Journal of School Psychology 71 (2018) 108–121

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Cognitive profile analysis in school psychology: History, issues, and continued concerns

Ryan J. McGill[a,*], Stefan C. Dombrowski[b], Gary L. Canivez[c]

[a] William & Mary, United States of America
[b] Rider University, United States of America
[c] Eastern Illinois University, United States of America

ARTICLE INFO

Action Editor: F Nicholas Benson

Keywords:
Cognitive profile analysis
Intelligence testing
Evidence-based assessment

ABSTRACT

Intelligence testing remains a fixture in school psychology training and practice. Despite their popularity, the use of IQ tests is not without controversy and researchers have long debated how these measures should be interpreted with children and adolescents. A controversial aspect of this debate relates to the utility of cognitive profile analysis, a class of interpretive methods that encourage practitioners to make diagnostic decisions and/or treatment recommendations based on the strengths and weaknesses observed in ability score profiles. Whereas numerous empirical studies and reviews have challenged long-standing assumptions about the utility of these methods, much of this literature is nearly two decades old and new profile analysis methods (e.g., XBA, PSW) have been proffered. To help update the field's understanding of these issues, the present review traces the historical development of cognitive profile analysis and (re)introduces readers to a body of research evidence suggesting new and continued concerns with the use of these methods in school psychology practice. It is believed that this review will serve as a useful resource to practitioners and trainers for understanding and promoting a countering view on these matters.

1. Introduction

Researchers have long debated how cognitive measures should be interpreted in clinical practice (Fiorello et al., 2007; Watkins, 2000) with some questioning whether they should be used at all (Gresham & Witt, 1997). Further complicating the matter are the numerous interpretive systems, heuristics, and complex software programs (e.g., cross-battery assessment, ipsative assessment, levels-of-analysis approach [i.e., Intelligent Testing], X-BASS), that are available for practitioners to use; many of which encourage users to engage in some variant of cognitive profile analysis (i.e., making inferences about strengths and weaknesses observed in an individual's profile of scores). Much of the debate, and subsequent contention, hinges on the empirical veracity of these interpretative practices and the relative value of profile analysis, in general, for diagnostic activities and treatment planning.

Numerous profile analysis procedures are described in test technical manuals, clinical guidebooks (Flanagan & Alfonso, 2017; Kaufman, Raiford, & Coalson, 2016), and texts devoted to cognitive assessment (Flanagan & Harrison, 2012; Groth-Marnat & Wright, 2016; Sattler, 2008). Thus, it is not surprising that surveys (e.g., Alfonso, Oakland, LaRocca, & Spanakos, 2000; Benson, Floyd, Kranzler, Eckert, & Fefer, 2018; Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000), have long indicated that these procedures are prevalent in school psychology training and practice. However, numerous empirical studies and reviews have challenged assumptions

* Corresponding author at: William & Mary School of Education, P.O. Box 8795, Williamsburg, VA 23188, United States of America.
E-mail address: rmcgill@wm.edu (R.J. McGill).

https://doi.org/10.1016/j.jsp.2018.10.007
Received 8 April 2018; Received in revised form 13 October 2018; Accepted 29 October 2018
0022-4405/ © 2018 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.

about the utility of these methods. Yet, despite the availability of a long standing body of empirical evidence advising clinicians to "just say no" to cognitive profile analysis methods (e.g., Macmann & Barnett, 1997; Watkins, 2000; Watkins & Kush, 1994), many practitioners remain devoted to their application. For example, in a national survey of assessment practices among school psychologists (N = 938) by Benson et al. (2018), 55.2% and 49.3% of the respondents reported engaging in subtest- and composite-level profile analyses respectively.

2. Purpose of the present review

It has been nearly 20 years since the assessment literature was substantively reviewed by Watkins (2000) to determine if cognitive profile analysis was an empirically supported practice. Highlighted by theoretical and empirical advances, the landscape of cognitive testing in school psychology has changed dramatically since the publication of that seminal critique. Modern approaches to profile analysis have supplanted older questionable methodologies (e.g., subtest pattern analysis). Accordingly, it has been suggested that these newer methods of test interpretation provide users with a psychometrically defensible means for evaluating and generating inferences from score profiles (Flanagan, Ortiz, & Alfonso, 2013; Kaufman et al., 2016). Given that much of the literature cited in previous reviews is dated (published prior to 2000), it may be tempting to disregard these findings as inapplicable. Accordingly, the purpose of the present review is to again critically evaluate the use of cognitive profile analysis in school psychology. Although a review of the previous literature is provided as a contextual backdrop to current debates on these issues, particular emphasis is placed on more recent psychometric evidence. It is believed that this review will serve as a useful counterpoint to the strong claims that are made in the professional literature regarding the utility of these interpretive methods.

3. Intelligent testing and the popularization of profile analysis methods

Whereas the exact genesis of cognitive profile analysis is difficult to discern, early researchers hypothesized that subtest scatter would be a useful predictor of pathology (Harris & Shakow, 1937) and formal methods for these types of analyses have been proposed in the clinical and school psychology literatures for well over 70 years. Rapaport, Gil, and Schafer (1945) proposed a process for evaluating intraindividual cognitive scatter in a two-volume series devoted to diagnostic testing. The Rapaport et al. system involved graphically plotting subtest scores and generating hypotheses about the presence of pathology based upon the visual inspection of peaks and valleys in an examinee's profile. As tests expanded, clinicians were provided with more scores and score comparisons to interpret, and psychologists began to speculate that variability between these indicators might be a sign of neurological and behavioral dysfunction.

Later, Kaufman (1979) articulated a method he called *Intelligent Testing* (IT) that blended clinical and psychometric approaches to test interpretation. According to Kaufman et al. (2016), he was motivated by a need to "impose some empirical order on profile interpretation; to make sensible inferences from the data with full awareness of errors of measurement and to steer the field away from the psychiatric couch" (p. 7). In the IT approach, users are encouraged to interpret test scores in a step-wise fashion beginning with the FSIQ and culminating at the subtest level. However, interpretation of the FSIQ is deemphasized and practitioners are encouraged to focus most, if not all, of their interpretive weight on the scatter and elevation observed in lower-order scores (e.g., composites/indexes and subtests). In some applications of the IT levels-of-analysis approach (i.e., Sattler, 2008), practitioners are even encouraged to evaluate an examinee's performance on individual items. Inferential hypotheses are then generated from these observations as well as the qualitative behaviors observed during the test administration. It was believed that this knowledge about cognitive strengths and weaknesses within a particular IQ test would help clinicians to develop more useful diagnostic and treatment prescriptions about individuals. Although Kaufman's text outlined the application of these procedures with the WISC-R, IT was designed as a systematic approach to test interpretation that could be readily applied to any cognitive measure.

3.1. Critiques of the IT approach (1990–2000)

During the 1990s a series studies called into question core features of the IT paradigm. McDermott, Fantuzzo, and Glutting (1990) surveyed the extant literature on subtest interpretation and concluded that there was little empirical support for intradindividual or interindividual interpretation of these metrics exhorting users to "just say no" to many of the practices involving subtests that were popular at that time. McDermott, Fantuzzo, Glutting, Watkins, and Baggaley (1992) followed this paper with a study focusing more specifically on the psychometric integrity of a particular variant of subtest analysis known as *ipsative* analysis. First proposed by Cattell (1944), ipsative analysis involves subtracting obtained scores from a reference anchor, usually the mean of the profile of scores or a global composite such as the FSIQ. The resulting deviation score is then interpreted as a relative strength or weakness for an individual. As noted by McDermott et al. (1992), "These ipsatized scores hold a certain intuitive appeal because, by removing the general ability component as reflected in one's average performance level, the consequent score profile appears to isolate and amplify the pattern of abilities peculiar to the child" (p. 505). However, in a series of analyses with the WISC-R normative sample they found that ipsative scores were significantly less reliable with lower internal consistency estimates and lower stability over time compared to normative scores. They concluded that, despite their intuitive appeal, ipsative assessment failed to convey useful information to examiners, findings that were replicated in subsequent investigations with other measures (Glutting, McDermott, Konold, Snelbaker, & Watkins, 1998).

Concurrent with the publication of the WISC-III, Kaufman (1994) published a revised text, which encouraged users to make a number of inferences regarding the patterns of scores at all levels of the test. Additionally, several approaches based on different

configurations of Wechsler subtests that were thought to be useful for the diagnosis of learning disabilities were outlined and practitioners were encouraged to search for pathognomonic meaning in the patterns of scores in these groupings. For instance, Kaufman (1994) noted that individuals with disabilities tended to score lower on the subtests comprising the SCAD profile (Symbol Search [S], Coding [C], Arithmetic [A], and Digit Span [D]). Additional profiles included the ACID profile, Bannatyne pattern (Bannatyne, 1968), and the Learning Disability Index (LDI). However, subsequent research studies found that the diagnostic accuracy of all of these profiles rarely exceed chance levels rendering them unsuitable for educational decisions (e.g., Smith & Watkins, 2004; Watkins, Kush, & Glutting, 1997; Watkins, Kush, & Schaefer, 2002).

Nevertheless, Kaufman (1994) has long argued that these limitations may be managed by skilled detective work and that the system itself involves a thoughtful integration of statistical criteria and clinical acumen. Macmann and Barnett (1997) used computer simulations to measure the impact of measurement error on the reliability of interpretations at various stages of the IT on the WISC-III. Results indicated that error rates for interpretations of VIQ-PIQ differences, composite/index score differences, and subtest profile patterns were substantial. Of the sample, 62.4% presented with at least one significant strength or weakness in a breakout composite score and the base rate was even higher at the subtest level. In consideration of these data, Macmann and Barnett concluded that clinicians can expect to generate at least one *meaningful* hypothesis for an individual using the IT protocol and that there was strong potential for clinical error and confirmation bias as additional data were collected. These findings were later replicated with a clinical sample. Watkins and Canivez (2004) examined the temporal stability of WISC-III ipsative subtest and composite strengths and weaknesses and found that strengths and weaknesses were replicated across test-retest intervals at chance levels. Watkins and Canivez (2004) also found low levels of longitudinal agreement for various levels of subtest scatter, IQ score differences (i.e., VIQ-PIQ), and composite score scatter. As a consequence, they suggested that the inferences generated from these patterns of scores are likely to be unreliable and thus invalid.

Kaufman et al. (2016) argue that these critiques were unduly harsh and fail to take into consideration that IT is a systematic yet flexible approach that emphasizes clinical insight and the application of relevant theory to test interpretation. Practitioners do not blindly interpret abnormal test findings absent base rate data and other pieces of information that implicate the significance of those findings. For example, there are a number of reasons why a profile of cognitive scores may change over time and an individual may present with different strengths and weaknesses across testing sessions. Further, Kaufman and Lichtenberger (2006) argue that validity studies using group data (e.g., Macmann & Barnett, 1997) may obscure important individual differences and clinicians should use their professional judgement to discern when these results may be applicable when interpreting the assessment data.

3.2. Shared professional myth

The debate about the utility of these methods and cognitive profile analysis in general culminated in a special issue of *School Psychology Quarterly* that was devoted to the topic in 2000. Whereas a survey of 354 nationally certified school psychologists revealed that 89% of respondents regularly used subtest profile analysis in clinical practice (Pfeiffer et al., 2000) and several studies provided anecdotal information supporting the presence of unique cognitive profiles in clinical groups, Watkins (2000) articulated a countering view. In a critique entitled “Cognitive Profile Analysis: A Shared Professional Myth,” Watkins surveyed the empirical literature to date, examining the clinical utility of various forms of profile analysis and highlighted nearly 20 years of consistently negative findings. Although much of the evidence cited focused on the integrity of subtest-level profiles, additional data were presented examining the diagnostic utility of composite scores. Results indicated that neither the presence of a relative cognitive weakness, normative cognitive weakness, or a combination of cognitive and achievement weaknesses was sufficient for identifying exceptional children in the normative data for several commercial ability measures. As a result, Watkins concluded that “psychologists should eschew interpretation of cognitive test profiles and must accept that they are acting in opposition to the scientific evidence if they engage in this practice” (p. 476).

4. Cognitive profile analysis 2.0

Despite the aforementioned negative findings, the majority of test technical manuals continue to describe the step-wise interpretive procedures inspired by IT. Nevertheless, the views of IT proponents have changed over time due to the lack of evidence supporting subtest analysis and ipsative assessment and a series of new approaches have emerged that encourage practitioners to focus exclusively on the normative interpretation of composite-level scores as measures of broad cognitive abilities (i.e., processing). Given that these scores are generally more reliable than subtests, it is believed that the profiles generated from these indices will demonstrate better clinical utility than the profiles of strengths and weaknesses derived from subtests. Although partially inspired by Kaufman's work, proponents of these methods were drawn specifically to advances in theory, particularly the Cattell-Horn-Carroll theory, which has come to dominate the cognitive testing landscape in school psychology since 2001 (Ortiz & Flanagan, 2009).[1]

[1] Beyond the diagnostic utility data provided by Watkins (2000), little information was available about the usefulness of composite-level profiles at the time of his critique. Previous work focused mostly on the utility of VIQ-PIQ differences. CHC did not emerge in the professional literature until the following year (McGrew & Woodcock, 2001).

4.1. Rise of CHC theory and its impact on test development and interpretation

The Cattell-Horn-Carroll theory of cognitive abilities (CHC; Schneider & McGrew, 2012) presently guides test development and interpretation in psychology and education. It conceptualizes cognitive abilities within a hierarchical taxonomy in which elements are stratified according to breadth. Over the course of the last decade, the CHC model has served as a de facto blueprint for some of the more frequently used tests in clinical practice and it undergirds several popular interpretive systems (e.g., Profiles of Strengths and Weaknesses [PSW; Flanagan, Fiorello, & Ortiz, 2010], Cross-Battery Assessment [XBA; Flanagan et al., 2013], and the Culture-Language Interpretive Matrix [C-LIM; Rhodes, Ochoa, & Ortiz, 2005]). These contemporary approaches to profile analysis differ significantly from earlier versions. As an example, Flanagan and Kaufman (2004) articulated a revised IT-based system for use with the WISC-IV that emphasized primary interpretation of composite scores, grounding interpretation firmly in CHC theory. For the first time, practitioners were encouraged to explicitly forego subtest-level interpretation and ipsative assessment was de-emphasized.[2] Two of these methods in particular (XBA, PSW) have been widely disseminated and promoted over the last decade.

4.2. Cross-battery assessment (XBA)

According to Flanagan et al. (2013), the Cross-Battery Assessment Approach (XBA) is a systematic, theory-based method of organizing and obtaining a more complete understanding of an individual's pattern of cognitive strengths and weaknesses. Although earlier versions of the XBA method were guided by Horn and Cattell's *Gf-Gc* theory (e.g., Flanagan & McGrew, 1997; McGrew & Flanagan, 1998), it has since been revised to align with the CHC model. When XBA was formulated, there was no single test that adequately measured all of the purported broad cognitive abilities in the Horn-Cattell model or the integrated *Gf-Gc* model (McGrew, 1997) which served as a forerunner to CHC. Thus, clinicians interested in obtaining estimates for these dimensions were required to administer more than one test battery and combine ("cross") the scores for clinical interpretation. Although selective testing and "crossing" batteries was common in neuropsychological assessment, the XBA approach was unique in that it was the first system to articulate an organizing framework for interpreting scores across different tests.

The XBA approach requires users to interpret subtests and composite scores based upon their hypothesized CHC classification, regardless of the label assigned to those tests by the publisher. The approach involves a series of steps beginning with the selection of a comprehensive ability measure as a core battery in the assessment. Users are encouraged to focus primary interpretation on the norm-based composite (Stratum II) scores thought to represent CHC broad ability dimensions. However, not all composites are regarded as "pure" measures of a broad ability. If a particular broad ability is not well represented in the core battery, users may supplement it with a score from another test or through combining two or more subtests from different test batteries to construct their own narrow or broad ability composites. For the latter procedure, the revised XBA text (Flanagan et al., 2013) comes with software that generates composites based on the combination of subtests that are inputted by the user.

Although Flanagan et al. (2013) suggested that that XBA "provides practitioners with the means to make systematic, reliable, and theory-based interpretations of any ability battery" (p. 1), critiques against the method have been levied. Watkins and colleagues (Glutting, Watkins, & Youngstrom, 2003; Watkins, Glutting, & Youngstrom, 2002; Watkins, Youngstrom, & Glutting, 2002) noted several substantive concerns with the XBA approach including, but not limited to, the perceived lack of appropriate psychometric evidence to support its use and the potential vulnerability to misuse by practitioners. Later, Schneider and McGrew (2011) published a white paper critiquing the approach used to combine subtests from different batteries to create composite scores. They concluded that "Under no circumstances should averaged pseudo-composite scores be entered into equations, formulas, or procedures that involve high-stakes and important decisions regarding individuals" (p. 15).[3]

In response, Flanagan and colleagues have provided spirited defenses of their methods (e.g., Ortiz & Flanagan, 2002a, 2002b; Flanagan & Schneider, 2016). The purpose of many of these rejoinders has been to correct widespread misconceptions about their model. For instance, the XBA method has been mischaracterized by some detractors as an ipsative approach to test interpretation, and the critiques levied at the creation of composites fail to take into account that Flanagan and colleagues have deemphasized this practice in more recent writings, encouraging practitioners to rely on normative composite scores whenever possible. Nevertheless, few empirical investigations examining the integrity of XBA procedures appear to have been conducted over the course of the last 15 years. Despite this evidentiary lacuna, the XBA method has been instrumental in promoting the idea that CHC taxonomy can be used as a kind of periodic table to classify tests from different cognitive batteries (Reynolds, Keith, Flanagan, & Alfonso, 2013).

4.3. Patterns of strengths and weaknesses (PSW)

In the late 1990s researchers began to propose alternative models for specific learning disability (SLD) identification based on the observation of unique patterns of individual strengths and weaknesses, generally referred to as PSW. Since 2001, several PSW type

[2] Since 2004, these procedures have been consistently recommended in popular resources such as the *Essentials* series (Flanagan & Alfonso, 2017; Lichtenberger & Kaufman, 2013); however, subtest analysis continues to be recommend in test manuals (e.g., Wechsler, 2014) and other clinical guidebooks (e.g., Groth-Marnat & Wright, 2016; Sattler, 2008).

[3] Current versions of XBA software (Flanagan, Ortiz, and Alfonso (2017)) no longer employ the simple averaging approach. Instead, composites are generated based on the correlation between the subtests. Although this approach is more psychometrically defensible, it is important to keep in mind that these scores are not linked to established norms.

models have been proposed in the literature. These include the Concordance/Discordance Model (C/DM; Hale & Fiorello, 2004), the Discrepancy/Consistency Model (D/CM; Naglieri, 2011), and the Dual/Discrepancy Consistency Model (D/DC; Flanagan et al., 2018). Although the models differ procedurally, they all share the same core assumption that SLD is marked by unexpected underachievement and corresponding weakness in specific cognitive abilities. Closely related procedures have been articulated in the professional literature for almost 20 years (e.g., Naglieri, 1999) and multiple jurisdictions now permit use of these methods as part of their regulatory criteria for SLD identification (Maki, Floyd, & Roberson, 2015). Not surprisingly, numerous procedural materials have also been developed to help guide model implementation and interpretation of these assessment data (see Ventura County SELPA, 2017 for an example of these materials). What follows is a brief description of the application of profile analysis within the various PSW approaches.

4.3.1. Concordance/discordance model (C/DM)

The C/DM approach (Hale & Fiorello, 2004) is atheoretical and suggests that SLD is demonstrated by a pattern of concordant and discordant relationships in an ability profile. Specifically, significant differences (discordances) should be observed between a processing strength and an achievement deficit as well as that same processing strength and a relevant processing weakness. Conversely, consistency (concordance) or no significant difference should be observed between the processing weakness and the achievement weakness. Potential concordances or discordances are determined to be statistically significant if they exceed critical values obtained using a standard error of the difference (SED) formula. Additionally, users may elect to supplement C/DM with additional neuropsychological assessment procedures to help generate inferential hypotheses about an individual's cognitive functioning. For example, in a chapter describing application of C/DM within the context of a broader *cognitive hypothesis testing* framework, Fiorello and Wycoff (2018) encourage practitioners to conduct a *demands analysis* to help identify the neuropsychological processes underlying task completion on cognitive tests.

4.3.2. Discrepancy/consistency model (D/CM)

The D/CM method (Naglieri, 2011) utilizes the same conceptual approach as C/DM to examine profile variability for the presence of SLD however it employs different criteria to identify a cognitive weakness. According to Naglieri (2011), dual criteria are applied to determine whether score reflects a legitimate cognitive weakness: (a) the score must represent a relative weakness (via ipsative analysis[4]) and (b) a normative weakness (e.g., standard score < 90). If a child presents with a cognitive weakness that is related to an achievement weakness in the presence of otherwise spared abilities, that may be regarded as a confirmatory PSW pattern. Although the D/CM method can be used with any cognitive test, it is most often associated with various iterations of the Cognitive Assessment System—an instrument that is purported to measure PASS/Luria (Planning, Attention, Simultaneous, and Successive) neurocognitive processes (Naglieri & Otero, 2018).

4.3.3. Dual/discrepancy consistency model (D/DC)

The D/DC model (D/DC; Flanagan et al., 2018) was originally proposed in the XBA literature but is conceptually different from that approach. D/DC specifically uses CHC as a theoretical foundation for creating an operational definition for SLD. Clinicians are encouraged to administer cognitive measures that correspond with the broad abilities posited in the CHC model and then compare those scores to scores obtained from a norm-referenced test of achievement. Briefly, the DD/C method of SLD identification requires users to (a) identify an academic weakness, (b) determine that the academic weakness is not primarily due to exclusionary factors, (c) identify a cognitive weakness, and (d) determine whether a student displays a PSW consistent with several criteria specific to the DD/C model (see Flanagan et al., 2018 for an in-depth description of these criteria). Although normative cutoffs (e.g., standard score < 90) have been provided in the D/DC literature, clinicians may use professional judgement in determining whether a child manifests a relevant weakness in cognition or achievement. To aide decision-making, Flanagan, Ortiz, and Alfonso (2017) have developed a cross-battery assessment software system (X-BASS) that contains a PSW score analyzer inspired by CHC theory.

5. Cognitive test interpretation: acting on evidence

As a result of long-standing training gaps, recent efforts have been directed at developing and promoting evidence-based assessment (EBA) in applied psychology (e.g., Youngstrom, 2013). EBA is an approach that uses scientific research to guide the methods and measures used in clinical assessment, providing concrete guidance on essential psychometric criteria for the appropriate use of assessment instruments and the relative value afforded by popular interpretive approaches and heuristics such as profile analysis. In particular, the EBA approach emphasizes the results obtained from construct validity and diagnostic validity studies. Consistent with this framework, a critical review of the psychometric literature is provided to illustrate new and continued concerns about the potential integrity of these procedures.

5.1. Potential reliability issues with stratum II scores and score comparisons

As profile analysis involves making inferences about an individual at any one point in time, it is important to demonstrate that

[4] We encourage readers to consider the aforementioned limitations associated with these analyses (e.g., McDermott et al., 1992) when engaging in these procedures as these issues may still apply to the use of ipsative assessment on modern tests.

cognitive scores are both stable and free of unacceptable levels of systematic measurement error. In contrast to subtest-level indices, composite scores often possess desirable estimates of internal consistency (covariation among items in a test). However, given the emphasis on identification of strengths and weaknesses in many interpretive approaches, it is critically important to establish the temporal stability of these indices. Cronbach and Snow (1977) noted that, "any long-term recommendations as to a strategy for teaching a student would need to be based on aptitudes that are likely to remain stable for months, if not years" (p. 161). Although test-retest studies are often reported in test technical and interpretive manuals, these studies often provide weak evidence for the longitudinal stability of cognitive scores due to the extremely short retest intervals. Given the fact that these scores may provide the basis for high-stakes educational decisions, it is important to identify the inferences that can be made confidently from these data when making long-term (i.e., 1–3 years) prescriptive statements about individuals.

Watkins and Smith (2013) evaluated the longitudinal stability of the WISC-IV with a sample of 344 students twice evaluated for special education eligibility across a mean retest interval of 2.84 years. Stability coefficients for the WISC-IV index scores ranged from 0.65 to 0.76 and 29% to 44% of the composite scores demonstrated differences of ≥10 standard score points over this period, results consistent with previous research (e.g., Watkins & Canivez, 2004). Although we stipulate that there is no established threshold at which a profile of scores can be regarded as "stable enough," probability theory can be used to determine how useful a cognitive profile may be for particular individuals.

In an example provided by a reviewer using the corrected stability coefficients from Watkins and Smith (2013), if a child at Time 1 has a WISC-IV profile of VCI = 81, PRI = 109, WMI = 79, and PSI = 74, there is a 98% probability that the PRI will remain the highest score at Time 2 but the PRI, is not, on average, expected to deviate from the other score as strongly at time 2 as it did at Time 1 because of regression to the mean. Thus, although the PRI was the highest index score at Time 1 by 28 points, at time 2 it will by the highest by the same margin only 6% of the time. Nevertheless, there is a 32% and 77% chance that the PRI will remain the highest score by 20 and 10 points respectively. This example shows that most profiles over time will likely resemble each other to a reasonable degree. However, most identified strengths and weaknesses will not, on average, be as extreme upon reevaluation. Watkins and Canivez (2004) illustrated in a roughly three-year longitudinal stability study that WISC-III subtest strengths and weaknesses and composite scores composed of between two and five subtests produced stability no better than chance. Some might argue that over such a time interval, cognitive abilities might be changed by interventions provided to the children with disabilities but empirical research showing such powerful and lasting effects on subtest and composite scores is lacking and thus likely not a plausible explanation for these results.

These issues with long-term stability have mostly been attributed to sources of error unique to the instrument; however, a series of recent studies have documented that transient error (i.e., other sources of error that are unique to each testing situation) may also contribute to score instability. In a sample of 2783 children evaluated by 448 regional school psychologists, McDermott, Watkins, and Rhoad (2014) found that assessor bias accounted for nontrivial amounts of variation in WISC-IV scores that had nothing to do with actual individual differences. Later, Styck and Walsh (2016) conducted a meta-analysis of the prevalence of examiner errors on the Wechsler Scales and found that, on average, 15% to 77% of the primary index scores reported in protocols was changed as a result of an error(s) committed by the examiner. Whereas examiner error is only one of many potential sources of assessor bias, these results suggest that the coefficients reported in test technical manuals may actually overestimate the reliability of cognitive scores as they do not capture variance that may be due to these other sources.

A remaining issue that has not been adequately addressed in the professional literature is the psychometric problems associated with the numerous score comparisons required in many of the 2.0 era methods. As noted by Canivez (2013), the same limitations associated with pairwise comparisons of subtests *also* apply to composite scores. Nevertheless, several software programs have been developed that now provide clinicians with a means for engaging in a host of XBA and PSW score comparisons (e.g., Dehn, 2012; Flanagan et al., 2017). For example, in the X-BASS, several indices (e.g., g-value, facilitating cognitive composite [FCC], inhibiting cognitive composite [ICC]) can be computed based on the normative scores that are specified by the user and the PSW analyzer requires users to verify that there are significant discrepancies between these metrics as part of establishing a confirmatory PSW pattern. Additionally, users can elect to substitute an *Alternative Cognitive Composite* (ACC) or individual Stratum II or Stratum I (i.e., subtests) scores for these analyses based on their professional judgement. Thus, if a clinician observes a potential cognitive weakness (ICC) such as a composite standard score < 85 and uses another composite standard score of 100 for the ACC comparison, the X-BASS will generate a critical value (the value is dependent on the scores used for the comparison, but is usually around 10 points). If the observed discrepancy between the two scores (15) exceeds the critical value, this particular comparison is considered to be statistically significant ($p < .05$).

If two scores are positively correlated, the reliability of their difference is always weaker than the average of the reliability coefficients of the two scores. For example, if a clinician conducts a pairwise comparison between a composite with a reliability coefficient of 0.88 and a composite with a reliability coefficient of 0.96, and the correlation between the two measures is 0.55, the reliability of the difference score will be 0.82. The 95% confidence interval for these analyses will be relatively large (SEM = 12.47)[5] indicating that it may be possible to obtain a statistically significant value in some circumstances where the confidence band may overlap above and below the critical value. It should be noted that this is only one of several pairwise comparisons that are required in the X-BASS program where this issue may be encountered. Additionally, given the reliance on statistical formulae in many of these approaches (e.g., C/DM, D/DC), it is unclear what safeguards are provided to protect users against the inflated Type I (false positive)

[5] Other score comparisons are likely to be less precise given that the coefficients used here exceed typical estimates for many composite and subtest scores.

error that will result as a product of making multiple score comparisons. For example, McGill, Styck, Palomares, and Hass (2016) note that use of the SED formula articulated in the C/DM model may produce statistically significant results from discrepancies between tests that are relatively small (i.e., 3–5 standard score points). These issues coupled with the lack of information regarding base rates, and the absence of reliability and validity evidence for many of these indices, suggest caution in using these technologies in practice until empirical evidence is provided to support their use.

5.2. Potential validity issues with stratum II score interpretation

There are several types of evidence that are necessary to establish the validity of an instrument. Broadly, this includes evidence based on test content, internal structure (i.e., factorial validity), and relations to other variables. Whereas each of these are important in their own right, establishing factorial validity (also known as structural validity) is especially important because it is used to establish the theoretical structure of an instrument and it provides the statistical rationale for the scores that are provided to users for interpretation. That is, if a factor that is thought to represent a legitimate Stratum II dimension (e.g., Visual Processing) is not located, the score representing that construct may be illusory. This is the foundational validity measure and supports all other attempts at establishing validity.

The approach to validating the structure of an instrument is predicated upon factor analysis. There are two general classes of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Although a more in-depth discussion of these techniques is beyond the scope of the present review, it is important to note that both methods are complementary in that when EFA and CFA results align, practitioners can have greater confidence in the interpretive approach that is suggested for a test and its scores (Brown, 2015; Carroll, 1998; Gorsuch, 1983).

Since 2000, a body of independent factor analytic research has emerged raising concern about the integrity of many of the structures/interpretive models proffered by the publishers of commercial ability measures. For example, Dombrowski, McGill, and Canivez (2017) used EFA to explore the latent structure of the WJ IV Cognitive and found insufficient evidence to support the seven-factor CHC structure posited by the publisher. Application of a seven-factor extraction to the normative data produced a host of complexly determined factors that did not cohere with publisher theory (i.e., theoretically inconsistent fusion of broad abilities) and factors that were deemed inadequate (i.e., defined by fewer than two indicators) which may be symptoms of overfactoring (Frazier & Youngstrom, 2007). Conversely, support for an alternative four-factor structure resembling the familiar Wechsler framework (VC, PR, WM, PS) was found. This alternative model was also found to best fit the WJ IV normative data in a subsequent CFA investigation, raising additional questions about the legitimacy of the CHC-inspired structure (Dombrowski, McGill, & Canivez, 2018). These findings may have broader implications beyond the WJ as that instrument is likely to serve as the preeminent reference measure for making refinements to the CHC model and providing support for clinical applications of CHC such as XBA and D/DC over the course of the next decade.

Even when posited Stratum II dimensions can be located, these indices often contain insufficient unique variance for confidant clinical interpretation (Watkins, 2017). Albeit simplified, all Stratum II factor scores contain different mixtures of true score variance attributable to general intelligence (*g*) and group factors that can be sourced to the target construct, which is of primary concern to practitioners who interpret these indices in practice. If a particular measure contains insufficient group factor variance apart from *g*, it creates an interpretive confound as clinicians are not able to disentangle these effects at the level of the individual. It is this very reason that Carroll (1993, 1995) *insisted* that higher- and lower-order variances be partitioned in factor analytic studies because it helps guide appropriate interpretation of composite and subtest scores. Unfortunately, this information is rarely provided in intelligence test technical manuals.

Canivez, Watkins, and Dombrowski (2017) used CFA to examine the structural integrity of the WISC-V. Results from the 16 primary and secondary subtests did not support the five-factor CHC-based model suggested by the test publisher. Instead a four-factor structure (VC, PR, WM, PS) consistent with the WISC-IV was found to best fit the normative data. More concerning, the Canivez et al. noted a discrepancy between the degrees of freedom reported in the technical manual and the models that were estimated by the publisher. Curiously, specification of the publisher's model resulted in a model specification error suggesting an improper solution.

Fig. 1 presents the decomposed sources of variance in the 16 WISC-V subtests based on the CFA results from Canivez et al. (2017), illustrating well the pervasive influence of general intelligence across all of the WISC-V subtests (except Coding). Conversely, the broad abilities accounted for relatively meager portions of subtest variance with the exception of the subtests associated with Processing Speed. As the subtests combine to form the Stratum II composite scores, general intelligence will also saturate those scales as well. The variance partitioning results suggest that practitioners who elect to interpret the WISC-V from the standpoint of the broad ability composite scores, may be at risk of overinterpreting or misinterpreting these indices. It is important to keep in mind that this issue is not obviated if one elects to forgo interpretation of the FSIQ score. Furthermore, this scale multidimensionality also suggests that conventional reliability indices such as coefficient alpha may overestimate the “reliability” of some Stratum II scales.

Chen, Hayes, Carver, Laurenceau, and Zhang (2012) stressed that for multidimensional constructs, the alpha coefficient is complexly determined and that omega-hierarchal (ω_H) and omega-hierarchical subscale (ω_{HS}) coefficients provide better estimates of the potential interpretive relevance of composite scores. The ω_H coefficient is the model-based estimate of the proportion of variance in the unit-weighted score for the general intelligence factor with variability of group factors removed, while the ω_{HS} coefficient is the model-based estimate of the proportion of variance in the unit-weighted score for a group factor with all other group *and* general factor variance removed (Reise, 2012). Thus, these indices are more consistent with how scores may be interpreted in practice. Omega estimates may be obtained from EFA or CFA studies. In terms of interpretive relevance, Reise, Bonifay, and Haviland (2013) suggest that omega-hierarchical coefficients should exceed 0.50, but 0.75 is preferred (Reise et al., 2013). Table 1 presents the omega

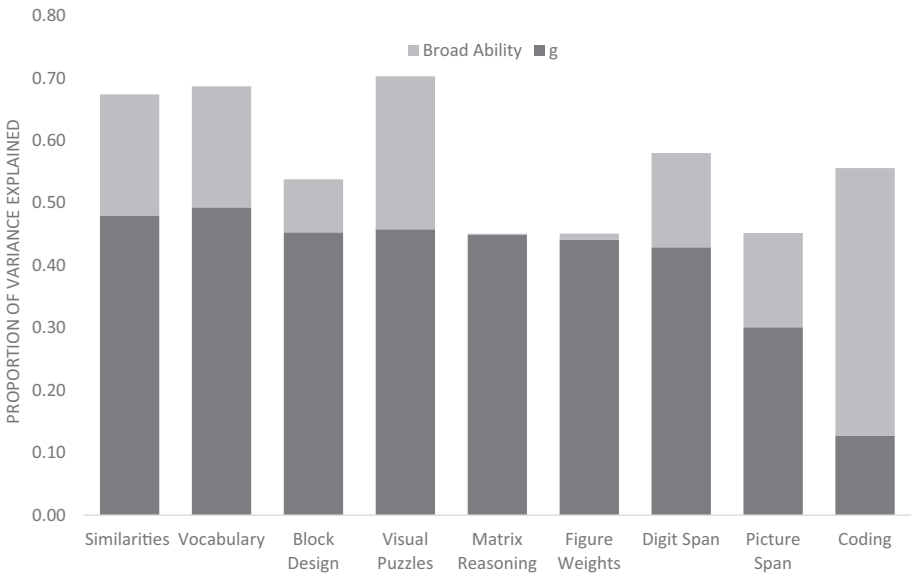

Fig. 1. Sources of variance in the WISC-V 10 primary subtests for the total standardization sample based on the CFA results by Canivez et al. (2017). *g* = general intelligence.

Table 1
Model-based reliability estimates for contemporary cognitive measures based upon suggested alignment with CHC theory.

Test	Source	Method	*g*	Gc	Gf	Ga	Gv	Gsm/wm	Glr	Gs
DAS-II	Canivez and McGill (2016)	EFA	**0.819**	0.266	0.066		0.140			
	Dombrowski, Golay, et al. (2018)	BSEM	**0.833**	0.269	0.005		0.118			
KABC-II	McGill and Dombrowski (2017)	EFA	**0.800**	0.197	*		*	0.393	0.236	
WAIS-IV	Gignac and Watkins (2013)	CFA	**0.840**	0.290	*		*	0.280		0.390
WISC-IV Spanish	McGill and Canivez (2016)	EFA	**0.833**	0.280	*		*	0.120		0.179
	McGill and Canivez (2017)	CFA	**0.870**	0.280	0.000		0.150	0.150		0.320
WISC-V	Dombrowksi, Canivez, Watkins, and Beaujean (2015)	EFA	**0.849**	**	*		*	0.190		0.162
	Canivez, Watkins, and Dombrowski (2016)	EFA	**0.828**	0.251	*		*	0.185		**0.505**
	Canivez et al. (2017)	CFA	**0.849**	0.200	*		*	0.182		**0.516**
WPPSI-IV	Watkins and Beaujean (2014)	CFA	**0.860**	0.270	0.050		0.120	0.110		0.330
WJ IV cognitive	Dombrowski, McGill and Canivez (2017)	EFA	**0.797**	**0.489**	*	**	*	0.213	*	0.355
	Dombrowski, McGill, and Canivez (2018)	CFA	**0.829**	0.373	*	*	*	0.211	*	0.282

Note. **Bold** denotes coefficients that suggest interpretation may be warranted. *g* = general intelligence, Gc = Crystallized Ability, Gf = Fluid Reasoning, Ga = Auditory Processing, Gv = Visual Processing, Gsm/wm = Short-Term Memory/Working Memory, Glr = Long-Term Storage and Retrieval, Gs = Processing Speed, BSEM = Bayesian structural equation modeling. Coefficients for general intelligence are omega hierarchical estimates. Coefficients for CHC-based factors are omega hierarchical subscale estimates.
*Dimension was located as part of a complexly determined factor, omega coefficients cannot be calculated.
**Dimension was not able to be located.

coefficients that have been reported in the assessment literature since 2014. Whereas the omega-hierarchical coefficients for general intelligence are strong suggesting that dimension can be interpreted with confidence, the omega-hierarchical subscale coefficients associated with various CHC-related broad ability indices were much weaker. Among these, only Processing Speed in the WISC-V appeared to provide enough unique measurement on its own to warrant confidant clinical interpretation.

These results are not unique to the WISC-V and similar EFA and CFA results have been observed in studies of the DAS-II (Canivez & McGill, 2016; Dombrowski, Golay, McGill, & Canivez, 2018), KABC-II (McGill & Dombrowski, 2017), WAIS-IV (Canivez & Watkins, 2010; Gignac & Watkins, 2013), WISC-IV Spanish (McGill & Canivez, 2016, 2017), WPPSI-IV (Watkins & Beaujean, 2014), SB5 (Canivez, 2008; DiStefano & Dombrowski, 2006), and WJ IV Cognitive (Dombrowski et al., 2017; Dombrowski, McGill, & Canivez, 2018). In reviewing these results it is important to keep in mind that the differences between the publisher-suggested factor structures and those reported by independent researchers are not always that substantial. For example, disagreements about what the WISC-V measures focus mostly on the inability to potentially replicate a single factor[6] (i.e., Gf) while the rest of the test was found to be

[6] Reynolds and Keith (2017) recently published a CFA of the WISC-V that supported an alternative structure featuring a previously unspecified correlated residual term between the Gf and Gv factors. However, research on other versions of the WISC-V has not supported this alternative structure (e.g., Watkins, Dombrowski, & Canivez, 2017).

structurally sound. On the other hand, evidence from independent studies of the SB5 (e.g., Canivez, 2008; DiStefano & Dombrowski, 2006) were unable to locate any of the group-specific factors posited by the test publisher. Even so, McGrew (2018) argues that these factor analytic studies are presently given too much weight in "academic school psychology" and that it is important to consider the results from other construct validity investigations (e.g., Meyer & Reynolds, 2017; Tucker-Drob, 2009) when making a determination as to whether a particular dimension is located well by a test. While it is beyond the scope of this paper to fully adjudicate this contention, it is important to keep in mind that CHC was developed largely on the basis of factor analytic modeling.

Resolving the discrepancy between these results and the results reported in technical manuals is further complicated by research practices that have been questioned in the literature. As an example, Beaujean (2016) reported that the test publisher employed a series of undisclosed constraints in order to fit a solution to the WISC-V data, potentially explaining why Canivez et al. (2017) were unable to replicate the results reported in the technical manual. This lack of disclosure is disconcerting and is a practice that at the very least should receive additional scrutiny by practitioners and trainers.

These results also have important implications for the predictive validity and explanatory validity of Stratum II scores. Although the relationships between CHC based cognitive abilities and achievement is well documented (e.g., Cormier, Bulut, McGrew, & Sing, 2017; McGrew & Wendling, 2010), many of the studies examine these effects *without* accounting for the influence of the general intelligence factor. In studies that have employed alternative designs that account for general factor effects, the predictive validity of these indices are comparatively more modest (e.g., Benson, Kranzler, & Floyd, 2016; Kranzler, Benson, & Floyd, 2015). For example, McGill and Busse (2015) examined incremental validity of WJ III Cognitive CHC composite scores in predicting achievement beyond the omnibus GIA score. Whereas the GIA accounted for 55% of the variance in Broad Reading, the CHC broad ability scores combined contributed only an additional 6% of reading achievement variance beyond *g*. However, it should be noted that there is not an established point at which small variance increments become meaningful and thus interpretation of these results is very much in the eye of beholder (Schneider, Mayer, & Newman, 2016). While some practitioners may find this additional portion of variance compelling, others may question the additional cost in time and resources that may be needed to obtain it in comparison to administering a more parsimonious test battery. As noted by Cucina and Howardson (2017), these fragile Stratum II effects have not been recognized consistently in the CHC literature.

Incremental validity studies illustrate well that some broad abilities consistently provide increments of achievement prediction beyond g that exceed zero, usually Gc, but others do not. Effects beyond prediction or explanation and more importantly, what differential treatment or instruction they might provide, is even more important for composite score interpretation, but presently lacking compelling empirical support. Other non-cognitive factors might also play a role in achievement and deserve even greater consideration in clinical decision making (i.e., motivation, persistence, curriculum, instruction). Clinicians choosing to interpret broad abilities must fully consider the empirical evidence and the potential strengths and limitations of these indices when rendering interpretive judgements about individuals (i.e., Schneider, 2013).

5.2.1. Caveat 1: potential interpretive issues for the ω_{HS} coefficient

In spite of these results, it is important to acknowledge several limitations that readers should bear in mind when interpreting the ω_{HS} coefficient. First, although ω_{HS} is frequently referred to in the literature as a "reliability" coefficient, it is actually a model-based estimate of the proportion of variance that would be obtained in a unit-weighted score composed of subtests associated with a group-specific factor. Thus, it reflects aspects of both reliability and validity and should not be regarded as a reliability coefficient in the traditional sense (Rodriguez, Reise, & Haviland, 2016). As a result, a low ω_{HS} coefficient does not imply that a target construct is not a viable psychological dimension or that the information furnished by that composite score should automatically be dismissed. Second, although embedded in the use and interpretation of ω_{HS} is the idea that there is a point at which there is "enough" unique measurement in a broad ability composite to warrant confidant clinical interpretation, the interpretive guidelines furnished by Reise et al. (2013) are subjective and a rational for these benchmarks has yet to be furnished. Relatedly, a reviewer illustrated well that even in an ideal measurement scenario in which a factor is defined by three or more salient subtest loadings, the corresponding ω_{HS} coefficient may still fail to exceed the guidelines suggested for interpretive relevance in the presence of a strong general factor. Whether this observation reflects the *actual* nature of relations between the general factor and group-specific factors or is an endemic limitation of the coefficient remains to be seen.

5.2.2. Caveat 2: clinical assessment does not occur in a vacuum

In response to this body of literature, proponents of profile analysis methods suggest that much of the work focusing on the dominance of the general factor results from methodological preferences that may stack the deck against broad cognitive abilities and fails to take into consideration that the goal of assessment is to understand how learning problems manifest in children and adolescents and to develop effective interventions—not prediction (Decker, Hale, & Flanagan, 2013). Further, Hale and Fiorello (2004) suggest that an examination of cognitive strengths and weaknesses is necessary for determining which interventions may be relevant for one profile versus another and that the construct of general intelligence has little practical relevance for treatment utility (Hale & Fiorello, 2004). However, Flanagan et al. (2013) explicitly warn against applying profile analytic methods in a cavalier fashion—-basing educational decisions solely on the mere observation of a unique PSW for an individual. Instead, practitioners are encouraged to enhance the ecological validity of assessment by integrating multiple sources of data that confirm their clinical hypotheses (e.g., Fiorello, Hale, & Wycoff, 2012). While such default statements are intuitively appealing, it remains to be seen whether they function as appropriate safeguards against the potential reliability and validity limitations raised in the present review. To date such suggestions of potential utility of unique PSWs and integration of multiple data sources confirming clinical hypotheses often do not include empirical evidence to support the practice. Thus, it may be best to consider the method as scientifically invalid until such

evidence is provided (McFall, 1991).

5.3. Diagnostic and treatment validity (utility) evidence

Although the results of reliability and validity studies are important, there is an additional research base that raises concern about the use of profile analyses in contemporary practice: diagnostic and treatment validity (utility) evidence. In essence, these types of investigations focus on the clinical *bottom line*. When asked to provide evidence to support the use of profile analysis, proponents of these approaches frequently cite studies highlighting unique cognitive PSWs (e.g., subtypes) in clinical groups. For example, Feifer, Nader, Flanagan, Fitzer, and Hicks (2014) utilized the D/DC approach to classify 283 elementary school students into five different SLD groups as well as a non-SLD comparison group. Participants were then administered a battery of cognitive-achievement tests and the achievement scores were regressed on the cognitive test scores for each group separately. Differential patterns of statistically significant relations with achievement scores were observed and this finding was interpreted as evidence of different cognitive profiles emerging based upon the presence of unique PSWs.

However, group differences are necessary but not sufficient for discriminating among individuals. In contrast, studies employing methods for estimating diagnostic accuracy (e.g., receiver operating characteristic curve [ROC] analyses) have found that (a) cognitive scatter identifies broader academic dysfunction at no better than chance levels (McGill, 2018), (b) specific cognitive weaknesses have low positive predictive values in identifying the presence of focal academic weaknesses (Kranzler, Floyd, Benson, Zaboski, & Thibodaux, 2016a), and (c) that PSW methods have low to moderate sensitivity and low positive predictive values for identifying SLD (Miciak, Taylor, Stuebing, & Fletcher, 2018; Stuebing, Fletcher, Branum-Martin, & Francis, 2012). It should also be noted that, despite intense speculation, specific cognitive profiles confirming or disconfirming the presence of a learning disorder have yet to be identified (Mather & Schneider, 2015).

Nevertheless, in a rejoinder to the results furnished by Kranzler et al. (2016a), Flanagan and Schneider (2016), argue there are many *potential* reasons why cognitive deficits may not lead to academic deficits for some individuals as "cognitive abilities are causally related to academic abilities, but the causal relationship is of moderate size, and only probabilistic, not deterministic" (p. 141). Put simply, most people with cognitive weaknesses are able to get through school just fine and most people with academic difficulties do not have a learning disorder. Yet, in the aggregate, having a cognitive weakness does increase the risk of having academic skill deficits.

Although it is frequently claimed in the professional and commercial literature that use of profile analytic methods such as PSW may be useful for diagnosis and treatment planning for individuals with academic weaknesses, a countering body of literature has emerged over the last five years documenting a host of psychometric and conceptual concerns about these methods (e.g., Miciak, Fletcher, Stuebing, Vaughn, & Tolar, 2014; Miciak, Taylor, Denton, & Fletcher, 2015; Miciak, Taylor, & Stuebing, 2016; Taylor, Miciak, Fletcher, & Francis, 2017; Williams & Miciak, 2018). Furthermore, a recent meta-analysis by Burns et al. (2016) found that the effect sizes associated with academic interventions guided by cognitive data (i.e., aptitude-by-treatment interaction [ATI]) were mostly small, with only the effects associated with interventions informed by phonological awareness providing moderate treatment effects. As a result, they concluded, "the current and previous data indicate that measures of cognitive abilities have little to no utility in screening or planning interventions for reading and mathematics" (p. 37).

PSW proponents often point to anecdotal case studies to justify the treatment validity of these procedures. For example, Fiorello, Hale, and Snyder (2006) administered a battery of cognitive tests to a child who presented with difficulties in word reading and attention. Based on the patterns of scores that was obtained, the authors recommended the consideration of a stimulant medication trial and that the child receives a phonics based reading intervention—a treatment package that, we suggest, could have been plausibly identified absent the cognitive assessment data. In summation, it is presently difficult to reconcile the laudatory ATI rhetoric with available research evidence, and this limitation has even been acknowledged by some proponents of these methods. As noted by Schneider and Kaufman (2017), "After rereading dozens of papers defending such assertions, including our own, we can say that this position is mostly backed by rhetoric in which assertions are backed by citations of other scholars making assertions backed by citations of still other scholars making assertions" (p. 8).

6. Conclusion

Since the formal inception of the field, numerous methods for cognitive profile analysis have been articulated in the school psychology literature and the dissemination of these methods in clinical training and practice continues to be widespread. As an example, in a recent survey by Benson et al. (2018), over 50% of practitioners reported using some variant of profile analysis on a routine basis. However, clinical tradition should not be confused with clinical validation (Lilienfeld, Wood, & Garb, 2007). The present review illustrates well that a host of psychometric and conceptual concerns have been raised in the professional literature regarding these methods for well over 30 years and these concerns have implications for every variation of these techniques. It is important to recognize that new interpretive approaches such as XBA and PSW, in its various forms, may simply be a re-parameterization of previous practices. That is, although these methods have more focal clinical applications and, as a result, their own unique strengths and limitations apart from the broader issues that have long been associated with profile analysis methods in general, these additional concerns may also complicate the use of these modern approaches to test interpretations. Clinicians and trainers must seriously consider the present information to make informed decisions about which interpretive methods/procedures have satisfactory replicated evidence to be used in practice.

However, making a determination about whether a particular interpretive approach is *research-based* is complicated as *research-*

based is a loaded term that does not always equate to *empirically-based*. School psychologists frequently receive information from books, book chapters, workshops, and advertisements that discuss the merits of these techniques and there are now several certification programs dedicated to advancing profile-analytic practices. Unfortunately, the issues raised in the present review are rarely discussed or presented in these venues. For example, in one particular PSW procedural manual (i.e., Ventura County SELPA, 2017), the reliability and validity issues outlined here are not discussed nor is any of the empirical literature related to these issues cited. Whereas Fiorello, Flanagan, and Hale (2014) suggest that these methods are "empirically supported" (p. 55), other scholars have raised serious questions about the evidential quality of this literature (e.g., Kranzler, Floyd, Benson, Zaboski, & Thibodaux, 2016b). For example, Hale et al. (2010) authored, what purported to be an expert consensus paper on SLD identification. Later, a response paper (Consortium for Evidence-Based Early Intervention Practices, 2010) called attention to the fact that approximately 73% of the works that were cited were non-empirical commentary articles, chapters, and books written by one or more of the authors. As a result, eminent assessment scholars have recently begun to question the direction that the field appears to be heading. For example, Naglieri (as cited in Kaufman et al., 2016) responded as follows when asked to summarize his perceptions about the state of "Intelligent Testing" in the field:

> I am convinced by recent papers showing WJ is a one-factor test, the lack of evidence for cross battery assessment, the expansion of the number of subtests in WISC-V, the over-emphasis on subtest analysis, the illusion that subtests can be combined across tests without regard for differences in standardization samples, the view that someone can look at a test and decide what it measures, the misuse of factor analysis of WISC subtests old and new to decide the structure of intelligence, and the over-interpretation of tests from a 'neuropsychological' perspective that our field has gone down a path that will not help children (p. 19).

Given these complexities, it is imperative that practitioners develop a skill set that helps them to discern when claims made in the assessment literature are credible and when potential conflicts of interest may or may not be present (Lilienfeld, Ammirati, & David, 2012; Schmidt, 2017; Truscott, Baumgart, & Rogers, 2004).

6.1. Implications for practice

As noted by Fletcher, Stuebing, Morris and Lyon (2013), "It is ironic that methods of this sort continue to be proposed when the basic psychometric issues are well understood and have been documented for many years" (p. 40). However, many of these investigations were conducted more than two decades ago; thus, it is possible that many practitioners (and trainers) may not be sufficiently aware of the concerns that have been raised in the empirical literature that may have implications for the use of these methods. As the lessons learned from the past fade, there may even be a temptation to engage in previously disavowed practices (i.e., subtest pattern analyses) As a result, it is important for clinicians to be cognizant of the accumulated body of countering evidence and the need for proponents to establish the reliability, validity, and diagnostic utility of these techniques *before* using such practices in clinical assessment (Glutting et al., 2003; McFall, 1991).

In closing, we recognize that school psychologists are always seeking better and sound methods to identify and help at-risk and children and adolescents. While cognitive profile analysis procedures are intuitively appealing and there have been *some* incremental advances in the theoretical and conceptual development of newer variations of these methods over the course of the last decade, replicated empirical evidence for the reliability, validity, diagnostic utility, and treatment utility of these methods remains less than compelling. As a result, despite the *perceived* value of the information afforded by these assessment practices, the bulk of available empirical evidence continues to support the recommendation against using cognitive profile analysis as a focal point for diagnostic and treatment decisions in clinical practice (Fletcher & Miciak, 2017).

References

Alfonso, V. C., Oakland, T. D., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review, 29*, 52–64. Retrieved from http://www.nasponline.org.

Bannatyne, A. (1968). Diagnosing learning disabilities and writing remedial prescriptions. *Journal of Learning Disabilities, 1*, 242–249. https://doi.org/10.1177/002221946800100403.

Beaujean, A. A. (2016). Reproducing the Wechsler intelligence scale for children-fifth edition: Factor model results. *Journal of Psychoeducational Assessment, 34*, 404–408. https://doi.org/10.1177/0734282916642679.

Benson, N., Floyd, R. G., Kranzler, J. H., Eckert, T. L., & Fefer, S. (2018, February). Contemporary assessment practices in school psychology: National survey results. *Paper presented at the meeting of the National Association of School Psychologists, Chicago, IL*.

Benson, N. F., Kranzler, J. H., & Floyd, R. G. (2016). Examining the integrity of measurement of cognitive abilities in the prediction of achievement: Comparisons and contrasts across variables from higher-order and bifactor models. *Journal of School Psychology, 58*, 1–19. https://doi.org/10.1016/j.jsp.2016.06.001.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford.

Burns, M. K., Peterson-Brown, S., Haegele, K., Rodriguez, M., Schmitt, B., Cooper, M., ... Hosp, J. (2016). Meta-analysis of academic interventions derived from neuropsychological data. *School Psychology Quarterly, 31*, 28–42. https://doi.org/10.1037/spq0000117.

Canivez, G. L. (2008). Orthogonal higher order factor structure of the Stanford-Binet Intelligence Scales-Fifth Edition for children and adolescents. *School Psychology Quarterly, 23*, 533–541. https://doi.org/10.1037/a0012884.

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.

Canivez, G. L., & McGill, R. J. (2016). Factor structure of the Differential Ability Scales-Second Edition: Exploratory and hierarchical factor analyses with the core subtests. *Psychological Assessment, 28*, 1475–1488. https://doi.org/10.1037/pas0000279.

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV). Exploratory and higher order factor analyses. *Psychological Assessment, 22*, 827–836. https://doi.org/10.1037/a0020429.

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2016). Factor structure of the Wechsler Intelligence Scale for Children-Fifth Edition: Exploratory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 28*, 975–986. https://doi.org/10.1037/pas0000238.

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler Intelligence Scale for Children-Fifth Edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 29*, 458–472. https://doi.org/10.1037/pas0000358.
Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies.* New York: Cambridge University Press.
Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research, 30*, 429–452. https://doi.org/10.1207/s15327906mbr3003_6.
Carroll, J. B. (1998). Human cognitive abilities: A critique. In J. J. McArdle, & R. W. Woodcock (Eds.). *Human cognitive abilities in theory and practice* (pp. 5–23). Mahwah, NJ: Lawrence Erlbaum.
Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*, 292–303. https://doi.org/10.1037/h0057299.
Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80*, 219–251. https://doi.org/10.1111/j.1467-6494.2011.00739.x.
Consortium for Evidence-Based Early Intervention Practices (2010). A response to the Learning Disabilities Association of America white paper on specific learning disabilities (SLD) identification. Retrieved from http://www.isbe.state.il.us/speced/pdfs/LDA_SLD_white_paper_response.pdf.
Cormier, D. C., Bulut, O., McGrew, K. S., & Sing, D. (2017). Exploring the relations between Cattell-Horn-Carroll (CHC) cognitive abilities and mathematics achievement. *Applied Cognitive Psychology, 31*, 530–538. https://doi.org/10.1002/acp.3350.
Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.
Cucina, J. M., & Howardson, G. N. (2017). Woodcock-Johnson–III, Kaufman adolescent and adult intelligence test (KAIT), Kaufman assessment battery for children (KABC), and differential ability scales (DAS) support Carroll but not Cattell-horn. *Psychological Assessment, 29*, 1001–1015. https://doi.org/10.1037/pas0000389.
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. https://doi.org/10.1002/pits.
Dehn, M. J. (2012). *Psychological processing analyzer [computer software].* Sparta, WI: Schoolhouse Educational Services, LLC.
DiStefano, C., & Dombrowski, S. C. (2006). Investigating the theoretical structure of the Stanford–Binet–Fifth Edition. *Journal of Psychoeducational Assessment, 24*, 123–136. https://doi.org/10.1177/0734282905285244.
Dombrowksi, S. C., Canivez, G. L., Watkins, M. W., & Beaujean, A. A. (2015). Exploratory bifactor analysis of the Wechsler Intelligence Scale for Children-Fifth Edition with the 16 primary and secondary subtests. *Intelligence, 53*, 194–201. https://doi.org/10.1016/j.intell.2015.10.009.
Dombrowski, S. C., Golay, P., McGill, R. J., & Canivez, G. L. (2018). Investigating the theoretical structure of the DAS-II core battery at school age using Bayesian structural equation modeling. *Psychology in the Schools, 55*, 190–207. https://doi.org/10.1002/pits.22096.
Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2017). Exploratory and hierarchical factor analysis of the WJ-IV cognitive at school age. *Psychological Assessment, 29*, 394–407. https://doi.org/10.1037/pas0000350.
Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2018). An alternative conceptualization of the theoretical structure of the WJ IV cognitive at school age: A confirmatory factor analytic investigation. *Archives of Scientific Psychology, 6*, 1–13. https://doi.org/10.1037/arc0000039.
Feifer, S. G., Nader, R. G., Flanagan, D. P., Fitzer, K. R., & Hicks, K. (2014). Identifying specific reading subtypes for effective educational remediation. *Learning Disabilities: A Multidisciplinary Journal, 20*, 18–30.
Fiorello, C. A., Flanagan, D. P., & Hale, J. B. (2014). Response to the special issue: The utility of the pattern of strengths and weaknesses approach. *Learning Disabilities: A Multidisciplinary Journal, 20*, 87–91.
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. https://doi.org/10.1080/09084280701280338.
Fiorello, C. A., Hale, J. B., & Snyder, L. E. (2006). Cognitive hypothesis testing and response to intervention for children with reading problems. *Psychology in the Schools, 43*, 835–853. https://doi.org/10.1002/pits. 20192.
Fiorello, C. A., Hale, J. B., & Wycoff, K. L. (2012). Cognitive hypothesis testing. In D. P. Flanagan, & P. L. Harrison (Eds.). *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 484–496). (3rd ed.). New York, NY: Guilford.
Fiorello, C. A., & Wycoff, K. L. (2018). Cognitive hypothesis testing: Linking test results to the real world. In D. P. Flanagan, & E. M. McDonough (Eds.). *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 715–730). (4th ed.). New York: Guilford.
Flanagan, D. P., & McGrew, K. S. (1997). A cross-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.). *Contemporary intellectual assessment:Theories, tests, and issues* (pp. 314–325). New York: Guilford.
Flanagan, D. P., & Alfonso, V. C. (2017). *Essentials of WISC-IV assessment.* Hoboken, NJ: Wiley.
Flanagan, D. P., Alfonso, V. C., Sy, M. C., Mascolo, J. T., McDonough, E. M., & Ortiz, S. O. (2018). Dual discrepancy/consistency operational definition of SLD: Integrating multiple data sources and multiple data gathering methods. In V. C. Alfonso, & D. P. Flanagan (Eds.). *Essentials of specific learning disability identification* (pp. 431–476). (2nd ed.). Hoboken, NJ: Wiley.
Flanagan, D. P., Fiorello, C. A., & Ortiz, S. O. (2010). Enhancing practice through application of Cattell-Horn-Carroll theory and research: A "third-method" approach to specific learning disability identification. *Psychology in the Schools, 47*, 739–760. https://doi.org/10.1002/pits.20501.
Flanagan, D. P., & Harrison, P. L. (Eds.). (2012). *Contemporary intellectual assessment: Theories, tests, and issues*(3rd ed.). New York: Guilford.
Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of the WISC-IV.* Hoboken, NJ: Wiley.
Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed.). Hoboken, NJ: Wiley.
Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2017). *Cross-battery assessment software system 2.0 (X-BASS).* Hoboken, NJ: Wiley.
Flanagan, D. P., & Schneider, W. J. (2016). Cross-battery assessment? XBA PSW? A case of mistaken identity: A commentary on Kranzler and colleagues' "classification agreement analysis of cross-battery assessment in the identification of specific learning disorders in children and youth". *International Journal of School and Educational Psychology, 4*, 137–145. https://doi.org/10.1080/21683603.2016.1192852.
Fletcher, J. M., & Miciak, J. (2017). Comprehensive cognitive assessments are not necessary for the identification and treatment of learning disabilities. *Archives of Clinical Neuropsychology, 32*, 2–7. https://doi.org/10.1093/arclin/acw103.
Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence, 35*, 169–182. https://doi.org/10.1016/j.intell.2006.07.002.
Fletcher, J. M., Stuebing, K. K., Morris, R. D., & Lyon, G. R. (2013). Classification and definition of learning disabilities: A hybrid model. In H. L. Swanson, & K. R. Harris (Eds.). *Handbook of learning disabilities* (pp. 33–50). (2nd ed.). New York: Guilford.
Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research, 48*, 639–662. https://doi.org/10.1080/00273171.2013.804398.
Glutting, J. J., McDermott, P. A., Konold, T. R., Snelbaker, A. J., & Watkins, M. W. (1998). More ups and downs of subtest analysis: Criterion validity of the DAS with an unselected cohort. *School Psychology Review, 27*, 599–612. Retrieved from http://www.nasponline.org.
Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multifactored and cross-battery assessments: Are they worth the effort? In C. R. Reynolds, & R. W. Kamphaus (Eds.). *Handbook of psychological and educational assessment of children: Intelligence aptitude, and achievement* (pp. 343–374). (2nd ed.). New York: Guilford.
Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
Gresham, F. M., & Witt, J. C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future directions. *School Psychology Quarterly, 12*, 249–267. https://doi.org/10.1037/h0088961.
Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment* (6th ed.). Hoboken, NJ: Wiley.
Hale, J., Alfonso, V., Berninger, V., Bracken, B., Christo, C., Clark, E., ... Yalof, J. (2010). Critical issues in response-to-intervention, comprehensive evaluation, and specific learning disabilities identification and intervention: An expert white paper consensus. *Learning Disabilities Quarterly, 33*, 223–236. https://doi.org/10.1177/073194871003300310.
Hale, J. B., & Fiorello, C. A. (2004). *School neuropsychology: A practitioner's handbook.* New York: Guilford.
Harris, A. J., & Shakow, D. (1937). The clinical significance of numerical measures of scatter on the Stanford-Binet. *Psychological Bulletin, 34*, 134–150. https://doi.org/

10.1037/h0058420.
Kaufman, A. S. (1979). *Intelligent testing with the WISC-R.* New York: Wiley.
Kaufman, A. S. (1994). *Intelligent testing with the WISC-III.* New York: Wiley.
Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). Hoboken, NJ: Wiley.
Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2016). *Intelligent testing with the WISC-V.* Hoboken, NJ: Wiley.
Kranzler, J. H., Benson, N., & Floyd, R. G. (2015). Using estimated factor scores from a bifactor analysis to examine the unique effects of the latent variables measured by the WAIS-IV on academic achievement. *Psychological Assessment, 27*, 1402–1406. https://doi.org/10.1037/pas0000119.
Kranzler, J. H., Floyd, R. G., Benson, N., Zaboski, B., & Thibodaux, L. (2016a). Classification agreement analysis of cross-battery assessment in the identification of specific learning disorders in children and youth. *International Journal of School and Educational Psychology, 4*, 124–136. https://doi.org/10.1080/21683603.2016.1155515.
Kranzler, J. H., Floyd, R. G., Benson, N., Zaboski, & Thibodaux, L. (2016b). Cross-battery assessment pattern of strengths and weaknesses approach to the identification of specific learning disorders: Evidence-based practice or pseudoscience? *International Journal of School and Educational Psychology, 4*, 146–157. https://doi.org/10.1080/21683603.2016.1192855.
Lichtenberger, E. O., & Kaufman, A. S. (2013). *Essentials of WAIS-IV assessment* (2nd ed.). Hoboken, NJ: Wiley.
Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing be- tween science pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology, 50*, 7–36. https://doi.org/10.1016/j.jsp.2011.09.006.
Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2007). Why questionable psychological tests remain popular. *Scientific Review of Alternative Medicine, 10*, 6–15.
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. https://doi.org/10.1037/h0088959.
Maki, K. E., Floyd, R. G., & Roberson, T. (2015). State learning disability eligibility criteria: A comprehensive review. *School Psychology Quarterly, 30*, 457–469. https://doi.org/10.1037/spq0000109.
Mather, N., & Schneider, D. (2015). The use of intelligence tests in the diagnosis of specific reading disability. In S. Goldtstein, D. Princiotta, & J. A. Naglieri (Eds.). *Handbook of intelligence: Evolutionary theory, historical perspective, and current concepts* (pp. 415–434). New York: Springer.
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. https://doi.org/10.1177/073428299000800307.
McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *Journal of Special Education, 25*, 504–526. https://doi.org/10.1177/002246699202500407.
McDermott, P. A., Watkins, M. W., & Rhoad, A. (2014). Whose IQ is it?—Assessor bias variance in high-stakes psychological assessments. *Psychological Assessment, 26*, 207–214. https://doi.org/10.1037/a0034832.
McFall, R. M. (1991). Manifesto for a science of clinical psychology. *The Clinical Psychologist, 44*(6), 75–88.
McGill, R. J. (2018). Confronting the base rate problem: More ups and downs for scatter analysis. *Contemporary School Psychology*. https://doi.org/10.1007/s40688-017-0168-4 Advance online publication.
McGill, R. J., & Busse, R. T. (2015). Incremental validity of the WJ III COG: Limited predictive effects beyond the GIA-E. *School Psychology Quarterly, 30*, 353–365. https://doi.org/10.1037/spq0000094.
McGill, R. J., & Canivez, G. L. (2016). Orthogonal higher order structure of the WISC-IV Spanish using hierarchical exploratory factor analytic procedures. *Journal of Psychoeducational Assessment, 36*, 600–606. https://doi.org/10.1177/0734282915624293.
McGill, R. J., & Canivez, G. L. (2017). Confirmatory factor analyses of the WISC-IV Spanish core and supplemental subtests: Validation evidence of the Wechsler and CHC models. *International Journal of School and Educational Psychology*. https://doi.org/10.1080/21683603.2017.1327831 Advance online publication.
McGill, R. J., & Dombrowski, S. C. (2017). Factor structure of the CHC model for the KABC-II: Exploratory factor analyses with the 16 core and supplemental subtests. *Contemporary School Psychology*. https://doi.org/10.1007/s40688-017-0152-z Advance online publication.
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*, 159–170. https://doi.org/10.1177/0731948715618504.
McGrew, K. S. (1997). Analysis of the major intelligence test batteries according to a proposed comprehensive Gf-GC framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.). *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151–179). New York: Guilford.
McGrew, K. S. (2018, April 12). *Dr. Kevin McGrew and Updates to CHC Theory* [video webcast]. Invited podcast presentation for school psyched! Podcast presented 12 April 2018. Retrieved from https://itunes.apple.com/us/podcast/episode-64-dr-kevin-mcgrew-and-updates-to-chc-theory/id1090744241?i=1000408728620&mt=2.
McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment.* Needham Heights, MA: Allyn and Bacon.
McGrew, K. S., & Wendling, B. J. (2010). Cattell-Horn-Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools, 47*, 651–675. https://doi.org/10.1002/pits.20497.
McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock Johnson III technical manual.* Itasca, Il: Riverside Publishing.
Meyer, E. M., & Reynolds, M. R. (2017). Scores in space: Multidimensional scaling of the WISC-V. *Journal of Psychoeducational Assessment.* https://doi.org/10.1177/0734282917696935 Advance online publication.
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*, 21–37. https://doi.org/10.1037/spq0000037.
Miciak, J., Taylor, W. P., Denton, C. A., & Fletcher, J. M. (2015). The effect of achievement test selection on identification of learning disabilities within a patterns of strengths and weaknesses framework. *School Psychology Quarterly, 30*, 321–334. https://doi.org/10.1037/spq0000091.
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*, 21–33. https://doi.org/10.1177/0734282916683287.
Miciak, J., Williams, J. L., Taylor, W. P., Cirino, P. T., Fletcher, J. M., & Vaughn, S. (2016). Do patterns of strengths and weaknesses predict differential treatment response? *Journal of Educational Psychology, 108*, 898–909. https://doi.org/10.1037/edu0000096.
Naglieri, J. A. (1999). *Essentials of CAS assessment.* New York: Wiley.
Naglieri, J. A. (2011). The discrepancy/consistency approach to SLD identification using the PASS theory. In D. P. Flanagan, & V. C. Alfonso (Eds.). *Essentials of specific learning disability identification* (pp. 145–172). Hoboken, NJ: Wiley.
Naglieri, J. A., & Otero, T. M. (2018). Refining intelligence with the planning, attention, simultaneous, and successive theory of neurocognitive processes. In D. P. Flanagan, & E. M. McDonough (Eds.). *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 195–218). (4th ed.). New York: Guilford.
Ortiz, S. O., & Flanagan, D. P. (2002a). Cross-battery assessment revisited: Some cautions concerning "some cautions" (part I). *Communiqué, 30*(7), 32–34.
Ortiz, S. O., & Flanagan, D. P. (2002b). Some cautions concerning "some cautions concerning cross-battery assessment" (part II). *Communiqué, 30*(8), 36–38.
Ortiz, S. O., & Flanagan, D. P. (2009). Kaufman on theory, measurement, interpretation, and fairness: A legacy in training, practice, and research. In J. C. Kaufman (Ed.). *Intelligent testing: Integrating psychological theory and clinical practice* (pp. 99–112). New York: Cambridge University Press.
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. https://doi.org/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.
Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667–696. https://doi.org/10.1080/00273171.2012.715555.
Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*, 129–140. https://doi.org/10.1080/00223891.2012.725437.
Reynolds, M. R., Keith, T. Z., Flanagan, D. P., & Alfonso, V. C. (2013). A cross-battery, reference variable, confirmatory factor analytic investigation of the CHC taxonomy. *Journal of School Psychology, 51*, 535–555. https://doi.org/10.1016/j.jsp.2013.02.003.

Reynolds, M. R., & Keith, T. Z. (2017). Multi-group and hierarchical confirmatory factor analysis of the Wechsler Intelligence Scale for Children-Fifth Edition: What does it measure? *Intelligence, 62*, 31–47. https://doi.org/10.1016/j.intell.2017.02.005.

Rhodes, R. L., Ochoa, S. H., & Ortiz, S. O. (2005). *Assessing culturally and linguistically diverse students: A practical guide.* New York: Guilford.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: calculating and interpreting statistical indices. *Psychological Methods, 21*, 137–150. https://doi.org/10.1037/met0000045.

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

Schmidt, F. L. (2017). Beyond questionable research methods: The role of omitted relevant research in the credibility of research. *Archives of Scientific Psychology, 5*, 32–41. https://doi.org/10.1037/arc0000033.

Schneider, W. J. (2013). What if we took our models seriously? Estimating latent scores in individuals. *Journal of Psychoeducational Assessment, 31*, 186–201. https://doi.org/10.1177/0734282913478046.

Schneider, W. J., & Kaufman, A. S. (2017). Let's not do away with comprehensive cognitive assessments just yet. *Archives of Clinical Neuropsychology, 32*, 8–20. https://doi.org/10.1093/arclin/acw104.

Schneider, W. J., Mayer, J. D., & Newman, D. A. (2016). Integrating hot and cool intelligences: Thinking broadly about broad abilities. *Journal of Intelligence, 4*(1), 1–25. https://doi.org/10.3390/jintelligence4010001.

Schneider, W. J., & McGrew, K. S. (2011). *"Just say no" to averaging IQ subtest scores (applied psychometrics 101 #10)*. Institute of Applied Psychometrics. Retrieved from http://www.iqscorner.com/2011/03/iap-applied-psychometrics-101-report-10.html.

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* (pp. 99–144). (3rd ed.). New York: Guilford.

Smith, C. B., & Watkins, M. W. (2004). Diagnostic utility of the Bannatyne WISC-III pattern. *Learning Disabilities Research and Practice, 19*, 49–56. https://doi.org/10.1111/j.1540-5826.2004.00089.x.

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*, 3–22. Retrieved from http://www.nasponline.org.

Styck, K. M., & Walsh, S. M. (2016). Evaluating the prevalence and impact of examiner errors on the Wechsler scales of intelligence: A meta-analysis. *Psychological Assessment, 28*, 3–17. https://doi.org/10.1037/pas0000157.

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*, 446–457. https://doi.org/10.1037/pas0000356.

Truscott, S. D., Baumgart, M. B., & Rogers, K. B. (2004). Financial conflicts of interest in the school psychology assessment literature. *School Psychology Quarterly, 19*, 166–178. https://doi.org/10.1521/scpq.19.2.166.33311.

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology, 45*, 1097–1118. https://doi.org/10.1037/a0015864.

Ventura County SELPA (2017). The Ventura County SELPA pattern of strengths and weaknesses model for specific learning disability eligibility procedural manual. Retrieved from http://wwwventuracountyselpa.com.

Watkins, M., Glutting, J., & Youngstrom, E. (2002). Cross-battery cognitive assessment: Still concerned. *Communiqué, 31*(2), 42–44.

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

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: From alpha to omega. *The Clinical Neuropsychologist, 31*, 1113–1126. https://doi.org/10.1080/13854046.2017.1317364.

Watkins, M. W., & Beaujean, A. A. (2014). Bifactor structure of the Wechsler preschool and primary scale of intelligence-fourth edition. *School Psychology Quarterly, 29*, 52–63. https://doi.org/10.1037/spq0000038.

Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of WISC-III subtest composite: Strengths and weaknesses. *Psychological Assessment, 16*, 133–138. https://doi.org/10.1037/1040-3590.16.2.133.

Watkins, M. W., Dombrowski, S. C., & Canivez, G. L. (2017). Reliability and factorial validity of the Canadian Wechsler intelligence scale for children-fifth edition. *International Journal of School and Educational Psychology*. https://doi.org/10.1080/21683603.2017.1342580 Advance online publication.

Watkins, M. W., & Kush, J. C. (1994). Wechsler subtest analysis: The right way, the wrong way, or no way? *School Psychology Review, 23*, 640–651. Retrieved from http://www.nasponline.org.

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997). Prevalence and diagnostic utility of the WISC-III SCAD profile among children with disabilities. *School Psychology Quarterly, 12*, 235–248. https://doi.org/10.1037/h0088960.

Watkins, M. W., Kush, J. C., & Schaefer, B. A. (2002). Diagnostic utility of the WISC-III learning disability index. *Journal of Learning Disabilities, 35*, 98–103. https://doi.org/10.1177/002221940203500201.

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler intelligence scale for children-fourth edition. *Psychological Assessment, 25*, 477–483. https://doi.org/10.1037/a0031653.

Watkins, M. W., Youngstrom, E. A., & Glutting, J. J. (2002). Some cautions concerning cross-battery assessment. *Communiqué, 30*(5), 16–20.

Wechsler, D. (2014). *Wechsler intelligence scale for children-fifth edition technical and interpretive manual.* San Antonio, TX: NCS Pearson.

Williams, J., & Miciak, J. (2018). Adopting costs associated with processing strengths and weaknesses methods for learning disabilities identification. *School Psychology Forum, 12*, 17–29. Retrieved from http://www.nasponline.org.

Youngstrom, E. A. (2013). Future directions in psychological assessment: Combining evidence-based medicine innovations with psychology's historical strengths to enhance utility. *Journal of Clinical Child and Adolescent Psychology,* (1), 139–159. https://doi.org/10.1080/15374416.2012.736358.