Archives of Assessment Psychology, Vol. 10, No. 1, (45-76)
Printed in U.S.A. All rights reserved

Interpreting Intelligence Test Scores in Forensic Mental Health Assessments: Conceptual and Psychometric Considerations for "Intelligent" Testing

James R. Andretta, Ph.D. and Ryan J. McGill, Ph.D.

Abstract

The assessment and interpretation of intelligence tests and intelligence test scores are featured in many forensic mental health assessments (FMHA). Given the role these instruments play in adjudicating legal decisions for individuals (i.e., Atkins cases), it is critical that the scores possess adequate reliability and validity. Unfortunately, a growing body of empirical literature has raised significant psychometric and conceptual concerns about the integrity and clinical utility of IQ test part- and subtest-level scores. Apprehension has also been raised regarding the use of cognitive profile analysis interpretive methods such as the popular "Intelligent Testing" (IT) framework, which emphasize primary interpretation of IQ test part- and subtest-level scores. Given the popularity of IT and other related profile analysis methods in clinical practice, concerns raised in the present review will be useful for psychologists and legal professionals tasked with evaluating the accuracy of expert witness testimony and FMHA reports featuring the interpretation of cognitive test scores. Implications and recommendations for advancing evidence-based assessment in the field of forensic psychology are discussed.

Introduction

While some attention has been given to the relevance of intelligence tests in forensic mental health assessments (FMHA), very little discussion has focused on establishing evidence-based guidelines for the interpretation of intelligence test scores and the use of various interpretive methods/systems when that relevance is certain. Where intelligence tests have been discussed by forensic assessment experts, there appears to be a consensus consistent with long-standing clinical doctrine in terms of the interpretation of IQ test scores (i.e., Kaufman, 1994; Sattler, 2018). That is, modern commercial ability measures provide users with a panoply of scores to interpret at various levels of the test. Practitioners are advised to (a) interpret the various global (i.e., FSIQ) and lower-order composite scores in a stepwise fashion, (b) evaluate for the presence of scatter within and between these indices to determine if said indices are "interpretable," and (c) attempt to derive pathognomonic meaning from the various peaks and valleys that are observed in an examinee's profile of scores (e.g., Heilbrun, DeMatteo, Holliday, & LaDuke, 2014; Melton et al., 2018). In a recent paper, Erickson, Salekin, Johnson, and Doran (2020) suggested that factor-level index scores from the WAIS-IV, particularly Verbal Comprehension and Working Memory, may be useful for predicting performance on the Standardized Assessment of Miranda Abilities (SAMA). Some have even encouraged practitioners to attempt to glean insight from performance on individual subtest-level scores in isolation and the qualitative behaviors observed during the administration of those indicators. For instance, Frumkin (2010) encouraged practitioners to consider the results from the Vocabulary subtest from the Wechsler Scales as part of a *Miranda* waiver analysis. Whereas these applications represent unique uses of individual test scores specific to FMHA, an array of systems and procedures have been developed to support clinicians as they interpret the wealth of scores provided by commercial ability measures as a matter of course. Whether a clinician elects to employ these steps selectively or in total, they all fall under the umbrella of a class of interpretive practices known as cognitive profile analysis. Although profile analysis methods and techniques may differ, they all share the same fundamental assumptions: (a) interpretation of FSIQ and the influence of general intelligence is deemphasized, and (b) primary interpretation should focus on the lower-order scores (e.g., broad ability composite and indexes, subtest-level scores) as representations of hypothesized cognitive processing abilities (McGill, Dombrowski, & Canivez, 2018). As will be discussed below, the history of profile analysis methods in applied psychology is long and not without considerable controversy (Davison & Kuang, 2000; Watkins, 2003).

The objective of the present position paper is to bring into balance the FMHA literature with the precedent evidence-base as it pertains to the use of cognitive profile analysis methods in clinical assessment. In so doing, focus was placed on the core aspects of the Intelligent Testing (IT) and cross-battery assessment (XBA) approaches that appear to be most popular among clinicians (Lockwood & Farmer, 2020; Pfeiffer et al., 2000). We set the stage with a review of the origins and applications of profile analysis. Next, we generated both a discussion on relevant case law germane to the use of intelligence tests in legal cases and a brief review of the literature on the use of intelligence tests across a variety of FMHA contexts, the frequency of intelligence testing in FMHAs, and best practices in intelligence testing as outlined in FMHA learned treatises (e.g., Heilbrun et al., 2014). We then summarized a respected body of literature which calls into question numerous aspects of popular profile analysis score interpretations and practices in general?. Finally, we use group-to-individual inference (G2i) theory to frame a conclusion on the use of intelligence test scores in FMHA, and provide recommendations for the interpretation of intelligence tests scores in both forensic report writing and testimony (i.e., framework and diagnostic) that are more consistent with the present status of the empirical literature on these matters.

Origins of Cognitive Profile Analysis and the Rise of Intelligent Testing (IT)

Although the genesis of cognitive profile analysis is difficult to discern, these techniques are not new and have been articulated in the forensic, clinical, and school psychology literatures for well over 70 years. Dating back to 1937, Harris and Shakow hypothesized that subtest scatter from the Stanford-Binet would be a useful predictor of learning difficulties. Later, Rappaport and colleagues (1945) developed diagnostic intelligence testing—a systematic multi-step approach that encouraged users to examine and compare scores from all levels of the Wechsler-Bellevue Scale and to generate diagnostic inferences from these clinical observations. Whereas the Rappaport system did not provide clinicians with any formal rules for interpreting test scores per se, practitioners were encouraged to engage in an open-ended form of subtest pattern analysis by inspecting for peaks and valleys in an examinee's score profile after visually plotting the scores. It was thought that particular patterns of peaks and valleys (i.e., idiographic strengths and weaknesses) could be linked to particular forms of pathology. That is, each form of pathology had its own cognitive profile signature and if that signature was present, it was very likely that an individual could be diagnosed with that particular condition with a high degree of accuracy (Glutting et al., 1997; McGill, 2018). More recently, Kaufman (1994) elaborated on the framework articulated by Rappaport and colleagues (1945), adding a bevy of multivariate statistical procedures for evaluating scatter and creating unique composite scores not available to users in test technical manuals. The so-called "Kaufman method," formally known as Intelligent Testing (IT), encouraged users to supplement their clinical acumen with additional psychometric approaches to test interpretation in order to generate more useful diagnostic inferences about an individual's cognitive test profile. According to Fletcher-Janzen (2009, p. 25) IT "demands that we tell a *story* [emphasis added] about an individual in the hope that the story makes sense and leads to an improvement in his or her quality of life." According to Kaufman, Raiford, and Coalson (2016), the IT approach was born out of a perceived need to "impose some empirical order on profile interpretations; 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). That is, IT was intended to be a sort of guardrail preventing clinicians from engaging in unconstrained clinical judgement, which was rampant at the time. A summary of the stepwise procedures from Lichtenberger and Kaufman (2013) for application of IT to the Wechsler Adult Intelligence Scale-Fourth Edition[1] are as follows: Step 1) report the person's FSIQ, index, and subtest scaled scores. Step 2) determine the best way to summarize overall ability (if the variability among the index scores is too great [23 points] invalidate the FSIQ and opt for the General Ability Index [GAI]). Step 3) determine if the difference between the GAI and the General Proficiency Index is unusually large (after first ascertaining whether those scores are unitary and can be interpreted in isolation). Step 4) interpret the four index scores or the five indexes from an alternative measurement model outlined in that text. Step 5) determine whether each index score is interpretable. Step 6) determine normative strengths and weaknesses in the index-level profile. Step 7) determine personal (ipsative) strengths and weaknesses in the index-level profile. Step 8) develop hypotheses about fluctuations in the person's index-level profile. Step 9) conduct planned pairwise comparisons between various subtest scores. Step 10) conduct planned pairwise comparisons between pseudo clusters[2] developed for the WAIS-IV in that text. In a departure from previous profiles analysis systems, IT encourages users to employ numerous informal *rules of thumb* to help to determine when a composite score is deemed to be interpretable or when a particular score should be supplemented with another. For example, in earlier versions of IT (e.g., Kaufman, 1994), clinicians were encouraged to interpret individual subtest-level scores if they met the following criteria: (a) the subtest must contain at least 25% specific[3] variance and (b) the level of specificity in the test must exceed the error variance for the measure. Where did this rule come from? According to Kaufman and Lichtenberger (2006), it was extrapolated from the results of a factor analysis on the Wechsler Intelligence Scale for Children by Cohen (1959) in which "Cohen *seemed* [emphasis added] to treat 25% as the amount of specific variance that is large enough (as long as it exceeds error variance) to warrant subtest specific interpretation" (p. 236). However, according to Cohen (1959), "no subtest at any age has as much as one-third of its variance attributable to specificity.

[1] The IT approach has been amended numerous times; however, a majority of the provisional steps outlined here remain popular among practitioners even if the nomenclature associated with the scores has changed or interpretive guidance has now de-emphasized some of the previous interpretive steps.

[2] A number of alternative composite scores not available in the WAIS-IV Technical Manual are provided in Lichtenberger and Kaufman (2013).

[3] Specific variance or subtest specificity refers to the proportion of reliable variance in a subtest that is unique to that subtest. It can also be thought of as what a subtest uniquely contributes to a test battery.

It is apparent that these specificities are quite inadequate to serve as a basis for subtest-specific rationale" (p. 290).

While Kaufman and colleagues have often illustrated use of IT methods with various iterations of the Wechsler Intelligence Scales, these interpretive methods can be applied to any IQ test. Thus, it is not surprising that IT or close approximations of IT (i.e., Sattler's levels-of-analysis approach [Sattler, 2018]) have been widely adopted by practitioners and trainers engaged in the clinical assessment of intelligence since their development (Dombrowski & McGill, 2019; Lockwood & Farmer, 2020; Sotelo-Dynega & Dixon, 2014) No doubt, much of the intuitive appeal of IT is the almost mystical characteristics it attributes to its users. For example, in describing core philosophical tenets of the IT approach, Fletcher-Janzen (2009) noted that it respects both the classical and romantic aspects of assessment with an overarching emphasis on clinical judgement permitting users to go *beyond* obtained test scores. To wit, "intelligent testing ascends to the concrete where all deductive and inductive judgements are guided by theory, translated by the clinician, and synthesized into an elegant whole" (p. 25). Over the course of the last 30 years, the step-by-step procedures championed by Kaufman and colleagues have become a veritable *sine qua non* for many of those engaged in the clinical assessment of intelligence worldwide. Whether referenced directly or not, various aspects of these procedures are described in virtually every test technical manual and interpretive guidebook (e.g., Flanagan & Alfonso, 2017; Sattler, 2018) that has been produced since their development. As an example, for many years Kaufman has co-edited the popular *Essentials* series published by John Wiley, and several of those volumes contain whole chapters illustrating the application of IT procedures to various tests. Furthermore, although not directly involved, proponents of more recent profile analysis systems inspired by the development of Cattell-Horn-Carroll theory (Schneider & McGrew, 2018) such as cross-battery assessment (XBA; Flanagan, Ortiz, & Alfonso, 2013) have referenced the importance of Kaufman's influence on their work (e.g., Ortiz & Flanagan, 2009).

Use of Profile Analysis Methods in Contemporary Practice

From a conceptual point of view, there is nothing inherently wrong with utilizing IT-derived profile analysis methods to aide in interpreting scores obtained within and/or between cognitive measures. However, it should be self-evident that the usefulness of those score interpretations rests on a series of fundamental psychometric assumptions, including but not limited to (a) intelligence test scales *actually* measure the abilities thought to underlie performance on the various scales, (b) if located, the scales measure those abilities with enough precision (reliability) to warrant confidant clinical interpretation, (c) if interpreting patterns of performance at any one point in time, those patterns are sufficiently stable to permit confidant clinical interpretation, and (d) unique patterns of performance observed among the scales have empirically established diagnostic and treatment implications for the test taker (Haynes, Smith, & Hundley, 2011). An identified shortcoming in any one of these areas will likely degrade the clinical utility of the inferences generated from these analyses and suggest that a clinician who engages in these analyses may be spending a great deal of time on assessment procedures from which their client is not likely to benefit (Kranzler et al., 2016; McGill, Styck, Palomares, & Hass, 2016; Nisbett, Zukier, & Lemley, 1981). Ensuring that these criteria are satisfied before engaging in profile analysis interpretations is also an ethical imperative. Lest we be accused of pontificating, the *Standards for Educational and Psychological Testing* (a.k.a., joint test standards; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) stipulate that when interpretation of profiles is suggested, relevant evidence in support of such interpretations should be provided *a priori* (Standard 1.14). Furthermore, such presentations of empirical evidence should "give due weight to all relevant findings in the scientific literature, including those inconsistent with the intended interpretation or use." (Standard 1.2, p. 23). Whereas test developers have the responsibility to furnish this information, the ultimate responsibility for evaluating the quality of the psychometric evidence provided rests with the test user. Unfortunately, nearly thirty years of negative research findings on IT and other related methods of cognitive profile analysis suggest that many, if not all, of those assumptions are presently violated (see McGill, Dombrowski, & McGill, 2018; Watkins, 2000, for authoritative reviews). Unfortunately, these research findings are rarely, if ever, cited in popular interpretive guidebooks and articles where these practices are disseminated (e.g., Groth-Marnat & Jordan Wright, 2016; Lichtenberger & Kaufman, 2013; Sattler, 2018). As a result, their potential impact on assessment practice in clinical settings has effectively been muted. Not surprisingly, contemporary surveys indicate that use of profile analysis methods in psychology training and practice is as popular as ever. For example, Lockwood and Farmer (2020) conducted a national survey of the cognitive assessment course in graduate training programs and found that nearly 70% of instructors emphasized index and composite score comparisons and over 33% continue to encourage ipsative[4] subtest-level comparisons despite compelling research evidence suggesting those procedures are psychometrically contraindicated (e.g., Macmann & Barnett, 1997). XBA and IT were the two most popular comprehensive interpretive systems that were emphasized in clinical training at 60% and 38% by the instructors that were surveyed respectively. Predictably, whereas only 38% of the clinicians surveyed in a national assessment of *actual* cognitive assessment practices reported consistently interpreting the FSIQ (10% reported never interpreting FSIQ under any circumstance), over 60% reported always interpreting broad ability scores and over 83% reported being able to translate cognitive assessment results into individualized interventions for examinees as would be suggested by conventional IT rhetoric (Sotelo-Dynega & Dixon, 2014). As noted by McGill and colleagues (2018) in a critical history of the use of cognitive profile analysis methods in school psychology, the gap between the popularity of profile analysis methods among practitioners and their scientific status in the empirical literature has likely never been greater. Although the overwhelming majority of these research findings have been produced from the school psychology literature, the advent of the evidence-based assessment (EBA; Youngstrom, Choukas-Bradley, Calhoun, & Jensen-Doss, 2015) movement in clinical psychology suggests that such discrepancies are not limited to that field and are also likely to be found in other disciplines where clinical assessment is frequently employed such, as FMHA.

Case Law on the Admissibility of Intelligence Scores and Score Interpretations

Broadly, the gatekeeping of expert evidence in legal cases, and thereby intelligence test scores and score interpretations more specifically, has been structured by two Supreme Court decisions. In *Frye v. United States*, expert evidence was limited to methods that met the standard of general acceptance in the field. The *Daubert v. Merrill Dow Pharmaceuticals, Inc*. (1993) court further narrowed the scope of expert evidence by imposing three benchmarks germane to validity and reliability: (a) the proffered theories and techniques are testable and subject to testing; (b) there

[4] In contrast to normative scores, ipsative assessment involves subtracting an observed score from a reference anchor (e.g., the mean of the profile of scores). The resulting deviation scores are interpreted as relative strengths and weaknesses for that particular examinee.

is available evidence of a tenable error rate; (c) and the theories and techniques have been subjected to some form of peer review. Both the *Frye* and *Daubert* standards have been described elsewhere in detail (see Faigman, Monahan, and Slobogin, 2014), and these discussions have included other relevant precedents (e.g., *Kumho Tire Ltd. v Carmichael*, 1999).

Turning to case law specific to intelligence test scores and score interpretations, a review of *Frye* and *Daubert* hearings indicated courts do not appear to have thoroughly vetted IT inspired test interpretations such as scatter analysis, or subtest- and index-level pattern analysis in any meaningful way. In fact, we were able to locate only one case in which the clinical interpretation of an examinee's cognitive profile was at issue in an evidentiary hearing. In *Baxter v. Temple* (2008), results of a neuropsychological evaluation were contested in a personal injury case based on the consequences of lead poisoning. Dr. Bruno–Golden employed a flexible assessment method called the Boston Process Approach (BPA) in evaluating the plaintiff. In so doing, Dr. Bruno–Golden relied upon an unplanned collection of composite and subtest scores from a variety of tests. Her goal was to assess various domains of cognitive functioning, including verbal memory, visual memory, planning, attention span, language, visual perception, academic performance, and self-control. A point of contention was that Dr. Bruno–Golden observed significant scatter between scores on two of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III) verbal subtests, and "therefore determined that she needed further verbal testing to clearly understand how the plaintiff functions in her left hemisphere" (*Baxter v. Temple*, p. 10). A 6-day *Daubert* hearing ensued before the Supreme Court of New Hampshire on neuropsychological methods. In the end, Dr. Bruno–Golden's testimony was allowed in a decision that cited American Psychological Association (APA) practice standards:

> Pursuant to the above language in the *APA Standards*, Dr. Bruno–Golden properly used literature to interpret the inter test scatter. In the context of a clinical evaluation, as opposed perhaps to a research study, Dr. Bruno–Golden did not need further literature to then validate her interpretation of the interaction between the WISC and the individual test examining the specific construct (p. 10).

The Baxter court cited three legal precedents in their allowance of Dr. Bruno–Golden's testimony. Relying on *Kumho Tire Ltd. v. Carmichael* (1999) at the national level, the Baxter court asserted the *Daubert* factors were intended to be, "helpful, not definitive" (p. 3). The court also relied upon two local legal precedents in New Hampshire specifically, where the court indicated that, "it would be unreasonable to conclude that the subject of scientific testimony must be known to certainty" (p. 3, citing *State v. Dahood*, 2002). Citing *State v. Langill* (2010), the Baxter court further opined, "for the testimony to be inadmissible, the flaws in application must so infect the procedure as to skew the methodology itself"…because, "the adversary process is available to highlight the errors and permit the fact-finder to assess the weight and credibility of the expert's conclusions" (*Baxter v. Temple,* 2008, p. 3).

Other courts have demonstrated similar levels of leniency in the admission of testimony involving intelligence test scores. In fact, we were unable to identify a case in which a court excluded intelligence test results or associated testimony based on the use of questionable assessment and/or interpretive methods. To provide a few examples of evidentiary leniency, the Supreme Court of Tennessee allowed the results of an incomplete WAIS-III administration in a non-contact room for the purposes of determining whether a petitioner was competent at the time that they elected to withdraw their motion for post-conviction relief (*Hugeley v. State,* 2011). A United States District Court in Oklahoma allowed for the analyses of a single verbal subscale score from the WAIS to establish the presence of verbal cognitive deficits in plaintiffs suing a company for damages due to lead exposure (*Palmer v. Asarco*, 2007). The Louisiana Court of Appeals permitted the state's "expert" psychologist to make a clinical determination of competence based on IQ test results that were obtained by a third-party technician, even though the witness was not present during the administration of the test and did not score the test themselves (*State v. Mullins*, 2014). A United States District Court in Georgia allowed for the interpretation of WAIS scores administered by an interpreter, even though the Technical Manual discourages such practices (*US v. Loaiza-Clavijo*, 2012). Finally, a United States District Court in Minnesota allowed for the results from an antiquated intelligence test published in the 1950s to be accepted into evidence (*Webb v. Ethicon*, 2015), even though professional standards dictate that clinicians, responsible for appraising an individual's intellectual functioning, *must* use measures with updated norms to protect against artifactual score fluctuations due to the Flynn Effect and other artifacts attributable to longitudinal measurement error (Beaujean, 2015; Gresham, 2009). In one of the more infamous examples of the *substantial* leeway afforded expert witnesses when interpreting test scores, an expert witness in *Bartlett v. New York State Board of Law Examiners* (1998) concluded that the plaintiff had a reading disability and thus was not eligible for long sought-after accommodations on the bar exam, even though they scored in the average range on reading tests that were administered by another psychologist and no substantial cognitive deficits were uncovered. In dismissing those test results, the witness suggested that it was possible that the plaintiff had developed so-called "compensatory strategies" as a result of their previous experience as a public-school teacher that could explain their intact reading scores. In finding for the appellant, it is clear that the court found the subjective impressions of the witness more persuasive than the actual test scores[5]. On the other hand, courts have not moved to suppress testimony challenging the validity of intelligence test results. For instance, in 2016, the Supreme Court of California qualified Dr. Lee Coleman to testify that, "IQ testing is not a reliable measure of intelligence" (p. 3) despite long-standing scientific consensus that such a position is not empirically supported (Brody, 1992; Deary, 2012; Hunt, 2011; Mackintosh, 2011). Put simply, these legal precedents illustrate well that judges are unlikely to engage in nuanced scrutiny of psychological assessment practices in their gatekeeping function (Gresham, 2009). Instead, the results of testimony are most likely to shape whether specific interpretive practices are given weight (Fisher, 2017). For example, Chorn and Kovera (2019) recently tested whether the reliability and validity of psychological testing underlying an expert's opinion influenced the judgements made by judges and mock jurors. Results indicated that scientific quality did not impact judges' admissibility decisions nor their perceptions of the scientific quality of the evidence presented. More concerning, informed cross-examinations did not help mock jurors better evaluate the validity and reliability of test results. These results illustrate well that it is critically important for fact finders to have a better understanding as to which scores and score interpretations have the necessary empirical support to be trusted in legal proceedings and those that should be treated with a higher level of discernment or, in some cases, disregarded entirely (Andretta, Morgan, Cantone, & Renbarger, 2019; Neal et al., 2019).

The Use of Intelligence Test Scores in Forensic Mental Health Evaluations

Beginning with relevance, *Atkins* evaluations are likely the only FMHA in which intelligence test scores are *directly* related to the psychological question at hand in legal cases:

[5] We are left to wonder if the test scores would have been given more pathognomonic weight by the expert witness for the appellant if they would have better supported the presence of a learning disability?

Does the defendant have an intellectual disability? However, even in *Atkins* evaluations, intelligence test scores are not ultimately dispositive of the legal issue in question. Instead, intelligence test scores comprise just part of a complex matrix of data (Cunningham, 2010). To provide an example, in 2010, intelligence test scores were the focal point of an *Atkins* case before the Tennessee Supreme Court: *Coleman v. State.* The Coleman court opined that the intellectual functioning of an individual can be determined by competent expert testimony about an individual's "functional intelligence quotient," (p. 2) including a quantitative estimate based on clinical expertise as opposed to *actual* test scores. In another *Atkins* case before a United States District Court in Oklahoma, it was concluded that intelligence test scores alone were not dispositive of an intellectual impairment as an intellectual disability diagnosis also requires that a significant impairment in adaptive functioning also be established during the fact-finding process (*Howell v. Workman*, 2011).

Across the various capacity related FMHAs (e.g., *Miranda*, competency, etc.), intelligence test scores are typically used to provide evidence of functional impairments that call into question one's functional legal capacity. Using a competency analysis framework, extremely low cognitive scores may be used to explain why an individual cannot benefit from oral instruction on abstract concepts (functional impairment), and would therefore be unable to appreciate instruction on why she should not meet with the district attorney without her attorney present (functional legal capacity). In terms of frequency of use, intelligence tests are among the most commonly applied assessment tools in FMHAs. To provide just a few examples, Lees-Haley, Smith, Williams, and Dunn (1996) reported that, among neuropsychologists, the Wechsler Intelligence Scales were the most commonly used assessment tool in personal injury cases. Borum and Grisso (1995) showed that both psychologists and psychiatrists frequently used intelligence tests in criminal responsibility cases (i.e., 34% and 40%, respectively). In a more recent study that included a variety of different FMHA types, the Wechsler Intelligence Scales were the second most frequently used tools across FMHAs (Archer, Buffington-Vollum, Stredny, & Handel, 2006).

Notwithstanding research indicating the frequent use of intelligence testing in forensic assessment, we were unable to locate any relevant discussion pertaining to evidence-based assessment in the FMHA literature. We were also unable to locate data on the specific interpretive practices employed by clinicians for IQ testing in FMHAs. although it stands to reason, given long-standing training patterns (e.g., Ready & Veague, 2014), that forensic clinicians likely engage in IT-inspired methods of test interpretation at a prevalence rate that is equivalent with clinical and school psychologists (i.e., Benson et al., 2019; Pfeidder et al. 2009; Sotelo-Dynega & Dixon, 2014). Guidance we were able to locate for interpreting intelligence test scores specifically in FMHAs was largely vague and declarative in nature (e.g., Heilbrun, Grisso, & Goldstein, 2009); although, some seminal FMHA texts (e.g., Heilbrun et al., 2014) feature sample reports where intelligence test scores are discussed in more depth. For instance, in writing about FMHAs broadly, Heilbrun et al. (2009) argued the following:

> If a traditional test such as the WAIS-III[6] is administered, it will provide information about some capacities (e.g., vocabulary, verbal reasoning, information processing) that is important in describing certain deficits that may interfere with the person's functional capacities to understand charges and assist counsel in her defense (p. 61).

[6] Although the WAIS-III was revised in 2008, the provisional interpretive guidance offered in most authoritative texts has not been modified.

More narrowly, in a frequently cited case-book of sample FMHA reports (Heilbrun et al., 2014), two *Miranda* waiver capacity reports include examples of how to analyze various capacities using WAIS-III scores. One report suggested focusing most of the interpretive weight on verbal comprehension scores in *Miranda* cases. To wit: Mr. Armstrong is someone who functions at the Extremely Low to Borderline range of intelligence. Although he does not meet criteria for mental retardation [sic], his Full-Scale IQ score of 70 is at the lower 2% compared to others his age. His verbal comprehension abilities, which are more relevant to one's interaction with law enforcement, are at the lower 9% range. Research is clear and showing a correlation between low intelligence and interrogative suggestibility (the extent to which an individual comes to accept information communicated during formal questioning is true). (p. 15).

Consistent with the IT approach, a second FMHA report included reporting the results of pairwise comparisons between various subscale scores: The WAIS-III consists of two major sections, one that taps verbal abilities associated with intelligence, and the second, which measures nonverbal skills associated with intellectual functioning…Mr. Lopez's scores on the nonverbal subtests are more heterogeneous. His performance on three Performance subtests reflects significant intellectual deficits, while three other subtest scores fall at or close to? the average range. Mr. Lopez's ability to acquire new perceptual learning, his ability to make use of subtle cues in order to establish cause-and-effect relationships, and his ability to concentrate on a perceptual task, would be surpassed by 95% to 98% of the general population. On a test requiring perceptual orientation skills, his score would be surpassed by 75% of the population. As measured by the WAIS-III, Mr. Lopez's ability to discriminate the essential from the unessential and to reason through a perceptual task were found to be his strongest intellectual skills. His scores on the subtests would be exceeded by 63% of the population (Heilbrun et al., 2014, pp. 28-29).

Finally, evaluation of test scatter and the identification of unique patterns of strengths and weaknesses have been detailed in an exemplar report on how to tailor FMHAs for learning disability evaluations: Sam has a number of processing strengths. He processes and retains visual information accurately, and he can demonstrate this ability as long as the assessment task requires recognition and not motoric reproduction of what he has seen. On two subtests, he accurately identified incomplete pictures (39th percentile rank) and easily recognized which pictures and a group of pictures he had seen before (99.7th percentile rank). This latter was an area of real strength and suggests that visual presentation of information could help enhance his recall of information…Sam's scores on the WJ–R Tests of Cognitive Ability fell roughly between the 15th (low average) and 85th (high average) percentiles for his age. Sam's Broad Cognitive Score (i.e., his overall score, of 17 fell in the low average range; however, this score is the result of significantly different domain scores, suggesting wide variability in his abilities (Melton et al., 2018, p. 700).

Not surprisingly, the interpretive narratives employed in these case studies are largely consistent with the steps and strategies previously outlined when describing the origins of IT. Unfortunately, there is now ample evidence to indicate that many of the popular interpretive strategies and heuristics that continue to be suggested in a majority of the professional literature presently lack sufficient reliability or validity to be used for high stakes decision-making in clinical settings (Canivez, 2013b; Freeman & Chen, 2019; Kranzler & Floyd, 2013; Youngstrom, 2008). We next survey seminal articles spanning the course of nearly 30 years and outline the long-standing limitations associated with IT and IT inspired interpretive procedures from this body of literature. As noted by Youngstrom, Choukas-Bradley, Calhoun, & Jensen-Doss (2015), thousands of articles and other evidential resources compete for our attention and clinicians do not have the time to sift through all of that information to find the one or two gems that are both scientifically valid and clinically relevant. Accordingly, it is out hope that the following may serve as a “go to” resource of sort for forensic assessment professionals and those responsible for evaluating the credibility of assessment results in legal proceedings particularly so that they may be better prepared to fully evaluate FMHAs that involve the use of cognitive profile analysis methods such as IT.

Intelligent Testing and Cognitive Profile Analysis*: Provisional Limitations*.

Many of the popular interpretive methods and heuristics employed by practitioners today were originally proffered at a time when psychologists did not have access to the technologies needed (i.e., statistical computing software) to fully investigate whether suggested scores and interpretations of scores had the psychometric integrity necessary for individual decision-making. Due to the fact that these rules often made intuitive sense, many of them were uncritically accepted and have been passed down to subsequent generations of practitioners and maintained through clinical lore (Greiffenstein, 2009; Lilienfeld, Ammirati, & David, 2012; Lilienfeld, Wood, & Garb, 2007). However, as noted by Watkins (2000), in psychological assessment, “personal explanations are an unreliable validation strategy (Ruscio, 1998) and neither popularity nor longevity of a clinical practice necessarily speaks to its verity” (p. 465). Moreover, the replication crisis in scientific psychology has illuminated the need to re-evaluate the evidence-base for widely accepted ideas and theories even if they are presently regarded within a particular field as sacrosanct (Oberauer & Lewandowsky, 2019).

Concerns Pertaining to the Interpretation of Subtest Scores

As previously mentioned, both the Rappaport diagnostic testing approach and earlier versions of IT suggest that clinicians should carefully inspect for various patterns and signs in subtest score profiles and, in some cases, interpret individual subtest scores in isolation. As a consequence, historical surveys have consistently documented the popularity of subtest-level interpretations among practitioners. For example, at the turn of the century, Pfeiffer and colleagues (2000) reported that 89% of the practitioners that were surveyed indicated they engaged in some form of subtest-level interpretation. However, such practices actually predate the Rappaport levels-of-analysis system. For example, David Wechsler was instrumental in advancing the idea that specific intellectual and psychodynamic trait functions could be assigned to Wechsler subtests (Wechsler, 1941). Nevertheless, despite a long-standing fascination with this body of doctrine, the empirical evidence for subtest analyses has consistently been negative (Bray, Kehle, & Hintze, 1998; Watkins, 2003).

Subtest Specificity. The theoretical bases for subtest pattern analysis rest on the assumption that a substantial part of a subtest's variance is associated with the specific measurement functions (i.e., traits) linked to those measures. In the 1950s, Cohen (1957, 1959) conducted a series of factor analytic studies that challenged the basis for this principle on the grounds of small specific variance for subtests from the Wechsler Scales. Subtest specificity refers to the amount of reliable variance that is unique to a particular subtest and can be calculated by subtracting a test's internal consistency reliability component from its communality, with the remainder constituting the test's specificity. According to Cohen (1959), "only this component can carry the variance necessitated by a hypothesized one-to-one correspondence of subtest to trait in any clinical rationale" (p. 290). In his analyses of the WISC, Cohen (1959) found that the overall median specificity value across the subtests was .18 and that no measure had as much as one-third of its variance attributable to specificity. He argued that specific values that low were inadequate to serve as the basis for subtest-trait linkages. Even so, Kaufman (1994) has consistently cited Cohen (1959) to support the use of a rule-of-thumb[7], widely used in the IT approach, to support the interpretation of subtest-level indicators.

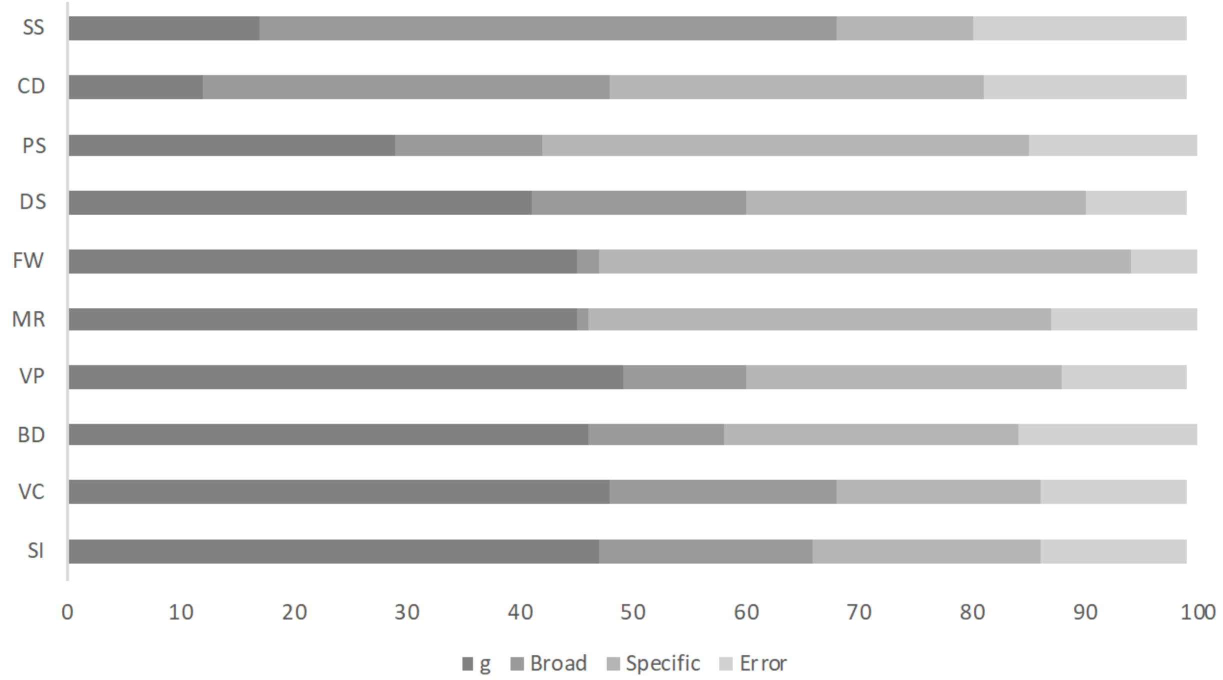

Figure 1. Decomposed sources of WISC-V core subtest variance. *g* = general intelligence, SI = Similarities, VC = Vocabulary, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, DS = Digit Span, PS = Picture Span, CD = Coding, SS = Symbol Search.

That is, a clinical hypothesis may be generated from an individual subtest if (a) at least 25% of the variance in that measure is specific and (b) the portion of specific variance exceeds the proportion

[7] This rule of thumb has also been utilized in other, non-IT, profile analysis interpretive resources (e.g., Flanagan, McGrew, & Ortiz, 2000; McGrew & Flanagan, 1998).

of error variance. Whereas we argue that this rule provides a relatively low-bar for clinical interpretation, many of the subtests from contemporary cognitive measures still fail to meet this standard. Figures 1 and 2 illustrate decomposed sources of WISC-V and WAIS-IV subtest variance based on the independent factor analytic results furnished by Canivez and Watkins (2010, 2016). Inspection of these graphic arrays indicate that, in both cases, ~50% of the core subtests from each measure are considered uninterpretable using Kaufman's guidelines and among the subtests that cleared those hurdles, *none* contained more than 47% specific variance. Thus, even in the most optimistic scenario, clinicians attempting to link Wechsler subtests to specific psychological attributes may be at-risk of overinterpreting those measures.

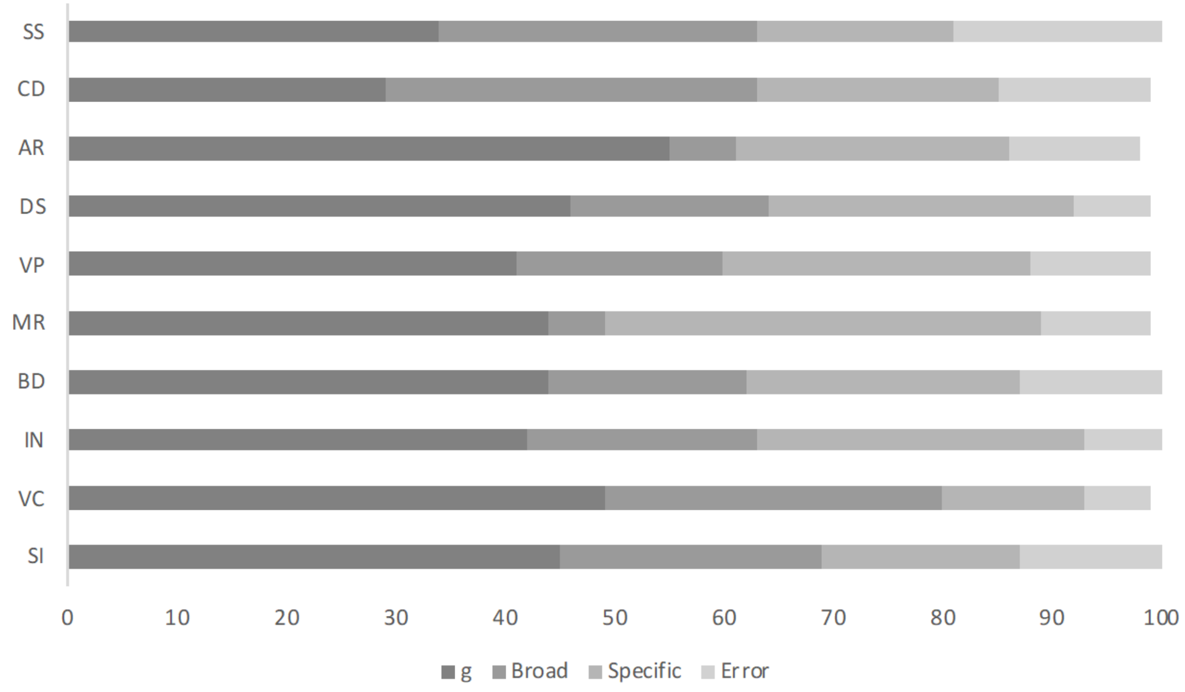

Figure 2. Decomposed sources of WAIS-IV core subtest variance. *g* = general intelligence, SI = Similarities, VC = Vocabulary, IN = Information, BD = Block Design, MR = Matrix Reasoning, VP = Visual Puzzles, DS = Digit Span, AS = Arithmetic, CD = Coding, SS = Symbol Search.

"Just Say No."

In perhaps the most influential critique of subtest pattern analysis, McDermott, Fantuzzo, and Glutting (1990) surveyed a body of empirical literature dismantling the fundamental assumption undergirding subtest pattern analysis. That is, that cognitive subtests measure specific abilities rather than an omnibus global dimension (i.e., *g*). They also identified a series of methodological flaws common in the profile analysis literature; in particular, the failure to compare hypothesized pathognomonic signs against a viable null hypothesis. Finally, they demonstrated that a number of hypothesized core WAIS-R diagnostic profiles were also common in the population, demeaning their potential diagnostic utility. Accordingly, McDermott and colleagues (1990) warned, "until preponderant and convincing evidence shows otherwise, we are compelled to advise psychologists to just say 'no' to subtest analysis" (p. 299). Later, McDermott and colleagues (1992) evaluated the technical properties of ipsative assessment and found that resulting deviation scores were substantially less reliable than their normative counterparts. As a result, the average long-term stability of an identified strength or weakness was only 19% and 21% respectively. More concerning, the predictive validity of ipsative measures was uniformly inferior in comparison to normative scores with the former conveying no uniquely useful information. Given the aforementioned limitations associated with the normative subtest interpretation approaches, it was concluded that no compelling justification could be made for the continued use of *any* form of interindividual or intraindividual form of subtest pattern analysis. Subsequent empirical research on the long-term stability of subtest strengths and weaknesses and the diagnostic accuracy of various subtest profiles has consistently produced negative findings.

For example, Watkins and Kush (1994) utilized cluster analysis to identify core profile subtypes in the WISC-R normative sample and found that 96% of the cases in sample of children and adolescents with exceptionalities were found to be probabilistically similar to those subtypes suggesting that those profiles likely reflected normal intellectual variation and not diagnostic acumen. Of the clinical cases that were unique, no subgroups could be formed, indicating that the subtest variability in that particular sample was likely random and uninterpretable.

Temporal Stability of Subtest-Level Strengths and Weaknesses.

For useful clinical insight to be gleaned from cognitive test scores at any one point in time, the scores that are the focal point of interpretation must possess adequate temporal stability or clinicians may be at risk of misidentifying the presence *or* absence of pathology (Haynes, Smith, & Hunsley, 2011). That is because assessment professionals are often tasked with rendering high-stakes diagnostic decisions about individuals via these data that are expected to remain valid for months if not years. In fact, some evaluative decisions (e.g., Atkins cases) can even carry the weight of a life or death outcome for an examinee. Said another way, performance on an intelligence test should reflect an enduring trait. The question as to what constitutes an acceptable level of reliability remains. Hunsley and Mash (2008) have developed what they define as a "good enough" criterion for assessing the reliability of psychological test scores. Using these guidelines, *adequate* to *excellent* test-retest reliability is defined as correlations meeting or exceeding .70 across days to years respectively. Unfortunately, the stability of cognitive subtest patterns has been shown numerous times to be woefully inadequate for individual decision-making. Watkins and Canivez (2004) obtained long term (M = 2.80 years) test-retest data on the WISC-III from special education evaluations for 579 students. Using kappa coefficients, they examined the degree to which various clinical observations (i.e., strengths and weaknesses in individual scores, intertest scatter, and pairwise differences between scores) for an individual were again observed at Time 2 testing. For the 12 individual subtests, 54 subtest composites, and 10 strengths and weakness categorizations that were analyzed, the median kappa coefficients ranged from -.01 to +.02. Such values reflect agreement at chance levels and indicate that, on average, virtually any unique subtest-level strength or weakness that is observed at Time 1 testing is unlikely to replicate at Time 2 testing (Cichetti, 1994). These results are not surprising given the outcomes previously reported in a comprehensive assessment of the reliability of interpretations for the IT approach to the WISC-III (Macmann & Barnett, 1997). Two independent samples of 5,000 cases were generated from computer simulations of WISC-III data. Of the 54 ipsative profile patterns posited by Kaufman (1994), the average number of significant test observations per case ranged from three to five. In total, 59% of the sample had at least one significant profile pattern in their test data. However, on average, only ~14% of the profile patterns identified as significant were likely to be replicated in a parallel form of the same test. More concerning, less than 30% of the significant test patterns were maintained over a three-week retest interval and the conditional probability of agreement for strengths and weaknesses was woefully insufficient (.20). Whereas practitioners are invited to speculate and generate numerous clinical hypotheses to explain significant deviations in unity in a profile of test scores, these results indicated that the "inferences that result are largely determined by chance" (Macmann & Barnett, 1997, p. 229). Given the endemic error rates observed throughout all levels of the Kaufman system, psychologists were encouraged to regard the assertion that such limitations are able to be overcome through skilled detective work (i.e., Kaufman, 1994) as a veritable myth.

On the Enduring Tradition of Subtest Pattern Analysis. Although subtest-level interpretive procedures continue to be described in seminal assessment texts (e.g., Flanagan & Alfonso, 2017; Lichtenberger & Kaufman, 2013; Sattler, 2018), contemporary guidance as matter of course has generally been tempered due to the aforementioned psychometric limitations associated with subtest pattern analysis. For example, Groth-Marnat and Jordan Wright (2016) warn that, "Clinicians should *never* interpret subtests merely by noting what seem to be high/low subtests and then listing the abilities provided in the subtest descriptions," doing so may result in "incorrect and even potentially damaging conclusions about the examinee" (p. 169). Yet, despite such exhortations, recent surveys indicate that subtest-level analyses remain popular in clinical practice (e.g., Kranzler, 2020). What factors are responsible for this seemingly glaring research to practice gap? While it is beyond the scope of the present discussion to fully adjudicate this matter, we suggest that one factor may be a tendency to disregard previous research on these matters as dated and thus, assume that the long-standing limitations associated with those methods are no longer applicable for contemporary measures (Glutting, Watkins, & Youngstrom, 2003; Lilienfeld, Wood, & Garb, 2003). However, McGill and colleagues (2018) noted that, over the course of the last 20 years, there does not appear to be any compelling research evidence to suggest the psychometric or conceptual concerns that have been raised regarding subtest pattern analysis have been overcome. As noted by Greenwald (1980), the arc of self-correction in our business will remain stunted as long as practitioners and trainers continue to be psychologically invested in these procedures. For these reasons, practitioners are encouraged to bear these limitations in mind when, if at all, employing these practices.

Concerns Pertaining to the Interpretation of Composite Scores and FSIQ

Given the persistent concerns associated with subtest pattern analysis and ipsative assessment, proponents of the intelligent testing approach to test score interpretations now argue that practitioners should focus most, if not all, of their interpretive weight on the index and composite score level of the test (CHC Stratum II) and apply the principles of profile analysis to those indicators (e.g., Decker, Hale, & Flanagan, 2013; Flanagan & Alfonso, 2017). As scores at that level of the test frequently have higher reliability coefficients than subtests and they are better linked to theoretically supported attributes in the assessment literature, it may be assumed that the pitfalls that plague the interpretation of subtest-level indicators no longer apply (Glutting, Watkins, & Youngstom, 2003). However, a growing body of empirical literature suggests that many of those previously identified psychometric shortcomings continue to hold for index and composite level indicators *and* that there are a number of unique concerns that pertain to those particular set of indicators (Kamphaus, Winsor, Rowe, & Kim, 2018).

Long-Term Stability of Composite Scores. Similar to subtest-level indicators, it is assumed that index-and composite-level indices are measuring psychological traits that are longitudinally stable. Despite this assumption, several studies have cast doubt on the relative stability of these scores as well. Ryan, Glass, and Bartels (2010) examined the test-retest reliability of WISC-IV scores in elementary and middle school children with a median retest interval of 11 months. They showed that although the reliability coefficients for the indexes and FSIQ composite were moderate to strong (.54 to .88), 42% of the sample had FSIQ scores that changed by 5 or more points at retest. While these may appear to be trivial distinctions to some, we encourage readers to think about the implications of these findings for individuals suspected of having an intellectual disability whose FSIQ scores reside at or near the diagnostic threshold for that condition and if those scores are subsequently contested as part of an *Atkins* hearing.

Later, in a more substantive investigation, Watkins and Smith (2013) evaluated the test-retest reliability of WISC-IV scores in a sample of 344 students with exceptionalities over an average interval of 2.84 years. 25% of the students earned an FSIQ score that differed by 10 or more points and, 29%, 39%, 37%, and 44% of the VCI, PRI, WMI, and PSI scores, respectively, fluctuated at that same level. As the reliability coefficients obtained by Ryan and colleagues (2010) and Watkins and Smith (2013) were both uniformly lower than those reported in the WISC-IV Technical Manual (Wechsler, 2003), it suggests that test technical manuals, which report results from much shorter interval periods, may overestimate the long-term retest reliability of IQ test scores. Interestingly, factors beyond systematic measurement error appear to play a role in these score differences. McDermott, Watkins, and Rhoad (2014) used multilevel linear modeling to assess four sources of variances in test-retest score differences in a sample of 2,783 children and found that examiner bias was responsible for significant and non-trivial portions of the score differences. That is, if Examiner A assesses a client at Time 1 and Examiner B tests that same client at Time 2 and, that client obtains two different IQ test score profiles across the assessment sessions, those fluctuations may have little to do with individual differences. To be fair, given the fact that these samples containing children receiving special education interventions in the public-school system, it is possible there may have been some developmental changes in the traits that could explain some of this variation. However, this hypothesis counters the long-standing negative research history associated with aptitude by treatment interactions and attempts at cognitive remediation in the schools (Burns et al., 2016; Elliott & Resing, 2015). As noted by McDermott and colleagues (2014), "the nontrivial and substantial amounts of assessor bias that plague almost all factor index and subtest scores effectively diminishes the legitimacy of analyses of score patterns, profiles, or assessments of relative intellectual strengths and weaknesses" (p. 212). Accordingly, clinicians must account for these confounds when attempting to glean insight from unique patterns among composite and index scores.

Construct Validity of Composite Scores. According to Price (2017), construct validity refers to the "appropriateness of inferences drawn from test scores regarding individual standings on a variable defined as a construct" (p. 138). A vital first step in the process of construct validation is establishing the structural validity of a test, usually through exploratory (EFA) and/or confirmatory factor analysis (CFA). The results of these analyses are vital as they provide the statistical rationale for test scores and can be used to determine the degree to which an instrument is aligned with the interpretive theory suggested by the test publisher (Brunner & Wilhelm, 2012, Kane, 2013). Not surprisingly, detailed procedures for establishing a test's internal structure are described in most test technical manuals. Interestingly, over the last 30 years, we are not aware of any manual that has failed to report factor analytic results that did not support the theoretical model posited for those tests. In the era of the so-called replication crisis, these results constitute a remarkable string of successes for psychological test validation (Henne & Ferguson, 2017).

Unfortunately, many of these models have not been replicated in independent factor analytic research raising concern about the tenability of proposed interpretive guidance for many IQ tests, in particular, the interpretation of hypothesized Stratum II index and composite scores. For example, Dombrowski, McGill, and Canivez (2017) evaluated the internal structure of the Woodcock-Johnson IV Tests of Cognitive Abilities using best practice EFA procedures and did not find any evidence to support the existence of seven Stratum II broad abilities as suggested by publisher *and* CHC theory. When attempting to force the extraction of seven factors, three of the seven factors were mathematically impermissible and two of the "viable" factors were complexly determined and could not be identified, all symptoms of over-extraction (Gorsuch, 2003). Instead, a more parsimonious four-factor model, consistent with previous Wechsler Theory, was found to best fit the normative data. Dombrowski and colleagues (2018) have replicated these results using CFA methods. While it may be argued that this is but an extreme example, it is important to keep in mind that even in the case where discrepancies between the models reported in test technical manuals and those identified in the empirical literature competing pertain to only one or two scores, that also has important consequences for the clinical interpretation of the test. For example, Canivez and Watkins (2016) used EFA and CFA to independently examine the structure of the WISC-V. Their results consistently supported a four-factor model consistent with previous Wechsler Theory rather than the five-factor CHC model preferred by the test publisher. However, as the four-factor model combines Fluid Reasoning and Visual-Spatial into a complexly determined Perceptual Reasoning dimension, these results suggested that 40% of the primary index-level scores on the test should be interpreted with caution. Even when theoretical models can be replicated, the resulting index- and composite-level scores often contain insufficient reliable variance to be interpreted independently (Watkins, 2017). To wit, Nelson, Canivez, and Watkins (2013) examined the internal structure of the WAIS-IV in a clinical sample using CFA. Whereas four theoretically consistent factors were able to be identified, *g* accounted for more total and common variance in the WAIS-IV measures than all of the subordinate index scores combined. Resulting indices of interpretive relevance (i.e., Rodriguez, Reise, & Haviland, 2016) suggest that only Processing Speed can be interpreted with confidence beyond FSIQ. These results are not limited to the WAIS-IV as they have been documented for every major commercial ability measure on the market today. In a review of this research, McGill, Dombroswki, and Canivez (2018), were only able to identify two instances in which a subscale score provided interpretive relevance beyond *g* (Processing Speed for the Wechsler Scales and Verbal Ability for the WJ-IV). More concerning, the replication rate in the scores associated with three out of the seven hypothesized CHC broad abilities was only 33% in the factor analytic literature that was surveyed. The only constructs that had a perfect replication rate across studies were general intelligence (i.e., FSIQ) and Working Memory/Short-Term Memory. Kranzler and Keith (1999) argue that the absence of structural validity precludes construct validity. One only needs to look to the incremental validity literature to see why this is the case. Briefly, incremental validity refers to the ability of a measure to explain or predict a phenomenon of interest above and beyond existing measures or available information (Haynes & Lench, 2003). If IQ tests are interpreted in the step-by-step IT fashion, it is assumed that subscale scores provide users with meaningful information beyond the more omnibus FSIQ. However, this is rarely, if ever, the case. To cite but one example, Canivez (2013a) used hierarchical multiple regression to evaluate whether the WAIS-IV index scores accounted for meaningful portions of achievement above and beyond that accounted for by the FSIQ. Whereas the FSIQ alone accounted for 42% to 76% of the variance in the achievement constructs examined, the index-level scores combined only provide small unique (1%-9%) contributions to predicting achievement scores. Across analyses, no individual index score predicted more than 2% achievement variance beyond *g*.

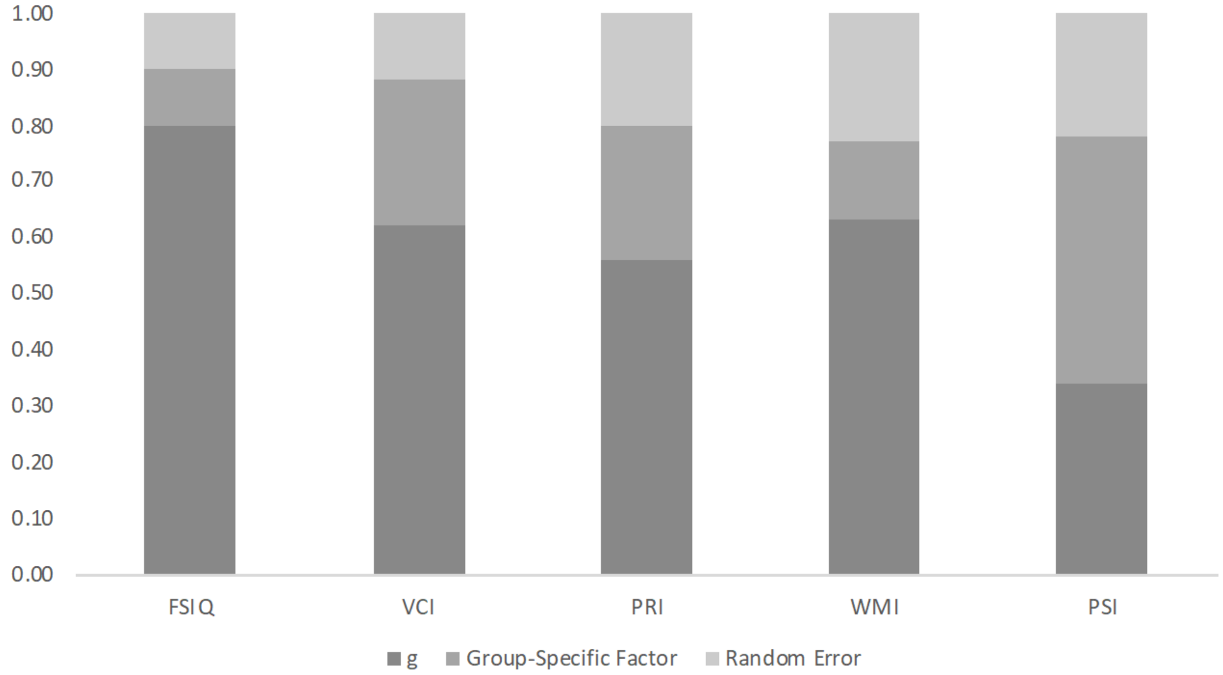

Figure 3. Sources of variance in a unit-weighted index or global composite score for the WAIS-IV.

Put simply, these structural and predictive validity results call into question the construct validity of many composite- and index-level scores. Given the large degree of *g* saturation and dimensional complexity at this level, primary interpretation of scores at that level may be misguided. Figure 3 illustrates the sources of variance in WAIS-IV global composite and index scores[9]. As can be seen, *g* has a pervasive influence on the index scores and only Processing Speed has enough unique variance to be interpreted with confidence. These results indicate that the confidence intervals for group-specific factor scores are likely misleading as they co-mingle the influence of *g* and the broad abilities. If one were to calculate a confidence interval based upon only the broad ability variance (the true focus when clinically interpreting those indices), the interval would likely be 20-30 points wide, a substantial deviation from indices calculated from coefficient alpha (Dunn, Baguley, & Brunsden, 2014). Such a wide interval suggests that the true score for an individual likely resides somewhere between one or more categorical levels of test classification (i.e., average, low average, etc.) in most circumstances for broad ability scores, and clinicians should

[9] To obtain these estimates, the correlation matrix for the 10 core WAIS-IV subtests for normative sample participants (*N* = 2,200) was extracted from the Technical Manual (Wechsler, 2008) and subjected to exploratory bifactor modeling procedures (four-factor extraction consistent with publisher theory) using the Omega function, available in the *Psych* package in **R** (R Developmental Core Team, 2020).

temper the clinical impressions generated from these indices accordingly (Han, 2013; Kranzler & Floyd, 2013).

Does Scatter Actually Matter? Evaluating variability across, between, and within composite and index-level scores from IQ tests has been a core feature of the IT tradition since its inception. Scatter analyses involves two core assumptions: (a) significant scatter between two or more indicators that make up a composite calls into question the validity[10] of that composite score, and (b) significant intertest scatter reflects a unique pattern of strengths and weaknesses that may serve as a potential marker of pathology. In fact, many practitioners continue to be taught to never interpret an FSIQ score unless the lower-order index scores are unitary (Lockwood & Farmer, 2020). Despite the ubiquity of scatter analysis, relatively little empirical research has been done over the last 30 years to examine whether or not the premises of these assessment procedures actually hold. Results from available research raise significant questions about the use of these procedures, in particular, when attempting to validate or invalidate the global FSIQ score.

In a 2007 special issue of *Applied Neuropsychology* dedicated to the topic, Daniel (2007) simulated WISC-IV assessment data and then separated the sample into different groups based on the level of scatter that was observed between the four index-level scores. An EFA was then run on each group to determine if an FSIQ score could be extracted. Across groups, results supported the presence of a general factor indicating, that the FSIQ was valid at all levels of scatter. McGill (2016) extended these findings by analyzing the structural and predictive validity of scores from the Kaufman Assessment Battery for Children-Second Edition (KABC-II) for normative sample participants who had clinically significant levels of scatter between their highest and lowest index scores. Across age groups, the hierarchical measurement model for the instrument was found to be invariant for individuals with and without significant scatter, calling into question the scatter hypothesis. Interestingly, the ability of the FSIQ to predict achievement was more robust in the scatter group than the non-scatter group. Taken together, these findings suggest that the construct validity of the FSIQ score is not violated when significant test scatter is observed (Watkins, Glutting, & Lei, 2007). As such, long-standing clinical doctrine to invalidate such indices in these circumstances should be disavowed until compelling empirical evidence emerges to support this belief.

Additionally, clinicians must keep in mind the importance of base rates when determining whether or not an observed difference is worth evaluating further as scatter and outlier scores are common in the population (Glutting, McDermott, Watkins, & Kush, 1997). For example, the Technical Manual for the WISC-IV provides users with cutoffs for determining whether pairwise differences in various scores are considered *statistically* significant as well as base rates for determining whether or not one should regard an observed difference as *clinically* significant. Typically, index-level comparisons of 15-17 points or more are considered to be clinically significant. However, by systematically evaluating each pairwise comparison in isolation, examiners substantially increase the chances of finding at least one significant finding due to inflated Type I error. In fact, this interpretive approach is akin to a specification search in statistical significance testing which is anathema. To wit, technical information, not available in the WISC-V manual, reported in Kaufman, Raiford, and Coalson (2016) indicate that score differences should not be considered unusual for most individuals unless they exceed 30 points or more. Thus, it is no surprise that diagnostic validity studies, evaluating the discriminant validity of test scatter,

[10] Contemporary interpretive manuals often use more ambiguous language such as the composite score is not considered to be “representative” of a particular ability for that examinee. Regardless of the nomenclature used, we suggest that this is question of validity.

have consistently shown that varying levels of scatter accurately diagnose high incidence disabilities such as specific learning disability at no greater than chance levels (e.g., McGill, 2018; Watkins, 2005). Before concluding on this topic, a brief discussion on the relevance of breakout scores and outliers for evaluations where intellectual disability (ID) is suspected is warranted as it is frequently assumed that if an examinee obtains a low average or better score on one of more part scores (i.e., subtests, indexes, lower-order composites), than a diagnosis of ID is not supported, regardless of the level of the FSIQ. Whereas the authoritative “Green Book” (American Association on Intellectual and Developmental Disabilities, 2010) continues to emphasize the use of FSIQ with no consideration given to the role of part scores in the classification process, the current version of the DSM (American Psychiatric Association, 2013) contains language that references discrete cognitive abilities and appears to implicitly provide support for the use of part scores in the classification process. Results from a recent survey on state identification criteria in the United States reveal that while the vast majority of states’ identification criteria continues to emphasize the use of FSIQ for documenting significant intellectual impairment, the use of part scores were referenced in 10% of the states that were surveyed (McNicolas et al., 2018). While it often assumed that individuals with ID present with unitary profiles of abilities, this is often not the case. To test the unitary assumption, Bergeron and Floyd (2006) examined how often children diagnosed with ID presented with a breakout score on one or more of the seven CHC broad ability cluster scores on the WJ-III. Results indicated that over 36% of the sample obtained at least one cluster score within the average range and over 80% obtained at least one score that was ≥ 80. The authors concluded, “it is important for practitioners to remember that an average or low average part score does not necessarily mean that the overall system or other cognitive abilities in the system are functioning adequately” (pp. 427-428). In a more comprehensive investigation, Bergeron and Floyd (2013) examined the prevalence rate of breakout scores with normative participants who were diagnosed with ID for three major intelligence tests (WISC-IV, DAS-II, KABC-II). It was found that 33% to 52% of the participants obtained at least one part score in the low average range across tests. Taken together, these results suggest that part score variation, per se, should not be used in isolation to rule out ID and cast additional doubt on using significant cognitive scatter as a prima facie justification for invalidating global composite scores.

Conclusion

According to Weiner (1989), an effective clinician will (a) know what their tests can do and (b) act accordingly and this axiom is outlined in virtually every ethical code that governs the practice of applied psychological assessment (Fisher, Brown, Barnett, & Wakeling, 2016). The present review outlines long-standing, replicated, empirical research results that raise fundamental questions about prevailing clinical assessment doctrine that may have substantive implications for conducting evidence-based FMHAs. In particular, all levels of the step-by-step “intelligent” testing procedures that continue to be endorsed in prominent interpretive guides (e.g., Flanagan & Alfonso, 2017; Sattler, 2018) that encourage primary interpretation of IQ tests beyond the first-step of interpreting the FSIQ. To be clear, we are not encouraging practitioners to limit their interpretations to only the FSIQ score as a matter of course when conducting FMHAs as there will be circumstances when interpretation of subscale scores may be justified. However, in doing so we encourage clinicians to adopt the Keep it Simple Scientific (KISS) approach articulated by Kranzler and Floyd (2013) when determining which scores can be interpreted beyond the FSIQ. This approach encourages selective and cautious interpretation of subscale scores as measures of discrete cognitive abilities only after taking into consideration the reliability of the score, the incremental utility of the score beyond general intelligence, and evidence supporting the validity of the score as representing a legitimate psychological dimension. Put simply, we are advocating for a more circumspect interpretation of these measures that better coheres with available research evidence. An exemplar interpretive report template for what that might look like is provided as an Appendix (https://osf.io/zjy6c/?view_only=b4f4ce28b15d4f0cb7f92c2eb76ae4ac).

Limitations

As is the case with any critical review, the present manuscript is not without limitations. First and foremost, our understanding of these matters is mostly fueled by psychometric research involving various iterations of the Wechsler Scales. Whereas those instruments have been regarded as the "gold standard" (Hartman, 2009) in the assessment literature, our field would be aided by additional research, in particular, research examining the long-term stability of scores on other measures. Second, virtually all of the discussions regarding EBA applications for intelligence testing have emerged from the school psychology literature and mostly involve discussions about implications for children and adolescents (e.g., Canivez, 2019). Thus, there is a need to examine the generalizability of these some of these findings with adult samples which are often the focal point of FMHAs. We should note that several of the psychometric limitations noted here have also been found for prominent commercial ability measures designed specifically for adults (e.g., Canivez & Kush, 2013; Nelson, Canivez, & Watkins, 2013). Finally, the literature on these matters is vast and we are keenly aware that practitioners often do not have the time or access to resources to sift through hundreds of articles to find the one or two studies that exemplify clinical gold (Youngstrom, 2013). To be fair, it is not that there is *no* evidence to support the use of IT procedures; however, it is our contention that the research base for profile analysis is presently less than compelling and does not meet the standard for even the weakest definition of evidence-based practice. Therefore, it is critical for clinicians and other stakeholders to consider the quality of the research design and whether such a design even matches the articulated research question(s) when evaluating the quality of available evidence in this literature (Beaujean, Benson, McGill, & Dombrowski, 2018).

Implications for Professional Practice

Despite the numerous psychometric and conceptual issues that have been associated with profile analysis techniques such as IT, proponents of these methods assert that these limitations may be able to overcome through skilled detective work and clinical acumen (i.e., Kaufman, Raiford, & Coalson, 2016). Despite the beguile of these entreaties, such prescriptive statements in psychological science are rarely justified and require strong forms of empirical evidence (Marley & Levin, 2011). As noted long ago by Matarazzo (1990), subjective clinical "impressions" of examiners are no longer considered a sufficient basis for framing interpretations of intelligence test scores. Furthermore, as all profile analytic techniques require practitioners to combine and integrate a considerable amount of information, it is worth considering whether clinicians are actually capable of rendering confident judgements about the psychological functioning of an individual formed from test score patterns or configural relations.

With regard to expert witness testimony on cognitive scores, group to individual (G2i; Faigman et al., 2014) theory may be useful for framing interpretation of IQ test scores in legal proceedings. G2i rationale makes an imperative distinction between *framework* and *diagnostic* testimonies. Whereas framework testimony refers to opinions about constructs developed through research on appropriate reference populations (e.g., a nationally representative normative sample), diagnostic testimony is an attempt to apply the constructs derived from population-level studies to the individual case. Accordingly, diagnostic testimony relies on the assumption that the underlying construct(s) have been validated in the reference population (Faigman et al., 2014). For example, a clinician may dismiss the critical implications suggested by the group studies cited in this review on the basis that these findings do not pertain to their client because they represent an exception to what has been reported in the empirical literature. That is, the findings from group research are instructive but they do not apply for certain individuals (i.e., Kaufman & Lichtenberger, 2006). While these arguments may be compelling, we suggest a higher level of judicial scrutiny should govern the admissibility of such conclusions given the fact that it is difficult, if not impossible, in most circumstances to empirically determine whether an individual *actually* represents an exception to the psychometric rule. Whereas clinicians often claim to base their diagnostic conclusions on the integration of most, if not all, of the data that is available to them, research on clinical decision making suggests that (a) the integration of data described by clinicians rarely, if ever, occurs (b) access to more data does not necessarily improve the accuracy of judgements, (c) clinicians have a difficult time disentangling contradictory information, and (d) most practitioners are prone to numerous biases and heuristics such as illusory correlation that complicate configural pattern analyses of data (Faust, 1989; Fischoff & Broomell, 2020; Zappala et al., 2018). The psychometric limitations noted in the present review further amplify these shortcomings. That is not to say that all of the diagnostic impressions generated from IT and related approaches will necessarily always be incorrect. However, it is important to consider the conditions under which clinical judgement is likely to be most accurate: (1) when operating in a "high validity" environment, and (2) when clinicians receive prolonged systematic feedback on the accuracy of their decisions (Kahneman & Klein, 2009). Unfortunately, those two elements are often missing in most practice environments (Gambrill, 2012). The present review suggests that forensic evaluators who attempt to glean additional insight form IQ test scores using prevailing methods of cognitive profile analysis, likely risk overinterpretation of test scores in most, if not all, circumstances. As a consequence, it is suggested that fact finders and other stakeholders involved in the legal system bear these, and the other limitations raised in the present review, when evaluating the credibility of expert witness testimony involving the clinical interpretation of cognitive test scores. Likewise, judges are also encouraged to better scrutinize the statistical inferences that are used to support profile analysis test interpretations, if available, in their gatekeeping of diagnostic testimony (Andretta et al., 2019).

About the Authors

James R. Andretta, Ph.D. is with Bridgetown Psychological, LLC. Correspondence concerning this article should be addressed to James R. Andretta, Bridgetown Psychological, PO Box 19924, Portland, OR, 97223. E-mail: andrettajames1(at)gmail.com.

Ryan J. McGill, Ph.D., BABA-D, NCSP is Associate Professor of School Psychology and Chair of the Department of School Psychology and Counselor Education at the William & Mary School of Education in Virginia.

References

American Association on Intellectual and Developmental Disabilities. (2010). *Intellectual disability: Definition, classification, and systems of supports* (11th ed.). Washington, DC: Author.

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Andretta, J. R., Morgan, G. B., Cantone, J. A., & Renbarger, R. L. (2019). Applying statistics to the gatekeeping of expert evidence: Introducing the Structured Statistical Judgement (SSJ). *Behavioral Sciences & the Law, 37,* 133-144.doi: 10.1002/bsl.2405

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment*, *87,* 84-94. doi: 10.1207/s15327752jpa8701_07

Atkins v. Virginia, 536 U.S. 304 (2002).

Bartlett v. New York State Bd. of Law Examiners, 156 F.3d 321, 2nd Cir., N.Y. (1998).

Baxter v. Temple, 157 N.H. 280 (2008). Beaujean, A. A. (2015). Adopting a new test edition: Psychometric and practical considerations. *Research and Practice in the Schools, 3,* 51-57.

Benson, N. F., Beaujean, A. A., McGill, R. J., & Dombrowski, S. C. (2018). Critique of the Core-Selective Evaluation Process. *The DiaLog, 47 (2),* 14-18.

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 National Survey. *Journal of School Psychology, 72,* 29-48. doi: 10.1016/j.jsp.2018.12.004

Bergeron, R., & Floyd, R. G. (2006). Broad cognitive abilities of children with mental retardation: An analysis of group and individual profiles. *American Journal on Mental Retardation,111,* 417-432.

Bergeron, R., & Floyd, R. G. (2013). Individual part score profiles of children with intellectual disability: A descriptive analysis across three intelligence tests. *School Psychology Review, 42,* 22-38. doi: 10.1080/02796015.2013.12087489

Borum, R., & Grisso, T. (1995). Psychological test use in criminal forensic evaluations. *Professional Psychology: Research and Practice*, *26*, 465-473. doi: 10.1037/0735-7028.26.5.465

Bray, M. A., Kehle, T. J., & Hintze, J. M. (1998). Profile analysis with the Wechsler Scales: Why does it persist? *School Psychology International, 19,* 202-220. doi: 10.1177/0143034398193002

Brody, N. (1992). *Intelligence*. San Diego, CA: Academic Press.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality, 80,* 796-846. doi: 10.1111/j.1467-6494.2011.00749.x

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

Canivez, G. L. (2013a). Incremental validity of WAIS-IV factor index scores: Relationships with WIAT-II and WIAT-III subtest and composite scores. *Psychological Assessment, 25,* 484-495. doi: 10.1037/a0032092

Canivez, G. L. (2013b). 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. (2019). Evidence-based assessment for school psychology: Research, training, and clinical practice. *Contemporary School Psychology, 23,* 194-200. doi: 10.1007/s40688-019-00238-z

Canivez, G. L., & Kush, J. C. (2013). WISC–IV and WAIS–IV structural validity: Alternate methods, alternate results. Commentary on Weiss et al. (2013a) and Weiss et al. (2013b). *Journal of Psychoeducational Assessment, 31,* 157-169. doi: 10.1177/0734282913478036

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. doi: 10.1037/a0020429

Canivez, G. L., & Watkins, M. W. (2016). Review of the Wechsler Intelligence Scale for Children-Fifth Edition: Critique, commentary, and independent analyses. In A. S. Kaufman, S. E. Raiford, & D. L. Coalson (Eds.), *Intelligent testing with the WISC-V* (pp. 683-702). Hoboken, NJ: Wiley.

Chorn, J. A., & Kovera, M. B. (2019). Variations in reliability and validity do not influence judge, attorney, and mock juror decisions about psychological expert evidence. *Law and Human Behavior*, 43, 542-557. doi: 10.1037/lhb0000345

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284-290. doi: 10.1037/1040-3590.6.4.284

Cohen, J. (1959). The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *Journal of Consulting Psychology, 23,* 285-299. doi: 10.1037/h0043898

Coleman v. State, 341 S.W.3d 221 (Tenn., 2011).

Cunningham, M. (2010). *Evaluation for capital sentencing*. New York: Oxford University Press.

Daubert v. Merrill Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Daniel, M. H. (2007). "Scatter" and the construct validity of FSIQ: Comment on Fiorello et al. (2007). *Applied Neuropsychology, 14,* 291-295. doi: 10.1080/09084280701719401

Davison, M. L., & Kuang, H. (2000). Profile patterns: Research and professional interpretation. *School Psychology Quarterly, 15,* 457-464. doi: 10.1037/h0088801

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology, 63,* 453-482. doi: 10.1146/annurev-psych-120710-100353

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

Dombrowski, S. C., & McGill, R. J. (2019). Book review: Assessment of Children: Cognitive Foundations and Applications-Sixth Edition. *Journal of Psychoeducational Assessment, 37,* 1048-1051. doi: 10.1177/0734282919830217

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. doi: 10.1037/pas0000350

Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2018). An alternative conceptualization of the theoretical structure of the Woodcock-Johnson IV Tests of Cognitive Abilities at school age: A confirmatory factor analytic investigation. *Archives of Scientific Psychology, 6,* 1-13. doi: 10.1037/arc0000039

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 399-412. doi: 10.1111/bjop.12046

Elliott, J. G., & Resing, W. C. M., (2015). Can intelligence testing inform educational intervention for children with reading disability? *Journal of Intelligence, 3,* 137-157. doi: 10.3390/jintelligence3040137

Erickson, S. L., Salekin, K. L., Johnson, L. N., & Doran, S. C. (2020). The predictive power of intelligence: Miranda abilities of individuals with Intellectual disability. *Law and Human Behavior, 44,* 60-70. doi: 10.1037/lhb0000356

Faigman, D., Monahan, J., & Slobogin, C. (2014). Group to Individual (G2i) Inference in Scientific Expert Testimony. *University of Chicago Law Review, 81*, 417-480.

Faust, D. (1989). Data integration in legal evaluations: Can clinicians deliver on their premises? *Behavioral Sciences and the Law, 7,* 469-483. doi: 10.1002/bsl.2370070405

Fischoff, B., & Broomell, S. B. (2020). Judgement and decision making. *Annual Review of Psychology, 71,* 331-355. doi: 10.1146/annurev-psych-010419-050747

Fisher, B. A. J. (2017). A new challenge for expert witnesses relying on subjective information. *Forensic Sciences Research, 2,* 113-114. doi: 10.1080/20961790.2017.1342587

Fisher, M., Brown, S., Barnett, G., & Wakeling, H. (2016). Forensic context testing. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 231-243). New York: Oxford University Press.

Flanagan, D. P., & Alfonso, V. C. (2017). *Essentials of WISC-V assessment*. Hoboken, NJ: John Wiley.

Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). *The Wechsler Intelligence Scales and Gf-Gc theory: A contemporary approach to interpretation*. Needham Heights, MA: Allyn & Bacon.

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

Fletcher-Janzen, E. (2009). Intelligent testing: Bridging the gap between classical and romantic science in assessment. In J. C. Kaufman (Ed.), *Intelligent testing: Integrating psychological theory and clinical practice* (pp. 15-29). New York: Cambridge University Press.

Freeman, A. J., & Chen, Y.-L. (2019). Interpreting pediatric intelligence tests: A framework from evidence-based medicine. In G. Goldstein, D. N. Allen, & J. DeLuca (Eds.), *Handbook of psychological assessment* (4th ed., pp. 65-101). San Diego, CA: Academic Press.

Frumkin, I. B. (2010). Evaluations of competency to waive Miranda rights and coerced or false confessions: Common pitfalls in expert testimony. In G. D. Lassiter & C. A. Meissner (Eds.), *Decade of behavior/Science conference grant. Police interrogations and false confessions: Current research, practice, and policy recommendations* (p. 191–209). Washington, DC: American Psychological Association. *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).

Gambrill, E. (2012). *Critical thinking in clinical practice: Improving the quality of judgements in decisions* (3rd ed.). Hoboken, NJ: John Wiley.

Glutting, J. J., McDermott, P.A., Watkins, M. W., Kush, J. C., & Konold, T. R. (1997). The base rate problem and its consequences for interpreting children’s ability profiles. *School Psychology Review, 26,* 176-188. Retrieved from http://www.nasponline.org

Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multifactored and cross-battery ability 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* (2nd ed; pp. 343-374). New York, NY: Guilford Press.

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & F. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 143-164). Hoboken, NJ: John Wiley.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist, 35,* 603-618. doi: 10.1037/0003-066X.35.7.603

Greiffenstein, M. F. (2009). Clinical myths of forensic neuropsychology. *The Clinical Neuropsychologist, 23,* 286-296. doi: 10.1080/13854040802104873

Gresham, F. M. (2009). Interpretation of intelligence scores in *Atkins* cases: Conceptual and psychometric issues. *Applied Neuropsychology, 16,* 91-97. doi: 10.1080/09084280902864329

Groth-Marnat, G., & Jordan Wright, A. (2016). *Handbook of psychological assessment* (6th ed.). Hoboken, NJ: John Wiley.

Han, P. K. J. (2013). Conceptual, methodological, and ethical problems in communicating uncertainty in clinical evidence. *Medical Care Research and Review, 70,* 14S-36S. doi: 10.1177/1077558712459361

Harris, A. J., & Shakow, D. (1937). The clinical significance of numerical measures of scatter on the Stanford-Binet. *Psychological Bulletin, 34,* 134-150. doi: 10.1037/h0058420

Hartman, D. E. (2009). Wechsler Adult Intelligence Scale IV (WAIS IV): Return of the gold standard. *Applied Neuropsychology, 16,* 85-85. doi: 10.1080/09084280802644466

Haynes, S. N., & Lench, H. C. (2003). Incremental Validity of New Clinical Assessment Measures. *Psychological Assessment, 15,* 456-466. doi: 10.1037/1040-3590.15.4.456

Haynes, S. N., Smith, G. T., & Hunsley, J. D. (2011). *Scientific foundations of clinical assessment*. New York: Routledge.

Heene, M., & Ferguson, C. J. (2017). Psychological science's aversion to the null, and why many of the things you think are true, aren't. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 34-52). West Sussex, UK: Wiley.

Heilbrun, K., DeMatteo, D., Holliday, S. B., & LaDuke, C. (Eds.). (2014). *Forensic mental health assessment: A casebook.* New York: Oxford University Press.

Heilbrun, K., Grisso, T., & Goldstein, A. (2008). *Foundations of forensic mental health assessment*. New York: Oxford University Press.

Howell v. Workman, Not Reported in F.Supp.2d (2011).

Hugueley v. State, Not Reported in S.W.3d (2011).

Hunsley, J., & Mash, E. J. (2008). Developing criteria for evidence-based assessment: An introduction to assessments that work. In J. Hunsley & E. J. Mash (Eds.), *A guide to assessments that work* (pp. 3-14). New York: Oxford University Press.

Hunt, E. (2011). *Human intelligence*. New York: Cambridge University Press. Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64,* 515-526. doi: 10.1037/a0016755

Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2018). A history of intelligence test interpretation. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed.; pp. 56-70). New York: Guilford Press.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1-73.doi: 10.1111/jedm.12000

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: John Wiley.

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

Keith, T. Z., & Kranzler, J. H. (1999). The absence of structural fidelity precludes construct validity: Rejoinder to Naglieri on what the cognitive assessment system does and does not measure. *School Psychology Review, 28,* 303-321. doi: 10.1080/02796015.1999.12085967

Kranzler, J. H., & Floyd, R. G. (2013). *Assessing intelligence in children and adolescents: A practical guide*. New York: Guilford.

Kranzler, J. H., Floyd, R. G., Benson, N., Zaboski, B., & Thibodaux, L. (2016). Classification agreement analysis of cross-battery assessment in the identification of specific learning disorders in children and youth. *International Journal of School & Educational Psychology, 4,* 124-136. doi: 10.1080/21683603.2016.1155515

Kranzler, J. H., Maki, K. E., Benson, N. F., Eckert, T. L., Floyd, R. G., & Fefer, S. A. (2020). How do school psychologists interpret intelligence tests for the identification of specific learning disabilities? *Contemporary School Psychology*. Advance online publication. doi: 10.1007/s40688-020-00274-0

Kumho Tire Company v. Carmichael, 526 U.S. 137 (1999).

Lees-Haley, P. R., Smith, H. H., Williams, C. W., & Dunn, J. T. (1996). Forensic neuropsychological test usage: An empirical survey. *Archives of Clinical Neuropsychology, 11,* 45-51. doi: 10.1093/arclin/11.1.45

Lictenberger, E. O., & Kaufman, A. S. (2013). *Essentials of WAIS-IV assessment* (2nd ed.). Hoboken, NJ: John Wiley.

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

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2007). Why questionable psychological tests remain popular. *Scientific Review of Alternative Medicine, 10,* 6-15.

Lockwood, A. B., & Farmer, R. L. (2020). The cognitive assessment course: Two decades later. *Psychology in the Schools, 57,* 265-283. doi: 10.1002/pits.22298

Mackintosh, N. J. (2011). *IQ and human intelligence* (2nd ed.). New York: Oxford University Press.

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

Marley, S. C., & Levin, J. R. (2011). When are prescriptive statements in educational research justified? *Educational Psychology Review, 23,* 197-206. doi: 10.1007/s10648-011-9154-y

Matarazzo, J. C. (1990). Psychological assessment versus psychological testing: Validation from the school, clinic, and courtroom. *American Psychologist, 45,* 999-1017. doi: 10.1037/0003-066X.45.9.999

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

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. doi: 10.1177/002246699202500407

McDermott, P. A., Watkins, M. W., & Rhoad, A. (2014). Whose IQ is it? Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment, 26,* 207-214. doi: 10.1037/a0034832

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

McGill, R. J. (2018). Confronting the base rate problem: More ups and downs for cognitive scatter analysis. *Contemporary School Psychology, 22,* 384-393.
doi: 10.1007/s40688-017-0168-4

McGill, R. J., & Dombrowski, S. C. (2019). Critically reflecting on the origins, evolution, and impact of the Cattell-Horn-Carroll (CHC) Model. *Applied Measurement in Education, 32,* 216-231. doi: 10.1080/08957347.2019.1619561

McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology, 71,* 108-121. doi: 10.1016/j.jsp.2018.10.007

McGill, R. J., Styck, K. S., Palomares, R. S., & Hass, M. R. (2016). Critical issues in specific learning disability identification: What we need to know about the PSW model. *Learning Disability Quarterly, 39,* 159-170. doi: 10.1177/0731948715618504

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reverence (ITDR): Gf-Gc cross-battery assessment*. Needham Heights, MA: Allyn & Bacon.

McNicholas, P. J., Floyd, R. G., Woods, I. L., Jr., Singh, L. J., Manguno, M. S., & Maki, K. E. (2018). State special education criteria for identifying intellectual disability: A review following revised diagnostic criteria and Rosa's Law. *School Psychology Quarterly, 33,* 75-82. doi: 10.1037/spq0000208

Melton, G. B., Petrila, J., Poythress, N. G., Slobogin, C., Otto, R. K., Mossman, D., & Condie, L. O. (2017). *Psychological evaluations for the courts: A handbook for mental health professionals and lawyers*. New York: Guilford Press.

Neal, T. M. S., Slobogin, C., Saks, M. J., Faigman, D. L., & Geisinger, K. F. (2019). Psychological assessments in legal contexts: Are courts keeping "junk science" out of the courtroom? *Psychological Science in the Public Interest, 20,* 135-164.
doi: 10.1177/1529100619888860

Nelson, J. M., Canivez, G. L., & Watkins, M. W. (2013). Structural and incremental validity of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV) with a clinical sample. *Psychological Assessment, 25,* 618-630. doi: 10.1037/a0032086

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13,* 248-277. doi: 10.1016/0010-0285(81)90010-4

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crises in psychology. *Psychonomic Bulletin and Review, 26,* 1596-1618. doi: 10.3758/s13423-019-01645-2

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.

Palmer v. Asarco Inc., 510 F.Supp.2d 519 (2007).

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

Price, L. R. (2017). *Psychometric methods: Theory into practice*. New York: Guilford Press.

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.

R Developmental Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ready, R. E. & Veague, H. B. (2014). Training in psychological assessment: Current, practices, of clinical psychology programs. *Professional Psychology: Research and Practice, 45*, 278-282. doi: 10.1037/a0037439

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98,* 223-237. doi: 10.1080/00223891.2015.1089249

Ruscio, J. (1998). The perils of post-hockery. *Skeptical Inquirer*, *22,* 297-308.

Ryan, J. J., Glass, L. A., & Bartels, J. M. (2010). Stability of the WISC-IV in a sample of elementary and middle school children. *Applied Neuropsychology, 17,* 68-72. doi: 10.1080/09084280903297933

Sattler, J. M. (2018). *Assessment of children: Cognitive foundations and applications* (6th ed.). La Mesa, CA: Sattler Publishing.

State v. Dahood, 148 N.H. 723, 727, 814 A.2d 159 (2002).

State v. Langill, 157 N.H. 77, ——, 945 A.2d 1 (2008).

State v. Mullins, Not Reported in So.3d (2014).

Sotelo-Dynega, M., & Dixon, S. G. (2014). Cognitive assessment practices: A survey of school psychologists. *Psychology in the Schools, 51,* 1031-1045. doi: 10.1002/pits.21802

United States v. Loaiza-Clavijo, Not Reported in Fed. Supp. (2012).

Watkins, M. W., Glutting, J. J., & Lei, P. W. (2007). Validity of the full scale IQ when there is significant variability among WISC-III and WISC-IV factor scores. *Applied Neuropsychology, 14,* 13-20. doi: 10.1080/0908428070128035

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. doi: 10.1080/02796015.1994.12085739

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

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

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

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

Watkins, M. W., & Smith, L. (2013). Long-term stability of the Wechsler Intelligence Scale for Children-Fourth Edition. *Psychological Assessment, 25,* 477-483. doi: 10.1037/a0031653

Webb v. Ethicon Endo–Surgery, Inc., Not Reported in Fed. Supp. (2015).

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children-Fourth Edition technical and interpretive manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale-Fourth Edition: Technical and interpretive manual*. San Antonio, TX: Pearson.

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment, 53,* 827-831. doi: 10.1207/s15327752jpa5304_18

Youngstrom, E. (2008). Commentary: Evidence-based assessment is not evidence-based medicine—Commentary on evidence-based assessment of cognitive functioning in pediatric psychology. *Journal of Pediatric Psychology, 33,* 1015-1020. doi: 10.1093/jpepsy/jsn060

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, 42,* 139-159. doi: 10.1080/15374416.2012.736358

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

Zappala, M., Reed, A. L., Beltrani, A., Zapf, P. A., & Otto, R. K. (2018). Anything you can do, I can do better: Bias awareness in forensic evaluators. *Journal of Forensic Psychology Research and Practice, 1,* 45-56. doi: 10.1080/24732850.2017.1413532