Evaluating The Impact of Examiner Variability on The Verbal Comprehension index of The WISC-V

Brooke A. Butcher
ebrooke40@gmail.com

Follow this and additional works at: https://huskiecommons.lib.niu.edu/allgraduate-thesesdissertations

Part of the Psychology Commons

Recommended Citation
https://huskiecommons.lib.niu.edu/allgraduate-thesesdissertations/7133

This Dissertation/Thesis is brought to you for free and open access by the Graduate Research & Artistry at Huskie Commons. It has been accepted for inclusion in Graduate Research Theses & Dissertations by an authorized administrator of Huskie Commons. For more information, please contact jschumacher@niu.edu.
ABSTRACT

EVALUATING THE IMPACT OF EXAMINER VARIABILITY ON THE VERBAL COMPREHENSION INDEX OF THE WISC-V

Brooke A. Butcher, M.A.
Department of Psychology
Northern Illinois University, 2023
Kara M. Styck, Director

The purpose of the current study was to: a) explore rater variability on the VCI, b) explore rater tendencies, and c) compare rater tendencies between graduate students and licensed and/or certified school psychologists. Considering the results of prior literature that has investigated scoring errors and inter-rater agreement on the WISC, it was hypothesized that examiners would be significantly different from one another (overall and between groups) in terms of their tendencies towards severity/leniency on the Vocabulary and Similarities subtests. Raters ($N = 10$) were comprised of two groups: a) graduate students in training to become school psychologists ($n = 5$) and b) licensed and/or certified school psychologists ($n = 5$). A total of 30 partially completed protocols (Similarities and Vocabulary subtests only) were scored by raters using a connected rating design. Results of the study revealed that rater effects were not present overall, nor between groups (i.e., licensed and/or certified school psychologists and graduate students) suggesting that raters did not significantly differ in their tendencies toward severity/leniency when scoring items on the subtests that comprise the VCI. School psychologists should continue to adhere to standardized administration and scoring rules as
outlined in the WISC-V administration and scoring manual. Practitioners should also continue consulting with colleagues regarding ambiguous examinee responses.
EVALUATING THE IMPACT OF EXAMINER VARIABILITY ON THE VERBAL COMPREHENSION INDEX OF THE WISC-V

BY

BROOKE A. BUTCHER
©2023 Brooke A. Butcher

A THESIS SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE
MASTER OF ARTS

DEPARTMENT OF PSYCHOLOGY

Thesis Director:
Kara M. Styck
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>viii</td>
</tr>
</tbody>
</table>

## Chapter

I. INTRODUCTION ...................................................................................................1

Background and Rationale for the Study .................................................1

II. REVIEW OF THE LITERATURE .....................................................................5

Historical and Contemporary Use of WISC Scores in High-Stakes
Decisions..........................................................................................5

The Role of the Verbal Comprehension Index in High-Stakes Decisions
for K–12 Students .................................................................8

Technical Adequacy of the Verbal Comprehension Index ....................9

   Internal Consistency Reliability.........................................................9

   Test-Retest Reliability .................................................................11

   Standard Error of Measurement......................................................14

Examiner Variability.............................................................................15

Examiner Variability on the Verbal Comprehension Index ..........16

Scoring Errors on the Verbal Comprehension Index .....................16
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence of Rater Drift</td>
<td>19</td>
</tr>
<tr>
<td>Interscorer Agreement on the Verbal Comprehension Index</td>
<td>21</td>
</tr>
<tr>
<td>The Current Study</td>
<td>22</td>
</tr>
<tr>
<td>Research Questions</td>
<td>23</td>
</tr>
<tr>
<td>Research Predictions</td>
<td>24</td>
</tr>
<tr>
<td>Question 1</td>
<td>24</td>
</tr>
<tr>
<td>Question 2</td>
<td>25</td>
</tr>
<tr>
<td>Question 3</td>
<td>25</td>
</tr>
<tr>
<td>Question 4</td>
<td>26</td>
</tr>
</tbody>
</table>

### III. METHODOLOGY

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td>27</td>
</tr>
<tr>
<td>Protocols</td>
<td>29</td>
</tr>
<tr>
<td>Measures</td>
<td>31</td>
</tr>
<tr>
<td>Demographic and Training Variables Questionnaire</td>
<td>31</td>
</tr>
<tr>
<td>WISC-V</td>
<td>31</td>
</tr>
<tr>
<td>Procedure</td>
<td>32</td>
</tr>
<tr>
<td>Rating Design</td>
<td>33</td>
</tr>
<tr>
<td>Independent Variable</td>
<td>34</td>
</tr>
<tr>
<td>Dependent Variables</td>
<td>36</td>
</tr>
<tr>
<td>Chapter</td>
<td>Page</td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
</tr>
<tr>
<td>Data Analysis Plan</td>
<td>38</td>
</tr>
<tr>
<td>Question 1</td>
<td>38</td>
</tr>
<tr>
<td>Question 2</td>
<td>39</td>
</tr>
<tr>
<td>Question 3</td>
<td>39</td>
</tr>
<tr>
<td>Question 4</td>
<td>40</td>
</tr>
<tr>
<td>IV. RESULTS</td>
<td>41</td>
</tr>
<tr>
<td>Question 1</td>
<td>41</td>
</tr>
<tr>
<td>Similarities</td>
<td>41</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>44</td>
</tr>
<tr>
<td>Question 2</td>
<td>48</td>
</tr>
<tr>
<td>Similarities</td>
<td>48</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>48</td>
</tr>
<tr>
<td>Question 3</td>
<td>49</td>
</tr>
<tr>
<td>Similarities</td>
<td>49</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>50</td>
</tr>
<tr>
<td>Question 4</td>
<td>50</td>
</tr>
<tr>
<td>V. DISCUSSION</td>
<td>52</td>
</tr>
<tr>
<td>Question 1</td>
<td>53</td>
</tr>
<tr>
<td>Question 2</td>
<td>55</td>
</tr>
<tr>
<td>Question 3</td>
<td>57</td>
</tr>
</tbody>
</table>
Question 4 ..........................................................................................................................58

Implications .....................................................................................................................59

Limitations and Future Directions .................................................................................61

Conclusion .......................................................................................................................62

REFERENCES ..................................................................................................................64

APPENDICES ..................................................................................................................75
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Distribution of VCI and FSIQ Scores</td>
<td>30</td>
</tr>
<tr>
<td>2. ICC Values for the Similarities Subtest</td>
<td>43</td>
</tr>
<tr>
<td>3. ICC Values for the Vocabulary Subtest</td>
<td>46</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>-----------------</td>
<td>------</td>
</tr>
<tr>
<td>1. Rating design</td>
<td>35</td>
</tr>
</tbody>
</table>
LIST OF APPENDICES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. GRADUATE STUDENT DEMOGRAPHIC AND TRAINING VARIABLES QUESTIONNAIRE</td>
<td>75</td>
</tr>
<tr>
<td>B. LICENSED AND/OR CERTIFIED PSYCHOLOGIST DEMOGRAPHIC AND TRAINING VARIABLES QUESTIONNAIRE</td>
<td>79</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

Background and Rationale for the Study

Psychological assessment is a foundational skill for school psychologists and remains an integral domain of school psychology practice despite the increasing emphasis on direct/indirect services for children, families, and schools in the National Association of School Psychologist’s (NASP) Practice Model over the past 40+ years (National Association of School Psychologists, 1978, 1984, 1992, 1997, 2000, 2010). Indeed, results of a recently published survey of school psychologists’ professional practices conducted by McNamara and colleagues (2019) indicated that school psychologists reported the “highest level of engagement” in assessment activities related to special education. Moreover, results of another recent national survey of school psychologists test use and assessment practices revealed that school psychologists’ reported spending up to two thirds of their work time devoted to special education eligibility procedures, more broadly speaking, due to federal and state “child find” mandates to identify students with disabilities in education who require legal protections (Benson et al., 2019).
According to the results of the latter survey conducted by Benson et al. (2019), the Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V; Wechsler, 2014) is the second most commonly used psychological test instrument and the most frequently administered test of cognitive abilities by school psychologists across the U.S. School psychologists reported that they administer a WISC-V an average of 3.5 times per month—accounting for 80% of all standardized tests administered in as much time (Benson et al., 2019). The WISC-V reports five index scores (Verbal Comprehension, Visual-Spatial, Fluid Reasoning, Working Memory, and Processing Speed) that were designed with the intention of measuring broad cognitive abilities defined by the Cattel-Horn-Carroll theory of intelligence (CHC; Schneider & McGrew, 2012) and each index score is derived from two subtests. Additionally, a Full-Scale IQ (FSIQ) score can be obtained from the seven primary subtests which was designed with the intention of measuring global intelligence.

WISC-V scores generally have good psychometric properties when the test is administered and scored according to standardized administration and scoring rules (e.g., Thorndike & Thorndike-Christ, 2010; Wechsler, 2014b). However, results of a recent meta-analysis investigating the prevalence and impact of examiner errors on the Wechsler Scales of Intelligence conducted by Styck and Walsh (2016) suggests that adhering to standardized administration and scoring rules may be difficult for the subtests that compose the Verbal Comprehension Index (VCI). On average, 77% of all protocols \( N = 91 \) examined in a subgroup of studies reviewed within the meta-analysis \( k = 3 \) contained examiner administration and/or scoring errors that resulted in a change to the VCI. Furthermore, across all published iterations of
the WISC\(^1\), results of independent research has indicated that the VCI contains the largest number of scoring errors (Alper, 2012; Belk et al., 2002; Brazelton et al., 2003; de Ávila Quevedo, 2011; Hopwood & Richard, 2005; Ryan & Schnakenberg-Ott, 2003) and the overall weakest evidence of reliability when compared to other index scores (e.g., Ryan & Schnakenberg-Ott, 2003; Canivez et al., 2017; Irwin, 1966).

The foregoing research is problematic given that WISC scores have been historically used by school psychologists to make high-stakes decisions regarding special education eligibility and test scores that are used to make these decisions must meet stringent criteria of technical adequacy (Bracken, 1987; Reschly et al., 2002). Moreover, although numerous studies have been published to date on the technical adequacy of the Wechsler scales (e.g., Bracken, 1987; Canivez & Watkins, 2016), there has been no attempt to adjust scores on the VCI to account for examiner variability that remains after scoring errors are addressed. As a result, special education eligibility determinations that rely upon scores from the Wechsler Scales of Intelligence may be impacted by examiner variability on the VCI. The resultant goals of the current study are to: a) estimate the magnitude of examiner variability on the VCI, b) explore rater tendencies (e.g., severity/leniency, halo/horns, and central tendency), and c) compare rater tendencies between graduate students and licensed and/or certified school psychologists. The following literature review explores the historical and contemporary uses of WISC scores for

\(^1\) Unless otherwise specified, “WISC” refers to all iterations of The Wechsler Intelligence Scale for Children.
high-stakes eligibility decisions, the technical adequacy and psychometric properties of the VCI, and the prevalence and impact of rater variability on VCI scores.
CHAPTER II

REVIEW OF THE LITERATURE

Historical and Contemporary Use of WISC Scores in High-Stakes Decisions

WISC scores have primarily been used to identify students with intellectual (ID) and specific learning disabilities (SLD) due to the central role that cognitive ability testing plays in operationalizing criteria for an education disability in these categories (Maki et al., 2015; Individuals with Disabilities Education Act, 1997/2013). The essential features of ID are subaverage general intellectual functioning and concomitant deficits in adaptive behavior when compared to one’s same age, gender, and socioculturally matched peers (IDEA § 300.8[c][6]; American Psychiatric Association, 2013). Subaverage general intellectual functioning was historically defined operationally as a global IQ score < 70 in the Diagnostic and Statistical Manual of Mental Disorders (4th ed.; DSM-4; American Psychiatric Association, 1994). However, the current Diagnostic and Statistical Manual of Mental Disorders (5th ed.; DSM-5; American Psychiatric Association, 2013) acknowledges the presence of measurement error in IQ scores and discourages strict use of this cut-score for diagnosing ID as a result. Shortly after the publication of the DSM-V, the use of a single fixed cut-score on an IQ test for defining ID in
death penalty cases was determined to be unconstitutional in the landmark U.S. Supreme Court case Hall v. Florida (2014). Hall was found eligible for the death penalty and ultimately sentenced to death in an earlier case, Hall v. Florida (2013), because he scored greater than 70 on multiple IQ tests over his lifetime despite his scores ranging between 60 and 80 across all administrations. The U.S. Supreme Court ultimately determined that the use of a cut-score to define ID violated the Eighth Amendment because this practice failed to account for a test score’s standard error of measurement (SEM). Though progress is still being made in regards to using a single fixed IQ score, the score still commonly acts as an incentive for early release from prison (if the inmate volunteers to be sterilized) and avoidance of the death penalty. Likewise, current special education federal regulations state that special education eligibility evaluations may “not use any single measure or assessment as the sole criterion for determining whether a child is a child with a disability and for determining an appropriate educational program for the child” (IDEA § 300.304[b][2]). Nevertheless, evidence from a recently published analog study suggests that school psychologists may be inaccurately overidentifying students as having ID by placing undue emphasis on low IQ test scores – even in the presence of average adaptive skills and low-average to average academic performance (Sullivan et al., 2019).

The essential feature for SLD is a deficit in one or more of the basic areas of academic functioning that cannot be explained by other factors (IDEA § 300.8[c][10][i]) and this is typically operationalized as “unexpected underachievement” (McGill et al., 2020). According to IDEA (2004), there are three families of procedures that can be used to identify achievement as unexpectedly low. The first family of procedures include ability-achievement discrepancy
methods which utilize a standard scale point difference, standard deviation difference, or difference defined by the regression formula to operationalize unexpected underachievement as a significant discrepancy between student ability and student achievement (Reschly & Hosp, 2004). The second family of procedures include response-to-intervention methods, which operationalize unexpected underachievement as a persistent pattern of non-response to increasingly intensive interventions (RtI; Zirkel, 2017). Lastly, achievement can be unexpectedly low if it corresponds with a pattern of cognitive processing strengths and weaknesses (PSW; Hale & Fiorello, 2004). A PSW approach to operationalizing unexpected underachievement generally involves analyzing subtest and/or index score scatter and variability to identify specific patterns of cognitive weaknesses that correspond to academic weaknesses while other cognitive abilities are “spared” (McGill et al., 2020). As a result, all of the preceding methods propose use of some type of cut-score in order to identify children with SLD. In particular, over 75 different cognitive profiles have been hypothesized from WISC scores to date and many school psychologists continue to use these profiles in special education eligibility evaluations (Kranzler et al., 2020; Maki & Adams, 2018). Although the WISC was developed with the intention of measuring cognitive ability, and not disability per se, special education eligibility determinations often heavily rely upon score interpretations from standardized tests of intelligence, such as the WISC (Flanagan et al., 2018; Hale et al., 2010; Kavale et al., 2005).
The Role of the Verbal Comprehension Index in High-Stakes Decisions for K–12 Students

The subtests that compose the VCI can indirectly, yet critically, influence special education eligibility determinations for K–12 students because subtest- and composite-level profile analyses are commonly interpreted and used in these high-stakes decisions (Benson et al., 2019; McGill et al., 2018; Pfeiffer et al., 2000). For example, test developers suggest that the FSIQ and GAI scores may be compared with one another to inform neurodevelopmental diagnoses. Given that the GAI weighs verbal abilities more heavily than the FSIQ, this comparison is hypothesized to inform potential language impairments. Specifically, if the GAI is lower than the FSIQ, it is hypothesized to be an indicator of language deficits or poor verbal problem-solving abilities (Wechsler, 2014b, p. 167). Researchers have also suggested that the FSIQ and GAI can be compared to inform ID/SLD, ASD, and ADHD diagnoses due to the reduced influence of processing speed and working memory in constructing the GAI (Akbar et al., 2013; Horowitz-Kraus et al., 2014; Kasper et al., 1968; Niileksela & Reynolds, 2014). In particular, a pattern of scores whereby the FSIQ is significantly lower than the GAI has been proposed to indicate poor memory and low achievement, both of which have been associated with neurodevelopmental and learning disorders (Wechsler, 2014b).

The GAI may be used as a substitute for the FSIQ when there are significant discrepancies between one or more index scores, which is a prevalent occurrence when an examinee has extremely low or high intellectual abilities (Wechsler, 2014b; Grégoire et al., 2011). The GAI can also directly affect ability-achievement discrepancy formulas because it is
sensitive to cases in which there is a great deal of subtest scatter, a considerable variation in subtest scores, that commonly occurs in children with extremely high or low abilities. In turn, the GAI score may be artificially inflated or depressed when compared to the FSIQ. Such measurement error can occur whilst interpreting scores from any test, not just the WISC or other intelligence tests. Consequently, children who are likely in need of special education supports could be disproportionately affected by the weaker reliability of the VCI (e.g., children with ID and SLD). This could also occur in an array of special education assessments because the GAI can be inserted into any ability-achievement discrepancy formula and therefore affect intra-cognitive score relations within PSW methods.

Technical Adequacy of the Verbal Comprehension Index

**Internal Consistency Reliability**

Internal consistency reliability describes the stability of an examinee's responses across a subset of items written with the intention of measuring a single construct. Generally speaking, all items within the subset should reflect the same underlying construct, so scores on those items should be correlated with one another (Price et al., 2015). The VCI internal consistency reliability coefficient for the standardization sample of the WISC-V was calculated using the formula recommended by Guilford (1954), Haertel (2006), and Nunnally and Bernstein (1994),
and then averaged across ages using Fisher’s $z$ transformation (Silver & Dunlap, 1987; Strube, 1988). WISC-V test developers reported the internal consistency reliability as above the “very good” range (above .90) of Cronbach's alpha coefficients put forth by DeVellis’ (2017, pp. 136-137) criteria (VCI = .92; Wechsler, 2014b). The internal consistency reliability coefficients for the Vocabulary and Similarities subtests were computed using the split-half method and then averaged using Fisher’s $z$ transformation (Silver & Dunlap, 1987; Strube, 1988). The subtests also yielded “very good” internal consistency reliability, according to criteria outlined by DeVellis (2017; Similarities subtest = .87, Vocabulary subtest = .87; Wechsler, 2014b).

The WISC-V Technical and Interpretive Manual (Wechsler, 2014b) is the sole source of information to date regarding the internal consistency reliability for the VCI on the WISC-V and the subtests from which it is derived; and, only two independent studies have explored the internal consistency of the prior iteration of the test. Ryan and colleagues (2009) explored the internal consistency reliability of the WISC-IV in a sample of 76 students attending kindergarten through eighth grades in the Midwest region of the U.S. Results of the study indicated slightly higher internal consistency reliability coefficients than those reported by test developers for the standardization sample (VCI = .96, Similarities = .93, Vocabulary = .91, Comprehension = .89). However, the sample was restricted in range to the upper tail of the test score distribution and the authors of the study reported that the inflated estimates were likely due to the above average ability of participants and the influence of private school curriculum and culture on participants’ scores (Ryan et al., 2009). In the second study (Krouse & Braden, 2011), internal consistency reliability was investigated in a sample of 128 deaf and hard-of-hearing children aged six to 16
years old from nine different U.S. states (California, Arizona, Texas, Illinois, Pennsylvania, Maryland, Massachusetts, New York, and Virginia). Results indicated higher internal consistency reliability coefficients than those reported by test developers for the standardization sample (VCI = .97, Similarities = .96, Vocabulary = .93, Comprehension = .91). However, it should be noted that the internal consistency reliability for special groups (e.g., ID/SLD) is often higher than the normative sample because the children in those samples are more consistent in the rate and quality of their responses (Wechsler, 2003). Ultimately, internal consistency reliability is likely an inflated estimate of “reliability” for the subtests that compose the VCI because they are rater-mediated and do not take differences across raters into account.

**Test-Retest Reliability**

The test-retest reliability for the WISC-V VCI and the subtests from which it is derived was obtained by administering the WISC-V twice across an average test-retest interval of 26 days (ranging between 9 and 82 days; Wechsler, 2014b). Test-retest reliability was calculated using Pearson’s product-moment correlation and then the average reliability coefficient across all ages was calculated using Fisher’s \( z \) transformation (Silver & Dunlap, 1987; Strube, 1988). The average corrected stability coefficient for the VCI and the subtests from which it is derived for the standardization sample fell within the “good to excellent” range, according to Portney and
Watkins (2015) criteria (.75 or higher considered to be “good to excellent”, .50 – .75 considered “moderate to good”, .25 – .50 considered “fair”, and .25 or below to be “little or no relationship”) for acceptable correlation standards (VCI = .94, Similarities = .88, Vocabulary = .90; Wechsler, 2014b).

Nevertheless, IQ scores are thought to be stable over time (Gottfried et al., 2006; Schneider et al., 2014) and children who are eligible for special education programs are not required to be re-administered intelligence tests to evaluate continued eligibility or need for services (Madaus & Shaw, 2006); students may be recertified as eligible for services triennially using previous evaluations’ records and current academic performance data (Madaus & Shaw, 2006). Therefore, long-term test-retest stability is especially important. However, WISC-V test developers have not reported the long-term stability of WISC-V test scores over the typical three-year re-evaluation period. There has been only one independently published study of the test-retest reliability of the WISC-V, but multiple published studies of the test-retest reliability of the WISC-IV exist.

Watkins and colleagues (2021) explored the long-term test-retest reliability of WISC-V scores in a clinical sample of 225 children and adolescents from the mid-Atlantic U.S. The average test-retest interval was two and a half years ($M = 2.6, SD = 0.9$). According to Portney and Watkins (2015) criteria for acceptable test-retest reliability coefficients, the results showed “little or no relationship” for the correlation of the Similarities subtest (.20) and “fair” reliability for the Vocabulary subtest (.38) and VCI (.40). Meanwhile, Ryan and colleagues (2010) investigated the test-retest reliability of WISC-IV scores in a sample of 43 elementary and
middle school children across a test-retest interval of approximately 11 months ($M = 10.88$, $SD = 1.22$). Results showed “moderate to good” reliability coefficients for the VCI (.75) and Similarities subtest (.63), “fair” reliability for the Comprehension subtest (.49), and “good to excellent” reliability for the Vocabulary subtest (.81). In another study, Lander (2010) explored the test-retest reliability of WISC-IV scores in a sample of 131 students from a Northeastern U.S. suburban school district with intervals being approximately three years apart ($M = 2.89$). The results of Lander’s (2010) study also indicated a “moderate to good” test-retest reliability coefficient for the VCI (.65), while individual subtests’ reliability coefficients ranged from “fair” to “moderate to good” (Similarities = .48, Vocabulary = .56, Comprehension = .55). Finally, Watkins and Smith (2013) investigated the test-retest reliability of WISC-IV scores in a sample of 344 elementary and middle school students located in the Southwest U.S. who were administered the WISC-IV twice as part of special education evaluations across an average test-retest interval of 2.84 years ($SD = 0.75$). Results were similar to those reported by Lander (2010) with “moderate to good” test-retest reliability indicated for the VCI (.72), while individual subtests’ reliability ranged from “fair” to “moderate to good” (Similarities = .58, Vocabulary = .68, Comprehension = .48). Overall, scales intended for individual diagnostic decisions demand much higher levels of test-retest reliability (Devellis, 2017) and these results do not come close to matching test developer data. Discrepancies between independent researchers’ results and test developer data may exist for a number of reasons including but not limited to: a) differing lengths in test-retest intervals, b) subjectivity in the scoring of responses, and c) rater variability.
Standard Error of Measurement

Test score reliability is important because it is used to estimate the standard error of measurement (SEM) for an observed test score. According to classical test theory (CTT), the SEM is a numerical expression of the uncertainty of a test result that is calculated from test score reliability using the following formula:

$$SEM = SD \sqrt{1 - r_{xx}}$$

There is no measurement error in a test score that is perfectly reliable (i.e., reliability coefficient, \(r_{xx}\), of 1) and the SEM for a test score that is perfectly unreliable (i.e., reliability coefficient, \(r_{xx}\), of 0) indicates that measurement error is at its maximum with the SEM equal to the standard deviation of the observed test score. The SEM for the VCI and the subtests from which it is derived is and has been historically computed from the internal consistency reliability of those test scores. Consequently, as internal consistency reliability increases, the SEM decreases and confidence in the observed test score increases (Wechsler, 2014b). This means that rater variability may not be adequately considered within the SEM. Rater variability adds “noise” to a measurement system that can impact score precision at a single assessment occasion and across multiple assessment occasions, depending on examiners’ rating tendencies and whether or not the same examiners rate examinees each time an assessment is administered (Hoyt, 2000).
Examiner Variability

The subtests that compose the VCI are the only subtests on the WISC that have potential for examiner variability. The term “examiner variability” generally refers to the “systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee” (Scullen et al., 2000, p. 957). Such variability jeopardizes the measure’s reliability, validity, and overall legitimacy (Messick, 1989). Examiner variability is also sometimes referred to as “rater variability,” “rater/observer effects” or “rater/observer bias.” Sources of examiner variability often studied in the literature include severity/leniency, halo/horns, and central tendency effects. The severity/leniency effect occurs when raters consistently score too high/low when compared to other raters (Eckes, 2009). Therefore, school psychologists who are severe raters will consistently score subtests lower than school psychologists who are not, which may ultimately influence individual decisions. Similarly, the halo/horns effect is a cognitive bias in which the examiner’s ratings are influenced by additional unrelated criteria, and not the examinees performance (Eckes, 2009). Examiners who experience the horns effect may assign consistently low scores across items for reasons unrelated to the individual's ability level (i.e., first impressions related to race, socioeconomic status, or perceived intellectual ability). The randomization of items on a given measure can assist in deterring the halo/horns effect, but this is not possible with items on standardized measures due to the strict administration procedures that ensure accuracy and utility. Lastly, examiners who exhibit the central tendency effect avoid the extreme high or low categories of a rating scale or rubric and
tend to use categories near the midpoint of the scale instead. Raters who display such a restriction in range may not give examinees proper credit.

Examiner Variability on the Verbal Comprehension Index

Examiner variability on the VCI and related subtest scores has been explored in a variety of ways (e.g., Alper, 2012; Beasley et al., 1988; Hajzler, 1987). For example, some studies have investigated scoring and administration errors (Alfonso et al., 1998; Belk et al., 2002) while other studies have explored interscorer agreement (Brannigan et al., 1976; Kasper et al., 1968; Ryan et al., 1983; Ryan & Schnakenberg-Ott, 2003). In addition, there is some evidence that suggests rater drift may occur, given that the prevalence of scoring errors differs between graduate students and licensed and/or certified psychologists (Loe et al., 2007; Mrazik et al., 2012; Oak et al., 2019).

Scoring Errors on the Verbal Comprehension Index

Adherence to standardized administration and scoring rules is crucial for acquiring valid scores on standardized assessments. Scaled scores, index scores, and the FSIQ can be greatly affected if examiners fail to adhere to the manuals’ scoring rules and procedures. Given the
implications for utilizing incorrect scores, various types of scoring errors committed on the WISC have been independently investigated. For example, some studies have investigated administration errors (e.g., failure to query, incorrectly applying basal/ceiling rules, use of incorrect starting points; Belk et al., 2002; Oak et al., 2019), while other studies explored computation errors (e.g., calculation of raw subtest scores, conversion of scaled scores to standard scores, transferring scores incorrectly; Alper, 2012; Belk et al., 2002; Linger et al., 2007; Loe et al., 2007). In addition, numerous studies included failure to record examinee responses verbatim as an error (Alper, 2012; Belk et al., 2002; de Ávila Quevedo, 2011; Oak et al., 2019) and some even determined the degree to which examiners assigned incorrect scores to rather ambiguous responses (Alfonso et al., 1998; Loe et al., 2007).

On prior iterations of the WISC, results of empirical research has demonstrated that anywhere from 16-96% of examiners make at least one error when scoring the Vocabulary subtest (Alper, 2012; Brazelton et al., 2003) and between 13-75% of examiners have been demonstrated to make at least one error when scoring the Similarities subtest (Brazelton et al., 2003; Van Noord & Prevatt, 2002). Moreover, examiners have been demonstrated to make more errors, on average, on the VCI subtests when compared to other subtests (Belk et al., 2002; Hall, 1999; Linger et al., 2007; Loe et al., 2007; Miller et al., 1970; Mrazik et al., 2012; Slate & Jones, 1990) and VCI scores corrected to account for examiner errors have been reported to change by as much as 29 standard score points (Alper, 2012).

During the norming phase of the WISC-V, examiners received information regarding common scoring errors prior to scoring any protocols and were coached intermittently over the
entire course of scoring. Examiners also received feedback on initial scoring errors and received additional training, as needed, before scoring additional protocols. Each protocol was independently scored by two examiners and any discrepancies were resolved by a third examiner – a level of oversight that is rarely mimicked in practice. Following all scoring and data entry, all verbal responses in the standardization sample were reexamined to correct for any scoring drift that occurred. Information regarding the prevalence and impact of these errors on the uncorrected protocols, however, is not available in the publication manual (Wechsler, 2014b). Despite test developers' superior supervision and pedagogic methods during the norming and standardization phase, they do not provide further guidance nor recommendations for practitioners.

When assigning scores on the VCI subtests, examinee responses can be categorized as standard (i.e., response samples located in the administration and scoring manual) or ambiguous (i.e., response samples not located in the administration and scoring manual) in nature. For example, examinee responses that are highly similar or exact in wording to the sample responses provided in the scoring manual are “easy” to score, while examinee responses that do not closely align with those in the scoring manual require clinical judgment and are therefore “difficult” to score. When examiners face the latter situation, they have to rely on clinical judgment to assign scores. Consequently, this allows for rater effects (e.g., severity/leniency) to be introduced. Such examiner tendencies are not scoring errors, per se, but they are similar in that the examinee’s observed scores are mediated by characteristics of the rater. However, the WISC is unlike other instruments used to measure cognitive abilities (e.g., Woodcock-Johnson IV Tests of Cognitive
Abilities [WJ-IV-COG]; Mather & Wendling, 2014) in that there are no recommendations provided by the test publishers about how to score ambiguous responses (e.g., “balance scores” like on the WJ-COG). When the different types of scoring errors are considered separately, results of extant research on scoring errors on the WISC suggests that examiners demonstrate the most difficulty with the assignment of “correct” scores for ambiguous responses, rather than scoring errors, per se (Alper, 2012; Belk et al., 2002; Brazelton et al., 2003; de Ávila Quevedo, 2011; Mrazik et al., 2012; Neitzel, 1998; Plumb & Charles, 1955; Sattler et al., 1970). Consequently, examiners’ individual tendencies to rate ambiguous responses high/low may partially explain why the subtests that comprise the VCI contain more examiner scoring errors when compared to any other WISC subtests (Styck & Walsh, 2016).

Evidence of Rater Drift

Research suggests that examiner characteristics (i.e., examiner experience level, rater severity/leniency) may directly affect the prevalence and impact of scoring errors on the WISC (McDermott et al., 2014; Styck & Walsh, 2016). Graduate student examiners are largely the focus of many studies due to their lack of practice with standardized administration and scoring rules (Alfonso et al., 1998; Alper, 2012; Belk et al., 2002; Brannigan et al., 1976; Erdodi et al., 2009; Hall, 1999; Linger et al., 2007; Loe et al., 2007; Miller et al., 1970; Mrazik et al., 2012; Sattler et al., 1970). Independently published studies have also compared the scoring of graduate
student examiners to licensed and/or certified school psychologists to explore the effect of repeated practice, a naturally occurring consequence of examiner experience, on scoring errors (Brazelton et al., 2003; Neitzel, 1998; Oak et al., 2019; Sherrets et al., 1979; Van Noord & Prevatt, 2002). The common notion of “practice makes perfect” should therefore be distinctly evident in that licensed and/or certified school psychologists make fewer scoring errors than graduate students. However, results of a recent meta-analysis that investigated the prevalence and impact of examiner errors on the WISC found that licensed and/or certified school psychologist examiners produce more than two times as many errors per protocol ($M = 7.7$) than graduate student examiners ($M = 3.4$; Styck & Walsh, 2016). These results suggest that licensed and/or certified school psychologists may have a tendency to stray from administration and scoring rules over time—an issue known as rater drift (Sherrets et al., 1979).

Rater drift has been suggested to occur among licensed and/or certified school psychologists on the WISC due to increased experience and practice with administration and scoring rules in the absence of corrective feedback unlike graduate student examiners who receive ample feedback on their adherence to administration and scoring rules as they practice administering and scoring the test. Such repeated practice may result in a pattern of behavior in which examiners rely on their memory rather than the administration and scoring manual when evaluating responses (Erdodi et al., 2009; Oak et al., 2019). Licensed and/or certified school psychologists are viewed as experts in their field by most individuals; hence, they do not commonly receive direct supervision nor constructive feedback on the administration and
scoring of test protocols, which is a customary occurrence for graduate students. This lack of ongoing supervision may potentially contribute to rater drift.

**Interscorer Agreement on the Verbal Comprehension Index**

The WISC-V Technical and Interpretive Manual (Wechsler, 2014b) states that interscorer agreement for the subtests that compose the VCI was obtained via random selection of 60 protocols from the normative sample. Nine examiners who were training to be doctoral-level clinical psychologists and had completed at least one semester in a course in psychological assessment independently scored each test protocol in the selected sample (Wechsler, 2014b). The examiners did not have any prior experiences or practice with the WISC-V scoring rules. Intraclass correlation coefficients (ICCs) provided by the test publishers revealed “good” reliability for the Similarities (.98) and Vocabulary subtests (.99), according to Portney and Watkins (2015) criteria for interpreting the ICC.

The aforementioned results appear extremely promising and are a drastic improvement from the dismal interscorer agreement demonstrated on prior iterations of the WISC (16.7-65.1% agreement) in which the VCI was composed of the Vocabulary, Comprehension, and Information subtests (Brannigan et al., 1976; Cuenot & Darbes, 1982). However, the intensive training provided to the nine doctoral-level clinical psychology students used as raters for the norming sample may not be representative of what one would call “typical” training on the
standardized administration and scoring rules of a new test. It is unclear whether the estimated interscorer agreement reported in the WISC-V Technical Manual would generalize to psychologists in applied settings who may not have received such in-depth training on the test.

Moreover, results of empirical research on the scoring variability for prior iterations of the WISC suggests that such positive results are unlikely among applied practitioners. Specifically, McDermott, Watkins, and Rhoad (2014) estimated the percent of variance attributed to examiners in a sample of 2,783 students twice administered the WISC-IV by 448 examiners as part of special education evaluations and reported that 10% of the variance in WISC-IV VCI scores was due to the examiner, rather than the examinee. These results indicate that a non-negligible amount of variance in VCI scores was attributable to who administered and scored the test. However, these analyses were limited in that they were unable to determine the causal mechanism behind the variance as being due to scoring errors (i.e., administration, failure to record, computation, etc.) or rater effects (e.g., differences in severity/leniency).

The Current Study

Scoring errors and rater effects on the subtests that compose the VCI of the Wechsler Intelligence Scales are well-documented (Belk et al., 2002; Erdodi et al., 2009; Loe et al., 2007; McDermott et al., 2014; Mrazik et al., 2012; Oak et al., 2019; Styck & Walsh, 2016; Van Noord & Prevatt, 2002). These errors have the potential to artificially inflate/depress VCI scores, which
can adversely impact diagnostic assessment decisions that heavily rely on IQ test scores (e.g., SLD, ID). Moreover, although much research has been conducted on prior iterations of the WISC, there is a dearth of information regarding the presence and impact of scoring errors and rater effects on the current version of the test, the WISC-V. The only information currently available comes from the WISC-V Technical and Interpretive Manual (Wechsler, 2014b) and was derived from intensive training and score monitoring procedures that do not reflect typical practice in applied settings. In addition, there is evidence that rater drift (e.g., Styck & Walsh, 2016) occurs postgraduate school (the time in which the most intensive supervision is provided) and psychologists who work in applied settings may be more prone to scoring errors and rater effects. Also, McDermott et al. (2014) was the only study to explore rater effects (as opposed to scoring errors) and they were unable to inspect the tendencies towards severity/leniency of specific raters or their impact on individual examinees scores. The current study aimed to close these gaps in the literature and the following research questions guided the present investigation.

**Research Questions**

The following research questions were investigated in the proposed study: (1) Will raters be significantly different from one another in terms of their severity/leniency? (2) Does rater severity/leniency significantly differ as a function of trainee status (i.e., graduate student vs.
licensed and/or certified school psychologist)? (3) How much will subtest (Vocabulary, Similarities) scores change as a result of rater variability in severity/leniency? (4) How much will VCI scores change as a result of rater variability in severity/leniency?

**Research Predictions**

**Question 1: Will raters be significantly different from one another in terms of their severity/leniency?**

Evidence from the scoring error (e.g., Alper, 2012; Belk et al., 2002; Brazelton et al., 2003; de Ávila Quevedo, 2011; Hall, 1999; Loe et al., 2007; Miller et al., 1970; Mrazik et al., 2012; Neitzel, 1998; Plumb & Charles, 1955; Sattler et al., 1970) and interscorer agreement literature on the WISC (e.g., Brannigan et al., 1976; Cuenot & Darbes, 1982) suggests that examiners differ with regard to how they score the VCI subtests. Therefore, it was predicted that raters would be significantly different from one another in terms of their severity/leniency. Specifically, licensed and/or certified school psychologist raters were predicted to have a larger \( Q \) statistic than graduate student raters, indicating greater variability in the severity/leniency of their scoring.
Question 2: Does rater severity/leniency significantly differ as a function of trainee status (i.e., graduate student vs. licensed and/or certified school psychologist)?

Results of a recent meta-analysis on scoring errors (Styck & Walsh, 2016) and rater drift studies (e.g., Brazelton et al., 2003; Neitzel, 1998; Oak et al., 2019; Sherrets et al., 1979; Van Noord & Prevatt, 2002) suggest that: a) there is a difference between graduate student trainees and licensed and/or certified school psychologists in their scoring habits and b) that graduate student trainees may differ less from one another than licensed and/or certified psychologists. Specifically, graduate student raters have been found to exhibit more severe rating behavior on the VCI subtests in that they were more likely to underestimate scores than overestimate scores (Styck & Walsh, 2016). Therefore, it was hypothesized that graduate student raters would be more severe in their scoring compared to licensed and/or certified school psychologist raters.

Question 3: How much will subtest (Vocabulary, Similarities) scores change as a result of rater variability in severity/leniency?

No other study has examined the change in corrected versus uncorrected scores for the subtests that comprise the VCI. However, evidence from the scoring error literature (e.g., Belk et al., 2002; Hall, 1999; Linger et al., 2007; Loe et al., 2007; Mrazik et al., 2012) suggests that
these scores may change by 0-14 scaled score points due to examiner error. Thus, it was expected that subtest scores would change to a similar magnitude as a result of rater variability in severity/leniency.

Question 4: How much will VCI scores change as a result of rater variability in severity/leniency?

Evidence from a recent meta-analysis that investigated examiner scoring errors on the Wechsler Scales of Intelligence (Styck & Walsh, 2016) found that VCI scores changed by a minimum of 1 standard score points on 77% of all protocols (N = 91) examined within a subgroup of studies (k = 3). Further evidence from the scoring error literature indicates that a change of approximately four standard score points has been demonstrated (Alper, 2012; Loe et al., 2007; Mrazik et al., 2012). Therefore, VCI scores were predicted to change to a similar magnitude as a result of rater variability in severity/leniency.
Raters (also referred to as examiners) were comprised of two groups: a) graduate students in training to become school psychologists and b) licensed and/or certified school psychologists. All raters were recruited via the Illinois School Psychologists email listserv, Northern Illinois University school psychology alumni listserv, Northern Illinois University school psychology graduate student email listserv, and recruited via the primary investigator’s current practicum supervisors, who were directly contacted by the primary investigator (PI). Due to time (i.e., how long it takes to score each protocol) and monetary constraints (i.e., cost per protocol and rater compensation for participation in the study), five licensed and/or certified school psychologist raters and five graduate student raters were recruited for participation in the study ($N = 10$). In order to be eligible for the study, licensed and/or certified school psychologist raters had: a) at least a Master’s or Specialist degree in School Psychology, b) a certification and/or license to practice school psychology, and c) be a currently practicing school psychologist in the United States for at least ten years. In order to be eligible for the study, graduate student raters must
have had: a) current enrollment in a school psychology graduate program and b) successfully completed a course in the theory and assessment of intellectual functioning (e.g., PSYC-640 in the Northern Illinois University graduate school catalog).

A sample of 10 raters were recruited from the northern Illinois region. Most examiners were White (80%) and female (80%). Similarly, graduate students raters \( (N = 5) \) were primarily White (80%) and female (100%). Graduate student examiners, on average, were approximately 24.8 years old \( (SD = 2.17) \) and 80% had obtained at least a Master’s degree in Psychology. Graduate students reported, on average, having scored approximately 16 WISC-Vs \( (\text{Median} = 5) \), and 80% reported having no additional training and/or supervision related to WISC-V scoring and interpretation. A majority of graduate students (80%) also reported that there was no specially appointed person to double check scores nor provide administrative supervision related to intelligence testing at their current clinical site.

Licensed and/or certified psychologists were also primarily White (80%) and female (60%). The average age of licensed and/or certified psychologist examiners was 46.8 years old \( (SD = 8.87) \). Nearly half (40%) of licensed and/or certified psychologist examiners had obtained a Doctorate in Psychology, while 60% reported having a Specialist degree in School Psychology. In addition, 60% of licensed and/or certified psychologist raters reported to have been practicing for more than 21 years, and the remaining 40% of licensed and/or certified psychologist examiners reported to have been practicing for between 10 and 20 years. All licensed and/or certified psychologist raters graduated from a NASP-approved school psychology program (60% graduated from an APA accredited and NASP-approved program) and held a state credential to
practice. One-fifth of all examiners (20%) held a national certification in school psychology credential (NCSP), 20% were licensed for independent practice, and 20% had an American Board of School Neuropsychology (ABSNP) credential. Licensed and/or certified psychologists reported, on average, having scored approximately 13 WISC-Vs (Median = 10), and 60% reported having no additional training and/or supervision related to WISC-V scoring and/or interpretation. All licensed and/or certified psychologist examiners (100%) also reported that there was no specially appointed person to double check scores nor provide administrative supervision related to intelligence testing at their current work site.

Protocols

All protocols (also referred to as examinees) came from real administrations from a university training clinic. If examiner handwriting was difficult to interpret, a second graduate student was consulted for assistance. Examinee responses were then typed onto real WISC-V protocol record forms using size 12 Calibri font to aid readability. Two graduate students independently checked all protocols for administration and/or scoring errors before any analyses were conducted. A third researcher (who is a Nationally Certified School Psychologist [NCSP] and teaches graduate cognitive ability coursework) served as the final determinant for any disagreements among the two graduate students. Examinee responses were also “cleaned” by removing administration and/or scoring errors prior to examiner scoring. The purpose of this was
to minimize distractions from any existing scoring and/or administration errors that were present on the record forms. Despite this procedure, some examiners still noted differences in how they would’ve administered the test (e.g., examiners noted different discontinue points, examiners noted that certain responses were spoiled, etc.).

Protocols (also referred to as examinees) were 60% male. The average age of examinees was 10 years 6 months (Median = 10 years 2 months, $SD = 4.5$, Range = 6 years 1 month–15 years 1 month). VCI standard scores ranged from 76 (Very Low) to 124 (Very High). FSIQ standard scores ranged from 54 (Extremely Low) to 108 (Average). Distribution of VCI and FSIQ scores were approximately normally distributed as seen below in Table 1.

Table 1
Distribution of VCI and FSIQ Scores

<table>
<thead>
<tr>
<th>Composite Score Range</th>
<th>WISC-V Qualitative Descriptor</th>
<th>VCI</th>
<th>FSIQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>69 and below</td>
<td>Extremely Low</td>
<td>0</td>
<td>4 (13%)</td>
</tr>
<tr>
<td>70-79</td>
<td>Very Low</td>
<td>2 (7%)</td>
<td>7 (23%)</td>
</tr>
<tr>
<td>80-89</td>
<td>Low Average</td>
<td>9 (30%)</td>
<td>5 (17%)</td>
</tr>
<tr>
<td>90-109</td>
<td>Average</td>
<td>14 (47%)</td>
<td>14 (47%)</td>
</tr>
<tr>
<td>110-119</td>
<td>High Average</td>
<td>3 (10%)</td>
<td>0</td>
</tr>
<tr>
<td>120-129</td>
<td>Very High</td>
<td>2 (7%)</td>
<td>0</td>
</tr>
<tr>
<td>130 and above</td>
<td>Extremely High</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Measures

**Demographic and Training Variables Questionnaire**

Raters first responded to a self-report survey regarding demographic information (i.e., gender, age, ethnicity, education) and school psychology related training information (e.g., year in graduate program [if applicable], approximate number of WISC-Vs scored, additional training or supervision related to WISC-V scoring/interpretation). Please see Appendix A to review the Demographic and Training Variables Questionnaire that was completed by graduate student examiners and Appendix B to review the Demographic and Training Variables Questionnaire that was completed by licensed and/or certified psychologist examiners.

**WISC-V**

The WISC-V (Wechsler, 2014c) is an individually administered test of general intelligence for children ages six years zero months through 16 years 11 months. The WISC-V is composed of 16 total subtests expressed as scaled scores ($M = 10$, $SD = 3$). There are seven primary subtests (Similarities [SI], Vocabulary [VC], Block Design [BD], Matrix Reasoning [MR], Figure Weights [FW], Digit Span [DS], and Coding [CD]) that produce the FSIQ and
three additional primary subtests (Visual Puzzles [VP], Picture Span [PS], and Symbol Search [SS]) used to produce the five-factor index scores (two subtests each for the VCI, Visual Spatial Index [VSI], Fluid Reasoning Index [FRI], Working Memory Index [WMI], and Processing Speed Index [PSI]). In addition, there are six secondary subtests (Information, Comprehension, Picture Concepts, Arithmetic, Letter-Number Sequencing, and Cancellation) that are used either for substitution in FSIQ estimation (when one primary subtest is spoiled) or in estimating the General Ability Index, Cognitive Proficiency Index, or three additional ancillary index scores (Quantitative Reasoning [QRI], Auditory Working Memory [AWMI], and Nonverbal [NVI]). The FSIQ and index scores are expressed as standard scores ($M = 100, SD = 15$). For this study, the subtests that comprise the VCI (VC, SI) were referenced.

Procedure

The proposed study has been approved by the Northern Illinois University Institutional Review Board. Each rater completed one Demographic and Training Variables Questionnaire and scored twelve WISC-V protocols. Graduate student raters completed the Demographic and Training Variables Questionnaire via a Google Form and scored their assigned, partially completed WISC-V protocols (subtests composing the VCI only) in paper-pencil format. A WISC-V scoring manual was provided to graduate student examiners to use in scoring the WISC-V protocols, if needed. Licensed and/or certified school psychologist raters completed the
Demographic and Training Variables Questionnaire via a Google Form and received all WISC-V protocols via USPS mail with a prepaid and addressed envelope for return. The Demographic and Training Variables Questionnaire was anticipated to take approximately five to ten minutes to complete and each WISC-V protocol was anticipated to take approximately 15 minutes to score. In total, it was expected to take each participant approximately three hours to complete all study materials.

After the raters completed and returned all documents, the PI, one additional graduate student, and the PI’s faculty mentor double-checked the scoring of each individual protocol to ensure correct scoring as outlined in the WISC-V administration and scoring manual (Wechsler, 2014a). All scoring errors were noted and flagged for sensitivity analyses. Scoring errors included giving too much or too little credit for a standard response (e.g., a two- or one-point response spoiled by a zero-point response but was still scored as two or one points).

Rating Design

A total of 30 participant protocols were scored by ten examiners using the incomplete rating design illustrated in Figure 1. Incomplete rating designs feature systematic connections, either direct and/or indirect, between participants, items, and raters so that each of these elements can be unambiguously compared (Engelhard, 1997). An anchor set of 10 WISC-V protocols was
scored by every rater to create a linking set of commonly administered protocols; raters scored an additional two randomly selected protocols for a grand total of 12 participant protocols scored by each rater. In order to ensure that protocols represented the range of verbal cognitive abilities measured by the WISC-V, the 10 WISC-V protocols in the linking set consisted of protocols from participants with “Average”, “Low Average”, and “High Average” verbal cognitive ability.

Independent Variable

The current study sought to estimate the magnitude of examiner variability on the VCI between graduate students and licensed and/or certified school psychologists. Five graduate student raters and five licensed and/or certified school psychologist raters were recruited for participation. In order to be eligible for the study, graduate student examiners must have had: a) current enrollment in a school psychology graduate program and b) successfully completed a course in the theory and assessment of intellectual functioning (e.g., PSYC 640: Theory and Assessment of Intellectual Functioning). In order to be eligible for the study, licensed and/or certified psychologist school examiners must have had: a) at least a Master’s or Specialist degree in School Psychology, b) a license and/or certification to practice school psychology in good standing with their respective license and/or certification board, and c) be a currently practicing school psychologist in the United States for a minimum of ten years.
<table>
<thead>
<tr>
<th>Participant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduate Students</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>9</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Figure 1. Rating design.
Dependent Variables

Examiners’ rater severity/leniency were estimated using many-facet Rasch measurement (MFRM; Linacre, 1989). MFRM allows “facets” such as examinee proficiency, item difficulty, and rater severity/leniency estimates to be directly compared on the same equal-interval log-odds (logit) scale. In a three-facet MFRM model, the probability that an examinee will receive a given score on a given item is a function of the examinee’s proficiency, the difficulty of the item, and the severity/leniency of the rater. This feature of MFRM is particularly useful in assessing the impact of examiner variability on WISC-V VCI subtests’ scoring because it permits the separate estimation of rater severity/leniency from item difficulty and examinee proficiency.

Overall model fit was assessed by inspecting standardized residuals and mean-square (MSQ) infit/outfit statistics. Standardized residuals represent the differences between the observed and expected scores. Large standardized residuals (those that exceed ± 2.0) indicate that the observed score deviates by two or more standard deviations from the expected score. Standardized residual values that exceed ± 2.0 may indicate severe inconsistency of examiner scoring (Eckes, 2009). Therefore, less than 5% of the standardized residuals should exceed ± 2.0 and less than 1% of standardized residuals should exceed ± 3.0 (Linacre, 2018; Myford & Wolfe, 2003). In addition to inspecting global model-data fit, rater fit was also examined using MSQ fit statistics. MSQ infit/outfit statistics are the average of the squared standardized
residuals across and within facets. MSQ infit statistics are weighted and provide an overall estimate of rater consistency across examinees and criteria. Conversely, MSQ outfit statistics are unweighted in that they are sensitive to unexpected scores (Eckes, 2009). MSQ fit statistics should be distributed evenly and have values of approximately 1.0, indicating good model-data fit (Eckes, 2015; Linacre, 2003). MSQ infit/outfit values >1.0 indicate that the differences between observed and expected scores are greater than predicted and that the data underfit (misfit) the MFRM model. MSQ infit/outfit values <1.0 indicate that the differences between observed and expected scores are lower than predicted and suggest that the data may overfit the MFRM model (Bond & Fox, 2007; Eckes, 2009). Linacre (2003) suggested that MSQ infit/outfit values between 0.50 and 2.0 indicate acceptable model-data fit, whereas other researchers have suggested more stringent criteria such as MSQ infit/outfit values between 0.7 and 1.3 (Bond & Fox, 2007; Wright & Linacre, 1994). MSQ infit/outfit values between 0.5 and 2.0 were considered acceptable due to the exploratory nature of the study.

Differences between rater severity/leniency were estimated using the Q statistic. Q, the homogeneity statistic, is an omnibus test of significant differences in rater severity/leniency across examiners (Eckes, 2009). A significant (e.g., p < .05) Q value indicates that at least two raters differed significantly in the severity/leniency of their scoring and the larger the Q value would represent increased separation in raters’ scoring. Differences between rater severity/leniency between two trainee status groups (e.g., graduate students and licensed and/or certified school psychologists) were also estimated using a differential group functioning
analysis (i.e., item × group (φij) or examinee × group (φni) interaction; Eckes, 2009). This was used to clarify whether or not graduate student and licensed and/or certified school psychologist participant’s scoring tendencies are uniform across items and/or examinees. Finally, rater impact was estimated by inspecting the difference between observed average and fair average scores. A fair average score is predicted by the MFRM model and is the score that an examinee would have received on an item of average difficulty from a rater of average severity/leniency (Eckes, 2009). If an observed score is higher than its associated fair average score, more lenient rating behavior was exhibited by an examiner. Conversely, if an observed score is lower than its associated fair average score, then more severe rating behavior was exhibited by an examiner.

Data Analysis Plan

Question 1: Will raters be significantly different from one another in terms of their severity/leniency?

In order to assess differences in severity/leniency across all raters, a chi-square-based statistic (e.g., the Q statistic) was calculated. The Q statistic was calculated for each group, independently, to explore significant differences in severity/leniency between each trainee group.
Question 2: Does rater severity/leniency significantly differ as a function of trainee status (i.e., graduate student vs. licensed and/or certified school psychologist)?

A differential group functioning analysis was used to determine if severity/leniency differs as a function of trainee status (i.e., graduate student vs. licensed and/or certified school psychologist). This analysis was supplemented with separate analyses such as each individual group’s $Q$ statistic and individual model-data fit statistics (e.g., MSQ infit/outfit statistics).

Question 3: How much will subtest (Vocabulary, Similarities) scores change as a result of rater variability in severity/leniency?

In order to determine how much subtest scores changed as a result of variability in severity/leniency, observed scaled scores were compared to fair average scaled scores.
Question 4: How much will VCI scores change as a result of rater variability in severity/leniency?

Observed standard scores and fair average standard scores were compared to determine how much VCI scores changed as a result of variability in severity/leniency.
CHAPTER IV

RESULTS

Question 1: Will raters be significantly different from one another in terms of their severity/leniency?

Similarities

Category score usage (i.e., a score of 0, 1, or 2) was as expected considering the basal/ceiling rules for the Similarities subtest. Responses for items one through five were primarily assigned a score of 2, given that it is required to advance on to subsequent items. On the contrary, category usage for the latter items on the Similarities subtest (e.g., items 17-23) contained some sparseness, as many participants had already met the discontinue criteria (three consecutive scores of ‘0’) and did not have an opportunity to respond to these items. Data demonstrated good overall fit to the MFRM model as evidenced by only 4.2% of standardized residuals $\geq 2$ and only 1.38% $\geq 3$. However, inspection of facet elements indicated one misfitting rater (MNSQ infit/outfit = 2.12), Licensed and/or Certified Psychologist Rater 3. This indicated that Rater 3’s scoring was too unpredictable to be explained by the MFRM model.
Further inspection indicated that Rater 3 assigned four ratings with standardized residuals $\geq 3$. However, only one of these residuals was due to a scoring error. The remaining residuals were due to idiosyncratic “student” performance patterns that were unexpected by the MFRM model (i.e., the student responded incorrectly on easy items despite high proficiency overall). However, no substantial improvements in model-data fit resulted in removing the misfitting rater. As a result, all raters were retained for further analyses. Inspection of facet elements also demonstrated two misfitting items (MNSQ infit/outfit $\geq 2$), Items 5 and 7. Given that the WISC-V is a standardized measure and removal of items would not occur on a standardized administration, all items were retained for further analyses. Nevertheless, overall model-data fit was good. Finally, average measures increased monotonically with rating categories indicating that higher ratings were associated with higher levels of latent verbal ability.

$Q$, the homogeneity statistic, was not significant ($p = .63$) indicating that raters did not significantly differ in their tendencies toward severity/leniency when scoring the Similarities subtest. Moreover, intraclass correlation coefficients ranged from 0.59 to 1.0 across Similarities items (see Table 2). Only one item (Item 17) had an intraclass correlation coefficient (ICC) value that was fair ($\leq 0.59$) as outlined in Cicchetti (1994). This value indicates a less than desirable reliability of ratings on the above-mentioned item, and that relative to other items, raters did not score as similarly to one another. Variability in ratings may have occurred for this item because the latter items on the Similarities subtest contained some sparseness (there was less of an opportunity for raters to score this item) and/or this item was simply more difficult for raters to score.
Table 2

ICC Values for the Similarities Subtest

<table>
<thead>
<tr>
<th>Item</th>
<th>n</th>
<th>k</th>
<th>ICC</th>
<th>ICC Value Qualitative Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>2</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>3</td>
<td>120</td>
<td>10</td>
<td>0.61</td>
<td>Good</td>
</tr>
<tr>
<td>4</td>
<td>120</td>
<td>10</td>
<td>0.94</td>
<td>Excellent</td>
</tr>
<tr>
<td>5</td>
<td>120</td>
<td>10</td>
<td>0.76</td>
<td>Excellent</td>
</tr>
<tr>
<td>6</td>
<td>120</td>
<td>10</td>
<td>0.94</td>
<td>Excellent</td>
</tr>
<tr>
<td>7</td>
<td>120</td>
<td>10</td>
<td>0.93</td>
<td>Excellent</td>
</tr>
<tr>
<td>8</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>9</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>10</td>
<td>120</td>
<td>10</td>
<td>0.99</td>
<td>Excellent</td>
</tr>
<tr>
<td>11</td>
<td>120</td>
<td>10</td>
<td>0.93</td>
<td>Excellent</td>
</tr>
<tr>
<td>12</td>
<td>120</td>
<td>10</td>
<td>0.85</td>
<td>Excellent</td>
</tr>
<tr>
<td>13</td>
<td>120</td>
<td>10</td>
<td>0.92</td>
<td>Excellent</td>
</tr>
<tr>
<td>14</td>
<td>120</td>
<td>10</td>
<td>0.97</td>
<td>Excellent</td>
</tr>
<tr>
<td>15</td>
<td>120</td>
<td>10</td>
<td>0.88</td>
<td>Excellent</td>
</tr>
<tr>
<td>16</td>
<td>120</td>
<td>10</td>
<td>0.96</td>
<td>Excellent</td>
</tr>
<tr>
<td>17</td>
<td>120</td>
<td>10</td>
<td>0.59</td>
<td>Fair</td>
</tr>
<tr>
<td>18</td>
<td>120</td>
<td>10</td>
<td>0.83</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

(Continued on following page)
Table 2 (continued)

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>120</td>
<td>10</td>
<td>0.79</td>
<td>Excellent</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>120</td>
<td>10</td>
<td>0.78</td>
<td>Excellent</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>120</td>
<td>10</td>
<td>0.83</td>
<td>Excellent</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>120</td>
<td>10</td>
<td>0.63</td>
<td>Good</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>120</td>
<td>10</td>
<td>0.91</td>
<td>Excellent</td>
<td></td>
</tr>
</tbody>
</table>

**Vocabulary**

Comparable to the Similarities subtest, category score usage was as expected given the basal/ceiling rules for the Vocabulary subtest. Responses for items one through four were all assigned a score of 1, given that it is required to advance on to subsequent items. Dissimilarly, category usage for the latter items on the Vocabulary subtest (e.g., items 21-29) contained some sparseness, as many participants had already met the discontinue criteria (three consecutive scores of ‘0’) and did not have an opportunity to respond to these items. Data demonstrated good overall fit to the MFRM model as evidenced by only 2.01% of standardized residuals ≥ |2| and only 1.21% ≥ |3|. However, inspection of facet elements indicated one misfitting rater, Licensed and/or Certified Psychologist Rater 2 (MNSQ infit/outfit = 9). This indicated that Rater 2’s scoring was too unpredictable to be explained by the MFRM model. Further inspection indicated
that Rater 3 assigned five ratings with standardized residuals ≥ 3. However, only one of these residuals was due to a scoring error. The remaining residuals were due to idiosyncratic “student” performance patterns that were unexpected by the MFRM model (i.e., the student responded incorrectly on easy items despite high proficiency overall). However, no substantial improvements in model-data fit resulted in removing the misfitting rater. As a result, all raters were retained for further analyses. Inspection of facet elements also demonstrated four misfitting items (MNSQ infit/outfit ≥ 2): Items 7, 8, 10, and 17. Given that the WISC-V is a standardized measure and removal of items would not occur on a standardized administration, all items were retained for further analyses. Nevertheless, overall model-data fit was good. Finally, average measures increased monotonically with rating categories indicating that higher ratings were associated with higher levels of latent verbal ability.

\[ Q, \text{ the homogeneity statistic, was not significant (} p = .16 \text{) indicating that raters did not significantly differ in their tendencies toward severity/leniency when scoring the Vocabulary subtest. Moreover, ICC values ranged from 0.52 to 1.0 across Vocabulary items (see Table 3). One item (Item 10) had an intraclass correlation coefficient (ICC) value that was fair (≤ 0.59) as outlined in Cicchetti (1994). This value indicates a less than desirable reliability of ratings on the above-mentioned item, and indicates that relative to other items, raters did not score as similarly to one another. Variability in ratings may have occurred for this item because it was simply more difficult for raters to score and/or examinees had such variable responses for the item.} \]
Table 3

ICC Values for the Vocabulary Subtest

<table>
<thead>
<tr>
<th>Item</th>
<th>$n$</th>
<th>$k$</th>
<th>ICC</th>
<th>ICC Value Qualitative Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>2</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>3</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>4</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>5</td>
<td>120</td>
<td>10</td>
<td>0.89</td>
<td>Excellent</td>
</tr>
<tr>
<td>6</td>
<td>120</td>
<td>10</td>
<td>0.71</td>
<td>Good</td>
</tr>
<tr>
<td>7</td>
<td>120</td>
<td>10</td>
<td>0.62</td>
<td>Good</td>
</tr>
<tr>
<td>8</td>
<td>120</td>
<td>10</td>
<td>0.63</td>
<td>Good</td>
</tr>
<tr>
<td>9</td>
<td>120</td>
<td>10</td>
<td>0.84</td>
<td>Excellent</td>
</tr>
<tr>
<td>10</td>
<td>120</td>
<td>10</td>
<td>0.52</td>
<td>Fair</td>
</tr>
<tr>
<td>11</td>
<td>120</td>
<td>10</td>
<td>0.94</td>
<td>Excellent</td>
</tr>
<tr>
<td>12</td>
<td>120</td>
<td>10</td>
<td>0.92</td>
<td>Excellent</td>
</tr>
<tr>
<td>13</td>
<td>120</td>
<td>10</td>
<td>0.95</td>
<td>Excellent</td>
</tr>
<tr>
<td>14</td>
<td>120</td>
<td>10</td>
<td>0.93</td>
<td>Excellent</td>
</tr>
<tr>
<td>15</td>
<td>120</td>
<td>10</td>
<td>0.95</td>
<td>Excellent</td>
</tr>
<tr>
<td>16</td>
<td>120</td>
<td>10</td>
<td>0.97</td>
<td>Excellent</td>
</tr>
<tr>
<td>17</td>
<td>120</td>
<td>10</td>
<td>0.89</td>
<td>Excellent</td>
</tr>
<tr>
<td>18</td>
<td>120</td>
<td>10</td>
<td>0.98</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

(Continued on following page)
Table 3 (continued)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>20</td>
<td>120</td>
<td>10</td>
<td>0.94</td>
<td>Excellent</td>
</tr>
<tr>
<td>21</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>22</td>
<td>120</td>
<td>10</td>
<td>0.86</td>
<td>Excellent</td>
</tr>
<tr>
<td>23</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>24</td>
<td>120</td>
<td>10</td>
<td>0.91</td>
<td>Excellent</td>
</tr>
<tr>
<td>25</td>
<td>120</td>
<td>10</td>
<td>0.85</td>
<td>Excellent</td>
</tr>
<tr>
<td>26</td>
<td>120</td>
<td>10</td>
<td>0.86</td>
<td>Excellent</td>
</tr>
<tr>
<td>27</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>28</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>29</td>
<td>120</td>
<td>10</td>
<td>1.0</td>
<td>Excellent</td>
</tr>
</tbody>
</table>
Question 2: Does rater severity/leniency significantly differ as a function of trainee status (i.e., graduate student vs. licensed and/or certified school psychologist)?

**Similarities**

Results indicated that rater severity/leniency did not significantly differ as a function of trainee status, $Q(1) = 0.2$, $p = .64$. This suggests that graduate students and licensed and/or certified psychologists did not differ in their tendencies towards assigning high/low scores.

**Vocabulary**

Results indicated that rater severity/leniency did not significantly differ as a function of trainee status, $Q(1) = 0.7$, $p = .41$. This suggests that graduate students and licensed and/or certified psychologists did not differ in their tendencies towards assigning high/low scores.
Question 3: How much will subtest (Vocabulary, Similarities) scores change as a result of rater variability in severity/leniency?

Similarities

There were 4.47 statistically distinguishable levels of examinee proficiency on the Similarities subtest and the reliability of the examinee separation index was .91 indicating that 91% of the variance in the examinee measures were not due to measurement error. Given the absence of systematic differences in examiner severity/leniency as outlined in the results for Questions 1 and 2, a paired-samples $t$-test was calculated to compare observed average scaled scores to fair average scaled scores. If examiner-level analyses were conducted, no significant differences would likely exist given that there was an absence of systematic differences in examiner severity/leniency. Therefore, the observed average mean was 9.27 ($SD = 2.72$), and the fair average mean was 8.33 ($SD = 5.57$). There was not a significant difference between observed average and fair average scaled scores, $t(29) = 1.331, p = 0.194, d = 0.21$. 
Vocabulary

There were 6.37 statistically distinguishable levels of examinee proficiency on the Similarities subtest and the reliability of the examinee separation index was .95 indicating that 95% of the variance in the examinee measures were not due to measurement error. Given the absence of systematic differences in examiner severity/leniency as outlined in the results for Questions 1 and 2, a paired-samples t-test was calculated to compare observed average scaled scores to fair average scaled scores. If examiner-level analyses were conducted, no significant differences would likely exist given that there was an absence of systematic differences in examiner severity/leniency. Therefore, the observed average mean was 9.23 ($SD = 3.10$), and the fair average mean was 7.87 ($SD = 7.26$). There was not a significant difference between observed average and fair average scaled scores, $t(29) = 1.325$, $p = 0.195$, $d = 0.24$.

Question 4: How much will VCI scores change as a result of rater variability in severity/leniency?

Given the absence of systematic differences in examiner severity/leniency as outlined in the results for Questions 1 and 2, a paired-samples t-test was calculated to compare observed average standard scores to fair average standard scores. If examiner-level analyses were
conducted, no significant differences would likely exist given that there was an absence of systematic differences in examiner severity/leniency. Therefore, the observed average mean was 95.97 ($SD = 14.14$), and the fair average mean was 89.90 ($SD = 32.31$). There was not a significant difference between observed average and fair average standard scores, $t(29) = 1.326$, $p = 0.195$, $d = 0.25$. 
CHAPTER V
DISCUSSION

Given that scoring errors and rater effects are well documented on previous iterations of the WISC (Belk et al., 2002; Erdodi et al., 2009; Loe et al., 2007; McDermott et al., 2014; Mrazik et al., 2012; Oak et al., 2019; Styck & Walsh, 2016; Van Noord & Prevatt, 2002) and have the potential to artificially inflate and/or depress intelligence test scores, the current study aimed to explore rater effects on the VCI of the WISC-V. Additionally, there is some evidence that rater drift may occur postgraduate school (Styck & Walsh, 2016) indicating that psychologists who work in applied settings may be more prone to scoring errors and rater effects when compared to graduate students. Consequently, the purpose of the current study was to: a) explore examiner variability on the VCI, b) explore examiner tendencies, and c) compare examiner tendencies between graduate students and licensed and/or certified school psychologists.
Question 1: Will raters be significantly different from one another in terms of their severity/leniency?

The WISC scoring error literature has consistently shown that the subtests that compose the VCI contain the most errors (e.g., Alper, 2012; Belk et al., 2002; Brazelton et al., 2003; de Ávila Quevedo, 2011; Hall, 1999; Loe et al., 2007; Miller et al., 1970; Mrazik et al., 2012; Neitzel, 1998; Plumb & Charles, 1955; Sattler et al., 1970) indicating that examiners differ in regard to how they score the VCI subtest items. Therefore, it was hypothesized that examiners would be significantly different from one another in terms of their tendencies towards severity/leniency on the Vocabulary and Similarities subtests. However, results of the current study revealed that examiners did not differ in regard to their severity/leniency when scoring Similarities and Vocabulary subtest items. Such a lack of significant findings may be due to a few factors. First, ambiguous responses occurred infrequently on any given protocol. The subtests that comprise the VCI are the only subtests on the WISC-V that allow for rater effects to occur because responses for items on these subtests can be categorized as standard or ambiguous in nature. Standard responses closely align with response samples located in the administration and scoring manual; these responses are considered “easy” to score given that they are highly similar or exact in wording to the sample responses provided in the scoring manual. On the other hand, ambiguous responses do not closely align with any response samples in the scoring manual. It is for this reason that these responses are considered “difficult” to score, and they often require a significant amount of clinical judgment when arriving at a score. Such room for
clinical judgment allows for the introduction of rater effects to occur. The current study used real protocols obtained from the university psychological services center. However, in using these protocols, it became evident that few ambiguous responses were present, allowing for less opportunity for rater effects to occur. Additionally, because all responses used in the study were recorded and written by different examiners, all notated responses may not necessarily be verbatim (e.g., examiners may not have recorded every utterance from an examinee or paragraphed a lengthy response). It is possible that there may truly be more ambiguous responses on any given protocol, however, the current study was only able to use what was already documented and previously recorded. Given that real examinee responses were used for the current study, future research may consider using fabricated responses and manipulate the number of ambiguous responses in an attempt to elicit examiner variability. In addition, because examiners in the study (i.e., licensed and/or certified psychologists and graduate students) did not administer these subtests themselves, there may be “hints” and/or “cues” on how to score certain items based on information existing on the protocol. For example, the WISC-V has a ceiling rule for the subtests that comprise the VCI stating that testing must discontinue after three consecutive scores of zero. These three scores of zero must then apparently be mapped onto the last three items that have recorded responses. It is possible that examiners may have used their knowledge of the ceiling rule to score the last three items, and they may not have actually taken the time to inspect the responses and score them appropriately. Additionally, examiners had unlimited time to score the responses on each protocol. This may also contribute as to why no differences in severity/leniency were detected. During real administrations of the WISC-V,
examiners score in the moment to determine if examinees should be allowed to move on to the next item or if testing should be discontinued (i.e., if the ceiling has been met) and have state-mandated deadlines for completing testing as well as other responsibilities competing for their time and attention. However, our study allowed examiners unlimited time to score examinee responses, so perhaps this lack of time constraint allowed examiners to score more thoroughly and accurately.

Finally, although there were no differences amongst raters in their overall tendencies to rate high/low, there was some exact disagreement between raters. This disagreement typically occurred within one point on the sample of protocols used in the current study. Though some exact disagreement was detected amongst raters’ scores, the absence of rater effects indicates that this disagreement was not systematic in the sense that individual raters did not demonstrate any tendencies to rate high/low when presented with an ambiguous examinee response.

Question 2: Does rater severity/leniency significantly differ as a function of trainee status (i.e., graduate student vs. licensed and/or certified school psychologist)?

Results of rater drift studies (e.g., Brazelton et al., 2003; Neitzel, 1998; Oak et al., 2019; Sherrets et al., 1979; Van Noord & Prevatt, 2002) and a recent meta-analysis by Styck and Walsh (2016) suggest that school psychology graduate students and licensed and/or certified psychologists differ in their scoring habits on the WISC. In particular, these studies suggest that
graduate students make significantly fewer scoring errors than licensed and/or certified psychologists. Therefore, it was hypothesized that rater severity/leniency would differ between groups (i.e., graduate student vs. licensed and/or certified school psychologist). Results of the current study revealed that rater severity/leniency did not significantly differ as a function of trainee status (i.e., graduate student vs. licensed and/or certified school psychologist). This may be due to the fact that raters likely had similar training in graduate school, given that they were recruited via an alumni listserv and word-of-mouth using a “snowball technique” through the PI’s practicum supervisors (many of whom are alumni). Nevertheless, it is not possible for all licensed and/or certified psychologists that participated in the study to have had the exact same cognitive assessment training in graduate school due to changes in course instructors and/or assignments as well as varying practicum experiences and supervisors, etc.. Moreover, no significant differences were found in rater tendencies towards severity/leniency (Question 1). Therefore, there was no group-level (i.e., licensed and/or certified psychologists and graduate student examiners) heterogeneity to detect.
Question 3: How much will subtest (Vocabulary, Similarities) scores change as a result of rater variability in severity/leniency?

The change in observed average scores versus fair average scores has not previously been explored for the Similarities and Vocabulary subtests. Therefore, evidence from scoring error literature (e.g., Belk et al., 2002; Hall, 1999; Linger et al., 2007; Loe et al., 2007; Mrazik et al., 2012) was referenced, and these studies suggest that scores may change by 0-14 scaled score points due to examiner errors. It was hypothesized that subtest scores would change to a similar magnitude due to examiner variability in severity/leniency. However, results of the current study revealed that there was not a significant difference between observed average and fair average scaled scores. Though not significant, observed average scaled scores were consistently higher than the fair average scaled scores produced by the MFRM model. Given that no significant differences in rater severity/leniency were identified amongst our sample of raters, these findings reflect differences in item difficulty rather than differences in rater severity/leniency. For example, most examinees were not administered many difficult items (i.e., they met the discontinue criteria well before reaching the last item on a given VCI subtest). However, the fair average score is computed from the average item difficulty across all items on the subtest and average rater severity/leniency for each examinee and their level of estimated ability. This ultimately depressed the fair average scores produced by the MFRM model because the average item difficulty across all items on the subtest was higher than the average item difficulty that
most examinees were actually administered, resulting in higher observed average scores produced by the examiners.

Lastly, a closer examination of the item difficulty measures suggests that it is possible that some items may not be rank ordered correctly in terms of difficulty. This could have affected how the fair average scores produced by the MFRM model compared to the observed average scores because of the scoring rules of the WISC. For example, once the basal is established for an examinee, all “easier” items (items ordered below the basal) are scored correctly. However, those same items were not necessarily the “easier” items according to the MFRM model. This suggests that Vocabulary and Similarities subtest sum scores might be inflated, regardless of between-rater differences in scoring.

Question 4: How much will VCI scores change as a result of rater variability in severity/leniency?

Results of a recent meta-analysis by Styck and Walsh (2016) found that VCI scores changed by a minimum of 1 standard score points on 77% of all protocols examined in a subgroup of studies. Evidence from scoring error literature (Alper, 2012; Loe et al., 2007; Mrazik et al., 2012) also demonstrated a change of approximately four standard score points when exploring observed versus corrected VCI standard scores. It was hypothesized that VCI
scores would change to a similar magnitude due to examiner variability in severity/leniency.

However, results of the current study revealed no significant differences in observed versus fair average VCI standard scores, on average. This result was not altogether unexpected, given that the subtests that compose the VCI (Similarities and Vocabulary) also showed no significant differences in observed versus fair average scaled scores. Nevertheless, observed average VCI scores were consistently higher than the fair average VCI scores produced by the MFRM model. However, these results are likely due to the item effects depressing the fair average Vocabulary and Similarities scaled scores previously described.

Implications

Results of the current study suggest that graduate students and practicing school psychologists share similar rating tendencies and score similarly on the subtests that compose the VCI. School psychologists should continue to adhere to standardized administration and scoring rules as outlined in the WISC-V administration and scoring manual. Practitioners should also continue consulting with colleagues regarding ambiguous examinee responses. However, one integral question still remains to be answered if results of the present study are generalizable to standard administrations of the WISC-V – if scoring errors do not account for the examiner variability identified in McDermott et al.’s (2016) study as described by Styck and Walsh (2016) and between-rater differences in scoring examinee responses is not systematic as indicated by
results of the present study, then what explains the variability in WISC scores that McDermott et al. reported as being due to the examiner? McDermott et al. suggested that various characteristics of the examiner, examinee, or examiner-examinee relationship may impact test scores. Though not examined in this study, the familiarity effect (e.g., examiner familiarity with the examinee and/or the examinee’s perceived ability level), examinee motivation, or violation of standardized administration rules may impact WISC test scores.

Researchers should continue to investigate the ways in which rater effects impact scores on intelligence tests, given that their use and interpretation are used in many high-stakes decisions (e.g., eligibility for special education supports and services, the death penalty, possibility for parole, etc.). Assessing the WISC-V in this context may provide important information regarding where errors and rater effects are more likely to occur, which in turn may assist in improving the external validity of the resultant scores, and consequently improve the accuracy of clinical inferences made. Though few ambiguous responses were identified on the subtests that comprise the VCI, variance in scores still remain, and this variance is not attributed to scoring errors nor rater effects.
Limitations and Future Directions

Undoubtedly, there were some limitations of the present study. First, all graduate students were drawn from a single graduate program. This limits the possibility for examiner variability in scoring to appear given that all students had the same training regarding intelligence test scoring and administration procedures and it is possible that their tendencies towards severity/leniency are also similar as a result. Additionally, some licensed and/or certified psychologists may have graduated from this same training program, given participant recruitment strategies and all licensed and/or certified psychologists practiced in the same geographical region. This may have further contributed to rater homogeneity. As a result, this sample of raters may not be representative of graduate students and licensed and/or certified psychologists trained differently or practicing in different regions. Second, the current study had no protocols from students whose general cognitive functioning abilities (e.g., FSIQ scores) were within the Extremely High range. Though examinees’ VCI scores were approximately normally distributed, examinees’ FSIQ scores were restricted to the Average to Below Average range of cognitive ability and this may have also restricted the quality of examinee responses on the Vocabulary and Similarities subtests. Lastly, ambiguous responses occurred infrequently on each protocol, which may have impacted the results such that fewer ambiguous responses should result in higher agreement between raters, and less opportunity for rater effects to appear. However, the protocols were based on real examinee responses from WISC-V administrations conducted at a university training clinic to aid the generalizability of study findings.
Consequently, this limitation is also a strength in that it is possible that ambiguous responses may simply be far and few between on standard WISC-V administrations. Future research should examine the impact of administration errors and familiarity effects on WISC-V Vocabulary and Similarities subtest scores as well as replicate this study with other samples of raters in an attempt to further pinpoint the source of examiner variability on WISC scores. Additional ancillary research on the administration of the WISC-V should continue to look at how results may differ for iPad/virtual administrations of the test. For example, iPad administration allows the examiner to select the rubric response that best matches the examinee’s response and does not require the examiner to dictate verbatim the examinee’s response. Therefore, the examinee’s responses are not recorded and the extent to which an error in scoring was committed is unknown. Being forced to score in the moment without an opportunity to review the score against the rubric might magnify examiner tendencies toward severity/leniency. Nevertheless, results of current research indicate that there are no differences in digital-paper score equivalencies (e.g., Daniel & Wahlstrom, 2019).

Conclusion

The current study investigated examiner variability (i.e., severity/leniency) on the subtests that compose the VCI of the WISC-V. It was hypothesized that rater effects would be present overall and between groups (i.e., licensed and/or certified psychologists and graduate
students). Additionally, it was hypothesized that a considerable difference in scores between observed average and fair average scores would exist, considering that the results of prior literature (e.g., Belk et al., 2002; Hall, 1999; Linger et al., 2007; Loe et al., 2007; Mrazik et al., 2012; Alper, 2012; Loe et al., 2007; Mrazik et al., 2012) indicate that differences in raw versus adjusted scaled and standard scores do exist after accounting for scoring errors. Results revealed that raters effects did not exist for the Similarities and Vocabulary subtests in the current study. Future research should consider administration errors and familiarity effects for producing the unaccounted examiner variance in WISC scores.
REFERENCES


Alper, J. (2012). Graduate students’ administration and scoring errors on the WISC-IV: Reducing inaccuracies with training and experience (Publication No. 1651829054) [Doctoral dissertation, St. John’s University]. ProQuest Dissertations and Theses Global.


Demographic and Training Variables Questionnaire

**Directions:** Please carefully check/fill-in your answers below.

1) How do you currently describe your gender identity?
   a) Male
   b) Female
   c) Transgender
   d) Other: _______________

2) What is your age in years?
   __________

3) Which categories describe you? Check all that apply to you:
   a) American Indian or Alaska Native
   b) Asian or Pacific Islander
   c) Black
   d) Hispanic, Latino, or Spanish Origin
   e) Middle Eastern or North African
   f) Native Hawaiian or Other Pacific Islander
   g) White
   h) Other: _______________

4) What is the highest degree or level of education you have completed?
   a) Bachelor’s degree (e.g., BA, BBA, BFA, BS)
   b) Some post undergraduate work
   c) Master’s degree (e.g., MA, MBA, MFA, MS, MSW)
   d) Specialist degree (e.g., EdS, SSP)
   e) Applied or professional doctorate degree (e.g., MD, DDC, DDS, JD, PsyD)
f) Doctorate degree (e.g., EdD, PhD)
g) Other: _______________

5) What degree are you currently seeking? Check all that apply to you:
   a) Master’s degree (e.g., MA, MBA, MFA, MS, MSW)
   b) Specialist degree (e.g., EdS, SSP)
   c) Doctorate degree (e.g., EdD, PhD)

6) What year are you in your graduate program?
   a) 1st year
   b) 2nd year
   c) 3rd year
   d) 4th year
   e) 5th year or higher

7) Approximately how many WISC-Vs have you scored while in your graduate program (this may include any protocols scored in coursework and your time in practicum, externship, internship, or other clinical experiences)? Please specify:
   ____________

8) Have you received any additional training or supervision related to WISC-V scoring/interpretation?
   a) Yes (e.g., additional training by Pearson, professional development/continued education hours, etc.)
   b) No

9) If you answered yes to the above question, please describe:
   ________________________________
10) In your clinical experiences (e.g., practicum, externship, internship, or other clinical experiences), is/was there a specially appointed person for providing administrative supervision for intelligence testing (e.g., double checking/scoring protocols)?
   a) Yes
   b) No

11) If you answered yes to the above question, please describe:

_______________________________

12) What strategies do you use to ensure adherence to standardized scoring rules for the WISC-V (e.g., double checking scores, using a ruler or straightedge when referencing tables, etc.) within your clinical experiences (e.g., practicum, externship, internship, or other clinical experiences)? Please describe:

_______________________________
APPENDIX B

LICENSED AND/OR CERTIFIED SCHOOL PSYCHOLOGIST DEMOGRAPHIC AND TRAINING VARIABLES QUESTIONNAIRE
Demographic and Training Variables Questionnaire

Directions: Please carefully check/fill-in your answers below.

1) How do you currently describe your gender identity?
   a) Male
   b) Female
   c) Transgender
   d) Other: _______________

2) What is your age in years?
   ______________

3) Which categories describe you? Check all that apply to you:
   a) American Indian or Alaska Native
   b) Asian or Pacific Islander
   c) Black
   d) Hispanic, Latino, or Spanish Origin
   e) Middle Eastern or North African
   f) Native Hawaiian or Other Pacific Islander
   g) White
   h) Other: _______________

4) What is the highest degree or level of education you have completed?
   a) Bachelor’s degree (e.g., BA, BBA, BFA, BS)
   b) Master’s degree (e.g., MA, MBA, MFA, MS, MSW)
   c) Specialist degree (e.g., EdS, SSP)
   d) Applied or professional doctorate degree (e.g., MD, DDC, DDS, JD, PsyD)
   e) Doctorate degree (e.g., EdD, PhD)
f) Other: _______________

5) In what year did you graduate from your school psychology graduate program?
_______________

6) Was your school psychology graduate program APA accredited and/or NASP approved at the time of your graduation?
a) Yes, NASP approved
b) Yes, APA accredited
c) Yes, APA accredited and NASP approved
d) No, neither APA accredited nor NASP approved

7) Which credentials/certifications do you hold? Check all that apply to you:
a) State credential
b) Nationally Certified School Psychologist (NCSP)
c) American Board of Professional Psychology (ABPP) Certified
d) Licensed for independent psychological practice (e.g., licensed psychologist)
e) Other: _______________

8) Approximately how many WISC-Vs do you score per year?
_______________

9) Have you received any additional training or supervision related to WISC-V scoring/interpretation?
a) Yes (e.g., additional training by Pearson, professional development/continued education hours, etc.)
b) No
10) If you answered yes to the above question, please describe:

_______________________________

11) Does your school district have a specially appointed person for providing administrative supervision for intelligence testing (e.g., double checking/scoring protocols)?
   a) Yes
   b) No

12) If you answered yes to the above question, please describe:

_______________________________

13) What strategies do you use to ensure adherence to standardized scoring rules for the WISC-V (e.g., double checking scores, using a ruler or straightedge when referencing tables, etc.)?  Please describe:

_______________________________