

Western Kentucky University

TopSCHOLAR®

---

Masters Theses & Specialist Projects

Graduate School

---

Spring 2020

## Reliability of Index and Subtest Discrepancy Scores from the KABC-II NU

Grant Hacherl

Western Kentucky University, ghacherl@gmail.com

Follow this and additional works at: <https://digitalcommons.wku.edu/theses>



Part of the [Cognition and Perception Commons](#), [Cognitive Psychology Commons](#), and the [School Psychology Commons](#)

---

### Recommended Citation

Hacherl, Grant, "Reliability of Index and Subtest Discrepancy Scores from the KABC-II NU" (2020). *Masters Theses & Specialist Projects*. Paper 3174.

<https://digitalcommons.wku.edu/theses/3174>

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact [topscholar@wku.edu](mailto:topscholar@wku.edu).

RELIABILITY OF INDEX AND SUBTEST DISCREPANCY SCORES FROM THE  
KABC-II NU

A Specialist Project  
Presented to  
The Faculty of the Department of Psychology  
Western Kentucky University  
Bowling Green, Kentucky

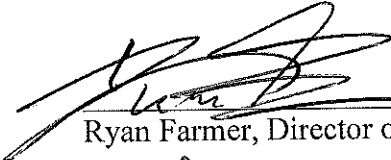
In Partial Fulfillment  
Of the Requirements for the Degree  
Specialist in Education


By  
Grant Hacherl


May 2020

RELIABILITY OF INDEX AND SUBTEST DISCREPANCY SCORES FROM THE KABC-II  
NU

Date Recommended 5/23/2019

  
Ryan Farmer, Director of Specialist Project

  
Thomas Gross

  
Sarah Ochs

Cheryl O. Oavo 2/17/2020  
Dean, The Graduate School Date

## CONTENTS

Literature Review .....	1
Previous Research on Discrepancy Scores .....	11
Purpose .....	13
Methods .....	14
Results .....	22
Discussion .....	28
Conclusion .....	34
References .....	35

LIST OF TABLES

Table 1. Demographics of Normative Sample ..... 15

Table 2. Subtest ICR and Intercorrelations for Ages 3 to 6 Sample ..... 18

Table 3. Subtest ICR and Intercorrelations for Ages 7 to 18 Sample ..... 19

Table 4. Index ICR and Intercorrelations of Ages 3 to 6 Sample ..... 20

Table 5. Index ICR and Intercorrelations of Ages 7 to 18 Sample ..... 20

Table 6. Summary of Discrepancy Score ICR ..... 23

Table 7. Summary of Discrepancy Score ICR ..... 24

Table 8. ICR of Subtest Discrepancy Scores for Ages 7 to 18 Sample ..... 25

Table 9. ICR of Index Discrepancy Scores for Ages 3 to 6 Sample ..... 27

Table 10. ICR of Index Discrepancy Scores for Ages 7 to 18 Sample ..... 27

RELIABILITY OF INDEX AND SUBTEST DISCREPANCY SCORES FROM THE  
KABC-II NU

Grant Hacherl

May 2020

46 Pages

Directed by: Dr. Ryan Farmer, Dr. Thomas Gross, and Dr. Sarah Ochs

Department of Psychology

Western Kentucky University

School psychologists often consider index- and subtest-level discrepancy scores from intelligence tests when making decisions regarding students' special education eligibility. Best practices for clinical decision-making indicate that scores may only be considered if they meet an established standard of reliability. Therefore, it is essential to assess whether an interpretation of discrepancy scores can be considered reliable. This research used data provided in the supplemental manual of the Kaufman Assessment Battery for Children, Second Edition Normative Update (KABC-II NU) to calculate internal reliability coefficients (ICR) for discrepancy scores for each of the sample age group batteries, ages 3-6 and ages 7-18. Subtest-level discrepancy score ICR for ages 3-6 ranged from .61 to .94 and index-level ICR ranged from .00 to .93. Subtest-level discrepancy score ICR for ages 7-18 ranged from .56 to .94 and index-level ICR ranged from .61 to .94. These scores are compared to established reliability standards and a discussion of implications for practitioners is provided.

## **Literature Review**

School psychologists report that they engage in special education eligibility evaluations more than any other professional task (Walcott, Charvat, McNamara, & Hyson, 2016). They are uniquely qualified to complete psychoeducational assessments and facilitate the development of individual education plans, and these decision-making processes are influenced by the quality of data informing the process (Hunsley & Mash, 2007, 2018; Marsh, De Los Reyes, & Lilienfeld, 2017; Vacha-Haase & Thompson, 2011). These decisions have a considerable impact on students' education and therefore it is essential that practitioners make accurate, informed decisions using reliable data.

### **Methods of Interpreting Intelligence Tests**

Standardized intelligence testing has been nearly synonymous with the identity of school psychologists for decades (Bardon, 1979, 1994; Fagan, 2014; Watkins, Crosby, & Pearson, 2001). Intelligence tests are used to confirm the presence of intellectual and developmental disabilities (IDD; McNicholas et al., 2018), specific learning disabilities (SLD; Maki, Floyd, & Roberson, 2015; NASP, 2016), and often to rule out cognitive deficits as part of a comprehensive assessment. In the 2015-2016 academic year, 34% of students between 3 and 21 served under the Individuals with Disabilities Education Act (IDEA) were identified as having a specific learning disability, equal to over 2.25 million students. An additional 7% of students were identified as having an intellectual disability, or approximately half a million, and raising the total number of students identified in these two disability categories to nearly three million (U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics [NCES], 2018).

Despite their frequency of use, there is considerable debate about how to interpret the multitude of scores obtained from modern intelligence tests (Beaujean & Benson, 2018; Beaujean, Benson, McGill, & Dombrowski, 2018; Canivez, 2013; Flanagan & Schneider, 2016; Fletcher & Miciak, 2017; Hale, Kaufman, Naglieri, & Kavale, 2006; Kranzler, Benson, & Floyd, 2016; Kranzler, Floyd, Benson, Zaboski, & Thibodaux, 2016; McGill & Busse, 2017; Watkins, 2000). For instance, emphasis on an overarching IQ (e.g., the Full Scale IQ) for IDD evaluations is generally supported by the literature (Canivez, 2013) and is consistent with clinical guidelines (e.g., American Association on Intellectual and Developmental Disabilities as reported in Schalock et al., 2010), and with state law (McNicholas et al., 2018). However, some experts continue to recommend alternative interpretation strategies, e.g., the variability hypothesis (McGill, 2016) and profile analysis (Hale et al., 2006; Kaufman, Raiford, & Coalson, 2015; Sattler, 2008).

The variability hypothesis is the notion that significant differences between cognitive ability scores render the overarching IQ invalid (McGill, 2016). The variability hypothesis is presented as fact in textbooks (Kaufman et al., 2015; Sattler, 2008) and test manuals (e.g., Wechsler, Raiford, & Holnack, 2014) despite conflicting evidence in the empirical literature (McGill, 2016; Schneider & Roman, 2017). This perspective often leads to the use of profile analysis. The profile analysis approach, in which scores from an intelligence test are examined for patterns, is commonplace among school psychology practitioners. Alfonso, Oakland, LaRocca, and Spanakos (2000) reported that 74% of school psychology training programs taught subtest-level interpretations with a moderate to great emphasis. More recent data from Cottrell and Barrett (2016) indicate that patterns of strengths and weaknesses are still a factor in the schools, with approximately half of



school psychologist practitioners in the survey reporting that they consider these differences for identification. In addition, profile analysis can still be found in intelligence test manuals (Reynolds & Kamphaus, 2015; Wechsler, Raiford, & Holdnack, 2014) and assessment guides (Flanagan & Alfonso, 2017; Kaufman et al., 2015; Mather & Wendling, 2015; Schrank, Decker, & Garruto, 2016).

For SLD, a number of competing interpretative approaches are available to school psychology practitioners, many of which assess for (a) differences between cognitive strengths and academic weaknesses and (b) consistency between cognitive deficits and academic deficits. These SLD identification strategies, generally known as pattern of strengths and weaknesses (PSW; Hale et al., 2006), like the variability hypothesis, rely heavily on cognitive profile analysis techniques, such as statistical discrepancies between scores generated from intelligence tests. State regulations and guidelines offer varied approaches that can be used to assess SLD, with nearly half of states allowing or supporting the use of PSW (Maki et al., 2015). Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer (2000) surveyed practitioners and reported that 70% found profile analysis strategies meaningful and as many as 89% conducted such analyses regularly. Subsequently, Decker, Hale, and Flanagan (2013) and Kranzler, Benson, and Floyd (2016) suggested that such strategies have likely increased in popularity with the increased emphasis on cross-battery assessment strategies (XBA; Flanagan, Ortiz, & Alfonso, 2013). However, supporting literature for cognitive profile interpretations is lacking (Beaujean et al., 2018; Canivez, 2013; Fletcher & Miciak, 2017; Kranzler, Benson, & Floyd, 2016; Kranzler, Floyd, et al., 2016; McGill & Busse, 2017).

## **Discrepancy Scores in Cognitive Profile Analysis**

Discrepancy scores have historically been utilized as a part of cognitive profile analysis (Canivez, 2013; Kaufman & Wechsler, 1979), especially as the impetus in interpretation has shifted to become more psychometric in nature (Kamphaus, Winsor, Rowe, & Kim, 2018). Although cognitive profile analysis has evolved over time, its basic components have remained consistent (Kaufman et al., 2015). Two types of scores produced from intelligence tests are typically used: subtests, which are collections of items that go together in a section within the overall test, and indices, which are collections of subtests that reportedly measure the same or similar constructs. When interpreting a cognitive profile using discrepancy scores, scores are derived from comparing one score (i.e., index or subtest) to another (i.e., index or subtest), with the intent being to identify patterns within the student's cognitive scores (Flanagan et al., 2013; Ortiz, Flanagan, & Alfonso, 2017; Sattler, 2008). Most commonly, quantification of differences is expressed by directly subtracting one score from another score. For example, a child with a standard score of 100 on the Knowledge/Gc Index and a standard score of 85 on the Planning/Gf Index of the Kaufman Assessment Battery for Children, Second Edition Normative Update (KABC-II NU; Kaufman & Kaufman, 2018) may subtract the Knowledge/Gc index from the Planning/Gf index and obtain a discrepancy score of 15 (i.e.,  $100 - 85 = 15$ ).

Instructions for completing such profile analyses are often explicit in test manuals (e.g., Kaufman, Kaufman, Drozdick, & Morrison, 2018, pp. 15, 24). The underlying rationale of completing these simple calculations is that differences in scores may reflect differences in cognitive abilities, and that differences in cognitive abilities are meaningful

and informative individual differences warranting consideration during differential diagnosis and treatment planning (Canivez, 2013; McGill, Dombrowski, & Canivez, 2018). These interpretations may lead to changes in instruction or more restrictive instructional placement, hence the need to determine if this procedure is well-supported (Gross et al., 2018; NASP, 2016).

### **Interpreting Scores from Intelligence Tests**

Due to the high volume of intelligence tests given by school psychologists (Benson, Floyd, Kranzler, Eckert, & Fefer, 2018) and the varied use of intelligence test results (Maki et al., 2015; McNicholas et al., 2018), evidence-based interpretations of the results are paramount. For any interpretation of an assessment score to be considered evidence-based, the score must first be reliable, valid, and demonstrate clinical utility (American Psychological Association Task Force on Evidence-Based Practice for Children and Adolescents, 2008; Canivez, 2013; Gross et al., 2018; Hunsley & Mash, 2007; Kranzler & Floyd, 2013). Reliability is the primary assessment characteristic for decision-making. It is defined by Rust and Golombok (1999) as “the extent to which the test measures anything at all” (p. 64) and by Price (2016) as “the degree to which scores are free from errors of measurement” (p. 203). The American Educational Research Association [AERA], the American Psychological Association [APA], and the National Council on Measurement in Education [NCME] (2014) published a set of standards for educational and psychological testing which discusses reliability and its importance at length, in which they define reliability as “consistency of scores across replications of a testing procedure” (p. 33). Both reliability and validity must always be considered when selecting a test and when interpreting results. Scores can only be considered meaningful

and clinically useful when they provide information that is reliable and can aid in decision-making (AERA et al., 2014; Hunsley & Mash, 2007, 2018). Extending Rust and Golombok's (1999) assertion, a score that is not reliable is not measuring anything at all.

Within the APA code of ethics (2017), psychologists may only use interpretations supported by research with well-established reliability. The NASP code of ethics (2010) mirrors these standards, requiring that assessment techniques must be research-based and all assessment instruments and strategies must be reliable. This means for any discrepancy score used by a school psychologist, it must meet all the above criteria. In the developing evidence-based assessment literature, reliability is a first step for establishing a score and interpretation as clinically useful (Haynes, Smith, & Hunsley, 2011; Hunsley & Mash, 2007)

These standards are ideal in theory; however, they provide no specific criteria for determining adequate reliability. This is done intentionally due to the wide variety of assessment purposes (e.g., screening, diagnostic, progress monitoring) used by school psychologists, because these different assessment outcomes may warrant varying levels of acceptable reliability (Beidas et al., 2015; Haynes et al., 2011; Hunsley & Mash, 2007, 2018). For example, a curriculum-based measure used as a screener may have a lower reliability because it is designed to be more sensitive to change. On the other hand, IQ is stable over time and requires a higher level of reliability. Though there are no universal criteria, there are recommendations in the literature regarding internal consistency reliability, which we were concerned with in this study (Hunsley & Mash, 2018; Kranzler & Floyd, 2013; Nunnally & Bernstein, 1994).

## **What is Internal Consistency Reliability?**

Internal consistency reliability (ICR) is a measure of how related scores are to each other. ICR coefficients are used to show the level of measurement error in a score on a scale of 0 to 1, with reliability increasing and measurement error decreasing as the coefficient approaches 1. It is essential to give substantial weight to this coefficient, because test scores are imperfect measures of attributes, and it informs school psychologist practitioners about some strengths and limitations of their instruments (Gambrill, 2012). The obtained scores on an intelligence test indicate one possible score in a distribution of all possible scores the individual could obtain (Price, 2016). For example, an obtained score of 83 may not be the true score of the individual; it is merely the score he or she obtained from this single administration. Error inherent in the measure or its administration could have produced a score of 82, 84, or even 90 in some cases. The true score is a hypothetical value, an unachievable number because it would require an infinitely large sample of independent test administrations (Price, 2016). Therefore, the resulting ICR coefficient should be interpreted as the ratio of estimated true-score to error variance.

The ICR is interpreted by considering the recommendations established in the literature. Nunnally (1978) and Nunnally and Bernstein (1994) suggested that a reliability coefficient of .80 or higher is the threshold for hypothesis generation and .90 or higher for clinical decision-making. He claimed .95 was the “desirable standard” (Nunnally, 1978, p. 246), but this level of reliability was rare in practice. Kranzler and Floyd (2013) also recommended the standard of .95 for scores stemming from intelligence tests, but later added that a criterion of .90 was more realistic in clinical practice (Floyd et al.,

2015). This standard is consistent with Reynolds and Livingston' (2014) recommendations in *Best Practices in School Psychology*.

Hunsley and Mash (2007, 2008, 2018) have offered guidelines for evaluating the adequacy of reliability based on these sources and others and concluded that clinical practice would need to rely on an approach wherein criteria are stringent enough to be meaningful, but lenient enough not to be disregarded by school psychology practitioners or to leave practice with too few options. They established guidelines for adequate, good, and excellent reliability. Adequate ICR ranges between .70 and .79; good ICR ranges between .80 and .89; and excellent ICR are coefficients of .90 or above (Hunsley & Mash, 2008, 2018). Given the often long-lasting impact of decisions stemming from intelligence test scores as well as the long-established psychometric properties of scores from intelligence tests, an excellent reliability (i.e., .90 or above) should be considered the minimum ICR for diagnostic interpretation for high-stakes clinical decisions (Beidas et al., 2015; DeVon et. al, 2007; Floyd et al., 2015; Kline, 2000; Nunnally & Bernstein, 1994; Ponterotto & Ruckdeschel, 2007; Reynolds & Livingston, 2014). Though, the good standard (i.e., ICR = .80 to .89), as the minimum standard for clinical use, could also be evaluated as it is appropriate for hypothesis generation. Reliability coefficients below .70 are considered too low for interpretation. However, it should be noted that these are not universally agreed upon and others may take issue with these guidelines (Charter, 2001; Streiner & Norman, 2008)

### **Interpretation of Cognitive Assessments**

The Evidence-Based Assessment movement in school psychology (e.g., Canivez, 2013; Gross et al., 2018; Hunsley & Mash, 2007; Kranzler & Floyd, 2013) prompts

school psychologists and practitioners to question whether they are engaging in interpretive practices that are guided by the best available evidence. Intelligence tests, such as the KABC-II NU (Kaufman & Kaufman, 2018), are developed over long periods of time using large standardization samples, which could lead many practitioners to assume that all scores produced are psychometrically sound. However, the overall score generated from the test alone cannot be the sole indicator of reliability. Individual subfactor scores must be determined to be reliable within each sample and for each purpose of interest if they are to be used in clinical practice (Price, 2016). For intelligence tests, in most cases, general intelligence composites, or IQs, produce the most reliable score within the assessment (Kranzler & Floyd, 2013). Several researchers (Canivez, 2013; Farmer & Floyd, 2018; Floyd, Farmer, Schneider, & McGrew, in press; Gross et al., 2018; Kranzler & Floyd, 2013) encourage practitioners to focus interpretation on the overall IQ as it is the most psychometrically sound, predictive, and pragmatic of the possible test results. Subsequently, index scores (i.e., scores such as the Knowledge/Gc from the KABC-II NU) often meet reliability criteria for interpretation but have varied validity and utility evidence (Kranzler & Floyd, 2013).

It is when the interpretation extends into *ipsative* analysis that score reliability may drop below expected criteria. Ipsative analysis is the process of comparing an index score to an index score or a subtest score to a subtest score for the same person within the same assessment. While reliability may be adequate for each of the indices or subtests individually, reliability may not be adequate when comparing the two scores. As Hunsley and Mash (2007) suggest, subtests and indices usually have much lower reliability coefficients, which means that interpretation based on these scores increases the

likelihood that the school psychology practitioner may be misinformed in their clinical judgment. Discrepancy score reliability is a function of (a) the score reliability of the two comparison scores and (b) the correlation between those two scores, and as such the reliability of each the scores being compared acts as an upper limit on the reliability of the newly derived discrepancy score (Price, 2016). Therefore, to obtain an adequately reliable discrepancy score, we must begin with highly reliable comparison scores.

While the overall IQ usually has a very high reliability (Kranzler & Floyd, 2013), the Successive Levels approach to cognitive interpretation (Sattler, 2008; cf. intelligent testing, e.g., Kaufman et al., 2015) interprets index and subtest scores. These score reliabilities are usually lower than that of the overall IQ before accounting for the reliability reduction introduced when comparing scores. A core component of the Successive Levels/Intelligent Testing approach to cognitive interpretation is the emphasis placed on the differences between scores. Discrepancy scores are used to determine when differences between indices or subtests are meaningful by assessing whether an examinee has a *relative* cognitive ability strength or weakness. This is usually determined when one index is significantly (as per the manual guidelines; e.g., Kaufman et al., 2018, pp. 24, 115) higher than a comparison index (e.g., the Knowledge/Gc index is higher than the Planning/Gf index). Discrepancy scores have been integrated into a number of intelligence tests (e.g., Kaufman & Kaufman, 2018; Reynolds & Kamphaus, 2015; Wechsler, 2014), and have been the focus of a number of peer-reviewed articles (Brown & Ryan, 2004; Charter, 2001, 2002; Glass, Ryan, & Charter, 2010; Glass, Ryan, Charter, & Bartels, 2009; Ryan & Brown, 2005). For instance, the manual for the KABC-II NU (Kaufman et al., 2018) discusses how they calculated critical values for discrepancy



scores. However, technical manuals, including the manual for the KABC-II NU, do not provide reliability data, pursuant to the AERA et al. standards (2014), for discrepancy scores.

### **Previous Research on Discrepancy Scores**

To date, several studies have examined the reliability of difference scores for a number of intelligence tests. Charter (2002) calculated the reliability coefficients of difference scores for the Wechsler Memory Scale, Third Edition's (WMS-III; Wechsler, 1997b) primary indices, which ranged from .00 to .87. Using the same criteria from Hunsley and Mash (2008), none of the difference scores of the indices met the .90 excellent threshold, and when disaggregated into the thirteen age groups of the assessment, only 19 of the 104 comparisons met the .80 good standard for hypothesis generation (Charter, 2002).

Two separate studies examined the reliability statistics of the Wechsler Adult Intelligence Scales, Third Edition (WAIS-III; Wechsler, 1997a), one with a clinical sample of 100 men in a substance abuse treatment program (Brown & Ryan, 2004) and one with data from the assessment manual (Charter, 2001). They found very similar results, with comparisons using the data in the manual resulting in subtest reliability coefficients ranging between .44 and .85 with only 12% meeting the .80 good criterion (Charter, 2001). Index comparison scores provided higher reliability coefficients across ages, ranging from .77 to .88 with 84% surpassing .80; however, none met the .90 excellent criterion for decision-making. Brown and Ryan (2004) found subtest reliability coefficients ranging from .34 to .85. Only 7 out of 55 subtests met the .80 good threshold.

For index comparison scores, two of the four produced scores greater than .80, ranging from .79 to .87.

Ryan and Brown (2005) computed reliability statistics in the same way for the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999). The WASI produces only two index scores, the Verbal Scale IQ and the Performance Scale IQ. The reliability difference scores between them at the 23 different age groups varied from .78 to .91. Subtest discrepancy scores produce reliability coefficients from .59 to .85, with 9 of the 12 comparisons having reliabilities greater than the .80 good criterion. At only two of the age groups did the index comparison meet the .90 excellent clinical decision-making standard and no subtest-level discrepancy scores met the .90 standard.

As for more recent tests, Glass and colleagues (2010) calculated the reliability of difference scores at both the index and subtest levels for the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV; Wechsler, 2008a) using the standardization sample provided in the test manual (Wechsler, 2008b). They found that none of the 66 subtest comparisons met the excellent standard of .90, with reliability coefficients ranging from .55 to .88. Twenty-three of 66 subtest comparisons met the good standard. There were only three index discrepancy scores possible, all of which met the good criterion for hypothesis generation but fell short of .90. Overall, this means that data derived from discrepancy scores from the WAIS-IV should not be used as a rationale for decision-making and only some of the comparisons meet the guidelines for hypothesis generation in accordance with evidence-based practice.

This trend was also observed in Glass and colleagues' (2009) evaluation of the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV; Wechsler, 2003).

None of the reliability coefficients for the difference scores of indices or subtests met the excellent .90 criterion, however, 5 of the 66 subtest comparisons and 33 of 36 index comparisons had reliability coefficients greater than the good criterion of .80. As is the case with the intelligence tests described here, discrepancy scores derived from intelligence assessments often demonstrate reduced reliability (compared to index scores) and thus may be unfit for clinical decision-making. Assessments may have subtest- and index-level comparisons that meet the threshold for hypothesis generation, but even this inferior level of reliability cannot be assumed for any test. Of the discrepancy score analyses of recent intelligence tests, the newly normed KABC-II NU (Kaufman & Kaufman, 2018) has yet to be examined.

### **Purpose**

The purpose of this study was to evaluate the reliability of discrepancy scores produced in the KABC-II NU (Kaufman & Kaufman, 2018). Two questions were asked:

- 1) What is the ICR of the discrepancy scores produced from the subtests on the KABC-II NU?
- 2) What is the ICR of the discrepancy scores produced from the indices on the KABC-II NU?

In relation to these questions, two hypotheses were considered:

- 1) Discrepancy scores produced from subtests on the KABC-II NU will not meet Hunsley and Mash's (2008, 2018) "good" or "excellent" reliability guidelines (i.e., ICR will be below .80).
- 2) Discrepancy scores produced from indices on the KABC-II NU will meet Hunsley and Mash's (2008, 2018) "good" reliability guideline (i.e., ICR between

.80 and .89; for hypothesis generation), but will not meet the “excellent” reliability guidelines (i.e., ICR .90 or above; for clinical decision-making).

## **Methods**

### **Measure**

**Kaufman Assessment Battery for Children, Second Edition.** The KABC-II NU (Kaufman & Kaufman, 2018) is an intelligence test consisting of 18 subtests, 6 second-order broad ability indices, and 3 global indices and serves as an update to the 2004 KABC-II (Kaufman & Kaufman, 2004). The standardization sample of the KABC-II NU included 700 children between the ages of 3 and 18 (Kaufman et al., 2018). There are 14 age groups, one for each year of age between 3 and 14, with one for ages 15 and 16, and one for ages 17 and 18. Each age group consisted of approximately 50 participants. The sample used United States Census data 1-year-period estimates (Ruggles, Genadek, Goeken, Grover, & Sobek, 2017) for stratification within each age group for gender, race/ethnicity, parent education level, and geographic region to assist in obtaining a nationally representative sample. This information is provided in Table 1.

The KABC-II NU (Kaufman & Kaufman, 2018) organizes the ages into two groups: one for all students between the ages of 3 and 6 years old and one for all students between the ages of 7 and 18 years old. Data from the ages 3 to 6 sample and the ages 7 to 18 sample consist of 17 and 16 subtests, respectively, each of which yield a scaled score with a mean of 10 and a standard deviation of 3.

Table 1

*Demographics of Normative Sample*

Demographics	Percentage of Each Age Group	
	3 – 6	7 – 18
Female	48.95%	50.39%
Race/Ethnicity		
African American	2.00%	5.98%
Asian	15.65%	13.74%
Hispanic	20.23%	23.11%
Other	8.08%	4.78%
White	54.05%	52.40%
Region		
Northeast	20.83%	20.30%
Midwest	17.50%	8.28%
South	41.43%	51.40%
West	20.73%	19.72%
Education		
Did Not Graduate High School	8.13%	11.96%
High School Diploma	22.20%	21.70%
Some College	34.38%	34.47%
College or Graduate Degree	35.30%	32.87%

For the ages 3 to 6 battery, the KABC-II NU has subtest normative data for Atlantis, Conceptual Thinking, Face Recognition, Number Recall, Gestalt Closure, Rover, Atlantis Delayed, Expressive Vocabulary, Verbal Knowledge, Rebus, Triangles, Block Counting, Word Order, Pattern Reasoning, Hand Movements, Rebus Delayed, and Riddles. Although the Story Completion subtest may be given at age 6, the manual supplement provides no intercorrelation data for this subtest at the ages 3 to 6 sample level and therefore cannot be included in the analyses for this age group (Kaufman et al., 2018). For the ages 7 to 18 sample, there are subtest normative data for Atlantis, Story Completion, Number Recall, Gestalt Closure, Rover, Atlantis Delayed, Expressive Vocabulary, Verbal Knowledge, Rebus, Triangles, Block Counting, Word Order, Pattern Reasoning, Hand Movements, Rebus Relayed, and Riddles (Kaufman et al., 2018).

The KABC-II NU utilizes a dual theoretical model grounded in the Cattell-Horn-Carroll Theory of Cognitive Abilities (CHC; Schneider & McGrew, 2018) and the Luria Model of Intelligence (Luria, 1973). The examiner can use either model with the KABC-II NU and therefore the names for each of the indexes will be presented together when possible. For example, CHC calls the index that measures short-term memory “Gsm” while Luria calls it “Sequential” and they both use the same subtests and arrive at the same score for the index. This will be referred to as “Sequential/Gsm to include both models of intelligence. One index, Knowledge/Gc, is found only in the CHC model for ages 7 to 18 but is included in both models for ages 3 to 6. These two models differ in their conceptualization of the overall score, also known as the primary g, which CHC calls “Fluid Crystallized Intelligence” and Luria calls “Mental Processing Index.” These two indices use different subtests to calculate their scores and will be evaluated as separate scores.

For the ages 3 to 6 battery, the KABC-II NU has index normative data for Nonverbal Index, Sequential/Gsm, Simultaneous/Gv, Learning/Glr, Knowledge/Gc, Delayed Recall, Fluid Crystallized Intelligence, and Mental Processing Index. For the ages 7 to 18 sample, there are index normative data for Nonverbal Index, Sequential/Gsm, Simultaneous/Gv, Learning/Glr, Planning/Gf, Knowledge/Gc, Delayed Recall, Fluid Crystallized Intelligence, and Mental Processing Index (Kaufman et al., 2018).

The manual supplement provides ICR coefficients for each subtest and index at all fourteen ages as well as estimates for the all students in the ages 3 to 6 sample and the ages 7 to 18 sample (Kaufman et al., 2018). ICR coefficients are split-half calculations

based on the standardization sample and were collected from the manual supplement's Table 3.1 (Kaufman et al., 2018). The median ICR of subtests across the ages 3 to 6 sample was .88 and for the ages 7 to 18 sample the median was .90. The median ICR of indices across the ages 3 to 6 sample was .95 and for the ages 7 to 18 sample the median was .95. Intercorrelation coefficients are only given for each age, without a total, and were collected from the manual supplement's Tables E.1 through E.14 (Kaufman et al., 2018). Due to the small sample size for each individual age norm block (Norfolk et al., 2015) and this project's intent to assess the aggregate reliability of discrepancy scores, results for each discrepancy score at all age points are not provided.

While subtest and index comparisons are commonly used in various interpretation strategies, the KABC-II NU recommends specific comparisons in the manual supplement (Kaufman et al., 2018). Other comparisons, however, are possible using external software; these comparisons will also be evaluated. There are 15 possible planned subtest comparisons and they will be evaluated in a group in addition to the aggregate analysis with the rest of the comparisons.

## **Procedure**

The ICR estimates were obtained from the KABC-II NU manual supplement (Table 3.1; Kaufman et al., 2018), as were the intercorrelation coefficients (Tables E.1 through E.14; Kaufman et al., 2018). ICR and intercorrelations for each subtest and index for both samples were reproduced in the following tables: Table 2 details the 17 subtests in the ages 3 to 6 sample, Table 3 details the 16 subtests in the ages 7 to 18 sample, Table 4 details the 5 second-order broad ability indices and 3 global indices in the ages 3 to 6 sample, and Table 5 details the 6 second-order broad ability indices and 3 global indices

Table 2

*Subtest ICR and Intercorrelations for Ages 3 to 6 Sample*

Subtest	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Atlantis	.96																
2. Conceptual Thinking	.29	.90															
3. Face Recognition	.29	.36	.83														
4. Number Recall	.35	.18	.32	.84													
5. Gestalt Closure	.26	.23	.42	.25	.85												
6. Rover	.00	.35	-	.13	.08	.86											
7. Atlantis Delayed	.47	.17	.45	.26	.20	.29	.83										
8. Expressive Vocabulary	.35	.47	.30	.40	.41	.22	.18	.89									
9. Verbal Knowledge	.39	.53	.35	.42	.48	.43	.36	.66	.93								
10. Rebus	.43	.39	.25	.30	.18	.29	.38	.43	.40	.97							
11. Triangles	.28	.45	.28	.29	.30	.36	.33	.41	.45	.47	.91						
12. Block Counting	.49	.26	.28	.18	.19	.24	.36	.27	.38	.31	.52	.91					
13. Word Order	.31	.44	.40	.51	.31	.07	.27	.49	.51	.41	.46	.26	.86				
14. Pattern Reasoning	.18	.51	.14	.19	.30	.30	.21	.45	.46	.39	.52	.44	.40	.90			
15. Hand Movements	.23	.26	.42	.25	.26	.38	.17	.27	.35	.36	.24	.26	.49	.29	.83		
16. Rebus Delayed	.44	.30	.36	.36	.15	.19	.38	.31	.41	.85	.47	.26	.38	.25	.35	.96	
17. Riddles	.32	.54	.33	.33	.44	.39	.26	.69	.66	.42	.54	.43	.55	.58	.41	.29	.87

*Notes.* Internal Consistency Reliability for each subtest is given along the diagonal. Intercorrelations between subtests are given below the diagonal. An intercorrelation for Face Recognition and Rover is not given because the ages at which they may be administered do not overlap.



Table 3

*Subtest ICR and Intercorrelations for Ages 7 to 18 Sample*

Subtest	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Atlantis	.96															
2. Story Completion	.27	.81														
3. Number Recall	.27	.16	.84													
4. Gestalt Closure	.31	.30	.18	.79												
5. Rover	.22	.35	.28	.27	.94											
6. Atlantis Delayed	.69	.28	.17	.30	.19	.90										
7. Expressive Vocabulary	.37	.39	.36	.44	.36	.32	.88									
8. Verbal Knowledge	.46	.42	.35	.39	.35	.37	.71	.94								
9. Rebus	.53	.40	.34	.31	.35	.44	.45	.48	.95							
10. Triangles	.24	.42	.18	.33	.39	.21	.35	.41	.39	.87						
11. Block Counting	.21	.36	.24	.30	.43	.16	.35	.37	.32	.54	.95					
12. Word Order	.26	.26	.55	.18	.30	.22	.35	.34	.36	.22	.22	.89				
13. Pattern Reasoning	.30	.47	.31	.33	.46	.23	.40	.47	.43	.50	.48	.30	.92			
14. Hand Movements	.24	.25	.34	.13	.30	.17	.31	.30	.26	.26	.27	.38	.34	.84		
15. Rebus Delayed	.52	.36	.28	.29	.32	.46	.42	.46	.88	.36	.27	.30	.39	.22	.96	
16. Riddles	.40	.46	.40	.43	.43	.33	.75	.74	.53	.43	.44	.39	.48	.32	.48	.90

*Notes.* Internal Consistency Reliability for each subtest is given along the diagonal. Intercorrelations between subtests are given below the diagonal.

Table 4

*Index ICR and Intercorrelations of Ages 3 to 6 Sample*

Index	1	2	3	4	5	6	7	8
1. Nonverbal Index	.94							
2. Sequential/Gsm	.58	.91						
3. Simultaneous/Gv	.94	.51	.95					
4. Learning/Glr	.55	.49	.51	.98				
5. Knowledge/Gc	.68	.55	.67	.46	.94			
6. Delayed Recall	.49	.39	.44	.75	.31	.92		
7. Fluid Crystallized Intelligence	.87	.77	.87	.74	.84	.61	.97	
8. Mental Processing Index	.88	.78	.88	.78	.70	.64	.97	.96

*Notes.* Internal Consistency Reliability for each index is given along the diagonal. Intercorrelations between indices are given below the diagonal.

Table 5

*Index ICR and Intercorrelations of Ages 7 to 18 Sample*

Index	1	2	3	4	5	6	7	8	9
1. Nonverbal Index	.95								
2. Sequential/Gsm	.40	.91							
3. Simultaneous/Gv	.75	.34	.95						
4. Learning/Glr	.49	.39	.39	.97					
5. Planning/Gf	.86	.33	.58	.46	.91				
6. Knowledge/Gc	.61	.45	.49	.57	.57	.96			
7. Delayed Recall	.42	.31	.36	.84	.40	.50	.95		
8. Fluid Crystallized Intelligence	.82	.66	.74	.76	.77	.83	.66	.98	
9. Mental Processing Index	.83	.68	.76	.76	.79	.69	.65	.98	.97

*Notes.* Internal Consistency Reliability for each subtest is given along the diagonal. Intercorrelations between indices are given below the diagonal.

in the ages 7 to 18 sample. Variations of subtests in which scores can be recorded without time points, Triangles and Pattern Reasoning, were not included in the analysis because the manual supplement does not provide data for intercorrelations (Kaufman et al., 2018). The reliability for Delayed Recall is given alongside the subtests in the manual supplement (Kaufman et al., 2018); however, it is a combination of two subtests—Atlantis Delayed and Rebus Delayed—and thus was included in the analysis as an index.

All other subtests and indices were evaluated. This resulted in 135 subtest-level comparisons for the ages 3 to 6 sample, 120 subtest-level comparisons for the ages 7 to 18 sample, 28 index-level comparisons for the ages 3 to 6 sample, and 36 index level comparisons for the ages 7 to 18 sample for a total of 319 comparisons.

KABC-II NU subtest and index ICR and intercorrelation coefficients were recorded in Microsoft Excel 2016 spreadsheets by the primary author and an undergraduate research assistant. Rates of coding agreement were reviewed and were in agreement for 99% of data. IF-THEN syntax was used to detect disagreements between the primary author and the undergraduate research assistant. Disagreements were reviewed and corrected by referencing the appropriate KABC-II NU table.

The authors of the KABC-II NU manual supplement (Kaufman et al., 2018) did not provide subtest and index intercorrelation coefficients by ages 3 to 6 and ages 7 to 18 subsamples. To calculate discrepancy score reliability estimates, average ICR and intercorrelations for ages 3 to 6 and ages 7 to 18 were necessary. Average subtest and index ICR are provided by both age groupings (Kaufman et al., 2018), but intercorrelation coefficients are only reported by individual age grouping (i.e, for 3-year-olds, 4-year-olds, etc.). Therefore, a Fisher Z transformation was used to convert the intercorrelations into Z values, then they were averaged together into the two age groups of 3 to 6 and 7 to 18 (Fisher, 1921). The inverse Fisher Z transformation returned the average back to an r value.

Once all data were in the same format using the two age groups of 3 to 6 and 7 to 18, Thorndike and Hagen's (1961) formula was used to calculate reliability coefficients of discrepancy scores:

$$r = \frac{\left\{ \left[ \frac{r_a + r_b}{2} \right] - r_{ab} \right\}}{(1 - r_{ab})}$$

In Thorndike and Hagen's (1961) formula,  $r$  is the reliability of the difference score;  $r_a$  and  $r_b$  are the ICR coefficients for each of the contrast scores; and  $r_{ab}$  is the intercorrelation between both contrasted scores. Put more simply, the reliability of the difference score is calculated by finding the mean ICR of the two contrasted scores, subtracting them from the intercorrelation of the two contrasted scores, and dividing it all by one minus the intercorrelation of the two contrasted scores.

These calculations were completed using the following formula:  $\left( \left( \frac{r_a + r_b}{2} \right) - r_{ab} \right) / (1 - r_{ab})$  where each variable was identified by a specific cell in the Microsoft Excel 2016 sheet. All index- and subtest-level discrepancy scores from the KABC-II NU (Kaufman & Kaufman, 2018) were calculated and compared to reliability guidelines, where .80 is good for hypothesis generation and .90 is excellent for clinical decision-making (Nunnally, 1994; Hunsley and Mash, 2008, 2018).

These data are organized by subtests and indices and are discussed in aggregate form. Measures of central tendency (mean, median, range, and standard deviation) were calculated as well. Fishers Z transformation and inverse were used to calculate the mean ICR of the discrepancy scores and Microsoft Excel 2016 functions were used to calculate median, maximum and minimum values for range, and standard deviation.

## **Results**

A summary of results with percentages of subtests and indices that meet each of the reliability guidelines are provided in Table 6.

Table 6

*Summary of Discrepancy Score ICR*

	Reliability Guidelines		Total Comparisons
	≥ .80	≥ .90	
3-6 Sample			
Subtest-Level	70%	1%	135
Recommended Subtests	64%	7%	14
Index-Level	64%	14%	28
7-18 Sample			
Subtest-Level	64%	15%	120
Recommended Subtests	44%	22%	9
Index-Level	81%	22%	36

*Notes.* ≥ .80 is considered good and used for hypothesis generation. ≥ .90 is considered excellent and used for clinical decision-making.

**Subtest-Level Comparisons**

Subtest discrepancy score ICR coefficients for the KABC-II NU (Kaufman & Kaufman, 2018) ages 3 to 6 sample are displayed in Table 7. Reliabilities of the comparisons ranged from .61 to .94 (*Mdn* = .83; *M* = .83; *SD* = .06), with the Expressive Vocabulary – Riddles comparison at the low-end and Atlantis – Rebus comparison at the high-end of the range. Of the 135 subtest comparisons, 81 had reliability coefficients between .80 and .90, and 13 had reliability coefficients of .90 or higher. Roughly 70% met the .80 criterion while less than 1% met the .90 criterion.

The subtest ICR coefficients for discrepancy scores in the ages 7 to 18 sample are displayed in Table 8. Comparisons ranged from .56 to .94 (*Mdn* = .84; *M* = .84; *SD* = .06), with the Expressive Vocabulary – Riddles comparison at the low-end and Block Counting – Rebus Delayed comparison at the high-end of the range. Of the 120 subtest comparisons, 72 had reliability coefficients between .80 and .90, and 18 had

Table 7

*ICR of Subtest Discrepancy Scores for Ages 3 to 6 Sample*

Subtest	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Atlantis	-															
2. Conceptual Thinking	.90	-														
3. Face Recognition	.85	.79*	-													
4. Number Recall	.85	.84	.76	-												
5. Gestalt Closure	.87	.84	.72	.79	-											
6. Rover	.91	.82*	-	.83	.84	-										
7. Atlantis Delayed	.80	.84	.69	.78	.80	.78	-									
8. Expressive Vocabulary	.88	.80	.80	.78	.78	.84	.83	-								
9. Verbal Knowledge	.91	.82	.81	.80	.79	.82	.81	.73*	-							
10. Rebus	.94*	.89	.87	.86	.89	.88	.84	.88	.92	-						
11. Triangles	.91	.83*	.82*	.82	.83	.82*	.80	.83	.85	.89	-					
12. Block Counting	.87	.87*	.82*	.85	.85	.85*	.80	.86	.87	.91	.81*	-				
13. Word Order	.87	.79	.74	.69*	.79	.85	.79	.76	.78	.86	.79	.84	-			
14. Pattern Reasoning	.92	.80	.84	.84	.82	.83	.83	.81	.84	.89	.80	.83	.80	-		
15. Hand Movements	.86	.82	.71	.78	.78	.75	.79	.81	.81	.84	.83	.82	.69	.81	-	
16. Rebus Delayed	.93	.90	.84	.84	.89	.89	.83	.89	.91	.77	.88	.91	.85	.91	.84	-
17. Riddles	.88	.75	.78	.78	.75	.78	.80	.61*	.70*	.86	.76	.81	.70	.72	.75	.88

*Notes.* \* denotes a recommended subtest comparison. A discrepancy score reliability for Face Recognition and Rover is not given because the ages at which they may be administered do not overlap.

Table 8

*ICR of Subtest Discrepancy Scores for Ages 7 to 18 Sample*

Subtest	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Atlantis	-														
2. Story Completion	.84	-													
3. Number Recall	.86	.79	-												
4. Gestalt Closure	.82	.71	.77	-											
5. Rover	.94	.81	.85	.82	-										
6. Atlantis Delayed	.77	.80	.84	.78	.90	-									
7. Expressive Vocabulary	.87	.75	.78	.70	.86	.84	-								
8. Verbal Knowledge	.91	.79	.83	.78	.91	.87	.69*	-							
9. Rebus	.90*	.80	.84	.81	.92	.87	.85	.89	-						
10. Triangles	.89	.72	.82	.74	.84*	.86	.81	.84	.85	-					
11. Block Counting	.94	.81	.86	.81	.90*	.91	.87	.91	.93	.80*	-				
12. Word Order	.90	.80	.70*	.81	.88	.87	.82	.87	.88	.85	.90	-			
13. Pattern Reasoning	.91	.74*	.83	.78	.87	.88	.83	.87	.89	.79	.87	.86	-		
14. Hand Movements	.87	.77	.76	.79	.84	.84	.80	.84	.86	.80	.86	.78	.82	-	
15. Rebus Delayed	.92	.82	.86	.82	.93	.87	.86	.91	.64	.87	.94	.89	.90	.87	-
16. Riddles	.88	.73	.78	.73	.86	.85	.56*	.69*	.84	.80	.87	.83	.83	.81	.87

*Note.* \* denotes a recommended subtest comparison.

reliability coefficients of .90 or higher. Roughly 75% met the .80 criterion while 15% met the .90 criterion.

**Recommended Comparisons.** The KABC-II NU supplemental manual (Kaufman et al., 2018) provides recommendations for subtests comparisons that may be more suited to meaningful interpretation than others. There are 15 possible comparisons recommended. However, as the KABC-II NU has both a 3 to 6 and a 7 to 18 form that feature different subtests, not all recommended comparisons are available for both age groupings. The data for the ages 3 to 6 sample produced 14 comparisons. ICR coefficients for these data ranged from .61 to .94 (*Mdn* = .82; *M* = .81; *SD* = .08). Of the 14 comparisons, nine met the .80 criterion, but only one of those comparisons met the .90 criterion. The data for the ages 7 to 18 sample produced nine comparisons. ICR coefficients for these data ranged from .56 to .90 (*Mdn* = .74; *M* = .78; *SD* = .11). Of the eight comparisons, four met the .80 criterion and two of those comparisons met the .90 criterion.

### **Index Level Comparisons**

The index ICR coefficients for discrepancy scores in the ages 3 to 6 sample are displayed in Table 9. Comparisons ranged from .00 to .93. (*Mdn* = .83; *M* = .80; *SD* = .26), with the Fluid Crystallized Intelligence – Mental Processing Index comparison at the low-end and Simultaneous/Gv – Learning/Glr comparison at the high-end of the range. Of the 28 index comparisons, 14 had reliability coefficients between .80 and .90, and 4 had reliability coefficients of .90 or higher. Approximately 64% met the .80 criterion while only 14% met the .90 criterion.



Table 9

*ICR of Index Discrepancy Scores for Ages 3 to 6 Sample*

Index	1	2	3	4	5	6	7
1. Nonverbal Index	-						
2. Sequential/Gsm	.82	-					
3. Simultaneous/Gv	.10	.86	-				
4. Learning/Glr	.91	.89	.93	-			
5. Knowledge/Gc	.81	.83	.83	.93	-		
6. Delayed Recall	.86	.86	.88	.80	.90	-	
7. Fluid Crystallized Intelligence	.65	.74	.68	.90	.72	.86	-
8. Mental Processing Index	.57	.70	.64	.86	.83	.83	.00

The index ICR coefficients for discrepancy scores in the ages 7 to 18 sample are displayed in Table 10. Comparisons ranged from .00 to .93. (*Mdn* = .88; *M* = .86; *SD* = .18), with the Fluid Crystallized Intelligence – Mental Processing Index comparison at the low-end and Simultaneous/Gv – Learning/Glr comparison at the high-end of the range. Of the 36 index comparisons, 21 had reliability coefficients between .80 and .90, and 8 had reliability coefficients of .90 or higher. Approximately 81% met the .80 criterion while 22% met the .90 criterion.

Table 10

*ICR of Index Discrepancy Scores for Ages 7 to 18 Sample*

Index	1	2	3	4	5	6	7	8
1. Nonverbal Index	-							
2. Sequential/Gsm	.88	-						
3. Simultaneous/Gv	.80	.89	-					
4. Learning/Glr	.92	.90	.93	-				
5. Planning/Gf	.51	.86	.84	.89	-			
6. Knowledge/Gc	.88	.88	.91	.92	.85	-		
7. Delayed Recall	.91	.90	.92	.75	.88	.91	-	
8. Fluid Crystallized Intelligence	.81	.84	.87	.90	.76	.83	.90	-
9. Mental Processing Index	.77	.82	.83	.88	.71	.89	.88	.00

## Discussion

Reliability is a foundational element of both score validity and diagnostic utility and is a central pillar to evidence-based assessment (Haynes et al., 2011; NASP, 2010). Despite this, published approaches to intelligence test interpretation recommend using discrepancy scores extensively (Flanagan et al., 2013; Kaufman et al., 2015; Sattler, 2008) and many practitioners continue to engage in discrepancy score analysis (Cottrell & Barrett, 2016; McGill et al., 2018; Pfeiffer et al., 2000). These procedures must be held to the overall standard of the profession and to the ethical codes of both APA (2017) and NASP (2010) which both state the need for the practitioner to take initiative in making certain their practice is evidence-based. Best practice in school psychology, and assessment practices in general, is to ensure assessment results meet a minimum standard of reliability before they are interpreted (AERA et al., 2014; American Psychological Association Task Force on Evidence-Based Practice for Children and Adolescents, 2008; Gross et al., 2018; Hunsley & Mash, 2007, 2018; Kranzler & Floyd, 2013; Reynolds & Livingston, 2014). Research exploring the reliability of difference scores from a variety of instruments (Brown & Ryan, 2004; Charter, 2001, 2002; Glass et al., 2010; Glass et al., 2009; Ryan & Brown, 2005) have found reliability coefficients similar to those observed from comparisons on the KABC-II NU (Kaufman & Kaufman, 2018).

Evaluating these results with our hypotheses, we see that the vast majority of ICR coefficients of discrepancy scores produced from the subtests on the KABC-II NU do not meet the excellent standard, and not all of them were able to meet the good criterion. Comparisons between index scores were better at meeting the excellent standard of .90, but still did not meet criterion for 25% percent of the comparisons in either age group.

Even when using the .80 criterion for index scores, all comparisons did not meet the threshold. These findings support the claim in the literature that all differences in scores within an intelligence test may not be suitable for interpretation and used for educational decision-making as they commonly are today (Canivez, 2013; Charter, 2002; Glass et al., 2009, Glass et al., 2010; Kranzler, Floyd, et al., 2016; McGill & Busse, 2017; Watkins, 2000). However, the purpose of the assessment and purpose of score use must also be considered (Charter, 2002; Haynes et al., 2011; Mash & Hunsley, 2008, 2018).

Authors who have explored discrepancy scores in the past have argued that some are adequate for hypothesis generation (Brown & Ryan, 2004; Charter, 2001, 2002; Glass et al., 2010; Glass et al., 2009; Ryan & Brown, 2005), which is consistent with the positions of various scholars who have promoted discrepancy-based interpretation strategies (e.g., Hale et al., 2006; Kaufman et al., 2015; Flanagan et al., 2013). While the scores may be appropriate for hypothesis generation, it is important to recognize that discrepancy scores in these interpretative strategies are not used merely for hypothesis generation. PSW is a clinical decision-making tool that uses discrepancy scores as a central component of its analyses, meaning the previous claims that hypothesis generation levels of reliability are acceptable for use are mostly invalid as that is not how discrepancy scores are utilized in practice (McGill, et al., 2018).

### **Limitations and Future Research**

While the .90 criterion for reliability we used in these analyses is established in the literature (Beidas et al., 2015; DeVon et. al, 2007; Floyd et al., 2015; Hunsley & Mash, 2008, 2018; Kline, 2000; Nunnally & Bernstein, 1994; Ponterotto & Ruckdeschel, 2007; Reynolds & Livingston, 2014), others have recommended a higher standard of .95

(Nunnally, 1978; Kranzler & Floyd, 2013). Charter (2001) offered a counter-argument, writing that reliability standards may be too high due to their relation to test length, as one could simply add more equivalent items until a higher level of reliability was achieved. Indeed, the reliability of a score that is too high indicates redundancy (Streiner, 2003). The more items used, in most cases, the higher the reliability. He argued that .90 may be too rigid to be a universal standard. Others have suggested .85 may be more appropriate for decision-making (Aiken, 1991; Rosenthal & Rosnow, 1991; Weiner & Stewart, 1984). As such, the choice to use Hunsley & Mash's (2008, 2018) model may also be contested (e.g., Charter, 2001; Streiner & Norman, 2008). We interpreted the results using Hunsley & Mash's criteria (2008, 2018), but also provided the reliabilities in Tables 7 – 10 so that these scores may be interpreted by other criteria deemed more fitting if one was to disagree with our rationale for using .80 and .90 as standards.

If discrepancy scores meet reliability criteria and are going to be considered for interpretation, a confidence interval (CI) must be calculated around the discrepancy score (Charter, 1999). While beyond the scope of this project, CI bounds may further limit or support the interpretability of some discrepancy scores and warrants further investigation based on recommendations by Charter (1999) and Charter and Feldt (2009). CI bounds provide the likely range in which a discrepancy score reliability coefficient may fall. As reliability decreases CI bounds increase, which means lower ICR values lead to larger bounds of the CI. The limits of the CI bounds would then be compared against established criterion for determining significance (Charter & Feldt, 2009), such as those corresponding to p-values of .05 and .10 criteria, which would provide an additional level of interpretation than this paper was intended to offer.

The intent of this paper was to evaluate ipsative comparisons, those within the test, and did not consider comparing scores across assessments (e.g., calculating a difference between a KABC-II NU composite score and a WISC-V composite score). Cross-battery assessment strategies (XBA; Flanagan et al., 2013; Ortiz et al., 2017) go beyond interpretations of the KABC-II NU itself, and thus these data cannot be generalized to cross-battery score comparisons. Likewise, these analyses do not include scores from achievement assessments. As such, these analyses are not intended to and cannot inform the ability-achievement discrepancy approach to diagnosing SLD (Flanagan & Alfonso, 2011; Hale et al., 2006; Kavale & Flanagan, 2007).

The KABC-II NU manual supplement (Kaufman et al., 2018) also suggests practitioners should consider comparison of subtest scores to mean subtest scores within indices (i.e. averaging the subtests that combine to form an index and comparing that number to a score of a specific subtest within the set). These types of comparisons are unique in that they reflect mean-to-subtest comparisons rather than subtest-to-subtest or index-to-index comparisons. As such, they require different data than was available in the manual supplement (Kaufman et al., 2018), and would require different methodology (e.g., calculation of stratified coefficient alpha and determination of intercorrelations from a sample of data). Further research should investigate the reliability of mean-to-subtest comparisons but doing so was outside the scope of this project.

This investigation used the standardization sample to evaluate reliabilities using data reported in the manual supplement (Kaufman et. al., 2018). While the sample used to norm the test was representative of the population with whom the assessment will be used (Ruggles et al., 2017), using only the publisher's data instead of the raw data prevents a

more in-depth evaluation of the relationships among the various demographic variables. Especially pertinent to special education, it may be beneficial for future research to examine whether the reliabilities are variant across subsamples (e.g., those with and without a diagnosis of SLD). Given that discrepancy scores are used more frequently in the presence of some referral concerns, these subsample analyses would provide more information about the reliability of discrepancy scores for their intended uses.

### **Implications for Practice**

Because there is so much weight placed on the interpretations of intelligence test results, practitioners must take care to ensure that they are making evidence-based decisions using data that are reliable, accurate, and have diagnostic utility (AERA et al., 2014; APA, 2017; Canivez, 2013; Gross et al., 2018; Hunsley & Mash, 2018; Kranzler & Floyd, 2013; NASP, 2010). Discrepancy scores, in general, may be adequate for hypothesis generation in many cases, yet their use in clinical decision-making should be avoided in all but the rare case (Brown & Ryan, 2004; Charter, 2001, 2002; Glass et al., 2010; Glass et al., 2009; Ryan & Brown, 2005). However, some of the most common uses of discrepancy scores are as part of larger interpretive systems (McGill et al., 2018).

If discrepancy score interpretations in isolation may introduce measurement error into clinical decisions, utilizing discrepancy scores as part of larger interpretive systems would also introduce error. While discrepancy scores may be appropriate for hypothesis generation based on reliability alone, they are not used as such in practice and thus warrant additional caution in their use. Despite this, not all discrepancy scores available from the KABC-II NU (Kaufman et al., 2018) had reliabilities below .80, and a few even met the excellent standard.

In some cases, it may be appropriate to interpret assessment results through discrepancy scores based on ICR. However, the practitioner must consider the specific comparison to evaluate whether it meets the appropriate threshold for their interpretive purposes. Tables 7 – 10 provide ICRs for subtests and indices across the two age groups which allows the practitioner to view the ICR of a potential discrepancy score. Guidelines discussed in this paper provide a basis for evaluating the reliability of discrepancy scores, but it is not possible to say that in every evaluation certain comparisons can be used and others cannot. As such, the responsibility for appropriately using discrepancy scores falls to the school psychology practitioner (AERA et al., 2014; APA, 2017; NASP, 2010). Some comparisons meet the hypothesis generation standard and may be considered to gain more information about the child while others meet the clinical decision-making standard and could be used for adapting a treatment plan. However, even when a certain discrepancy score meets the reliability standard for the intended purpose, caution must be taken when extrapolating test scores and making meaning of differences, as validity must also be considered for every score used.

Because the burden falls on the practitioner to correctly interpret and utilize test results, universities and training programs must change to better support the practitioner and root their trainings in evidence-based practice. A change in policy at the district, state, or national levels to encourage a greater focus on reliability in assessment interpretations may also be necessary to see a paradigm shift away from the procedures of the past (Cottrell & Barrett, 2016; Alfonso et al., 2000) to a future that is oriented towards better and more reliable clinical decision-making.

The focus of this discussion has been entirely on reliability, as is warranted by the data available from this study. However, reliability is a necessary, but not sufficient, requirement for score use. Practitioners also should ensure that a test is both valid and useful in addition to reliable prior to their use.

### **Conclusion**

Practitioners bear the responsibility of ensuring their assessment practice meets standards of evidence-based practice and they cannot rely on interpretive strategies that have not been evaluated. This study examined the reliability of discrepancy scores from the KABC-II NU (Kaufman & Kaufman, 2018) and found ICR to be lacking in the majority of comparisons when evaluated for purposes of decision-making, “excellent” criteria, and about one third of comparisons for hypothesis generation, “good” criteria (Hunsley & Mash, 2018). It is recommended that practitioners either eschew or individually evaluate the evidence supporting specific discrepancy scores produced from the KABC-II NU (Kaufman & Kaufman, 2018) when making educational decisions. If practitioners still wish to interpret discrepancy scores, it is recommended this be done with extreme caution and careful consideration of the established reliability, validity, and utility of the particular score they wish to interpret.



## References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. Washington, DC: Author.
- American Psychological Association Task Force on Evidence-Based Practice for Children and Adolescents. (2008). *Disseminating evidence-based practice for children and adolescents: A systems approach to enhancing care*. Washington, DC: American Psychological Association.
- Alfonso, V. C., LaRocca, R., Oakland, T. D., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review, 29*(1), 52-64.
- Bardon, J. L. (1979). How best to establish the identity of professional school psychology. *School Psychology Digest, 8*, 162–167.
- Bardon, J. L. (1994). Will the real school psychologist please stand up: is the past a prologue for the future of school psychology? The identity of school psychology revisited. *School Psychology Review, 23*, 584–588.
- Beaujean, A. A., & Benson, N. F. (2018). Theoretically-consistent cognitive ability test development and score interpretation. *Contemporary School Psychology, 1*-12.  
doi: 10.1007/s40688-018-0182-1
- Beaujean, A. A., Benson, N. F., McGill, R. J., & Dombrowski, S. C. (2018). A misuse of IQ scores: Using the Dual Discrepancy/Consistency Model for identifying

- specific learning disabilities. *Journal of Intelligence*, 6(36), 1: 1-25. doi: 10.3390/jintelligence6030036
- Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., ... & Mandell, D. S. (2015). Free, brief, and validated: standardized instruments for low-resource mental health settings. *Cognitive and Behavioral Practice*, 22(1), 5-19. doi: 10.1016/j.cbpra.2014.02.002
- Benson, N., Floyd, R. G., Kranzler, J. H., Eckert, T. L., & Fefer, S. (February 2018). *Contemporary assessment practices in school psychology: National survey results*. Paper Presented at the annual convention of the National Association of School Psychologists, Chicago, IL.
- Brown, K. I., & Ryan, J. J. (2004). Reliabilities of the WAIS-III for discrepancy scores: Generalization to a clinical sample. *Psychological Reports*, 95(3), 914-916. doi: 10.2466/pr0.95.3.914-916
- Canivez, G. L. (2013). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwann (Eds.), *The Oxford handbook of child psychological assessment* (pp. 84-112). New York, NY: Oxford University Press.
- Cottrell, J. M., & Barrett, C. A. (2016). Defining the undefinable: operationalization of methods to identify specific learning disabilities among practicing school psychologists. *Psychology in the Schools*, 53(2), 143-157. doi: 10.1002/pits.21892
- Charter, R. A. (1999). Testing for true score differences using the confidence interval method. *Psychological Reports*, 85, 808.

- Charter, R. A. (2001). Discrepancy scores of reliabilities of the WAIS-III. *Psychological Reports, 89*(2), 453-456. doi: 10.2466/PR0.89.6.453-456
- Charter, R. A. (2002). Reliability of the WMS-III Discrepancy Comparisons. *Perceptual and motor skills, 94*(2), 387-390. doi: 10.2466/pms.2002.94.2.387
- Charter, R. A., & Feldt, L. S. (2009). A comprehensive approach to the interpretation of difference scores. *Applied Neuropsychology, 16*, 23-30. doi: 10.1080/09084280802644110
- 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*(3), 300-313. doi: 10.1002/pits.21675
- DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., ... & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing scholarship, 39*(2), 155-164. doi: 10.1111/j.1547-5069.2007.00161.x
- Fagan, T. K. (2014). Trends in the history of school psychology in the United States. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology VI* (pp. 383-400). Bethesda: National Association of School Psychologists.
- Farmer, R. L. & Floyd, R. G. (2018). The use of intelligence tests in the identification of children and adolescents with intellectual disability. In D. P. Flanagan and E. M. McDonough (Eds.), *Contemporary intellectual assessment* (4th ed.). New York, NY: Guilford Press.
- Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron, 1*, 3-32.

- Flanagan, D. P., & Alfonso, V. C. (2017). *Essentials of WISC-V assessment*. John Wiley & Sons.
- Flanagan, D. P., Alfonso, V. C. (2011). *Essentials of specific learning disability identification*. Hoboken, NY: John Wiley & Sons.
- Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3<sup>rd</sup> ed.). Hoboken, NJ: Wiley & Sons.
- Flanagan, D. P., & Schneider, W. J. (2016). Cross-Battery Assessment? XBA PSW? A case of mistaken identity: A commentary on Kranzler and colleagues' "Classification agreement analysis of Cross-Battery Assessment in the identification of specific learning disorders in children and youth". *International Journal of School & Educational Psychology*, 4(3), 137-145. doi: 10.1080/21683603.2016.1192852
- Fletcher, J. M., & Miciak, J. (2017). Comprehensive cognitive assessments are not necessary for the identification and treatment of learning disabilities. *Archives of Clinical Neuropsychology*, 32(1), 2-7. doi: 10.1093/arclin/acw103
- Floyd, R. G., Farmer, R. L., Schneider, W. J., & McGrew, K. S. (in press). Theories and measurement of intelligence. In L. M. Glidden (Ed.), *APA Handbook of Intellectual and Developmental Disabilities*. Washington, D.C: American Psychological Association.
- Floyd, R. G., Shands, E. I., Alfonso, V. C., Phillips, J. F., Autry, B. K., Mosteller, J. A., & Irby, S. (2015). A systematic review and psychometric evaluation of adaptive behavior scales and recommendations for practice. *Journal of Applied School Psychology*, 31(1), 83-113. doi: 10.1080/15377903.2014.979384

- Gambrill, E. D. (2012). *Critical thinking in clinical practice: Improving the quality of judgments and decisions*. (3rd ed.). Hoboken, NJ: John Wiley & Sons.
- Glass, L. A., Ryan, J. J., Charter, R. A., & Bartels, J. M. (2009). Discrepancy score reliabilities in the WISC-IV standardization sample. *Journal of Psychoeducational Assessment*, 27(2), 138-144. doi: 10.1177/0734282908325158
- Glass, L. A., Ryan, J. J., & Charter, R. A. (2010). Discrepancy score reliabilities in the WAIS-IV standardization sample. *Journal of Psychoeducational Assessment*, 28(3), 201-208. doi: 10.1177/0734282909346710
- Gross, T. J., Farmer, R. L., & Ochs, S. E. (2018). Evidence-based assessment: Best practices, customary practices, and recommendations for field-based assessment. *Contemporary School Psychology*, doi: 10.1007/s40688-018-0186-x
- Hale, J. B., Kaufman, A., Naglieri, J. A., & Kavale, K. A. (2006). Implementation of IDEA: Integrating response to intervention and cognitive assessment methods. *Psychology in the Schools*, 43(7), 753–770. doi: 10.1002/pits.20186
- Haynes, S. N., Smith, G. T., Hunsley, J. D. (2011). *Scientific foundations of clinical assessment*. New York City: Routledge.
- Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology*, 3(1), 29-51. doi: 10.1146/annurev.clinpsy.3.022806.091419
- Hunsley, J., & Mash, E. J. (2008). *A guide to assessments that work*. London: Oxford University Press.
- Hunsley, J. D., & Mash, E. J. (2018). *A guide to assessments that work* (2nd ed.). London: Oxford University Press.

- Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2018). A history of intelligence test interpretation. In D. P. Flanagan and E. M. McDonough (Eds.) *Contemporary Intellectual Assessment: Theories, Tests and Issues* (3rd ed.). New York: Guilford Press.
- Kaufman, A. S. & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children, second edition* [Measurement instrument]. Bloomington, MMN: NCS Pearson.
- Kaufman, A. S. & Kaufman, N. L. (2018). *Kaufman Assessment Battery for Children, second edition, normative update* [Measurement instrument]. Bloomington, MMN: NCS Pearson.
- Kaufman, A. S., Kaufman, N. L., Drozdick, L.W., & Morrison, J. (2018). *Kaufman Assessment Battery for Children, second edition, normative update, manual supplement*. Bloomington, MN: NCS Pearson.
- Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2015). *Intelligent testing with the WISC-V*. John Wiley & Sons.
- Kaufman, A. S., & Wechsler, D. (1979). *Intelligent testing with the WISC-R*. New York: John Wiley & Sons.
- Kavale, K. A., & Flanagan, D. P. (2007). Ability—achievement discrepancy, response to intervention, and assessment of cognitive abilities/processes in specific learning disability identification: Toward a contemporary operational definition. In *Handbook of response to intervention* (pp. 130-147). Boston, MA: Springer.
- Kline, P. (2000). *Handbook of psychological testing*. New York: Routledge.
- Kranzler, J. H., Benson, N., & Floyd, R. G. (2016a). Intellectual assessment of children and youth in the United States of America: Past, present, and future. *International*

*Journal of School & Educational Psychology*, 4(4), 276-282. doi:  
10.1080/21683603.2016.1166759

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

Kranzler, J. H., Floyd, R. G., Benson, N., Zaloski, B., & Thibodaux, L. (2016b). Cross-Battery Assessment pattern of strengths and weaknesses approach to the identification of specific learning disorders: Evidence-based practice or pseudoscience? *International Journal of School & Educational Psychology*, 4(3), 146-157. doi: 10.1080/21683603.2016.1192855

Luria, A. R. (1973). *The working brain: An introduction to neuropsychology*. London: Penguin Books.

Maki, K. E., Floyd, R. G., & Roberson, T. (2015). State learning disability eligibility criteria: A comprehensive review. *School Psychology Quarterly*, 30(4), 457-469. doi: 10.1037/spq0000109

Marsh, J. K., De Los Reyes, A., & Lilienfeld, S. O. (2018). Leveraging the Multiple Lenses of Psychological Science to Inform Clinical Decision Making: Introduction to the Special Section. *Clinical Psychological Science*, 6(2), 167-176. doi: 10.1177/2167702617736853

Mather, N., & Wendling, B. J. (2015). *Essentials of WJ IV tests of achievement*. Hoboken, NJ: John Wiley & Sons.

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

- McGill, R. J., & Busse, R. T. (2017). When theory trumps science: A critique of the PSW model for SLD identification. *Contemporary School Psychology, 21*(1), 10-18.  
doi: 10.1007/s40688-016-0094-x
- 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.
- 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*(1), 75-82. doi: 10.1037/spq0000208
- National Association of School Psychologists. (2010). *Principles for professional ethics*. National Association of School Psychologists.
- National Association of School Psychologists. (2016). *School Psychologists' Involvement in Assessment* (Position Statement). Bethesda, MD: Author.
- Norfolk, P. A., Farmer, R. L., Floyd, R. G., Woods, I. L., Hawkins, H. K., & Irby, S. M. (2015). Norm Block Sample Sizes: A Review of 17 Individually Administered Intelligence Tests. *Journal of Psychoeducational Assessment, 33*(6), 544–554. doi: 10.1177/0734282914562385
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Ortiz, S. O., Flanagan, D. P., & Alfonso, V. C. (2017). *Cross battery assessment software system 2.0* [computer software]. Hoboken, NJ: Wiley.



- 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*(4), 376-385. doi: 10.1037/h0088795
- Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and motor skills, 105*(3), 997-1014. doi: 10.2466/pms.105.3.997-1014
- Price, L. R. (2016). *Psychometric methods: Theory into practice*. New York, NY: Guilford Publishing.
- Reynolds, C. R., & Kamphaus, R. W. (2015). *Reynolds Intellectual Assessment Scales, Second Edition*. Lutz, FL: Psychological Assessment Resources.
- Reynolds, C. R., & Livingston, R. B. (2014). A psychometric primer for school psychologists. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology* (6th ed., pp. 281–300). Bethesda, MD: National Association of School Psychologists.
- Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.
- Ruggles, S., Genadek, K., Goeken, R., Grover, J., & and Sobek, M. (2017). *Integrated Public Use Microdata Series: Version 7.0 [Data set]*. Minneapolis: University of Minnesota. doi: 10.18128/D010.V7.0
- Rust, J. & Golombok, S. (1999). *Modern psychometrics: The science of psychological assessment* (2nd ed.). New York, NY: Routledge.

- Ryan, J. J., & Brown, K. I. (2005). Enhancing the clinical utility of the WASI: Reliabilities of discrepancy scores and supplemental tables for profile analysis. *Journal of Psychoeducational Assessment, 23*, 140-145. doi: 10.1177/073428290502300203
- Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego, CA: Author.
- Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73-164). New York: Guilford Press.
- Schneider, W. J., & Roman, Z. (2017). Fine-tuning cross-battery assessment procedures: after follow-up testing, use all valid scores, cohesive or not. *Journal of Psychoeducational Assessment, 36*, 34–54. doi: 10.1177/0734282917722861
- Schrank, F. A., Decker, S. L., & Garruto, J. M. (2016). *Essentials of WJ IV cognitive abilities assessment*. Hoboken, NJ: Wiley.
- Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. J., Buntinx, W. H., Coulter, D. L., Craig, E. M., ... & Shogren, K. A. (2010). *Intellectual disability: Definition, classification, and systems of supports*. Washington, DC: American Association on Intellectual and Developmental Disabilities.
- Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment, 80*(1), 99-103. doi: 10.1207/S15327752JPA8001\_18

- Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). New York: Oxford University Press.
- Thorndike, R. L., & Hagen, E. (1961). *Measurement and evaluation in psychology and education* (2nd ed.). Oxford, England: Wiley.
- U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. (2018, April). Children and youth with disabilities. Retrieved from [https://nces.ed.gov/programs/coe/pdf/coe\\_cgg.pdf](https://nces.ed.gov/programs/coe/pdf/coe_cgg.pdf)
- Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement & Evaluation in Counseling and Development, 44*(3), 159-168. doi: 10.1177/0748175611409845
- Walcott, C. M., Charvat, J., McNamara, K. M., & Hyson, D. M. (2016). School psychology at a glance: 2015 member survey results. In *Special session presented at the annual meeting of the National Association of School Psychologists, New Orleans, LA*.
- Watkins, M. W. (2000). Cognitive profile analysis: A shared professional myth. *School Psychology Quarterly, 15*(4), 465-479. doi:10.1037/h0088802
- Watkins, M.W., Crosby, E. G., & Pearson, J. L. (2001). Role of the school psychologist: perceptions of school staff. *School Psychology International, 22*(1), 64–73. doi: 10.1177/01430343010221005
- Wechsler, D. (1997a). *WAIS-III Administration and scoring manual*. San Antonio, TX: The Psychological Association.
- Wechsler, D. (1997b). *Wechsler Memory Scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.

- Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (2008a). *WAIS-IV administration and scoring manual*. San Antonio, TX: Pearson.
- Wechsler, D. (2008b). *WAIS-IV technical and interpretive manual*. San Antonio, TX: Pearson.
- Wechsler, D. (2014). *Wechsler Intelligence Scale for Children, Fifth Edition*. Bloomington, MN: Pearson.
- Wechsler, D., Raiford, S. E., & Holdnack, J. A. (2014). *WISC-V technical and interpretive manual*. Bloomington, MN: Pearson.
- Weiner, E. A., & Stewart, B. J. (1984). *Assessing individuals: Psychological and educational tests and measurement*. Boston, MA: Little, Brown.