Spring 2020

# Reliability of Index and Subtest Discrepancy Scores on The WJ-IV Cognitive

Kacy Stinson
*Western Kentucky University*, kacy.stinson825@topper.wku.edu

# RELIABILITY OF INDEX AND SUBTEST DISCREPANCY SCORES ON THE WJ-IV COGNITIVE

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
Kacy Stinson

May 2020

RELIABILITY OF INDEX AND SUBTEST DISCREPANCY SCORES ON THE WJ-IV COGNITIVE

Date Recommended 4/10/2020

Sarah Ochs, Ph.D, Director of Specialist Project

Ryan Farmer, Ph.D.

Carl Myers

Carl Myers, Ph.D.

Cheryl D Davis Digitally signed by Cheryl D Davis
Date: 2020.05.15 08:19:56 -05'00'

Dean, The Graduate School Date

# CONTENTS

## LIST OF TABLES

# RELIABILITY OF INDEX AND SUBTEST DISCREPANCY SCORES ON THE WJ-IV COGNITIVE

Kacy Stinson May 2020 36 Pages

Directed by: Sarah Ochs, Ryan Farmer, and Carl Myers

Department of Psychology Western Kentucky University

Many guidebooks and publication manuals for intelligence tests recommend that practitioners calculate subtest and index level discrepancy scores along with many other types of scores. The reliability of the discrepancy scores, however, are not provided in the test manuals. The purpose of the study was to determine the reliability of the discrepancy scores in the Woodcock Johnson Test of Cognitive Abilities (WJ IV COG, 2014). Using data from the WJ IV technical manual, the reliability of the discrepancy scores were examined at both the subtest and index level. The results were compared to the reliability criteria recommended by the assessment community for both hypothesis generation and clinical decision making. At the subtest-level, the reliability coefficients ranged from .61 to .93. The composite-level reliability coefficients ranged from .59 to .93. Limitations and future directions are discussed.

# Introduction

School psychologists are among the few individuals who are qualified to conduct special education evaluations and interpret those data in the schools (National Association of School Psychologists [NASP], 2009; Ysseldyke et al., 2006). The NASP survey of professionals in 2010 (Castillo, Curtis, & Gelley, C. 2012) and 2015 (Walcott, Charvat, McNamar, & Hyson, 2016) identified that the most frequent role of a school psychologist was evaluating students to determine special education eligibility. Similarly, intelligence tests are among the most frequently used assessment strategies by practitioners (Benson et. al., 2019; Sotelo-Dynega & Dixon, 2014). Since the purpose of these assessments are typically to determine special education eligibility, the outcomes have a significant impact on students, parents, and other stakeholders in the child's life. That being said, it is critical for school psychologists to make data-based decisions using the most reliable testing methods (NASP, 2009; 2010).

## The Status of Intelligence Testing in School Psychology

Intelligence tests are used primarily to determine if an individual meets the criteria for an intellectual or developmental disability (Schalock et al., 2010; American Psychiatric Association, 2013; Farmer & Floyd, 2018; Kranzler, 2016; McNicholas et al., 2018). However, intelligence tests are also frequently used by practitioners to help determine whether or not an individual has a specific learning disability (SLD; Kranzler, Floyd, & Benson, 2016; Maki, Floyd, & Roberson, 2015), to identify gifted students (McClain & Pfeiffer, 2012), as a component of emotional disturbance and other IDEA classifications (Sotelo-Dynega & Dixon, 2014), to aid in treatment selection (Flanagan, Ortiz, & Alfonso, 2013), as part of overall case conceptualization (Floyd, Farmer, Schneider, & McGrew, in press), and various other disorders based on DSM and other available guidelines.

Within the schools, school psychologists report that they allocate a majority of their time to conducting assessments which often include the use of intelligence tests (Castillo et al., 2012; Walcott et al., 2016). Although the use of these tests are frequent and used for many purposes, there has been a longstanding debate in the field of school psychology regarding test interpretation on cognitive assessments (McGill et al., 2018). Various interpretation strategies are often described in textbooks (Flanagan & Alfonso, 2017; Kaufman, Raiford, & Coalson, 2016; Sattler, 2018) and within test administrative manuals (e.g., WJ IV COG; Schrank, & Dailey, 2014) which recommend several scores for practitioners to interpret, including discrepancy scores. Some test developers recommend interpreting multiple scores such as the aggregate score or full scale intelligence quotient (FSIQ), index scores, index score differences, and more. These different uses typically rely on varying interpretation strategies, including emphasis on general intelligence (e.g., General Intellectual Ability; e.g., Canivez, 2013; Kranzler & Floyd, 2013), interpretation of second-order composites based on Cattell-Horn-Carroll (CHC) factors (e.g., Comprehension Knowledge), interpretation of subtests and items (for review, Sattler, 2018), as well as comparisons between CHC factors (e.g., Comprehension Knowledge minus Working Memory) and subtests within (e.g., Oral Vocabulary minus Number Series) and across (e.g., Oral Vocabulary minus Number Series) CHC factors (Flanagan et al., 2013).

Assessing patterns within an individual's profile, profile analysis, typically includes evaluating the level of CHC factors as well as the individual's unique pattern of strengths and weaknesses across those factors, and sometimes across subtests (McGill et al., 2018). While measures of single scores (i.e., those representing general intelligence, individual CHC factors, or subtests) can be interpreted as indicators of a single construct, comparisons must be interpreted as the difference, or discrepancy, between the two scores—or two constructs. The implication of the former is that the construct (e.g., general intelligence) has predictive or classification validity of some kind, while the latter implies that the discrepancy between the two constructs (e.g., the difference between working memory and comprehension knowledge) is meaningful for clinical decision making (Canivez, 2013; McGill et al., 2018).

Although these various types of score interpretations are recommended by test developers, others (e.g., Beaujean, Benson, McGill, & Dombrowski, 2018; Kranzler, Floyd, Benson, Zaboski, & Thibodaux, 2016a; Kranzler, Floyd, Benson, Zaboski, & Thibodaux, 2016b; McDermott, Fantuzzo, & Glutting, 1990; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992; McGill & Busse, 2017; McGill, Dombrowski, & Canivez, 2018; Watkins, 2002; Watkins, Kush, & Glutting, 1997; Watkins, Kush, & Schaefer, 2002; Zaboski, Kranzler, & Gage, 2018) have highlighted concerns about over-interpretation of intelligence test scores and patterns among scores, including those strategies that rely on the direct comparison of two scores. Despite the fact that concerns have been raised regarding the over-interpretation of intelligence test scores, there are many school psychologists who continue to interpret multiple scores within intelligence tests. Specifically, Sotelo-Dynega and Dixon (2014) surveyed 323 school psychologists and found that they frequently used a systematic interpretive framework (81.6%), interpretive recommendations outlined in examiner’s manuals (67.8%), Sattler’s

"Assessment of Children" framework (51.2%), and Cross-Battery Assessment (24.4%) strategies for interpretation.

### What Are Discrepancy Scores?

There has been a long history of the use of discrepancy scores when interpreting the results of cognitive assessments. Discrepancy scores are typically produced by simply subtracting two scores from each other to find the discrepancy, or difference, between the two scores. From there, the cognitive skills are compared and interpreted for clinical use. The WISC-V publishers, for example, suggest that inferences can be drawn from the different types of scores that can be derived from their cognitive assessment such as diagnosing attention disorders, learning disorders, and autism spectrum disorder (Beaujean & Benson, 2018). This framework is known as the discrepancy/consistency model and it is used to find cognitive weaknesses by demonstrating a relative weakness in a score through the use of an ipsative analysis or identifying a normative weakness.

Although intelligence test scores have been interpreted in this manner, there are concerns regarding the over interpretation of intelligence test scores. McGill et al. (2018) stated that when two scores on an intelligence measure are positively correlated, the reliability of the difference score is not as strong. Additionally subtest level indices do not have the same stability and are not free of systematic measurement error in comparison to composite scores; thus, further compromising the reliability of the subtest level indices. When inferences are drawn from looking at the patterns of scores or discrepancy scores as recommended by alternative interpretive strategies (e.g., Patterns of Strengths and Weaknesses, Cross-battery assessments), the interpretation typically has significantly lower internal consistency reliability and lower stability over time (McGill et al., 2018). In contrast, when the overall score from a cognitive assessment, a score representing psychometric *g*, is used for interpretation, it typically has higher reliability (see Kranzler & Floyd, 2013) and stability (Canivez, 2013).

When interpreting scores from cognitive assessments, it is essential that the score meets the minimum reliability requirements necessary for score interpretation. Hunsley and Mash (2008) note that clinicians should follow a "good enough" principle when determining if an instrument or assessments psychometric values are appropriate for clinical use. A measure is deemed "good enough" for clinical use if it falls into one of the three categories: adequate, good, or excellent. Adequate internal consistency reliability on a measure indicates that α value is between .70-.79, good is between .80-.89, and excellent is greater than or equal to .90.

## Evidence-Based Assessment and Psychometric Criteria

As previously mentioned, practitioners conduct assessments for a wide variety of reasons (e.g., diagnosis, overall case conceptualization, treatment planning). The results of an assessment can have a significant and lasting impact on the lives of others, therefore it is essential that practitioners use strategies with sufficient evidence in order to provide the highest quality of care.  Evidence-Based Assessments (EBA) are viewed as problem-specific approaches that utilize assessments that are psychometrically sound, are appropriate for the context, and the data derived from the assessments are used to guide treatment (Hunsley & Mash, 2005). Having the acceptable evidence of reliability is the first step to determining that a score is appropriate to make clinical decisions. When conducting an EBA it is important that the measures being used are psychometrically strong by ensuring appropriate evidence of reliability, validity, and clinical utility.

## The Standards for Assessment

NASP (2009) indicates the need for school psychology practitioners to be knowledgeable about assessment, citing *The Standards for Educational and Psychological Assessment* (i.e., *The Standards*; American Education and Research Association [AERA], American Psychological Association [APA], & National Council for Measurement in Education [NCME], 1999) as a guiding document. The Standards (AERA et al., 1999) were updated in 2014 (AERA et al., 2014) but are largely consistent in that they mandate that scores must first be reliable and valid for the intended use. The purpose of the Standards are to provide guidelines on appropriate testing practices for individuals who conduct psychological assessments. Some of the areas that the Standards address are test development, evaluation of results, test selection, and score validity and reliability (AERA et al., 2014).

Within the Standards, validity is defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11) and reliability is defined as the "consistency of the scores across instances of the testing procedure" (AERA et al., 2014, p. 33). Validity is a crucial prerequisite to score use in high-stakes decisions. When a score has validity, it is said to be measuring the phenomenon it was designed to measure (Salkind, 2013) and can be interpreted with meaning. A prerequisite to validity, however, is reliability (Price, 2016).

Reliability can be defined as "the extent to which the tests measures anything at all" (Rust & Golombok, 1999, p. 64). Knowledge of reliability allows us to determine how consistent, replicable, or error-free a score is (AERA et al., 2014; Price, 2016; Salkind, 2013).  A score that is not sufficiently consistent (i.e., reliable) would also be insufficiently valid. It is the responsibility of the assessor to ensure adequate reliability and validity of the tests or measures they intend on using before conducting assessments on individuals (AERA et al., 2014). If the reliability and validity of the tests or measures being used meet the criteria for hypothesis generation or clinical decision making and has been normed around the individual that is being assessed, then the clinician's ability to interpret the data and make decisions increases (AERA et al., 2014).

The most commonly reported estimate of reliability is the internal consistency reliability (ICR), which is also known as split halves or coefficient alpha (Hogan, Benjamin, & Brezinski, 2000). Put simply, this form of reliability looks at the agreement between different items on a test by taking half of the items and correlating them with one another, then adjusting it to determine the reliability for the all of the items on the test (AERA et al., 2014). As such, when multiple items are combined into a single score, it is important to understand how much of the variance in that combined score stems from a single construct, as opposed to other constructs or error.

If a score contains a large amount of variance that stems from other constructs or error, then it is considered to be unfit to use for high-stakes decision making (Beidas et al., 2015; Henson, 2001; Hunsley & Mash, 2005; Hunsley & Mash, 2007; Hunsley & Mash, 2008; Mash & Hunsley, 2005; NASP, 2009; Reynolds & Livingston, 2014). ICR is most typically evaluated using Cronbach alpha (Hogan et al., 2000). Cronbach (or coefficient) alpha ($\alpha$; Cronbach, 1951) can be thought of as a ratio of explained true score variance to total variance (Price, 2016). Despite the need for sufficiently reliable scores, the NASP position paper (2009), the Standards (AERA et al., 2014), nor the NASP (2010) or APA ethical codes (2017) define what is sufficient in terms of reliability.

While the *Standards* (AERA et al., 2014) do not explicitly specify reliability cutoffs, a number of researchers suggest that scores used for diagnostic or classification must have an internal consistency reliability of 0.90 or higher to be interpretable and guide clinical decision making (Beidas et al., 2015; Hunsley & Mash, 2007; Hunsley & Mash, 2008; Kranzler & Floyd, 2013; Nunnaly, 1978; Nunnaly & Bernstein, 1994). IQ scores typically have reliability composites of 0.95 (Kranzler & Floyd, 2013) or higher which meets and exceeds the criteria to interpret and guide clinical decisions. Although we know that the most reliable and valid score obtained from an IQ test is the intelligence composite, many advocates (e.g., Flanagan, Ortiz, & Alfonso, 2013; Hale et al., 2006; Kaufman et al., 2016) continue to recommend that practitioners interpret less psychometrically sound scores from intelligence tests.

When a vast amount of scores are suggested for clinical use by test developers, it can lead to school psychologists differing in how they use and interpret scores derived from cognitive assessments. Benson, Floyd, Kranzler, Eckert, and Fefer (2018) surveyed school psychologists (N = 938) and gathered information regarding which type of cognitive analyses they engaged in. Results from this study revealed that 55.2% engage in subtest-level profile analyses and 49.3% use composite-level profile analyses in their practice.

When school psychologists interpret information differently based on the outcomes of an assessment, it can be confusing for the consumers of the information (Beaujean & Benson, 2018). For this reason, it is important for school psychologists to know which score from assessments are the most reliable and valid to use for clinical decision making in order to follow best practices. Outcomes from assessments conducted by practitioners have a significant and lasting impact on the lives of others. With this in mind, it is critical for practitioners to use evidence-based practices thus adhering to the highest standards in regards to score interpretation. Doing so could help elevate some of the inconsistencies in how we use data from tests to make clinical decisions.

The National Association of School Psychologists (NASP) and American Psychological Association (APA) dictate, as part of their professional codes of conduct, that practitioners should only use and interpret scores from tests when test publishers or independent researchers have established the score's basic psychometric properties, including reliability and validity (APA, 2017; NASP, 2010). While not explicitly referenced by either document, *The Standards for Educational and Psychological Assessment* (AERA et al., 2014) functions as a comprehensive professional document establishing expected guidelines for practitioners' use of assessment in clinical and educational settings. Assessment practices, like any other professional practice, must be substantiated with empirical evidence and meet said minimum criteria before they are implemented in standard practice (Kratochwill & Shernoff, 2004). Thus, it is important to understand what the literature says about the psychometric qualities of discrepancy scores and how they compare to community standards, such as those suggested by Hunsley and Mash (2008).

**Discrepancy Score Reliability: What we know.**

Currently interpretive manuals offer multiple modalities of score interpretation for intelligence tests. As previously mentioned, a number of people (Beidas et al., 2015; Hunsley & Mash, 2007; Hunsley & Mash, 2008; Kranzler & Floyd, 2013; Nunnaly, 1978; Nunnaly & Bernstein, 1994) have established reliability criteria deemed acceptable for practitioners to use for hypothesis generation and clinical decision making. When conducting assessments that are used for high-stakes decisions (i.e., intelligence tests), it is best to only interpret numbers that meet the .90 criteria for clinical decision making.

The reliability of difference scores have been examined by several individuals for a variety of intelligence tests. A summary of results from previous research on discrepancy score reliability are outlined in Table 1. Charter (2002) examined the Wechsler Memory Scale, Third Edition (WMS-III; Wechsler, 1997b) primary indexes, which ranged from .00 to .87. In this study, the threshold criteria of .90 was not met for any of the difference score indexes. Additionally, only 19 of the 104 comparison scores met the suggested .80 criteria. Another study that calculated the reliability coefficients of difference scores was Brown and Ryan (2004) who examined the reliability of the Wechsler Adult Intelligence Scales, Third Edition (WAIS-III; Wechsler, 1997a). The results from this study revealed that subtest reliability coefficients ranged from .34 to .85 with only 7 of the 55 subtests meeting the .80 threshold. The index comparison scores revealed that only two of the four had a score that was greater than .80. The ranges for the index comparison scores were between .79 and .87.

Glass, Ryan, Charter, and Bartel’s (2009) study examined the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV; Wechsler, 2003). The results of this study found none of the reliability coefficients for the difference scores of indexes or subtests met the .90 criteria. Additionally, this study revealed that only five of the 66 subtest comparisons had reliability coefficients greater than .80 while 33 of the 36 index comparisons had reliability coefficients greater than .80.

Similar findings to this were found in Glass, Ryan, and Charter (2009) who examined the reliability of difference scores for the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV; Wechsler, 2008a) at the index and subtest level. Of the 66 subtest comparisons, none of the subtests met the .90 standard. The reliability coefficients fanged from .55 to .88 and only 23 of the 66 subtest comparisons met the .80 standard. For the WAIS-IV there were only three index discrepancy scores possible to interpret and all of them met the criteria for hypothesis generation. It is notable, however, that the discrepancy scores did not meet the desired .90 criteria.

In a preprint by Farmer and Kim (2020, January 13) an examination of the reliability of difference scores was conducted for the Reynolds Intellectual Abilities Scale, Second Edition (RIAS-2; Reynolds & Kamphaus, 2015a) and Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V; Wechsler, 2014a) at the index and subtest level. At the index level on the RIAS-2, of the 10 comparisons, four of the 10 comparisions met the .90 cirteria while another one of the 10 met the .80 standard. At the subtest-level on the RIAS-2, three of the 28 comparisons met the .90 reliability criteria while another 15 of the 28 possible comparisons met the .80 criteria. On the WISC-V, 31 of the 55 possible comparisons at the index-level met the 0.80 reliability criteria and none of the comparisons met the .90 criteria. At the subtest level on the WISC-V, only 32 of the 120 possible comparisons met the .80 reliability criteria and none of the 120 comparison possibilities met the .90 reliability criteria.

When making clinical decisions, it is essential that the reliability meets the requirements of best practice. Looking at previous research using the recommended reliability criteria recommended by Beidas (2015), Hunsley and Mash (2007; 2008), Kranzler and Floyd (2013), and Nunnaly (1978), few studies found discrepancy scores that met the criteria for clinical decision making. Some assessments do have subtest and index level comparisons that meet the criteria for hypothesis generation, but not all of the comparisons within an assessment allow for hypothesis generation or clinical decision making when using the recommended reliability criteria.

**Table 1**

*Results From Previous Research on Discrepancy Score Reliability*

| | | Indicies | | Subtest | |
|---|---|---|---|---|---|
| Study | Test | .80 | .90 | .80 | .90 |
| Charter, 2002 | WMS-III | 18% | 0% | N/A | N/A |
| Brown & Ryan, 2004 | WAIS-III | 84% | 0% | 12% | 0% |
| Glass, Ryan, Charter, & Bartels, 2009 | WISC-IV | 92% | 0% | 8% | 0% |
| Glass, Ryan, Charter, & Bartels, 2009 | WAIS-IV | 100% | 0% | 35% | 0% |
| Farmer, R. L., & Kim, S. Y., 2020 | RIAS-2 | 10% | 40% | 54% | 11% |
| Farmer, R. L., & Kim, S. Y., 2020 | WISC-V | 56% | 0% | 27% | 0% |

*Note*. WMS-III = Wechsler Memory Scale Third Edition, WAIS-III = Wechsler Adult Intelligence Scale Third Edition, WISC-IV = Wechsler Intelligence Scale for Children Fourth Edition, WAIS-IV = Wechsler Adult Intelligence Scale Fourth Edition, RIAS-2 = Reynolds Intelligence Abilities Scale Second Edition

## Purpose of the Study

Discrepancy scores are core elements of commonly used intelligence test interpretation strategies (Beaujean & Benson, 2018; Flanagan et al., 2013; Flanagan et al., 2017; McGill et al., 2018). Furthermore, many textbooks (e.g., Sattler, 2018) encourage their use, and many practitioners report using them as part of more comprehensive interpretive strategies (Sotelo-Dynega & Dixon, 2014). Despite their endorsements and ongoing use, technical manuals do not provide psychometric evidence supporting discrepancy scores in direct contrast to AERA et al. (2014) standards. Given the need for clinical decisions to be based on reliable test scores (e.g., Hunsley & Mash, 2008), test users should have access to the reliability data, at the very least, supporting all scores for a given test. While the Woodcock-Johnson Tests of Cognitive Abilities, Fourth Edition (WJ IV COG; Schrank, & Dailey, 2014) does not directly produce ipsative discrepancy scores via its electronic scoring software, WJ IV COG scores are often used as part of profile analytic approaches such as cross battery assessment. The purpose of this study was to determine the reliability of discrepancy scores in the WJ IV COG (Schrank, McGrew, & Mather, 2014), to document those scores in the research literature, and to evaluate those scores in terms of Hunsley and Mash's (2008) model regarding the good enough principle. It was hypothesized that the results of the current study would yield results comparable to previous studies on discrepancy score reliability where few comparisons at the subtest and index level meet the excellent reliability criteria and approximately half meet the good criteria at the subtest and index level.

The following research questions were addressed:

1) Do discrepancy scores produced from subtests and indices on the WJ IV COG

meet adequate (i.e., between .70 and .80) reliability standard?

2) Do discrepancy scores produced from subtests and indicies on the WJ IV COG meet good (i.e., between .80 and .90) reliability standard?

3) Do discrepancy scores produced from subtests and indicies on the WJ IV COG meet excellent (i.e., ≥ .90) reliability standard?

## Method

### Measure

The Woodcock-Johnson Test of Cognitive Abilities, Fourth Edition (WJ IV COG; Schrank, McGrew, & Mather, 2014) is an intelligence test consisting of 18 subtests. The standardization sample of the WJ IV COG included a total of 7,416 individuals from 46 U.S. States and the District of Columbia. In this sample 664 were preschool aged or two to five years old, 3,891 were in grades kindergarten through 12th grade, 775 were undergraduate and graduate students, and 2,086 were adults (McGrew, LaForte, & Schrank, 2014). For the WJ IV COG, a stratified sampling design was used to randomly select participants for the norming sample that were representative of the U.S. population. This sample consisted of individuals ages 24 months to 90 years and older based on the 2010 Census data. In each age category, approximately 51% of the participants were male and 49% were female, except for the college group and the 65 and older group which was approximately 43% male and 57% female. Additionally, 63.7% of the sample were non-Hispanic white. Demographic data are available in Table 2.

Within the WJ IV COG, the reliability statistics were calculated for all tests and age ranges. The internal consistency reliabilities for the untimed tests and subtests that contained dichotomously scored items used the split-half procedure and were corrected via the Spearman Brown formula (Nunnally & Bernstein, 1994). For the tests that contained multiple-point scoring items and speeded tests, the Rach model was used to estimate necessary statistics (e.g., standard error of measurement) for the calculation of reliability. With the Rasch model the standard error of measurement that is associated with the ability estimate for each individual in the norming

**Table 2**

*Select Demographics of the Woodcock-Johnson IV Standardization Sample*

| | N | % |
|---|---|---|
| Gender and Race | | |
| Female | 3,835 | 51.7% |
| White | 4,813 | 64.9% |
| Black | 1,034 | 13.9% |
| AIANAT | 42 | 0.6% |
| ASIPAC | 293 | 4.0% |
| Other | 11 | 0.1% |
| White, Hispanic | 1,052 | 14.2% |
| Black, Hispanic | 22 | 0.3% |
| AIANAT, Hispanic | 10 | 0.1% |
| ASIPAC, Hispanic | 17 | 0.2% |
| Other, Hispanic | 122 | 1.6% |
| Age Groups | | |
| Preschool | 664 | 9.0% |
| K-12 | 3,891 | 52.5% |
| College | 775 | 10.5% |
| 18-24 | 874 | 11.8% |
| 25-44 | 1,083 | 14.6% |
| 25-64 | 596 | 8.0% |
| 65+ | 307 | 4.1% |
| Total | 7,416 | |

*Note*. Data extracted from Tables 3-2 through 3-5 from McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). *Technical Manual. Woodcock Johnson IV*. Rolling Meadows, IL: Riverside. AIANAT, American Indian or Alaska Native; ASIPAC, Asian, Native Hawaiian, or Other Pacific Islander. The participants in the college sample were also included in the adult samples (e.g., 18-24) as appropraite.

sample is attainable. The observed score variance and the mean-square error values were obtained for each individual in the norming sample within a technical age group.

Cluster score reliabilities and tests that contain subtests, such as Oral Vocabulary and Phonological Processing, on the WJ IV COG were calculated using the Mosier's (1943) unweighted equation. For the subtests found within tests, the split-half or the Rasch reliability procedures was used to calculate the reliability of each subtest that make up the test. For speeded tests, a test-retest study was conducted where individuals were administered the same form of the speeded test one day after they were administered the original test. Correlations between the two administrations were calculated and correction was used on the correlation.

## Procedure

Reliability data were collected from the WJ IV COG technical manual from Table B-1 and Table 4-3. Intercorrelations were also collected from the same manual from tables E-1 through E-6. This analysis included 18 subtests which were: Oral Vocabulary, Number Series, Verbal Attention, Letter-Pattern Matching, Phonological Processing, Story Recall, Visualization, General Information, Concept Formation, Numbers Reversed, Number-Pattern Matching, Nonword Repetition, Visual-Auditory Learning, Picture Recognition, Analysis-Synthesis, Object-Number Sequencing, Pair Cancellation, and Memory of Words. The following indexes were also included in the analysis: Comprehension-Knowledge (Gc), Fluid Reasoning (Gf), Short-term Working Memory (Gwm), Processing Speed (Gs), Auditory Processing (Ga), Long-Term Retrieval (Glr), Visual Processing (Gv), Quantitative Reasoning, and Auditory Memory Span.

## Analysis

The discrepancy scores from the WJ IV COG were calculated and compared to the reliability guidelines established by the assessment community (e.g. Hunsley & Mash, 2018; Nunnaly, 1978), which are .90 and above for excellent, .80 to .90 for good, and .70 to .80 for adequate. Reliability estimates of .69 and below are considered inadequate. Thorndike and Hagan's (1969) formula was used to calculate the reliability coefficients of discrepancy scores:

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

In the formula, above *r* represents the reliability of the difference score, $r_a$ and $r_b$ represent the internal consistency reliability (ICR) for the contrast scores, and $r_{ab}$ represents the intercorrelation between the contrast scores. The calculations were computed using Microsoft Excel 2016 through the use of the following formula:

$$=(((r_{a}+r_{b})/2)-r_{ab})/(1-r_{ab})$$

With this formula each variable was identified by a specific cell in the Microsoft Excel spreadsheet. The information is displayed in tables organized by type of score (e.g., subtest versus index) and the primary analysis focuses on the total sample. The mean, median, standard deviation, and range of discrepancy scores were calculated and reviewed. The mean and standard deviation was calculated by first transforming the reliability coefficients to *Z* scores via the Fisher *Z'* transformation (Nunnally & Bernstein, 1994), completing the calculations, and inverting the transformation to return the scores to reliability coefficients. The transformation and calculations were performed in Excel 2016 using established functions. Additionally, the count and percentage of discrepancy scores meeting each guideline (.70, .80, and .90) of the tripartite model (Hunsley & Mash, 2008) is presented to address the research questions.

# Results

## Subtest-Level Comparisons

A summary of the subtest score reliability coefficients for the WJ IV COG can be found in Table 3. The subtest-level comparisons had 153 possible comparisons and the reliability coefficients ranged from .61 to .93. Of these comparisons, 7% met the .90 reliability criteria, 57% met the .80 criteria, 31% met the .70 criteria, and the remaining 5% of the comparisons fell below the .70 criteria level. When looking at the 7% of comparisons that met the .90 criteria, nine of the eleven or 82% of those comparisons included Visual-Auditory Learning. In regards to the 5% of comparisons that fell below the .70 criteria, four of the seven or 57% of those comparisons  included Memory of Words.

## Index Level Comparisons

A summary of the Index score reliability coefficients for the WJ IV COG can be found in Table 4. For the Index Level Comparisons, there were 55 possible comparisons and the reliability coefficients ranged from .59 to .93. Out of the 55 possible comparisons, 20% met the .90 criteria, 60% met the .80 criteria, 15% met the .70 criteria, and the remaining 5% fell below the .70 criteria level. In regards to the 20% that met the .90 criteria, eight of the eleven comparisons, or 73%, included Long Term Retrieval. When looking at the comparisons that fell below the .70 criteria, two of the three comparisons, or 67%, included Fluid Reasoning.

**Table 3**

*Subtest Statistics for Total Reliability of Discrepancy Scores*

| | Subtest | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | Oral Vocabulary | | | | | | | | | | | | | | | | | |
| 2. | Number Series | .80 | | | | | | | | | | | | | | | | |
| 3. | Verbal Attention | .75 | .80 | | | | | | | | | | | | | | | |
| 4. | Letter-Pattern Matching | .83 | .84 | .83 | | | | | | | | | | | | | | |
| 5. | Phonological Process. | .68 | .75 | .70 | .78 | | | | | | | | | | | | | |
| 6. | Story Recall | .84 | .87 | .84 | .88 | .83 | | | | | | | | | | | | |
| 7. | Visualization | .80 | .80 | .81 | .82 | .75 | .82 | | | | | | | | | | | |
| 8. | General Info. | .61 | .85 | .79 | .84 | .77 | .86 | .80 | | | | | | | | | | |
| 9. | Concept Formation | .83 | .85 | .85 | .88 | .79 | .89 | .79 | .85 | | | | | | | | | |
| 10. | Numbers Reversed | .80 | .81 | .74 | .81 | .74 | .87 | .79 | .82 | .84 | | | | | | | | |
| 11. | Number-Pattern Matching | .81 | .78 | .78 | .73 | .77 | .85 | .79 | .84 | .85 | .79 | | | | | | | |
| 12. | Nonword Repetition | .83 | .88 | .79 | .88 | .80 | .88 | .82 | .85 | .88 | .85 | .85 | | | | | | |
| 13. | Visual-Auditory Learning | .89 | .92 | .89 | .91 | .84 | .92 | .84 | .90 | .91 | .86 | .90 | .92 | | | | | |
| 14. | Picture Recognition | .75 | .79 | .75 | .74 | .73 | .74 | .63 | .75 | .77 | .73 | .74 | .75 | .77 | | | | |
| 15. | Analysis-Synthesis | .85 | .85 | .83 | .86 | .81 | .88 | .78 | .85 | .85 | .84 | .83 | .90 | .92 | .73 | | | |
| 16. | Object-Number Sequencing | .80 | .84 | .70 | .81 | .72 | .85 | .77 | .82 | .83 | .77 | .79 | .81 | .89 | .72 | .84 | | |
| 17. | Pair Cancellation | .86 | .88 | .85 | .79 | .80 | .90 | .84 | .87 | .89 | .86 | .75 | .89 | .93 | .80 | .89 | .85 | |
| 18. | Memory of Words | .76 | .82 | .66 | .80 | .65 | .80 | .73 | .79 | .80 | .72 | .79 | .76 | .84 | .69 | .79 | .68 | .84 |

**Table 4**

*Index Statistics for Total Reliability of Discrepancy Scores*

| | Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | GIA | | | | | | | | | | |
| 2. | GF GC | .73 | | | | | | | | | |
| 3. | Comprehension Knowledge | .83 | .64 | | | | | | | | |
| 4. | Fluid Reasoning | .77 | .60 | .82 | | | | | | | |
| 5. | Short Term Working Memory | .77 | .84 | .84 | .84 | | | | | | |
| 6. | Cognitive Processing Speed | .87 | .91 | .90 | .90 | .87 | | | | | |
| 7. | Auditory Processing Speed | .81 | .85 | .84 | .85 | .80 | .89 | | | | |
| 8. | Long Term Retrieval | .93 | .92 | .91 | .91 | .90 | .93 | .90 | | | |
| 9. | Visual Processing | .81 | .83 | .83 | .82 | .81 | .85 | .80 | .81 | | |
| 10. | Quant. Reasoning | .76 | .77 | .87 | .59 | .83 | .89 | .87 | .91 | .80 | |
| 11. | Auditory Memory Span | .83 | .86 | .84 | .86 | .76 | .88 | .72 | .89 | .79 | .87 |

## Discussion

School psychologists allocate a significant portion of their time conducting assessments which frequently include the use of intelligence tests (Castillo et al, 2012; Walcott et al, 2016). For intelligence tests, there are numerous interpretive strategies recommended in textbooks (Flanagan & Alfonso, 2017; Kaufman et al., 2016; Sattler, 2018) and in test manuals (e.g., Schrank, McGrew, & Mather, 2014). It is the duty of the practitioner to ensure that the diagnostic measures used have adequate reliability and validity prior to conducting assessments and interpreting data from those assessments (AERA et al., 2014). Establishing acceptable evidence of reliability is the first step to determining if a score can be used to make clinical decisions. A measure is considered to be psychometrically strong and appropriate for clinical use when the measure has adequate evidence of reliability and validity.

The current study sought to answer the following research questions:

1) Do discrepancy scores produced from subtests and indices on the WJ IV COG meet adequate (i.e., between .70 and .80) reliability standard?

2) Do discrepancy scores produced from subtests and indicies on the WJ IV COG meet good (i.e., between .80 and .90) reliability standard?

3) Do discrepancy scores produced from subtests and indicies on the WJ IV COG meet excellent (i.e., ≥ .90) reliability standard?

It was hypothesized that very few discrepancy scores would meet the excellent reliability standard necessary for clinical decision making and that over half would meet the good standard for hypothesis generation at both the subtest and index level. Results indicate that at the subtest level 7% of the comparisons met the excellent reliability standard, followed by 57% at the good reliability standard, and 31% at the adequate reliability standard. At the index level 20% of the comparisions met the excellent reliability standard, 60% met the good reliability standard, and 15% met the adequate reliability standard. The results of this study support the hypothesis that very few discrepancy scores meet the excellent reliability criteria and approximately half of the comparisons met the good reliability criteria at both the subtest and index level. Previous research on difference scores (Brown & Ryan, 2004; Charter, 2001, 2002; Glass et al., 2009) reported reliability coefficients comparable to the comparisons I found in the WJ IV COG.

Most of the reliability coefficients produced from the discrepancy scores at the subtest and index level on the WJ IV COG do not meet the .90 reliability standard for clinical decision making (Beidas et al., 2015; Hunsley & Mash, 2007; Hunsley & Mash, 2008; Kranzler & Floyd, 2013; Nunnaly, 1978; Nunnaly & Bernstein, 1994). More than half of the comparison scores had reliability coefficients that met the criteria for hypothesis generation (i.e. .80-.89). At both the index and subtest level, a small portion of the comparisons fell within the .70-.79 range, which is a less restrictive reliability criteria. Additionally, there were still comparisons that fell below the .70 range. These results support prior research that suggests that discrepancy score interpretation on intelligence tests may not be appropriate for making high stakes decisions (Canivez, 2013; Charter, 2002; Glass et al., 2009, Glass et al., 2010; Kranzler, Floyd, et al., 2016; Watkins, 2000).

## Implications

School psychologists report that their most common role is evaluating students for special education (Castillo et. al., 2012; Walcott et. al., 2016) and that IQ tests are the most frequently used assessment strategies (Benson et. al., 2019; Sotelo-Dynega and Dixon, 2014). Since IQ tests are often used by practitioners when conducting special education evaluations, it is critical that they make data-based decision using both reliable and valid interpretive strategies in order to make evidence-based decisions (NASP, 2009; 2010). Discrepancy scores have shown to be adequate at times for hypothesis generation, but should be approached with caution and used rarely when making high stakes clinical or educational decisions. The current study focused solely on reliability and data available in the WJ IV COG technical manual. Based on current data, discrepancy score interpretation should be avoided as a tool used in isolation. Instead, school psychologists should include other data points or assessment methods in conjunction with discrepancy score interpretation when making high stakes clinical decisions. It is notable that the other assessment methods used in conjunction with discrepancy score interpretation do not increase the reliability of the discrepancy score. The addional data points that are used should be both reliable and valid tools that are appropriate for the referral concern. These data points should also be used to corroborate other data, thus supporting or refuting the hypothesis formed based on referral concern and guiding clinical decision making. Knowledge of reliability is critical in order to determine whether or not a score can or should be used for interpretation. It is the responsibility of the school psychologist to determine whether or not discrepancy scores are appropriate for either hypothesis generation or clinical decision making by assessing both the reliability and validity of a score in order to assess if the score can be used in clinical practice (AERA et. al., 2014; NASP, 2010).

## Limitations and Future Research

While this research adds to the literature on cognitive assessment scores, it does have limitations. The scope of this study centered on individual difference scores, thus the results of this study do not apply to the reliability of composites that are created by combining multiple tests. The current study also did not examine the reliability of profiles which is often used as a part of cognitive profile analysis. Scores produced on the WJ IV COG via the electronic scoring software do not directly produce ipsative discrepancy scores. The reliability of scores produced through the use of ipsative analysis was not investigated through the current study, although the scores from the WJ IV COG are often used as a part of profile analysis. It is notable that the AERA et al. (2014) Standards do not have established numerical guidelines for reliability standards, thus the .90 criteria for reliability which is established in the literature was used as the criteria for clinical decision making (Bedias et al, 2015; Hunsley & Mash, 2007; Hunsley & Mash, 2008; Kranzler & Floyd, 2013; Nunnaly, 1978; Nunnaly & Berstein, 1994). Some have recommended a higher standard of .95 for reliability (Nunally, 1978; Kranzler & Floyd, 2013). The selection of using .90 as the standard for clinical decision making is a limitation to the current study and can be refuted since there is no established numerical guideline for reliability standards. In future research, researchers should consider examining the reliability of composites that are created by combining multiple tests, profiles, and scores produced through the use of ipsative analysis. In clinical practice, these interpretive strategies are often used when conducting psychoeducational assessments and such practices are recommended for use by test publishers and in text books. For this reason, further investigation is warranted.

## 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 Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing.* American Educational Research Association.

American Educational Research Association, & American Psychological Association. (1999). National Council of Measurement in Education. *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

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

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. Washington, DC: Author.

Beaujean, A. A., & Benson, N. F. (2018). Theoretically-consistent cognitive ability test development and score interpretation. *Contemporary School Psychology*, *23*(2), 126-137.

Beaujean, A., Benson, N., McGill, R., & Dombrowski, S. (2018). A misuse of IQ scores: Using the dual discrepancy/consistency model for identifying Specific Learning Disabilities. *Journal of Intelligence*, *6*(36).

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 Behavioral Practice*, 22, 5–19.

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.

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.

Canivez, G. L. (2013). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. *The Oxford handbook of child psychological assessment*, (pp. 84-112). New York, NY: Oxford University Press. https:doi.org/10.1093/oxfordhb/9780199796304.013.0004

Castillo, J., Curtis, M., & Gelley, C. (2012) Professional practice school psychology 2010 – part 2: School psychologists' professional practices and implications for the field. *Communiqué*, *40*(8), 4–6.

Charter, R. A. (2002) Reliability of the WMS-III Discrepancy Comparisons. *Perceptual and Motor Skills*, *94*(2), 387-390.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

Famer, R. L. & Floyd, R. G. (2018). The use of intelligence tests in the identification for 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.

Farmer, R. L., & Kim, S. Y. (2020). Difference score reliabilities within the RIAS-2 and WISC-V. *Psychology in the Schools*. Advanced online publication. https://onlinelibrary.wiley.com/doi/abs/10.1002/pits.22369

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

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

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, DC: American Psychological Association.

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.

Glass, L. A., Ryan, J. J., & Charter R. A. (2009). Discrepancy Score Reliabilities in the WAIS-IV Standardization Sample. *Journal of Psychoeducational Assessment*, *28*(3), 201-208.

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.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*(3), 177-190.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and psychological measurement*, *60*(4), 523-531.

Hunsley, J., & Mash, E. J. (2005). Introduction to the special section on developing guidelines for the evidence-based assessment (EBA) of adult disorders. *Psychological assessment*, *17*(3), 251 –255.

Hunsley, J. D., & Mash, E. J. (2008). *A guide to assessments that work* (2nd ed.). London: Oxford University Press.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology*, *3*, 29-51.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R*. New York: Wiley.

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

Kranzler, J. H. (2016). Current practices and future directions for the assessment of child and adolescent intelligence in schools around the world. *International Journal of School and Educational Psychology*, *4*(4), 213-214.

Kranzler, J. H., Benson, N., & Floyd, R. G. (2016). 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.

Kranzler, J. H., Floyd, R. G., Benson, N., Zaboski, B., & Thibodaux, L. (2016a). Classification agreement analysis of Cross-Battery Assessment in the identification of specific learning disorders in children and youth. *International Journal of School & Educational Psychology*, *4*(3), 124-136.

Kranzler, J. H., Floyd, R. G., Benson, N., Zaboski, 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.

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

Kratochwill, T.R., & Shernoff, E.S. (2004). Evidence-based practice: Promoting evidence-based interventions in school psychology. *School Psychology Review, 33*(1), 34-48.

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

Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: issues and challenges. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 362–379.

McClain, M., & Pfeiffer, S. (2012). Identification of gifted students in the United States today: A look at state definitions, policies, and practices. *Journal of Applied School Psychology*, *28*(1), 59–88.

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*(3), 290-302.

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. *The Journal of Special Education*, *25*(4), 504-526.

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.

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.

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). *Technical Manual. Woodcock-Johnson IV*. Rolling Meadows, IL: Riverside.

McNicholas, P. J., Floyd, R. G., Woods, I. L., 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.

National Association of School Psychologists. (2009). *School Psychologists' Involvement in Assessment* (Position Statement). Bethesda, MD: Author.

National Association of School Psychologists. (2010). *Principles for professional ethics*. Retrieved from https://www.nasponline.org/standards-and-certification/professional-ethics.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Price, L. R. (2016*). Psychometric methods: Theory into practice.* New York, NY: Guilford Publications.

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.

Rust, J. & Golombok, S. (1999). *Modern psychometrics: The science of psychological assessment* (2nd ed.).: New York, NY: Routledge.

Salkind, N. J. (2013). *Tests and measurements for people who (think they) hate tests & measurements.* Thousand Oaks, CA: Sage Publications.

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

Schrank, F. A., & Dailey, D. (2014). *Woodcock-Johnson Online* [Online format]. Rolling Meadows, IL: Riverside.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV Tests of Cognitive Abilities.* Rolling Meadows, IL: Riverside.

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.

Sotelo-Dynega, M., & Dixon, S. G. (2014). Cognitive assessment practices: A survey of school psychologists. *Psychology in the Schools*, *51*(10), 1031-1045.

Walcott, C. M., Charvat, J., McNamara, K. M., & Hyson, D. M. (2016). *School psychology at a glance: 2015 member survey results*. Special session presented at the annual meeting of the National Association of School Psychologists, New Orleans, LA.

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997). Discriminant and predictive validity of the WISC-III ACID profile among children with learning disabilities. *Psychology in the Schools, 34*(4), 309-319.

Watkins, M. W., Kush, J. C., & Schaefer, B. A. (2002). Diagnostic utility of the learning disability index. *Journal of Learning Disabilities*, *35*(2), 98-103.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scales, Third Edition administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008a). *Wechsler adult intelligence scale–Fourth Edition (WAIS–IV).* San Antonio, TX: NCS Pearson, https://doi.org/10/1037/t15169-000

Wechsler, D. (2003). *Wechsler intelligence scale for children-Fourth Edition (WISC-V).* San Antonio, TX: The Psychological Corporation. https://doi.org/10.1037/t15174-000

Wechsler, D. (1997b). *Wechsler memory scale, third edition.* San Antonio, TX: The Psychological Corporation.

Ysseldyke, J., Morrison, D., Burns, M. K., Ortiz, S., Dawson, P., Rosenfield, S., et al. (2006). *School psychology: A blueprint for training and practive III.* Bethesda MD: National Association of School Psychologists.

Zaboski, B. A., Kranzler, N. G., & Gage, N. A. (2018). Meta-analysis of the relationship between academic achievement and broad abilities of the Catell-Horn-Carroll theory. *Journal of School Psychology, 71,* 42-56.