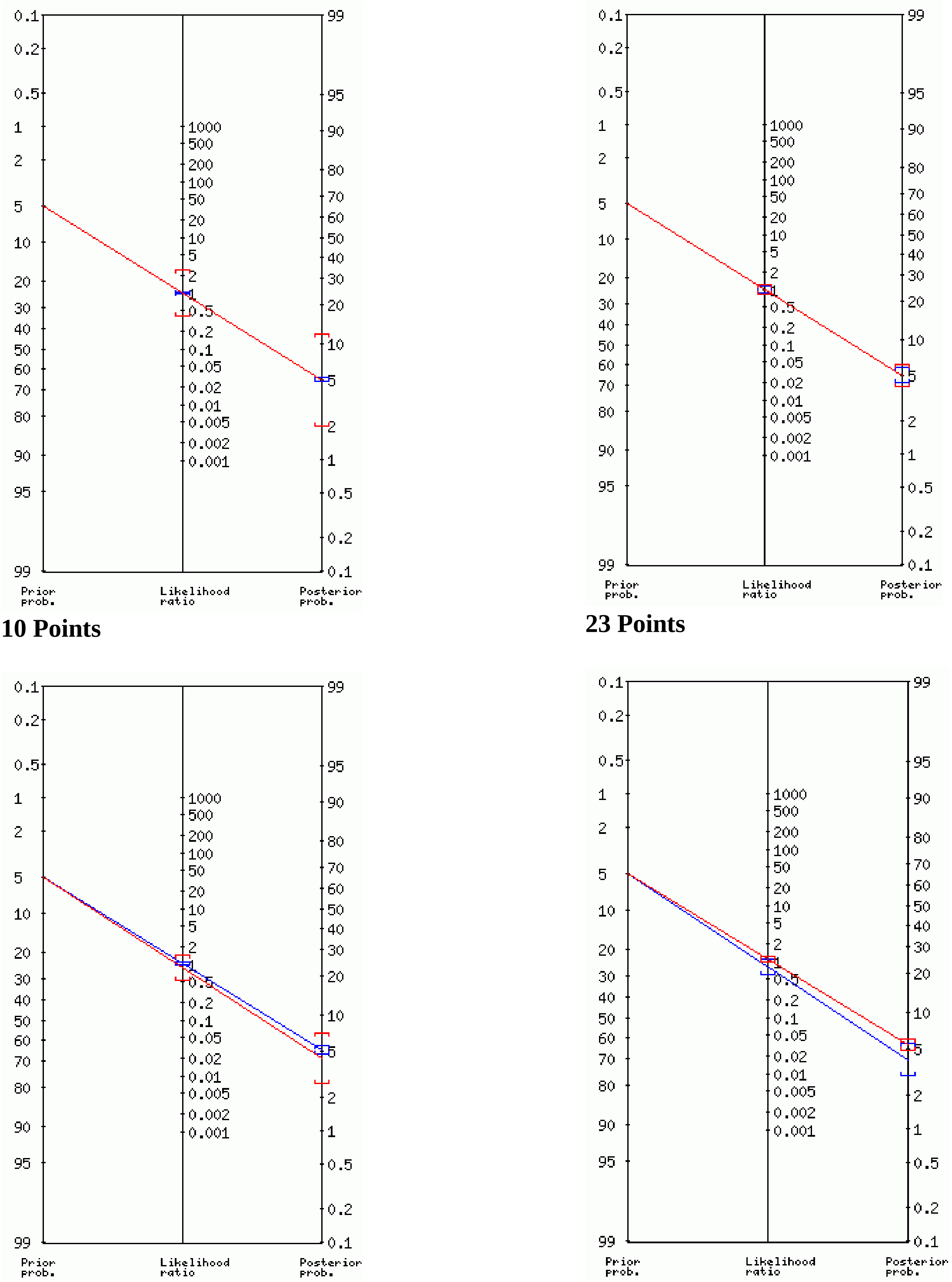


Figure 2. ROC graph illustrating the comparisons of true-positive and false-positive rates from individuals with and without LD diagnoses form the KABC-II normative sample when all possible cut scores were used.



10 Points

23 Points

15 Points

30 Points

Figure 3. Probability nomogram used to combine prior probability with likelihood ratios to estimate revised, posterior probability of LD diagnosis using different levels of cognitive profile scatter as a positive or negative test.