

1-1-2015

# Exploring the item difficulty and other psychometric properties of the core perceptual, verbal, and working memory subtests of the WAIS-IV using item response theory

Sara Ann Schleicher-Dilks

*Nova Southeastern University*, [ssdilks@gmail.com](mailto:ssdilks@gmail.com)

This document is a product of extensive research conducted at the Nova Southeastern University [College of Psychology](#). For more information on research and degree programs at the NSU College of Psychology, please [click here](#).

Follow this and additional works at: [http://nsuworks.nova.edu/cps\\_stuetd](http://nsuworks.nova.edu/cps_stuetd)



Part of the [Psychology Commons](#)

## Share Feedback About This Item

---

### NSUWorks Citation

Schleicher-Dilks, S. (2015). Exploring the item difficulty and other psychometric properties of the core perceptual, verbal, and working memory subtests of the WAIS-IV using item response theory. .

Available at: [http://nsuworks.nova.edu/cps\\_stuetd/87](http://nsuworks.nova.edu/cps_stuetd/87)

This Dissertation is brought to you by the College of Psychology at NSUWorks. It has been accepted for inclusion in College of Psychology Theses and Dissertations by an authorized administrator of NSUWorks. For more information, please contact [nsuworks@nova.edu](mailto:nsuworks@nova.edu).

**EXPLORING THE ITEM DIFFICULTY AND OTHER PSYCHOMETRIC  
PROPERTIES OF THE CORE PERCEPTUAL, VERBAL, AND WORKING  
MEMORY SUBTESTS OF THE WAIS-IV USING ITEM RESPONSE THEORY**

**By:**

**Sara Ann Schleicher-Dilks, M.S.**

A Dissertation Presented to the Center for Psychological Studies  
of Nova Southeastern University  
in Partial Fulfillment of the Requirements  
for the Degree of Doctor of Philosophy

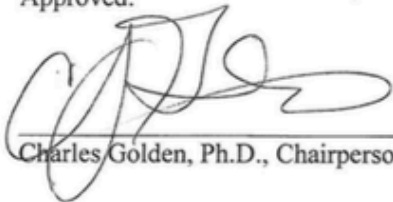
NOVA SOUTHEASTERN UNIVERSITY


2014

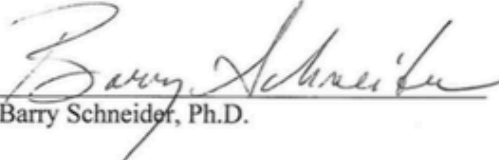
**APPROVAL**

This dissertation was submitted by Sara Ann Schleicher-Dilks under the direction of the Chairperson of the dissertation committee listed below. It was submitted to the Center for Psychological Studies and approved in partial fulfillment of the requirements for the Degree of Philosophy in Clinical Psychology at Nova Southeastern University.

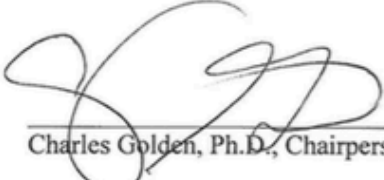
12/11/14  
Date of Defense

Approved:  
  
Charles Golden, Ph.D., Chairperson

  
Ryan Black, Ph.D.

  
Barry Schneider, Ph.D.

1/5/15  
Date of Final Approval

  
Charles Golden, Ph.D., Chairperson

### **Acknowledgments**

Though a dissertation has only one official author, no dissertation is a solitary work. Those people who helped shape and support this writer and those who helped refine and focus the writing were essential to the completion of this project.

To my family. Your love, encouragement, patience and generosity made my educational pursuits possible. Thank you Mom, Dad, Linda, Aunt Pat and Aunt Sharon.

To my friends. You acted as a baseline, cheerleading squad, sounding boards, advice columnists, shoulders to lean on, stand up comics, cat sitters, chauffeurs and pals to crack up with. Thank you Marcy, Jessica, Traci, Rachel, Rebecca, Johanna, Jada and Sean. Sean, thank you also for your help with the databases.

To my committee chair, Dr. Golden. Thank you for challenging me to do more than I thought possible.

To Dr. Black. Thank you for the many math lessons, theoretical discussions, and for introducing me to item response theory.

To all my committee members, Drs. Golden, Black and Schneider. Thank you for shaping the focus of this project through your mentorship and coursework.

### Table of Content

LIST OF TABLES.....	5
LIST OF FIGURES.....	8
ABSTRACT.....	9
CHAPTER I: STATEMENT OF THE PROBLEM.....	13
CHAPTER II: REVIEW OF THE LITERATURE.....	16
WAIS-IV Subtest Development.....	20
Influence of Classical Test Theory (CTT).....	33
Item Response Theory (IRT) and the Standard Dichotomous Rasch Model...	36
Current Research.....	40
CHAPTER III: METHODS.....	42
Participants.....	42
Measures.....	43
Procedure.....	49
Hypotheses.....	57
CHAPTER IV: RESULTS.....	59
Preliminary Analysis.....	59
Hypothesis 1.....	60
Hypothesis 2.....	129
CHAPTER V: DISCUSSION.....	162
Ceiling Rules.....	162
Basal Rules.....	170
Limitations.....	178
Future Studies.....	179
Conclusions.....	180
REFERENCES.....	184

**List of Tables**

Table 1. *Descriptive Statistics of WAIS-IV Subtests*..... 59

Table 2. *Block Design Andrich Thresholds*..... 63

Table 3. *Andrich Thresholds After Combining Response Categories*..... 64

Table 4. *Observed Averages After Combining Items' Response Categories*..... 65

Table 5. *Sample Expectation After Combining Items' Response Categories*..... 66

Table 6. *Block Design Item Fit Statistics*..... 67

Table 7. *Dimensionality of the Block Design Subtest*..... 68

Table 8. *Block Design Item Difficulty*..... 75

Table 9. *Similarities Items' Andrich Thresholds*..... 79

Table 10. *Similarities Andrich Thresholds After Combining Items' Response Categories* ..... 80

Table 11. *Similarities Average & Expected Ability Levels After Combining Response Categories*..... 81

Table 12. *Similarities Item Fit Statistics with Misfitting Item Removed*..... 83

Table 13. *Dimensionality of the Similarities Subtest*..... 85

Table 14. *Similarities Item Difficulty*..... 92

Table 15. *Vocabulary Items' Andrich Thresholds After Combining Select Items' Response Categories*..... 94

Table 16. *Vocabulary Observed Averages and Sample Expectations*..... 95

Table 17. *Vocabulary Item Fit Statistics*..... 96

Table 18. *Dimensionality of the Vocabulary Subtest*..... 96

Table 19. *Vocabulary Item Difficulty*..... 101

Table 20. <i>Digits Forward Items' Andrich Thresholds</i> .....	104
Table 21. <i>Average &amp; Expected Ability Levels for the Digits Forward Task</i> .....	105
Table 22. <i>Digits Forward Item Fit Statistics with Misfitting Items Removed</i> .....	107
Table 23. <i>Dimensionality of the Digits Forward Task</i> .....	109
Table 24. <i>Digits Forward Item Difficulty</i> .....	114
Table 25. <i>Digits Backward Items' Andrich Thresholds</i> .....	116
Table 26. <i>Average &amp; Expected Ability Levels for the Digits Backward Task</i> .....	116
Table 27. <i>Digits Backward Item Fit Statistics with Misfitting Items Removed</i> ...	117
Table 28. <i>Dimensionality of the Digits Backward Task</i> .....	118
Table 29. <i>Digits Backward Item Difficulty</i> .....	122
Table 30. <i>Digit Sequencing Items' Andrich Thresholds</i> .....	123
Table 31. <i>Average &amp; Expected Ability Levels for the Digit Sequencing Task</i> ...	124
Table 32. <i>Digit Sequencing Item Fit Statistics</i> .....	124
Table 33. <i>Dimensionality of the Digit Sequencing Task</i> .....	125
Table 34. <i>Digit Sequencing Item Difficulty</i> .....	129
Table 35. <i>Matrix Reasoning Observed Averages and Sample Expectations</i> .....	130
Table 36. <i>Matrix Reasoning Item Fit Statistics</i> .....	131
Table 37. <i>Dimensionality of the Matrix Reasoning Subtest</i> .....	132
Table 38. <i>Matrix Reasoning Item Difficulty</i> .....	142
Table 39. <i>Visual Puzzles Observed Averages and Sample Expectations</i> .....	138
Table 40. <i>Visual Puzzles Item Fit Statistics and Misfitting Item Removed</i> .....	139
Table 41. <i>Dimensionality of the Visual Puzzles Subtest</i> .....	140

Table 42. <i>Visual Puzzles Item Difficulty</i> .....	144
Table 43. <i>Information Observed Averages and Sample Expectations</i> .....	146
Table 44. <i>Item Fit Statistics for Information</i> .....	147
Table 45. <i>Dimensionality of the Information Subtest</i> .....	149
Table 46. <i>Item Difficulty of the Information Subtest</i> .....	153
Table 47. <i>Arithmetic Observed Averages and Sample Expectations</i> .....	154
Table 48. <i>Arithmetic Item Fit Statistics with Misfitting Item Removed</i> .....	155
Table 49. <i>Dimensionality of the Arithmetic Subtest</i> .....	156
Table 50. <i>Arithmetic Item Difficulty</i> .....	160



**List of Figures**

Figure 1. <i>Block Design Item Coverage</i> .....	73
Figure 2. <i>Similarities Item Coverage</i> .....	89
Figure 3. <i>Vocabulary Item Coverage</i> .....	99
Figure 4. <i>Digits Forward Item Coverage</i> .....	112
Figure 5. <i>Digits Backward Item Coverage</i> .....	120
Figure 6. <i>Digit Sequencing Item Coverage</i> .....	127
Figure 7. <i>Matrix Reasoning Item Coverage</i> .....	134
Figure 8. <i>Visual Puzzles Item Coverage</i> .....	142
Figure 9. <i>Information Item Coverage</i> .....	153
Figure 10. <i>Arithmetic Item Coverage</i> .....	158

**EXPLORING THE ITEM DIFFICULTY AND OTHER PSYCHOMETRIC PROPERTIES OF THE CORE PERCEPTUAL, VERBAL, AND WORKING MEMORY SUBTESTS OF THE WAIS-IV USING ITEM RESPONSE THEORY**

by

**Sara Ann Schleicher-Dilks, M.S.**

Nova Southeastern University

**ABSTRACT**

The ceiling and basal rules of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV; Wechsler, 2008) only function as intended if subtest items proceed in order of difficulty. While many aspects of the WAIS-IV have been researched, there is no literature about subtest item difficulty and precise item difficulty values are not available.

The WAIS-IV was developed within the framework of Classical Test Theory (CTT) and item difficulty was most often determined using  $p$ -values. One limitation of this method is that item difficulty values are sample dependent. Both standard error of measurement, an important indicator of reliability, and  $p$ -values change when the sample changes.

A different framework within which psychological tests can be created, analyzed and refined is called Item Response Theory (IRT). IRT places items and person ability onto the same scale using linear transformations and links item difficulty level to person ability. As a result, IRT is said to produce sample-independent statistics.

Rasch modeling, a form of IRT, is one parameter logistic model that is appropriate for items with only two response options and assumes that the only factors affecting test performance are characteristics of items, such as their difficulty level or their relationship to the construct being measured by the test, and characteristics of

participants, such as their ability levels. The partial credit model is similar to the standard dichotomous Rasch model, except that it is appropriate for items with more than two response options.

Proponents of standard dichotomous Rasch model argue that it has distinct advantages above both CTT-based methods as well as other IRT models (Bond & Fox, 2007; Embretson & Reise, 2000; Furr & Bacharach, 2013; Hambleton & Jones, 1993) because of the principle of monotonicity, also referred to as specific objectivity, the principle of additivity or double cancellation, which “establishes that two parameters are additively related to a third variable” (Embretson & Reise, 2000, p. 148). In other words, because of the principle of monotonicity, in Rasch modeling, probability of correctly answering an item is the additive function of individuals’ ability, or trait level, and the item’s degree of difficulty. As ability increases, so does an individual’s probability of answering that item. Because only item difficulty and person ability affect an individual’s chance of correctly answering an item, inter-individual comparisons can be made even if individuals did not receive identical items or items of the same difficulty level. This is why Rasch modeling is referred to as a *test-free measurement*.

The purpose of this study was to apply a standard dichotomous Rasch model or partial credit model to the individual items of seven core perceptual, verbal and working memory subtests of the WAIS-IV: Block Design, Matrix Reasoning, Visual Puzzles, Similarities, Vocabulary, Information, Arithmetic Digits Forward, Digits Backward and Digit Sequencing.

Results revealed that WAIS-IV subtests fall into one of three categories: optimally ordered, near optimally ordered and sub-optimally ordered. Optimally ordered

subtests, Digits Forward and Digits Backward, had no disordered items. Near optimally ordered subtests were those with one to three disordered items and included Digit Sequencing, Arithmetic, Similarities and Block Design. Sub-optimally ordered subtests consisted of Matrix Reasoning, Visual Puzzles, Information and Vocabulary, with the number of disordered items ranging from six to 16.

Two major implications of the result of this study were considered: the impact on individuals' scores and the impact on overall test administration time. While the number of disordered items ranged from 0 to 16, the overall impact on raw scores was deemed minimal. Because of where the disordered items occur in the subtest, most individuals are administered all the items that they would be expected to answer correctly. A one-point reduction in any one subtest is unlikely to significantly affect overall index scores, which are the scores most commonly interpreted in the WAIS-IV. However, if an individual received a one-point reduction across all subtests, this may have a more noticeable impact on index scores. In cases where individuals discontinue before having a chance to answer items that were easier, clinicians may consider testing the limits. While this would have no impact on raw scores, it may provide clinicians with a better understanding of individuals' true abilities. Based on the findings of this study, clinicians may consider administering only certain items in order to test the limits, based on the items' difficulty value.

This study found that the start point for most subtests is too easy for most individuals. For some subtests, most individuals may be administered more than 10 items that are too easy for them. Other than increasing overall administration time, it is not clear what impact, of any, this has. However, it does suggest the need to reevaluate

current start items so that they are the true basal for most people.

Future studies should break standard test administration by ignoring basal and ceiling rules to collect data on more items.

In order to help clarify why some items are more or less difficult than would be expected given their ordinal rank, future studies should include a qualitative aspect, where, after each subtest, individuals are asked describe what they found easy and difficult about each item. Finally, future research should examine the effects of item ordering on participant performance. While this study revealed that only minimal reductions in index scores likely result from the prematurely stopping test administration, it is not known if disordering has other impacts on performance, perhaps by increasing or decreasing an individual's confidence.

**Keywords:** WAIS-IV, Wechsler Adult Intelligence Scale, item difficulty, Rasch model, item response theory, partial credit model, psychometric properties

## CHAPTER I

### Statement of the Problem

The ceiling and basal rules of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV; Wechsler, 2008a) only function as intended if subtest items proceed in order of difficulty. If items are not ordered hierarchically, assumptions about individuals' performance on items not administered because the discontinue rule had been met, for example, could not be made. Individuals may indeed be able to correctly answer items after the ceiling rule if those items were easier than ones they had already been administered. Despite their importance, however, precise item difficulty values are not known. These values are not published in the Technical and Interpretive Manual of the WAIS-IV and a review of the literature yielded no results about the difficulty of WAIS-IV items.

The statistical procedures used to determine item difficulty may decrease the likelihood that all items are ordered according to difficulty. The WAIS-IV was developed within the framework of Classical Test Theory (CTT) and item difficulty was most often determined using  $p$ -values (J.J. Zhu, personal communication, July 10, 2013). One limitation of this method is that item difficulty values are sample dependent. Both standard error of measurement, an important indicator of reliability, and  $p$ -values change when the sample changes (Embretson & Reise, 2000). This makes it difficult to generalize item difficulty values to different samples. Additionally, in CTT, items are assumed to exist upon an interval scale because they are normally distributed. Actual empirical data to support this assertion is lacking. Finally, standard scores for many tests are normalized using either percentile matching or nonlinear transformations, which can

change the distance between items and create floor and ceiling effects (Embretson & Reise, 2000).

A different framework within which psychological tests can be created, analyzed and refined is called Item Response Theory (IRT). IRT places items and person ability onto the same scale using linear transformations and links item difficulty level to person ability. As a result, IRT is said to produce sample-independent statistics (Crocker & Algina, 2008; Hambleton & Jones, 1993).

The standard dichotomous Rasch model, a form of IRT, is one parameter logistic model that assumes that the only factors affecting test performance are characteristics of items, such as their difficulty level or their relationship to the construct being measured by the test, and characteristics of participants, such as their ability levels. When items have two or more possible response options and different thresholds, test performance is

mathematically represented as  $\frac{e^{\sum_{k=0}^x (\theta_j - \tau_{ki})}}{\sum_{x=0}^m e^{\sum_{k=0}^x (\theta_j - \tau_{ki})}}$  where  $m$  is the maximum score for the

item,  $\theta_j$  is the ability level of an individual,  $\beta_i$  is the difficulty level of the item and  $\tau_{ki}$  is the threshold of the rating scale of the item (Bond & Fox, 2007). Using the logistic Rasch model, item difficulty values and person ability levels are calculated and reported in logits. A logit scale arguably converts item difficulty and person ability levels to interval level scores and places both of these values on the same continuum.

Similar to the standard dichotomous Rasch model, a partial credit model can be used to examine items with more than two answer options. In the partial credit model, the probability of an individual correctly answering an item is mathematically represented as

$\frac{e^{\sum_{k=0}^x (\theta_j - \tau_{ki})}}{\sum_{x=0}^m e^{\sum_{k=0}^x (\theta_j - \tau_{ki})}}$  where  $m$  is the maximum score for the item,  $\theta_j$  is the ability level of

an individual,  $\beta_i$  is the difficulty level of the item and  $\tau_{ki}$  is the threshold of the rating scale of the item (Bond & Fox, 2007; Embretson and Reise, 2000).

The purpose of this study was to apply a standard dichotomous Rasch model or partial credit model to the individual items of seven core perceptual, verbal and working memory subtests of the WAIS-IV: Block Design, Matrix Reasoning, Visual Puzzles, Similarities, Vocabulary, Information and Arithmetic as well as Digit Sequencing from the Digit Span subtest. The core processing speed subtests were not included in this study because speeded tests where item difficulty is not designed to increase as the task progresses are not appropriate for IRT (Furr & Bacharach, 2013). The goal of this study was to achieve an interval level scale, and in so doing, use a Rasch model to evaluate the extent to which items are ordered along a hierarchy.



## CHAPTER II

### Review of the Literature

The WAIS-IV is based upon previous versions, giving the test a psychometric, theoretical, developmental and clinical history dating back to 1939 when David Wechsler first introduced the Wechsler-Bellevue Intelligence Scale (Wechsler, 1939). Wechsler noted that most adult intelligence test had been developed for children, making both the tasks and the norms inadequate for use with adults (Wechsler, 1955b). In order to fulfill a perceived need for adult measures of intelligence, Wechsler created the Wechsler-Bellevue Intelligence Scale based on his conceptualization of intelligence (Wechsler, 1944). Wechsler viewed intelligence as a multi-faceted construct comprised of logical behavior and reasoning as well as personality traits such as motivation. Additionally, in order to be considered intelligent, this behavior or reasoning must be purposefully and knowingly applied to a goal (Wechsler, 1975). Thus, Wechsler's first intelligence scale included a variety of tasks by measuring logical behavior and reasoning using 10 subtests, each with a distinct and overt goal. The 10 subtests consisted of Information, Similarities, Comprehension, Block Design, Picture Completion, Picture Arrangement, Object Assembly, Digit Span, Digit-Symbol Test and Arithmetic (Wechsler, 1944). The test was intended for individuals between the ages of 16 and 60.

To create the 10 original subtests, Wechsler used items from existing measures, such as the Army testing program, games, such as a children's toy called color cubes, and social media, such as a cartoon strip from a popular magazine (Tulsky, Saklofske, & Zhu, 2003). The subtests that comprised Wechsler's first intelligence scale, the Wechsler-Bellevue Intelligence Scale, were chosen after reviewing current measures being used at

the time, conducting studies on criterion validity of subtests Wechsler considered including, conducting testing trials with the completed intelligence test, and reviewing opinions of clinicians who had used to measure (Wechsler, 1944). Vocabulary was not included as a subtest because, although Wechsler recognized the relationship between vocabulary and intelligence, he had concerns that the test would be biased by individuals' educational opportunities and cultural differences (Wechsler, 1939a). As a result, although this subtest was included in the test, it was not mandatory.

Shortly after the development of the Wechsler-Bellevue Intelligence Scale, Wechsler developed another intelligence test, the Wechsler Mental Ability Scale, Form B, for the United States Army. In 1946, following World War II, the test was renamed the Wechsler-Bellevue Intelligence Scale, Form II (Wechsler, 1946), the Vocabulary subtest was added to the core battery and the test was made available for use by clinicians in the general population (Tulsky, et al., 2003). It was the first version of the Wechsler-Bellevue Intelligence Scales to include Vocabulary as a core subtest that was necessary to calculate full-scale intelligence (FSIQ) because Wechsler realized its relationship to intelligence may be more important than its potential bias (Tulsky, et al., 2003). The test contained 11 core subtests that were considered counterparts to the original test, with the exception of the additional required Vocabulary subtest, and took approximately 75 minutes to administer (Wechsler, 1946). Little other information is available on the Wechsler-Bellevue Intelligence Scale, Form II.

In 1955, Wechsler created the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955b) as a revision to the Wechsler-Bellevue Intelligence Scale. The WAIS was created in response to several practical and statistical difficulties, including restricted

range and vague items on the Wechsler-Bellevue Intelligence Scale. The WAIS included an increased upper age limit of 75, compared to the 60-year age limit of the Wechsler Bellevue Intelligence Scale. Also of note is that the WAIS had discontinuation but not reversal rules. It contained the 11 core subtests of Information, Comprehension, Arithmetic, Similarities, Digit Span, Digit Symbol, Picture Completion, Block Design, Picture Arrangement and Object Assembly and each was classified as verbal or performance, creating Full Scale IQ (FSIQ), Verbal IQ (VIQ) and Performance IQ (PIQ) scores (1955b).

The Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981) was a revision to the WAIS that was published in 1981, after Wechsler's death, though he was still cited as the author (Sattler & Ryan, 2009). It contained approximately 2/3 of the items from the Wechsler-Bellevue Intelligence Scale (McNemar, 1956). The WAIS-R contained 11 core subtests classified as either verbal or performance tasks. Information, Comprehension, Arithmetic, Digit Span, Similarities, and Vocabulary comprised the verbal subtests. The perceptual subtests included Picture Arrangement, Picture Completion, Block Design, Object Assembly, and Digit Symbol Coding. The three indices, FSIQ, VIQ, and PIQ, remained the same (Wechsler, 2008b).

In 1997, The Psychological Corporation published another updated version of the Wechsler Adult Intelligence Scale, the WAIS-III (Wechsler, 1997b), which was comprised of 11 core subtests, which were the same as the WAIS-R except for the addition of Matrix Reasoning and the removal of Object Assembly as core subtests. In addition to three primary indices, Verbal IQ (VIQ), Performance IQ (PIQ) and Full Scale IQ (FSIQ), four additional indices were also created: Verbal Comprehension, Perceptual

Organization, Working Memory, and Processing Speed (Sattler & Ryan, 2009).

Additionally, the WAIS-III subtests contained more items than the WAIS-R. Because of the newly created basal rules, however, many individuals did not receive the additional items because they did not need to be administered items preceding the start point. As a result, the overall administration time for the WAIS-III was consistent with the administration time of the WAIS-R (Wechsler, 1997b) of approximately 80 minutes (Wechsler, 2008c).

The next and most recent revision, the WAIS-IV (Wechsler, 2008), took four years to develop and finalize (Climie & Rostad, 2011) and was published in 2008. This current version contains 10 core subtests and takes an average of 67 minutes to administer (D. Wechsler, 2008a). The core subtests of the WAIS-IV consist of three perceptual subtests, Block Design, Matrix Reasoning, and Visual Puzzles, three core verbal subtests, Vocabulary, Similarities, and Information, two core working memory tasks, Arithmetic and Digit Span, and two core processing speed subtests, Coding and Symbol Search. Approximately 47% of the items on the WAIS-IV were taken from the WAIS-III (Sattler & Ryan, 2009). Consistent with the WAIS-III, the WAIS-IV subtests have a basal of 2. When individuals do not receive full credit on the first two items administered, items are administered in reverse until two perfect, consecutive scores are obtained.

Differences between the WAIS-III and the WAIS-IV include the addition of Visual Puzzles as a core perceptual subtest and the addition of Symbol Search as a core processing speed subtest. Picture Completion was removed as a core performance subtest and Comprehension was removed as a core verbal subtest. Both subtests were retained as

optional. Lastly, Picture Arrangement and Object Assembly were not included in the WAIS-IV. In addition to being essential in calculating FSIQ, the core subtests combine by type to create four other indices: Verbal Comprehension, Perceptual Reasoning, Processing Speed and Working Memory. Changes to the WAIS-IV, including the way the test subtests were combined to create overall index scores, were due largely to attempts by test developers to align the WAIS-IV with more current conceptualizations of intelligence as a four-factor construct (Benson, Hulac, & Kranzler, 2010).

### *WAIS-IV Subtest Development*

#### *Block Design*

The origins of the Block Design subtest can be dated back to 1911 when a children's toy was modified and included in testing batteries as a performance based task. In 1914, Frances M. Maxfield began using the blocks in a test at the University of Pennsylvania where he assembled a model and then asked individuals to reproduce it. This type of task quickly became popular and a similar Cube Construction test was created for the Army Beta Tests. Several years later in 1923, as part of his dissertation, Samuel C. Kohs replaced the examiner's model with a stimulus book of the designs individuals needed to construct (Tulsky, et al., 2003). In developing the Wechsler-Bellevue Intelligence Scale, Wechsler chose seven items from Kohs' Block Design task. While the cubes Wechsler used were multi-colored, only stimulus designs that were red and white were included in the Block Design subtest of Wechsler's original intelligence scale (Tulsky, et al., 2003; Wechsler, 1944).

The Block Design subtest on Wechsler-Bellevue Intelligence Scale included four items using four blocks, two items using nine blocks and one item using 16 blocks. Item

time limits ranged from 75 seconds to 195 seconds and whole points between 0 and 3 were awarded based on design accuracy and completion time. The Wechsler Adult Intelligence Scale contained 10 items, six of which contained designs with four blocks and four of which contained nine blocks. Items were worth 0, 4, 5 or 6 points, with values of 5 or 6 being given for quickly completed accurate designs. The discontinue rule was three consecutive scores of 0 (Wechsler, 1955b).

The WAIS-R contained nine Block Design items, seven of which were retained from the WAIS and two of which were newly created. The first five items contained designs requiring four blocks while the last five items contained designs that needed to be made with nine blocks, which was a slight alteration from the previous version of the test. The WAIS-R was also the first version of the test to include blocks with only red and white coloring. The discontinue rule remained the same, three consecutive scores of 0 (Wechsler, 1981).

When the items for Block Design for the WAIS-III were created, all nine items from the WAIS-R were retained and five new items were created, including reversal items that contained designs made with only two blocks. This subtest also contains 30, 60, or 120-second time limits, depending on the item. Reversal items and the first core subtest item are worth 0, 1 or 2 points and permit participants two trials to complete the task, though a correct attempt on a second trial results in a score of 1. The other nine items are worth 0, 4, 5, 6, or 7 points depending upon design accuracy and completion time. Item 4 is the recommended starting point for most adults. The discontinue rule of three consecutive scores of 0 remained consistent with the WAIS-R (Wechsler 1997a).

No literature was available on the difficulty level of the items on the Block Design subtest of the WAIS-III.

Block Design on the WAIS-IV is comprised of 14, ten of which were retained from the WAIS-III. One change from the WAIS-III was the inclusion of fewer items that award bonus points for items completed quickly. The WAIS-IV contains six of these items while the WAIS-III contained nine. In addition, the number of items with two trials permitted was shortened from six on the WAIS-III to five on the WAIS-IV (Wechsler, 2008c). Fewer items requiring nine blocks were included in the WAIS-IV compared to the WAIS-III, thus resulting in fewer items with a 120 second time limit and fewer items where a maximum of seven points are available. No literature on the difficulty level of the items on the Block Design subtest of the WAIS-IV was available.

### *Matrix Reasoning*

The Matrix Reasoning subtest was added as a subtest to the WAIS-III to replace the Object Assembly subtest used in previous versions of the Wechsler Intelligence Scales. Wechsler had long recognized the difficulties of the Object Assembly subtest, including a restricted range and the effects of learning which drastically altered retest performance (Tulsky, et al., 2003; Wechsler, 1939a). The Matrix Reasoning subtest was chosen because it was believed to be a relatively culture-free test without a motoric component that is a fairly good measure of fluid intelligence (Tulsky, et al., 2003; Wechsler 1997a). It was based on similar already existing matrix problem solving and serial reasoning tasks (Tulsky, et al., 2003). On the WAIS-III, this subtest contained four types of items, pattern completion, classification, analogy reasoning and serial reasoning (Tulsky et al., 2003; Wechsler 1997a).

The Matrix Reasoning subtest of the WAIS-III contained 26 items, including three sample items and three reversal items. Reversal and core subtest items were worth either 0 or 1 point. The discontinue rule was four consecutive scores of 0 or four scores of 0 on five consecutive items. It is not clear why this discontinue rule permits discontinuation after four scores of 0 on five consecutive items, something that is not done on any other subtest of the WAIS-III (Wechsler 1997a). No literature is available on the item difficulty of the Matrix Reasoning subtest of the WAIS-III.

According to the WAIS-IV technical and interpretive manual, the type of tasks on the Matrix Reasoning subtest was decreased from four to two to allow for sufficient learning and teaching. The number of sample items was decreased from three to two. Though the WAIS-IV technical and interpretive manual does not say why the number of sample items was decreased, it may be because there are only two types of problems, classification and analogy reasoning tasks, rather than the four types that were included on the WAIS-III. Twelve items were retained and fourteen new items were added to create a total of 26 items with possible scores of 0 or 1 (Wechsler, 2008b; Wechsler, 2008c). The discontinue rule was changed to three consecutive scores of 0 (Wechsler, 2008c), though it is unclear why. No literature is available on the item difficulty of the Matrix Reasoning subtest of the WAIS-IV.

### *Visual Puzzles*

The Visual Puzzles subtest contains 26 items that were created for the WAIS-IV, as it is a new subtest. It is thought to assess nonverbal fluid reasoning, mental transformations, analysis and synthesis, spatial ability, visual-perceptual discrimination and speed of visual-perceptual processing (Sattler, 2009). There is a demonstration and a



sample item, four reversal items and 22 core items. The reversal and first three core items have a time limit of 20 seconds while the remaining 19 subtests have a 30-second time limit. All reversal and core items are worth either 0 or 1 point. The discontinue rule is three consecutive scores of 0 (Wechsler, 2008b; Wechsler, 2008c). There is no literature available on the item difficulty of the Visual Puzzles subtest.

### *Similarities*

It is thought that Wechsler was familiar with the idea of an analogy-based test through his work with instruments that included such tasks during his work in the military (Tulsky, et al., 2003). Some of the original items on the Wechsler-Bellevue Intelligence Scale were taken directly from the Stanford-Binet Intelligence Test (Terman, 1916) and other similar tests (Tulsky, et al., 2003), though Wechsler created many items, as he found most current versions too difficult (Wechsler, 1944). The Wechsler-Bellevue Intelligence Scale contained 12 Similarities items worth 0, 1, or 2 points (Wechsler, 1944), while the WAIS contained 13 Similarities items with those point values (1955b) and the WAIS-R contained 14 items worth 0, 1, or 2 points (Wechsler 1997a).

The WAIS-III contains 11 of the same items from the WAIS-R, although two of these 11 items were made reversal items since they were answered correctly by nearly everyone except individuals diagnosed with mental retardation (Wechsler 1997a). This was the first time reversal items were included in the Similarities subtest. The WAIS-III technical manual states that three WAIS-R items were dropped due to poor psychometric characteristics or item bias. No further explanation is provided and no literature was available on the difficulty level of the items on the Similarities subtest of the WAIS-III. Eight new items were created for the WAIS-III, for a total of 19 items, five of which are

reversal items. While the reversal items are worth 0 or 1 point, the start point and proceeding items are worth 0, 1, or 2 points. Scoring criteria for the Similarities subtest of the WAIS-III was revised (Wechsler 1997a). Little detail is provided in the WAIS-IV technical manual about how these point values were determined and validated, except that "results were then subjected to psychometric analysis and clinical and bias expert review" (p. 30).

The WAIS-IV contains a total of 18 items, 12 of which are new and six of which were retained from the WAIS-III. A sample item and reversal items that permit corrective feedback were also added (Wechsler, 2008c), though it is unclear why these changes were made. The scoring of the items on Similarities on the WAIS-IV was changed so that all items are worth 0, 1, or 2 points (Wechsler, 2008a), though it is not clear why this change was made. The scoring criteria for the items was again revised (Wechsler, 2008c). Little detail is provided in the WAIS-IV technical manual about how these point values were revised, except that "results were then subjected to psychometric analysis and clinical and bias expert review" (p. 30). No literature was available on the difficulty level of the items on the Similarities subtest of the WAIS-IV.

### *Vocabulary*

Some researchers propose that the development of the Vocabulary subtest was influenced by Wechsler's time as military examiner, when he used a similar subtest while using other measures to assess intelligence (Tulsky, et al., 2003). The original Wechsler-Bellevue Intelligence Scale contained an optional Vocabulary subtest that included 42 items worth 0, 1, or ½ points. These items were chosen after Wechsler randomly selected approximately 100 words from a dictionary, tested them using a pilot study, and retained

the ones that could differentiate individuals into two groups (Tulsky, et al., 2003). One of the only mentions of item difficulty of this subtest is made by Wechsler in his book on intelligence, when he stated that the order of item difficulty was consistent for individuals in the New York City area, although all individuals tended to think some items were easier and some were harder (Wechsler, 1944, p. 99).

Although it is unclear why, none of the items from the optional Vocabulary subtest of the Wechsler-Bellevue Intelligence Scale were retained for the WAIS. It is also unclear how the words were chosen and validated, but Wechsler claimed that the items were of approximately equivalent difficulty as those on the original test (Wechsler, 1958). The WAIS contained 40 items worth 0, 1, or 2 points and introduced a stimulus card with the words individuals needed to define. Other than these changes, the administration and scoring processes have remained consistent through the current WAIS-IV (Tulsky, et al., 2003). The WAIS-R Vocabulary subtest introduced two additional items and removed seven of the items from the WAIS-R, for a total of 35 items (Tulsky, et al., 2003; Wechsler 1997a).

Twenty of the 40 items from the WAIS and the two additional WAIS-R items were retained and nine new items were added to create the Vocabulary subtest for the WAIS-III, for a total of 33 items, three of which are less frequently administered because they are reversal items (Tulsky, et al., 2003; Wechsler 1997a). All items are worth 0, 1, or 2 points (Wechsler 1997a). No literature was available on the difficulty level of the items on the Vocabulary subtest of the WAIS-III. The discontinue rule for this subtest for the WAIS-III was six consecutive scores of 0 (Wechsler 1997a). It is unclear how this number was determined.

The WAIS-IV contains 30 Vocabulary items, 21 of which were retained from the WAIS-III and six of which are new. The WAIS-IV technical manual does not state why only 21 of the WAIS-III Vocabulary items were retained. No literature was available on the difficulty level of the items on the Vocabulary subtest of the WAIS-IV. Of the 30 items, four are reversal items, three of which are picture items, in which individuals provide the name of a pictured object rather than the definition of a word (Wechsler, 2008b). The scoring on the WAIS-IV subtest was changed so that reversal items are worth 0 or 1 point and core items are worth 0, 1, or 2 points (Wechsler, 2008a), though it is unclear why this change was made. The discontinue rule was changed from six consecutive scores of 0 to three consecutive scores of 0 to reduce overall administration time (Wechsler, 2008b).

### *Information*

The development of the Information subtest for the Wechsler-Bellevue Intelligence Scale was thought to be influenced by a range of information test (Whipple, 1909) created by Guy Montrose Whipple that included 100 words from varying fields of knowledge such as history, golf and French, about which participants were asked questions (Frank, 1983). This test was later shortened, refined for group administration, and included in test batteries used by groups such as the Army Alpha testing program, which is thought to be where Wechsler was first exposed to a test of this kind. The main change Wechsler made when he adapted the test for individual usage in the Wechsler-Bellevue Intelligence Scale was removing multiple choice options and requiring individuals to produce their own answers (Tulsky, et al., 2003). Seventy-five items were tested and revised, primarily upon their ability to discriminate between intelligent and

less intelligent individuals, to 25 for the Information subtest of the Wechsler-Bellevue Intelligence Scale (Wechsler, 1944). Answers were worth 0, 1, or 2 points. The WAIS and WAIS-R Information subtests contained 29 items (Norman & Wilensky, 1961; Wechsler 1997a).

One study identified several items on the WAIS's Information subtest that were more difficult for schizophrenics compared to healthy individuals as well as several items that were easier for schizophrenics (Norman & Wilensky, 1961). The items with which schizophrenic individuals struggled were found to accurately differentiate individuals into either a schizophrenic group or a healthy control group. Another study that compared the percentage of individuals who correctly answered items found that several items were biased towards or against Canadians (Bornstein, McLeod, McClung, & Hutchison, 1983). Specifically, the study found that American political and history questions (e.g., How many United States senators are there? Who was Louis Armstrong?) were more difficult for Canadians while several literature and physics questions (e.g., Who wrote Hamlet? What is the boiling point of water?) were easier. This was the only literature available about item difficulty of the Information subtests of the WAIS and WAIS-R.

The WAIS-III Information subtests contained 28 items, 19 of which were retained from the WAIS-R. Of the total 28 total items, four were reversal items. According to the WAIS-III administration and scoring manual, the items not retained from the WAIS-R were removed because they had become outdated, though some researchers note that several items were found to be biased (Tulsky, et al., 2003). No mention of statistical analyses to determine the item's validity can be found in the technical manual or existing

literature. Each item on the Information subtest of the WAIS-III was worth either 0 or 1 point. As with other verbal subtests, scoring criteria was revised. An additional change to the WAIS-III Information subtest was the provision of additional sample responses to aid clinicians in scoring answers (Wechsler 1997a). No literature was available on the difficulty level of the items on the Information subtest of the WAIS-III.

The Information subtest on the WAIS-IV was revised to contain 15 items from the WAIS-III. Additionally, 11 new items were created for a total of 26 items, two of which are reversal items. While responses remained worth either 0 or 1 point, scoring criteria were revised for the WAIS-IV (Wechsler, 2008b). No literature was available on the difficulty level of the items on the Information subtest of the WAIS-IV.

### *Digit Span*

The Digit Span subtest dates back to Sir Francis Galton in the 1800s and was commonly used by psychologists around the time the Wechsler-Bellevue Intelligence Scale was published in 1939 from existing versions of the subtest being employed by psychologists (Tulsky, et al., 2003). Wechsler was the first to combine the forward and backward tasks into a single test (Tulsky, et al., 2003), which he did for two main reasons: (a) alone, each test had a limited range of items; and (b) to de-emphasize memory as an intelligence factor (Wechsler, 1944). Wechsler had concerns about even including this subtest in his test because of the similar tasks had previously been shown to have little correlation with intelligence. The task was included because Wechsler believed task had clinical significance, including possibly identifying those at the extreme low end of intelligence (Wechsler, 1944).

The Wechsler-Bellevue Intelligence Scale contained seven items, each with one trial per item for digits forward and digits backward. The number of digits ranged from two to nine. Seven items for digits forward and backward from the WAIS-R were retained for the WAIS-III (Wechsler 1997a). Each trial was worth 0 or 1-point for a maximum of 2-points per item. Each subtest was discontinued when a participant receives a score of 0 on both trials of an item. No literature was available on the difficulty level of the items on the Digit Span subtest of the WAIS-III.

The WAIS-IV is the first version of the test to include Digit Sequencing in the overall Digit Span subtest score calculation (Wechsler, 2007b; Wechsler 2008a). According to the WAIS-IV technical and interpretive manual, this task was developed to increase the working memory demands of the subtest overall. It is similar to digits forward and backward except it requires individuals to sequence numbers in ascending order. In addition to this change, an additional 2-digit item was added to digits backward and the item that required repeated 9 digits backward was eliminated (Wechsler, 2008a). Eleven of the digits forward and five of the digits backward items from the WAIS-III were retained. The additional items were modified to include more of the numbers from 0 to 9. All items still contain two distinct trials with the same number of digits. The scoring and discontinuation rules remain consistent with the WAIS-III (Wechsler 1997b; Wechsler 2008b). No literature was available on the difficulty level of the items on the Digit Span subtest of the WAIS-IV.

### *Arithmetic*

Mental arithmetic tasks have long been used in intelligence tests and were commonly included in test batteries around the time Wechsler developed his first

intelligence scale. It is unclear if Wechsler found or created the ten items included in the Arithmetic subtest of the Wechsler-Bellevue Intelligence Scale. The time limit for items ranged from 15 to 120 seconds. While the WAIS retained six of the items from the original version, the WAIS-R only retained one item from the WAIS (Tulsky, et al., 2003).

The WAIS-III retained 14 of the items from the WAIS-R and included six new items for a total of 20 items. According to the WAIS-III administration and scoring manual, the new items were added to increase the range of scores and to decrease the dependency on time bonuses. Several of the items retained were also reworded and contained updated prices in an effort to be more current (Wechsler 1997a). Items one through six have a 15-second time limit, items seven through 11 have a 30-second time limit and items 12 through 18 have a 60-second time limit. The first 18 items are worth 0 or 1 point while the last two items are worth 0, 1, or 2 points depending on response time and have a 120-second time limit. The discontinue rule is four consecutive scores of 0 (Wechsler, 1997b). The literature provided no information on how item difficulty on the Arithmetic subtest of the WAIS-III was determined.

The WAIS-IV Arithmetic subtest contains 12 items that require the same mathematical calculation(s) as the WAIS-III but have slightly altered content to be clearer, more contemporary and more applicable across cultures. Ten additional items were created, for a total of 22 items on this subtest. Items with time bonuses were eliminated to decrease the emphasis of quick task completion (Wechsler, 2008b). No literature was available about how item difficulty on the Arithmetic subtest of the WAIS-IV was determined.



### *Picture Arrangement*

While this subtest is not included in the WAIS-IV, it was part of all previous versions of the tests and it is one of the few subtests about which there is published literature on item difficulty. One of the first picture arrangement tests was developed by a French psychologist in 1914. It is likely that Wechsler first had practical experience with this test during his time in the military. The Wechsler-Bellevue Intelligence Scale contained six items, which were taken from existing measures and created from several cartoon strips from a popular magazine (Tulsky, et al., 2003). Items had a one or two-minute time limit. Arrangement on the WAIS consisted of eight items, with the last two items having point bonuses for quick responses. In addition, the first two items had two trials and were worth 2 or 4 points depending on how quickly the task was completed and on what trial number the task was completed. Three items had alternate, lower scoring, acceptable arrangements worth two points instead of four (Wechsler, 1955a). The WAIS-R Picture Arrangement subtest contained ten items (Wechsler 1997a).

For the WAIS-III, five of the 10 Picture Arrangement items were retained and six new items were created, for a total of 11 items, including one item with two trials. Five of the items were worth either 0 or 2 points while the other six items were worth 0, 1, or 2 points. The first item was worth 1 point if it is correctly completed on the second trial. The other items were scored 1 point if they were arranged in an acceptable but less than 2-point response pattern. Both trials of the first item had a 30-second time limit. The second item had a 45-second time limit. Items three and four had a 60-second time limit. Items five and six had a 90-second time limit. Items seven through 11 had a 120-second time limit. The discontinue rule was four consecutive scores of 0, starting with item two

(Wechsler 1997a). One reason for this subtest's removal from the WAIS-IV may have been to reduce the motor demands of the overall test (Climie & Rostad, 2011). Although literature exists suggesting that the items on the Picture Arrangement subtest of the WAIS-R and WAIS-III, the WAIS-IV technical and interpretive manual does not mention this as a reason for the subtest being removed.

One study explored the order of item difficulty of the Picture Arrangement subtest of the WAIS-R in a sample of traumatically brain injured individuals and found that items two and four were answered correctly less frequently than item three and five (Heath & Leathem, 1998). Results of the study also revealed that an approximately equal number of participants provided the 1 and 2-point responses to item two, which may have explained why this item appeared more difficult. Participants in this study also failed to provide a 2-point response to the 10<sup>th</sup> and final item, indicating it may have been too difficult.

An analysis of the Picture Arrangement subtest of the WAIS-III determined that the items were disordered (Costello & Connolly, 2005). Specifically, the third item was found to be approximately as difficult as the sixth item and the fifth item was found to be too easy. Item difficulty was determined by the percentage of individuals who obtained a 2-point score on the item. A follow-up study confirmed that the items are disordered and found that items three, five and nine are more or less difficult than their placement suggested (Ryan & Lopez, 1999).

### ***The Influence of Classical Test Theory (CTT)***

Like many psychological tests of its time, the objective of the WAIS developers has been to establish the validity and reliability of the scores within the framework of

classical test theory (CTT). In CTT, a person's observed score on a test is a function of the person's true score, plus error, which is expressed mathematically as  $X_o = X_t + X_e$ , where  $X_o$  stands for an individual's observed, or test, score,  $X_t$  represents an individual's true score and  $X_e$  designates error (Furr & Bacharach, 2008).

Reliability is defined as the extent to which the differences in the observed scores are a function of the differences in the true scores. Validity, on the other hand, is defined as the extent to which interpretations and applications of a test's scores are supported by evidence and theory (AERA, APA, & NCME, 1999, p.9). According to the WAIS-IV technical and interpretive manual, the average reliability across all age groups for all subtests ranges from .78 to .94. The index score reliability coefficients ranges from .90 to .98. Test-retest analyses were used to assess reliability of the processing speed subtests while split-half reliability and internal consistency reliability were used to determine the reliability of the remaining subtests (Sattler & Ryan, 2009; Wechsler, 2008b). Average internal consistency reliability coefficients are as follows: .87 for Similarities, .94 for Vocabulary, .93 for Information, .87 for Block Design, .90 for Matrix Reasoning, .89 for Visual Puzzles, .93 for Digit Span and .88 for Arithmetic (Sattler & Ryan, 2009).

According to the WAIS-IV technical and interpretive manual, criterion validity was assessed by comparing the WAIS-IV subtests with a number of neuropsychological measures, including the WAIS-III, Wechsler Individual Achievement Test (Wechsler, 1992), Wechsler Intelligence Scale for Children – Fourth Edition (Wechsler, 2003), the Delis-Kaplan Executive Functioning System (Delis, Kaplan, & Kramer, 2001) and the Repeatable Battery for the Assessment of Neuropsychological Status (Randolph, 1998).

Similar to the WAIS-III, content validity of the WAIS-IV was assessed through a review of the literature on both the test itself and the concept of intelligence, expert reviews and empirical analyses (Wechsler, 2008b). Criterion validity was again assessed by analyzing results of factory analysis and subtest inter-correlations (Wechsler, 2008b). Construct validity was also again analyzed by assessing results of factor analysis and subtest inter-correlation studies (Wechsler, 2008b).

***Limitations of CTT.*** Empirical evidence suggests that the WAIS-IV has adequate reliability and validity as defined by CTT. However, there are several limitations of CTT that restrict the way the psychometric properties of test scores can be assessed. First, the results of CTT are heavily dependent upon the sample. For example, in CTT the standard error of measurement is a mathematical representation of expected changes in scores due to error and is an important indicator of reliability. Standard error is consistent among scores for the same population and changes when the population changes.

Another example of the sample-dependent nature of CTT is the use of *p*-values, or the proportion of people correctly answering an item, to determine item difficulty (Embretson & Reise, 2000), as is commonly done in CTT. Using this methodology means that a sample of individuals with above average intelligence will produce lower difficulty values than a sample of individuals with below average intelligence (Hambleton & Jones, 1993). Additionally, while the overall order of the items is adequate for different samples, the distance between more difficult items may be greater for a sample with lower abilities than it is for a sample with above average abilities (Embretson and Reise, 2000).

It is not uncommon for test developers who create tests within the framework of CTT to assume that scores exist upon an interval scale because they are normally distributed, or forced to be normally distributed (Embretson & Reise, 2000). The standard scores for many tests are normalized using either percentile matching or nonlinear transformations, both of which change the distance between scores, potentially changing the scale from interval to ordinal. Even scores that arise from a normal distribution (e.g., Full Scale IQ) cannot be assumed to be on an interval level scale without empirical evidence that the scores have interval scale properties. If the WAIS-IV truly existed upon an interval scale, the difference between an FSIQ of 75 and 95 would be the same as the difference between an FSIQ of 115 and 135 and this difference would have to be same among individuals of all ages.

Lastly, when considering the mixed format of the WAIS-IV, another significant limitation of CTT is the inequity of test items that is created by varying the number of response options and the weights or values of the different response options, as occurs in some subtests (Embretson and Reise, 2000).

#### ***Item Response Theory (IRT) and the Standard Dichotomous Rasch Model***

***Item Response Theory (IRT).*** Item response theory (IRT) assumes that test items measure a latent trait, which individuals possess to varying degrees (Crocker & Algina, 2008). It has been used to enhance and refine the items of existing measures. For example, item content and order on the Stanford-Binet Intelligence Scales – Fifth Edition (Roid, 2003) and the SAT, a common college entrance examination, have been revised based on results from IRT analysis (Furr & Bacharach, 2013).

IRT modeling has advantages that allow it to overcome many of the previously

discussed limitations with CTT. One of the main advantages of employing IRT is that it places items onto an interval scale, “using justification in the measurement model” (Embretson & Reise, 200, p. 32). While the difficulty of items in CTT is dependent upon the sample being tested, IRT can produce unbiased estimates of item and test characteristics from heterogeneous samples because the person and item characteristics are independent of the sample being used (Crocker & Algina, 2008; Hambleton & Jones, 1993). Item difficulty levels are linked to person ability levels and placed along the same scale using linear transformations.

***Standard Dichotomous Rasch Model.*** The standard dichotomous Rasch model (Rasch, 1960), a type of IRT, is a one-parameter logistic model that examines item and person characteristics, which can be analyzed to glean information about overall test characteristics (Furr & Bacharach, 2013). For the standard dichotomous Rasch rating scale, several assumptions must be met. First, items that are more difficult are assumed to require a higher trait level in order to be correctly answered. Second, items that comprise the measure being examined must be locally independent. Local independence refers to the fact that items are independent of one another (Bond & Fox, 2007). In other words, item content is distinct and correctly answering one item does not aid an individual in correctly answering another item (Baghaei, 2008). Lastly, the standard dichotomous Rasch model assumes that measures are unidimensional (Embretson & Reise, 2000).

The standard dichotomous Rasch model can be employed with binary (e.g., yes or no) response categories. The probability of a participant receiving a score of 1 on an item is mathematically represented as  $P(X_{is} = 1 | \theta_s, \beta_i) = \frac{e^{(\theta_s - \beta_i)}}{1 + e^{(\theta_s - \beta_i)}}$  where  $X_{is}$  represents an

individual's response,  $\theta_s$ , stands for an individual's ability level,  $\beta_i$  refers to the difficulty of the item and  $e$  is the symbol for the base of the natural logarithm (Embretson and Reise, 2000).

Proponents of standard dichotomous Rasch model argue that it has distinct advantages above both CTT-based methods as well as other IRT models (Bond & Fox, 2007, Embretson & Reise, 2000; Furr & Bacharach, 2013; Hambleton & Jones, 1993) because of the principle of monotonicity, also referred to as specific objectivity, the principle of additivity or double cancellation, which “establishes that two parameters are additively related to a third variable” (Embretson & Reise, 2000, p. 148). In other words, because of the principle of monotonicity, in Rasch modeling, probability of correctly answering an item is the additive function of individuals' ability, or trait level, and the item's degree of difficulty. As ability increases, so does an individual's probability of answering that item. Because only item difficulty and person ability affect an individual's chance of correctly answering an item, inter-individual comparisons can be made even if individuals did not receive identical items or items of the same difficulty level. This is why Rasch modeling is referred to as a *test-free measurement*. This is not true for other IRT models (Embretson & Reise, 2000).

In addition to estimating individuals' trait levels, one of the primary applications of the standard dichotomous Rasch model is analyzing the characteristics of test items, including if they are measuring the intended trait, how difficult they are and how well they differentiate individuals with varying levels of ability (Bond & Fox, 2007; Embretson & Reise, 2000; Furr & Bacharach, 2013). By analyzing how individuals with different levels of a trait respond to items, the standard dichotomous Rasch model

provides an in depth analysis of item difficulty that is not possible using CTT (Crocker & Algina, 2008).

Because the standard dichotomous Rasch model assumes that the items on a test increase in difficulty, it is generally not an appropriate analysis for speeded tests, in which task difficulty is mainly a product of the imposed time limit and in which items remain approximately equally difficult (Bond & Fox, 2007; Embretson & Reise, 2000; Furr & Bacharach, 2013).

**Partial Credit Model.** For subtests with items that have two or more possible response options and different thresholds, a partial credit model can be employed. This model is mathematically represented as  $\frac{e^{\sum_{k=0}^x (\theta_j - \tau_{ki})}}{\sum_{x=0}^m e^{\sum_{k=0}^x (\theta_j - \tau_{ki})}}$  where  $m$  is the maximum score

for the item,  $\theta_j$  is the ability level of an individual,  $\beta_i$  is the difficulty level of the item and  $\tau_{ki}$  is the threshold of the rating scale of the item (Bond & Fox, 2007; Embretson and Reise, 2000). Just as with the standard dichotomous Rasch model, the partial credit model can also be used to examine other test and participant characteristics.

**Two-Parameter Model.** Another type of IRT is a two-parameter logistic model, which takes into account two item characteristics: item difficulty and item discrimination. Discrimination refers to differences in an item's ability to differentiate individuals based on trait levels. The item discrimination value is mathematically represented as alpha subscript  $i$  ( $\alpha_i$ ), in which  $i$  stands for item. Discrimination is a problem because it "indicates that groups cannot be meaningfully compared on the item" (Furr & Bacharach, 2014, p. 408). Including discrimination as a second parameter allows for more accurate estimations of item difficulty when item discrimination values vary (Bond & Fox, 2007).



*Three-Parameter Model.* A three-parameter model, which takes into account item difficulty, item discrimination, and guessing can also be employed to account for correct answers due to chance. Introducing this additional parameter, however, has the potential to produce an ordinal rather than interval scale (Embretson & Reise, 2000).

### *Current Research*

Because the WAIS-IV was designed with a more modern understanding of intelligence than previous versions, a more modern approach to understanding item difficulty is merited. The first hypothesis was that the items that comprise Block Design, Similarities, Vocabulary, Digits Forward, Digits Backward and Digit Sequencing do not proceed in order of difficulty. The second hypothesis was that the items that comprise Matrix Reasoning, Visual Puzzles, Information and Arithmetic do not proceed in order of difficulty.

Several subtests may be more likely to contain disordered items because of the subtest or item characteristics, including scoring criteria, time limits, and item content. For example, the scoring system for Similarities permits scores of 0, 1, or 2 based upon guidelines in the administration and scoring manual, sample answers provided in the stimulus book and subjective clinical judgment. Answers to Similarities from the WAIS-IV were assigned their point values from results of scoring studies conducted as part of the development of the WAIS-IV. Little detail is provided in the WAIS-IV technical manual about how these point values were determined and validated, except that "results were then subjected to psychometric analysis and clinical and bias expert review" (p. 30). However, some information about the determination of point values is provided in the WAIS-III technical manual, which explains that the entire WAIS-III development team

assigned 0, 1, or 2-point codes to each response from the standardization protocols. No other information is provided.

Because the time limit on the items of Visual Puzzles goes from 20 seconds on item seven to 30 seconds on item 8, this may alter the difficulty of the items. Slower individuals may need only slightly extra time to solve a problem that is only slightly more difficult than the proceeding problem. Thus, individuals may perform better on several items because of the increased time limit.

Digit Sequencing is unique because some of the numbers in the sequence are repeated. It is posited that this could make some of the items easier or more difficult. Because of this, it was hypothesized that the items that comprise Digit Sequencing do not proceed in order of difficulty.

### CHAPTER III

#### Methods

##### *Participants*

This study involved analysis of archival data from two databases. Participants consisted of adults referred for neuropsychological evaluation at the Neuropsychology Assessment Center at Nova Southeastern University and adults who have volunteered to participate in research and receive a full neuropsychological evaluation. Informed consent was obtained from each participant. Participants were previously administered a battery of neuropsychological tests by doctoral level clinical psychology students, who were trained in the administration, scoring, and interpretation of each test. Each student completed supervised training prior to test administration. Licensed clinical psychologists reviewed all testing results. While each participant received a comprehensive neuropsychological battery, for the purposes of the current study, only variables from the Wechsler Adult Intelligence Scale-IV were used in analyses.

Criteria for inclusion in the current study consisted of being at least 16 years of age and being referred for neuropsychological evaluation or agreeing to participate in a research study and complete a battery of neuropsychological measures. Exclusion criteria included being below the age of 16, not speaking English fluently, or not completing the core subtests of the WAIS-IV. There were no other exclusionary criteria.

According to information completed to comply with the Institutional Review Board (IRB), participants who were volunteers donated approximately 12 hours of their time, reviewed and signed an IRB approved consent form detailing the procedures, risks and benefits of participating in this study, and received a copy of a report describing their

test results. Individuals who sought neuropsychological services as clients received a copy of a report describing their results as well as a feedback session with their student clinician to review the report.

Participants consisted of 300 individuals ages 16 to 97 ( $M = 33.95$ ,  $SD = 14.48$ ), predominantly female (55%) and right-handed (86%), with a mean education of 14.28 years ( $SD = 2.35$ ). Participants were predominantly of Caucasian ethnicity (59.7%), with 19.7% endorsing Hispanic ethnicity, 9% endorsing African American ethnicity, 4.0% endorsing Haitian ethnicity, and 7.0% reporting “Other” ethnicity.

Average FSIQ was 99.01 ( $SD = 16.61$ ), average VCI was 103.45 ( $SD = 17.75$ ), average PRI was 99.40 ( $SD = 15.55$ ), average WMI was 95.26 ( $SD = 15.71$ ) and average PSI was 96.99 ( $SD = 15.95$ ).

Participants included 183 individuals (61%) with psychological diagnoses and 117 individuals (39%) without diagnoses. Individuals with only one diagnoses comprised 33% of the sample, while 21% of the sample had two diagnoses, 5.7% had three diagnoses and 1% had four diagnoses. Learning Disorders (21.2%) were the most commonly occurring diagnoses, followed by Anxiety Disorders (14%), then Cognitive Disorders, Personality Disorders and Diagnosis Deferred (8% each). Other disorders included V-Codes (7.3%), Mood Disorders (6.3%) and Adjustment Disorder (4.7%).

### *Measures*

#### *WAIS-IV*

Using 10 core subtests, the WAIS-IV measures global intellectual and cognitive functioning in adolescents and adults between the ages 16 to 90. To measure specific domains, the core subtests comprise indices, which include the Perceptual Reasoning,

Verbal Comprehension, Processing Speed, and Working Memory Indices. Standardized index scores have a mean of 100 and a standard deviation of 15, while subtest scores have a mean of 10 and a standard deviation of 3 (Wechsler, 2008a).

Raw scores from the Block Design, Matrix Reasoning, Visual Puzzles, Similarities, Information, Vocabulary, and Arithmetic subtests as well as the Digits Forward, Digits Backward and Digit Sequencing task of the Digit Span subtest of the WAIS-IV will be examined. Scores were entered only for those items to which participants actually responded. This is not consistent with the traditional scoring method, in which participants who correctly answer the first two items of any subtest are assigned points for the items that occur before the start point. Thus, scores for items that occur before the start point were only entered in analysis when individuals were administered those items due to the reversal rule of the WAIS-IV.

Because the processing speed subtests, Coding and Symbol Search, are not appropriate for IRT, they were not included in analyses.

**Block Design.** The Block Design subtest from the WAIS-IV has been found to measure visual-spatial organization, specifically, the analysis and synthesis of visual stimuli as well as abstract conceptualization (Sattler & Ryan, 2009). The subtest consists of 14 items and is part of the Perceptual Reasoning Index.

During the administration of Block Design, participants are asked to assemble two, four, or nine blocks, which are red on some sides, white on some sides, and half red and half white on some sides, to match a picture stimulus. The Sample and items 1 through two consists of a design made with two blocks. Items 3 through 10 contain designs made with four blocks while items 11 through 14 contain designs made with nine blocks.

Participants watch a demonstration item, are administered the Sample item and then begin the subtest with item five. Administration of this subtest is discontinued after participants receive two consecutive scores of zero.

Items one through four are worth 0, 1 or 2 points each. Items five through eight are scored 0 for either incorrect designs or designs completed after the time limit and four points for designs made correctly within the time limit. On items nine through 14, participants receive 0 points for incorrect designs or designs completed after the time limit and four, five, six, or seven points for designs made correctly within the time limit. Scores for correctly made designs within the time limit on items nine through 14 are dependent upon participants' item completion time. Raw scores are calculated by summing the total number of points received on all items. Raw scores range from 0 to 66 (Wechsler, 2008c).

***Matrix Reasoning.*** Matrix Reasoning can be described as a nonverbal task of fluid reasoning that involves classification ability, induction, and spatial ability (Sattler & Ryan, 2009). The Matrix Reasoning subtest is part of the Perceptual Reasoning Index. It consists of 26 items, in which participants view an incomplete matrix or series of visual stimuli and selects one of four response options to complete the matrix or pattern.

Participants are given two Sample items and then begin with item four. Participants receive 0 points for incorrect answers and 1 point for correct answers. Administration of this subtest is discontinued after three consecutive incorrect (i.e., 0-point) answers. Raw scores are calculated by summing the points received on all items. Raw scores range from 0 to 26 (Wechsler, 2008c).

**Visual Puzzles.** Visual Puzzles is a "spatial visual-perceptual reasoning" (Sattler & Ryan, 2009, p. 105) task that requires analysis and synthesis, nonverbal reasoning, and visual-perceptual discrimination. It is a 26-item timed Perceptual Reasoning Index subtest that requires participants to view a completed puzzle and select three response options that will combine to make the completed picture.

After a demonstration item, participants are given a sample item and then start with item five. If all three response options selected are correct and chosen within the time limit, individuals receive one point for the item; otherwise, individuals receive 0 points. Items one through seven have a 20-second time limit and items eight through 26 have a 30-second time limit. Raw scores are calculated by summing the total number of points received and range from 0 to 26 (Wechsler, 2008c).

**Similarities.** The Similarities subtest is comprised of 18 items and requires participants to describe how two words are alike. It has been shown to measure abstract concept formation and associative thinking ability (Sattler & Ryan, 2009) and is one of the subtests that comprises the Verbal Comprehension Index.

Participants are first administered a sample item and then begin with item four. Items are scored according to guidelines in the WAIS-IV Administration and Scoring Manual (Wechsler, 2008), which lists possible responses and their point-values (0, 1, or 2). The manual also specifies when participants' responses should be queried for further clarification. Subtest administration is discontinued after three consecutive 0-point responses are provided. Raw scores are calculated by summing the total number of points received and range from 0 to 36 (Wechsler, 2008c).

***Vocabulary.*** The Vocabulary subtest is comprised of 30 items and requires participants to verbally define words. As part of the Verbal Comprehension Index, it is thought to measure lexical knowledge, long-term memory, verbal comprehension and verbal expressive ability (Sattler & Ryan, 2009).

Subtest administration begins with item five. Participants are provided with corrective feedback if they do not provide responses that receive perfect scores (i.e., 2) on both items five and six. After item 6, corrective feedback is not provided. Items are scored according to guidelines in the WAIS-IV Administration and Scoring Manual (Wechsler, 2008a), which lists possible responses and their point-values. Items one through three are scored as 0 or 1 point while items 4 through 30 are scored as 0, 1, or 2 points. The WAIS-IV Administration and Scoring Manual (Wechsler, 2008a) also specifies when participants' responses should be queried. Administration continues until three consecutive 0-point responses are given. Raw scores range from 0 to 57 (Wechsler, 2008c).

***Information.*** The last score subtest of the Verbal Comprehension Index, Information, contains 26 items. It requires participants to answer questions about a broad range of knowledge and has been described as measuring verbal comprehension, factual knowledge, and expressive language abilities (Sattler & Ryan, 2009).

Answers are scored as either 0 or 1 according to guidelines in the WAIS-IV Administration and Scoring Manual (Wechsler, 2008a). Subtest administration begins with item three. Participants are provided with corrective feedback if they do not provide responses that receive perfect scores on items three and four. Participants who do not obtain perfect scores on items three and four are administered items one and two in



reverse order until they receive two consecutive perfect scores. Participants who receive two points each for their responses to items five and six are also given full credit (i.e., two total points) for items one and two. Administration continues until three consecutive 0-point responses are made. Raw scores are calculated by summing the total number of points received. Total raw scores range from 0 to 26 (Wechsler, 2008c).

***Digits Forward.*** The forward repetition trial of the Digit Span task requires individuals to repeat from two to eight digits. The first item is comprised of two trials, each containing a series of two digits. After the second item, the number of digits in the sequences increases by one per item. Thus, the eight total items contain sixteen trials of a series of digits ranging from two digits long to nine digits long. Each trial is scored zero for incorrectly recalled sequences or one for correctly recalled sequences. Therefore, items are scored as zero, one, or two points. All participants receive two sample items and then begin with item 1. Administration continues until participants receive a score of zero on both trials of an item. Raw scores for sequencing Digit Span are calculated by summing the total number of points received. Scores range from 0 to 16 (Wechsler, 2008c).

***Digits Backward.*** The backward repetition portion of the Digits Span subtest requires participants to reverse sequence a series of digits ranging in length from two to eight. There are two sample trials followed with a string of two digits. The first and second scored items contain two digits and then increase by one digit up to eight digits. Scoring is the same as for the Digits Forward task.

***Digit Sequencing.*** The Sequencing task in the Digit Span subtest requires participants to verbally sequence series of digits in ascending order. Items proceed and are scored in the same manner as items from Digits Forward and Digits Backward.

***Arithmetic.*** The Arithmetic subtest, part of the Working Memory Index, contains 22 items and requires participants to mentally solve arithmetic word problems under timed conditions. It has been described as a measure of quantitative reasoning, mental computation, attention, and auditory sequential processing (Sattler & Ryan, 2009). Items answered correctly within the 30-second time limit receive 1-point; otherwise, responses are scored 0. The examiner is allowed to repeat the word problem one time, if asked, though time is not stopped during the rereading. After a sample item, administration begins with item six. Participants who do not score 1-point on items six and seven are administered items one through five in reverse order until they receive two consecutive perfect scores. Participants who receive 2-points each for their responses to items six and seven are also given full credit (i.e., five total points) for items one through five. Subtest administration is discontinued after three consecutive 0-point answers. Raw scores range from 0 to 22 (Wechsler, 2008c).

### ***Procedure***

***Data Collection.*** For the purposes of this study, data were derived from two archival databases, which consisted of psychological evaluations of adults referred to the Neuropsychological Assessment Center at Nova Southeastern University and from volunteer research participants. Doctoral level clinical psychology practicum students, under the supervision of a licensed clinical psychologists at Nova Southeastern University, administered all of the measures. All students completed Nova Southeastern

University Citi training. Multiple measures were administered as part of the complete battery, but only the Wechsler Adult Intelligence Scale-IV will be included in this analysis.

***Institutional Review Board Requirements.*** Before any data were analyzed, approval was obtained from the Institutional Review Board (IRB) at Nova Southeastern University to conduct archival research. All data were de-identified in order to maintain confidentiality.

### ***Statistical Analyses***

***Preliminary Analyses.*** Demographic variables of age, ethnicity, gender, education and handedness were collected for each participant. Information about individuals' performance on subtests and indices was also collected. Participant diagnoses were recorded, as were the distribution of diagnoses for the sample. Descriptive statistics for the raw data of all subtests was also analyzed.

In order to evaluate the standard dichotomous Rasch and partial credit models, the following key statistics were evaluated: (1) expected order of endorsement of ordered response options as a function of trait levels (i.e., ordered thresholds), (2) item fit with respect to the inlier and outlier patterns, (3) person fit, (4) dimensionality, (5) person ability, (6) person reliability, (8) item reliability, (8) item coverage, (9) discrimination (estimated outside of the Rasch model), and (10) order of item difficulties. SPSS Version 21 in addition with Winsteps were employed for analysis of data in the study.

***Ordered Andrich Thresholds, Observed Averages and Sample Expectations.*** One of the first steps in using a Rasch or partial credit model is to explore items' Andrich thresholds. Andrich thresholds are calculated by creating probability curves for each

response option against the range of possible trait levels. Andrich thresholds are the points at which the probability curves for adjacent categories overlap. In order for Andrich thresholds to be calculated, participants must provide several responses to each response category; otherwise, the measurement model is unable to produce stable item calibrations (Bond & Fox, 2007).

Andrich thresholds refer to the ability level needed for participants to have a 50% chance of picking or generating one response versus picking or generating an adjacent response (Embretson & Reise, 2000). For example, in a rating scale where 0 is equivalent to an incorrect answer, 1 is equivalent to a partially correct answer and 2 is equivalent to a correct answer, thresholds exist between 0 and 1 and 1 and 2. Multiple Andrich thresholds only exist for items with more than two possible responses (e.g., incorrect, partially correct, correct), referred to as polytomous rating scales. The standard dichotomous Rasch model and the partial credit model assume that items have ordered Andrich thresholds.

Ordered Andrich thresholds ensure that the probability of obtaining a higher score on an item requires a higher trait level. For example, Andrich thresholds are ordered when the ability level needed for a participant to have a 50% chance of generating or picking a response option worth fewer points (e.g., a 1 point “partially correct” response) is less than the ability level needed for participants to have a 50% chance of generating or picking a response option worth more points (e.g., a 2 point “correct” response). Disordered thresholds indicate that the rating scale is not performing in the way the Rasch or partial credit model expect (Bond & Fox, 2007).

Disordered thresholds are indicative of a problem with the way in which

individuals receive credit on an item (Tennant, 2004). For example, disordered thresholds can occur when there is a mix of individuals with high trait levels being more likely to obtain a lower score. When disordered thresholds exist, it is necessary to adjust the rating scale. Adjusting disordered thresholds often requires combining response categories (Embretson & Reise, 2000). When the threshold between lower categories (e.g., between 0 and 1) is greater than the threshold between higher response categories (e.g., between 1 and 2), these response categories may be combined. The combined responses are then assigned the lower point value.

Observed averages are the average ability levels required for participants in this sample to generate or pick a response in a certain category. Sample expectations are estimates of the average ability level needed for participants to make a response in a certain category (e.g., correct or incorrect) (Bond & Fox, 2007). The standard dichotomous Rasch and partial credit models assume that observed averages and sample expectations increase as category, or answer, value increases (Embretson & Reise, 2000). This means that both the observed averages and sample expectations for wrong responses (i.e., those worth 0 point) should be lower than the observed averages and sample expectations for partially correct responses (i.e., those worth 1 point), which should be lower than the observed averages and sample expectations for correct answers (i.e., those worth 2 points).

***Item Fit.*** Item outfit mean square equals the sum of the standardized residuals squared divided by the number of subjects, where the standardized residual equals the residual divided by the square root of the variance. It approximates a chi square distribution (Linacre, 2014e). Mathematically, this is represented as  $Infit = \text{sum}$

$[(\text{residual}^2 / \text{model variance}) * \text{model variance}] / \text{sum}(\text{model variance}) = \text{average}$

$[(\text{standardized residuals})^2 * \text{model variance}] = \text{model variance-weighted mean-square.}$

Outfit is the chi square divided by degrees of freedom and is mathematically represented

as  $\text{outfit} = \text{sum}(\text{residual}^2 / \text{model variance}) / (\text{count of residuals}) = \text{average}$

$[(\text{standardized residuals})^2] = \text{chi-square/degrees of freedom} = \text{mean-square.}$

Infit Mean square infit and outfit values quantify the extent to which items conform to measurement model expectations (Bond & Fox, 2007). Items with both mean square infit and outfit values greater than 1.3 indicate response patterns that are “too haphazard” (Bond & Fox, 2007, p. 240). When both values are above 1.3, stable item statistics cannot be calculated because participants responded in extremely unexpected ways to these items (Bond & Fox, 2007).

***Person Fit.*** Person fit, which is similar to item fit, determines how well participants’ responses aligned with expectations of the standard dichotomous Rasch or partial credit models’ expectations. As with misfitting items, unstandardized mean square infit and outfit values are used to identify misfitting participants. These values are calculated in the same manner as item fit values, described above (Bond & Fox, 2007).

When both infit and outfit mean square values are greater than 1.3, participants are considered misfitting, indicating that they responded in unexpected ways (Bond & Fox, 2007). When a large number of participants are misfitting, employing a standard dichotomous Rasch or partial credit model is not useful because the model is unable to predict participant performance.

***Dimensionality.*** Dimensionality refers to how many latent constructs a test measures. The standard dichotomous Rasch model and partial credit model assume that

tests measure only one underlying construct (Bond & Fox, 2007; Embretson & Reise, 2007), which is called unidimensionality.

First, the total raw explained variance was examined. Raw explained variance is an unstandardized value that quantifies how much of participants' performance can be attributed to their ability and item difficulty. In standard dichotomous Rasch modeling and the partial credit model, the raw explained variance is calculated in the same manner as would be done during an exploratory factor analysis. The two components entered into this analysis are participants' ability levels and item difficulty.

The next step in assessing dimensionality is to look at the unexplained variance to determine whether there is another factor, in addition to item difficulty and person ability, that may be systematically affecting task performance (Linacre, 2014a). When another factor is systematically and significantly affecting task performance, this is indicative of multidimensionality (Bond & Fox, 2007).

To determine whether another factor can explain the unexplained, or residual, variance, a principal component analysis is performed on the residual variance. The principal component analysis identifies groups of items that account for portions of the residual variance. These groups of items are referred to as factors. The first factor is the one that explains the greatest portion of the residual variance. While there may be a number of these factors, Winsteps software includes up to five factors. Fewer factors are included if there are no items comprising additional factors (Linacre, personal communication, June 2, 2014).

Because the first factor explains the most residual variance, examining the residual variance requires comparing the unexplained variance accounted for by the first

factor with the total residual variance.

Finally, the raw unexplained variance accounted for by all factors is examined. Guidelines suggest that when the raw values of all factors are less than 2, a measure is unidimensional (Linacre, 2014a).

***Person Ability.*** Person ability is estimated via a Rasch or partial credit model using joint maximum likelihood estimation, which optimizes the model to data fit through an iterative calculus based approach. Similar to multiple regression, where the estimate method is ordinary least squares, the Rasch model and partial credit model uses joint maximum likelihood estimation, where it maximizes model to data fit.

Person ability is the estimate of an amount of attribute a person has, for example, average verbal abstraction abilities. It is reported on the logit scale, where a lower value indicates a lower level of a trait, such as verbal abilities, and a higher value represents a higher level of a trait (Furr & Bacharach, 2008). A wide range of person ability helps ensure the accurate calculation of item statistics, such as item difficulty (Embretson & Reise, 2000).

***Person Reliability.*** Person reliability is akin to Cronbach's alpha in CTT. It can be interpreted in a manner similar to which test score reliability would be interpreted. Person reliability is dependent upon the test having adequate coverage (discussed below) and including individuals with a wide range of trait levels (Bond & Fox, 2007). Values above .70 are considered acceptable (Furr & Bacharach, 2013). Person reliability can also be assessed by examining chi-square values. Significant chi-square values indicate a significant departure of the data (i.e., person responses) from the model (Bond & Fox, 2007).



***Item Reliability.*** Item reliability is an indication of the replicability of raw scores (Bond & Fox, 2007). Item reliability ranges from 0 to 1. Higher values indicate greater reliability, which suggests that the range of item difficulty was adequate (Furr & Bacharach, 2013), there was an adequate range of item difficulty and item statistics produced from these analyses are highly likely to be replicated (Linacre, 2014b).

***Item Coverage.*** Item coverage was examined next. Item coverage is the extent to which items span the entire continuum of participants' ability levels. Item difficulty and person ability levels are reported in logits. By using a logit scale, it has been argued that what often are ordinal level raw scores are converted to interval level scores, whereby both individual person ability levels and item difficulties can be placed on the same continuum, as shown in the item to person map (Embretson & Reise, 2000).

On the item to person map, individuals are typically represented as symbols such as dots or X's and item numbers designate items. Items plotted higher on the graph are more difficult while individuals plotted higher on the graph have higher trait levels. Poor coverage can indicate that there is not a large enough spread of items to adequately assess individuals of all trait levels (Bond & Fox, 2007; Crocker & Algina, 2008). Some items may be redundant (i.e., there are too many items with the same difficulty value) or there may be a dearth of items (i.e., items drastically jump in difficulty level).

***Item Discrimination.*** Item discrimination refers to an item's ability to differentiate individuals of varying ability levels. It is indicative of the relationship between the item and the underlying construct being measured (Furr & Bacharach, 2013).

Ideally, in Rasch modeling, all item discrimination values are assumed to be equal to one (Bond & Fox, 2007). That is, all items are assumed to differentiate individuals

equally, given that items' level of difficulty. Because item discrimination values are assumed to be equal to one, they are not calculated in Rasch modeling. However, it is empirically unlikely that all item discrimination values will actually equal one (Linacre, 2014b). Instead, guidelines suggest that all item discrimination values should be between 0.5 and 1.5 (Linacre, 2009). An item with a discrimination outside these values indicates that the item does not distinguish between individuals with high and low ability levels as would be expected given that item's level of difficulty (Linacre, 2014b).

*Item Difficulties.* Similar to multiple regression, where the estimate method is ordinary least squares, the Rasch model uses joint maximum likelihood estimation, where it maximizes model to data fit. Item difficulty is estimated via a Rasch model using joint maximum likelihood estimation, which optimizes the model to data fit through an iterative calculus based approach. Item difficulty represents how difficult it is to answer an item correctly (or partially correctly), is estimated in logits and exists on a true interval scale, in which the intervals between difficulty levels have a consistent value (Bond & Fox, 2007).

### *Hypotheses*

The first hypothesis was that the items that comprise Block Design, Similarities, Vocabulary, Digits Forward, Digits Backward and Digit Sequencing do not proceed in order of difficulty. The second hypothesis was that the items that comprise Matrix Reasoning, Visual Puzzles, Information and Arithmetic do not proceed in order of difficulty. To evaluate the first hypothesis, the standard dichotomous Rasch model was employed because it is appropriate for items with only two response options. To evaluate the second hypothesis, a partial credit model was employed because it is appropriate for

items with more than two response options. The standard dichotomous and partial credit models use joint maximum likelihood estimation, which maximizes model to data fit.

Item difficulty is estimated via the standard dichotomous Rasch and partial credit models using joint maximum likelihood estimation, which optimize the model to data fit through an iterative calculus based approach and calculates item difficulty values in logits.

## CHAPTER IV

## Results

*Preliminary Analysis*

Descriptive statistics for the WAIS-IV subtests are displayed in Table 1. Results revealed that none of the subtests had significantly skewed or kurtotic distributions, as would be indicated by values greater than  $\pm 2.0$  (Myers, Well, & Larch, Jr., 2010).

Table 1  
*Descriptive Statistics of WAIS-IV Subtests*

Subtest	<i>M</i> (Raw)	<i>SD</i>	Min. Score (Raw)	Max. Score (Raw)	Skewness	Kurtosis
Block Design	38.30	13.24	3	66	-.04	-.88
Matrix Reasoning	18.52	4.70	2	26	-1.13	.77
Visual Puzzles	14.53	4.75	5	25	.11	-.93
Similarities	25.84	5.75	6	35	-.86	.55
Vocabulary	37.52	11.11	4	56	-.65	-.08
Information	14.28	5.26	0	25	-.29	-.55
Arithmetic	12.58	3.68	3	22	.22	-.70
Digits Forward	9.60	2.51	0	16	.02	.46
Digits Backward	8.51	2.51	0	16	.35	.36
Digit Sequencing	8.74	2.38	0	15	-.30	.96

The range of raw scores for Block Design, Digits Forward, Digits Backward, Matrix Reasoning and Arithmetic extended to the maximum possible score. The range of

raw scores for Information and Digits Forward, Digits Backward and Digit Sequencing went as low as the minimum possible score. Block Design had the highest average raw score, likely because the maximum possible score for this subtest (i.e., 66) is higher than other subtests. Of all subtests, Arithmetic had the lowest average raw score, likely because the maximum possible total raw score for this subtest (i.e., 22) is lower than any other subtest. Block Design and Vocabulary had noticeably higher standard deviations than the other subtests included in analyses.

### *Hypothesis 1*

The first hypothesis asserted that the items that comprise the Block Design, Similarities, Vocabulary, Digits Forward, Digits Backward and Digit Sequencing subtests of the WAIS-IV do not proceed in order of difficulty. In order to analyze this hypothesis, performance on these subtests was examined using a partial credit model, which estimates the item difficulty and person ability levels in logits, which are arguably on the interval level (Bond & Fox, 2007; Embretson and Reise, 2000).

For subtests where the items have two or more possible response options and different thresholds, a partial credit model was used to examine test performance, which

is mathematically represented as  $\frac{e^{\sum_{k=0}^x (\theta_j - \tau_{ki})}}{\sum_{x=0}^m e^{\sum_{k=0}^x (\theta_j - \tau_{ki})}}$  where  $m$  is the maximum score for

the item,  $\theta_j$  is the ability level of an individual,  $\beta_i$  is the difficulty level of the item and  $\tau_{ki}$  is the threshold of the rating scale of the item (Bond & Fox, 2007; Embretson & Reise, 2000). A partial credit model was employed for Block Design, Similarities and Vocabulary.

### ***Block Design***

Block Design items one through four occur before the start point. Because of this, there were a limited number of responses provided to items one through three. As a result, the partial credit model was unable to produce item and person statistics for these items (Linacre, 1997a). As a result, items one through three were excluded from analyses. This produced a total of 11 items available for analyses.

Examining the responses to the items of the Block Design subtest also revealed that four had few people who provided responses that fell in the 0 category. The number of responses to these categories were so few that it was inadvisable to include this items without first combining the 0 and 1 response categories (Bond & Fox, 2007). This was done by recoding 1 point responses into 0 point responses using SPSS. This increased the frequency of responses and allowed for stable item statistics to be calculated.

### ***Ordered Andrich Thresholds, Observed Average and Sample Expectations***

One of the first steps in using a Rasch or partial credit model is to explore items' Andrich thresholds. Andrich thresholds are calculated by creating probability curves for each response option against the range of possible trait levels. Andrich thresholds are the points at which the probability curves for adjacent categories overlap. In order for Andrich thresholds to be calculated, participants must provide several responses to each response category; otherwise, the measurement model is unable to produce stable item calibrations (Bond & Fox, 2007).

Andrich thresholds refer to the ability level needed for participants to have a 50% chance of picking or generating one response versus picking or generating an adjacent response (Embretson & Reise, 2000). For example, in a rating scale where 0 is

equivalent to an incorrect answer, 1 is equivalent to a partially correct answer and 2 is equivalent to a correct answer, thresholds exist between 0 and 1 and 1 and 2. Multiple Andrich thresholds only exist for items with more than two possible responses (e.g., incorrect, partially correct, correct), referred to as polytymous rating scales.

The standard dichotomous Rasch model and the partial credit model assume that items have ordered Andrich thresholds. Ordered Andrich thresholds ensure that the probability of obtaining a higher score on an item requires a higher trait level. For example, Andrich thresholds are ordered when the ability level needed for a participant to have a 50% chance of generating or picking a response option worth fewer points (e.g., a 1 point “partially correct” response) is less than the ability level needed for participants to have a 50% chance of generating or picking a response option worth more points (e.g., a 2 point “correct” response).

Disordered thresholds indicate that the rating scale is not performing in the way the Rasch or partial credit model expect (Bond & Fox, 2007). They are indicative of a problem with the way in which individuals receive credit on an item. For example, disordered thresholds can occur when there is a mix of individuals with high trait levels being more likely to obtain a lower score. When disordered thresholds exist, it is necessary to adjust the rating scale so that it aligns with model expectations. Adjusting disordered thresholds often requires combining response categories (Embretson & Reise, 2000). When the threshold between lower categories (e.g., between 0 and 1) is greater than the threshold between higher response categories (e.g., between 1 and 2), these response categories may be combined. The combined responses are then assigned the lower point value.

Because items five through eight are dichotomous, there are not multiple Andrich thresholds for these items. Because Block Design items nine through 14 contain at least three possible response categories, they are polytymous and therefore Andrich thresholds were examined for these items.

The results displayed in Table 2 show the Andrich thresholds for Block Design items with five response options. These results revealed that only item 10 had ordered thresholds. Items 9, 11, 12, 13 and 14 have disordered thresholds, with the threshold between the 5 and 6 response categories being lower than the threshold between the 4 and 5 categories.

Table 2  
*Block Design Andrich Thresholds*

Item	Andrich Threshold			
	Between 0 & 4	Between 4 & 5	Between 5 & 6	Between 6 & 7
9	-1.29	-.04	-.87	2.20
10	-6.46	-.19	.03	3.08
11	-1.33	.42	-.94	1.85
12	-2.25	.14	-1.13	3.24
13	-2.45	.37	-1.01	3.09
14	-1.69	.24	-.67	2.12

To eliminate the disordered thresholds, response categories 4, 5 and 6 were combined for items 9, 11, 12, 13 and 14 by recoding responses using SPSS. The resulting categories for these items were 0, 4 (the combination of all 4, 5 and 6 responses) and 7.

In the same way that items are intercorrelated in multiple regression analyses, Rasch modeling assumes that the items that comprise a task are intercorrelated. As a result, combining response categories for some items will change the Andrich thresholds



of the other items. Therefore, after combining response categories for items 9 and 11 through 14, data were reanalyzed using Rasch measurement model software.

The thresholds for item 10, which has a 5-point rating scale, are as follows: results revealed that the threshold between 0 and 4 was -2.37, the threshold between 4 and 5 was -1.19, the threshold between 5 and 6 was -.11 and the threshold between 6 and 7 was 3.67.

Results for items with a 3-point rating scale (items 9 and 11 through 14) are displayed in Table 3. These results reveal that all thresholds are ordered.

Table 3  
*Andrich Thresholds After Combining Response Categories*

Item	Andrich Threshold	
	Between 0 & 4	Between 4 & 7
9	-1.79	3.17
11	-2.24	2.24
12	-3.05	3.05
13	-2.84	2.84
14	-1.77	1.77

Next, both observed averages and sample expectations were examined. Observed averages are the average ability levels required for participants in this sample to generate or pick a response in a certain category. Sample expectations are estimates of the average ability level needed for participants to make a response in a certain category (e.g., correct or incorrect) (Bond & Fox, 2007).

The partial credit model assumes that observed averages and sample expectations increase as category, or answer, value increases (Embretson & Reise, 2000). This means that both the observed averages and sample expectations for wrong responses (i.e., those worth 0 point) should be lower than the observed averages and sample expectations for

partially correct responses (i.e., those worth 1 point), which should be lower than the observed averages and sample expectations for correct answers (i.e., those worth 2 points).

Table 4 shows the observed averages for each response option of each item of the Block Design subtest. Results revealed that all observed averages are ordered, suggesting that higher ability levels are needed for participants to provide a response in a category worth greater points.

Table 4  
*Observed Averages After Combining Items' Response Categories*

Item	<u>Observed Average</u>					
	0	2	4	5	6	7
4	-5.69	-2.43	-	-	-	-
5	-2.86	-	1.50	-	-	-
6	-2.54	-	1.41	-	-	-
7	-3.18	-	1.49	-	-	-
8	-3.13	-	1.67	-	-	-
9	-.83	-	2.44	-	-	4.86
10	-.31	-	1.51	2.63	3.81	5.64
11	.57	-	2.57	-	-	4.94
12	1.05	-	3.01	-	-	5.40
13	1.90	-	3.22	-	-	4.48
14	2.37	-	3.79	-	-	5.40

Table 5 shows the sample expectations for each response option for each item of the Block Design subtest. Results revealed that all sample expectations are ordered.

Table 5  
*Sample Expectation After Combining Items' Response Categories*

Item	Sample Expect					
	0	2	4	5	6	7
4	-5.06	-2.51	-	-	-	-
5	-3.40	-	1.52	-	-	-
6	-3.81	-	1.44	-	-	-
7	-3.44	-	1.50	-	-	-
8	-2.72	-	1.64	-	-	-
9	-1.07	-	2.49	-	-	4.56
10	-.33	-	1.45	2.63	3.94	5.61
11	.78	-	2.49	-	-	4.66
12	1.34	-	2.88	-	-	5.26
13	1.88	-	3.20	-	-	5.40
14	2.39	-	3.74	-	-	5.59

**Item Fit**

Next, item fit was examined. Item outfit mean square equals the sum of the standardized residuals squared divided by the number of subjects, where the standardized residual equals the residual divided by the square root of the variance. It approximates a chi square distribution (Linacre, 2014e). Mathematically, this is represented as  $Infit = \frac{\sum [(residual^2 / model\ variance) * model\ variance]}{\sum (model\ variance)} = \text{average} [(standardized\ residuals)^2 * model\ variance] = \text{model variance-weighted mean-square}$ . Outfit is the chi square divided by degrees of freedom and is mathematically represented as  $outfit = \frac{\sum (residual^2 / model\ variance)}{(count\ of\ residuals)} = \text{average} [(standardized\ residuals)^2] = \text{chi-square/degrees of freedom} = \text{mean-square}$ .

Infit mean square infit and outfit values quantify the extent to which items conform to measurement model expectations (Bond & Fox, 2007). Items with both mean square infit and outfit values greater than 1.3 indicate response patterns that are “too haphazard” (Bond & Fox, 2007, p. 240) and stable item statistics cannot be calculated because participants responded in extremely unexpected ways to these items (Bond & Fox, 2007).

Table 6 shows that mean square infit values ranged from .71 to 1.23 while mean square outfit values ranged from .30 to 9.90. No items were misfitting.

Table 6  
*Block Design Item Fit Statistics*

Item	Infit Mean Square	Outfit Mean Square
4	.71	.30
5	1.18	1.02
6	1.17	7.87
7	.90	9.90
8	.74	.63
9	1.23	1.69
10	1.11	1.02
11	.78	.78
12	.78	.71
13	1.05	1.07
14	.96	.90

### *Dimensionality*

When using a standard dichotomous Rasch or partial credit model, it is important

to assess the dimensionality of a measure. Dimensionality refers to how many latent constructs a test measures. Both models used in these analyses assume that tests measure only one underlying construct (Bond & Fox, 2007; Embretson & Reise, 2007), which is called unidimensionality.

Raw explained variance, which is an unstandardized value that quantifies how much of participants' performance can be attributed to their ability and item difficulty, was examined first. The raw explained variance is calculated in the same manner as would be done during an exploratory factor analysis. The two components entered into this analysis are participants' ability levels and item difficulty. Results (Table 7) revealed that the raw variance explained by person ability and item difficulty was 73.5%, which can be interpreted as "good" unidimensionality (Fisher, 2007).

Table 7  
*Dimensionality of the Block Design Subtest*

	Raw Variance	Percentage
Total Variance Explained	29.4	73.5%
Variance Explained by Persons	14.4	36.0%
Variance Explained by Items	15.0	37.5%
Total Unexplained Variance	11.0	26.5%
Unexplained Variance Accounted for by 1 <sup>st</sup> Factor	1.6	14.9%
Unexplained Variance Accounted for by 2 <sup>nd</sup> Factor	1.4	12.7%
Unexplained Variance Accounted for by 3 <sup>rd</sup> Factor	1.3	11.6%
Unexplained Variance Accounted for by 4 <sup>th</sup> Factor	1.3	11.5%
Unexplained Variance accounted for by 5 <sup>th</sup> Factor	1.1	10.2%

The next step in assessing dimensionality is to look at the unexplained variance to

determine whether there is another factor, in addition to item difficulty and person ability, that may be systematically affecting task performance (Linacre, 2014a). When another factor is systematically and significantly affecting task performance, this is indicative of multidimensionality (Bond & Fox, 2007).

To determine whether another factor can explain the unexplained, or residual, variance, a principal component analysis is performed on the residual variance. The principal component analysis identifies groups of items that account for portions of the residual variance. These groups of items are referred to as factors. The first factor is the one that explains the greatest portion of the residual variance. While there may be a number of these factors, Winsteps software includes up to five factors. Fewer factors are included if there are no items comprising additional factors (Linacre, personal communication, June 2, 2014).

Because the first factor explains the most residual variance, examining the residual variance requires comparing the unexplained variance accounted for by the first factor with the total residual variance. Results revealed that the unexplained variance accounted for by the first factor was 1.6 while total raw residual variance was 11. This means that 14.5% [ $1.6/11$ ] of the total unexplained variance was accounted for by the first factor. This is just below the suggested cutoff of 15% (Linacre, 2013), demonstrating the unidimensionality of the Block Design subtest.

Finally, the raw unexplained variance accounted for by all factors was examined. Guidelines suggest that when the raw values of all factors are less than 2, a measure is unidimensional (Linacre, 2014a). Results revealed that the raw unexplained variance of factors one through five ranged from 1.1 to 1.6, indicating that Block Design is

unidimensional.

### ***Person Fit***

Person fit was evaluated next, to determine how well participants' responses aligned with expectations of the standard dichotomous Rasch or partial credit models' expectations. As with misfitting items, unstandardized mean square infit and outfit values are used to identify misfitting participants. These values are calculated in the same manner as item fit values, described above (Bond & Fox, 2007).

When both infit and outfit mean square values are greater than 1.3, participants are considered misfitting, indicating that they responded in unexpected ways (Bond & Fox, 2007). When a large number of participants are misfitting, employing a standard dichotomous Rasch or partial credit model is not useful because the model is unable to predict participant performance.

Results revealed that mean square person infit values ranged from .10 to 6.00 ( $M = .96$ ,  $SD = .86$ ) and outfit mean square values ranged from .05 to 9.90 ( $M = .84$ ,  $SD = 1.67$ ). Results also showed that 8.66% of participants ( $n = 26$ ) responded in unexpected ways. However, this did not prevent stable estimates of item difficulty and person ability from being made.

### ***Person Ability***

Person ability levels were examined next. Person ability is estimated via a Rasch or partial credit model using joint maximum likelihood estimation, which optimizes the model to data fit through an iterative calculus based approach. Similar to multiple regression, where the estimate method is ordinary least squares, the Rasch model and partial credit model uses joint maximum likelihood estimation, where it maximizes model

to data fit.

Person ability is the estimate of an amount of attribute a person has, for example, average perceptual reasoning abilities. It is reported on the logit scale, where a lower value indicates a lower level of a trait, such as verbal abilities, and a higher value represents a higher level of a trait (Furr & Bacharach, 2008). A wide range of person ability helps ensure the accurate calculation of item statistics, such as item difficulty (Embretson & Reise, 2000).

Results revealed that ability levels in this sample ranged from -7.48 to 10.00 ( $M = 1.32$ ,  $SD = 2.57$ ), with fewer individuals having extreme ability levels. With respect to variability, the first quartile was 0.00, the second quartile was 1.52, the third quartile was 3.20, and the inter-quartile range (IQR) was 3.20 [3.20 - 0.00]. These results indicate that there was a large range of ability represented in the sample.

### ***Person Reliability***

Person reliability was examined next. Person reliability is analogous to classical test theory reliability as would be typically estimated using Cronbach's alpha. It can be interpreted in a manner similar to which test score reliability would be interpreted. Person reliability values above .50 are generally acceptable (Linacre, 2014c). Log-likelihood chi square values can also be examined to explore reliability. Significant chi square values represent significant departures from actual performance and model estimates (Linacre, 2014d).

The person reliability was .85, suggesting that the range of ability in the sample was sufficient, the items were sufficiently difficult, there were a suitable number of response options and the length of the task was appropriate (Linacre 2014b). Additionally,



the non-significant chi-square value of 1991.68,  $p > .99$  indicates a good fit of the persons and the items to the model (Embretson & Reise, 2000).

### ***Item Reliability***

Item reliability, which refers to the reproducibility of raw item scores, was examined next (Linacre, 199 a). Item reliability values fall between 0 and 1, with higher scores indicating greater reliability (Bond & Fox, 2007). The item reliability of .99 revealed that the sample was sufficiently large, there was an adequate range of item difficulty and the item statistics produced from these analyses are highly likely to be replicated if given to a sample with a similar range of ability levels (Linacre, 2014b).

### ***Item Coverage***

Item coverage was examined next. Item coverage is the extent to which items span the entire continuum of participants' ability levels. Item difficulty and person ability levels are reported in logits. By using a logit scale, it has been argued that what often are ordinal level raw scores are converted to interval level scores, whereby both individual person ability levels and item difficulties can be placed on the same continuum, as shown in the item to person map (Figure 1) (Embretson & Reise, 2000). Person abilities appear to the left of the graph and item difficulties appear on the right.

Figure 1 shows that there was adequate coverage for most of the sample. Item difficulty levels were fairly well matched to participants' ability levels. However, there were several gaps in coverage. There were no items to differentiate individuals whose ability levels were very below average (ability level of -7), below average to average (ability levels between -2 and 0), slightly above average (ability levels between 1 and 3), and extremely above average (ability levels between 6 and 9).

Figure 1. *Block Design Item Coverage*

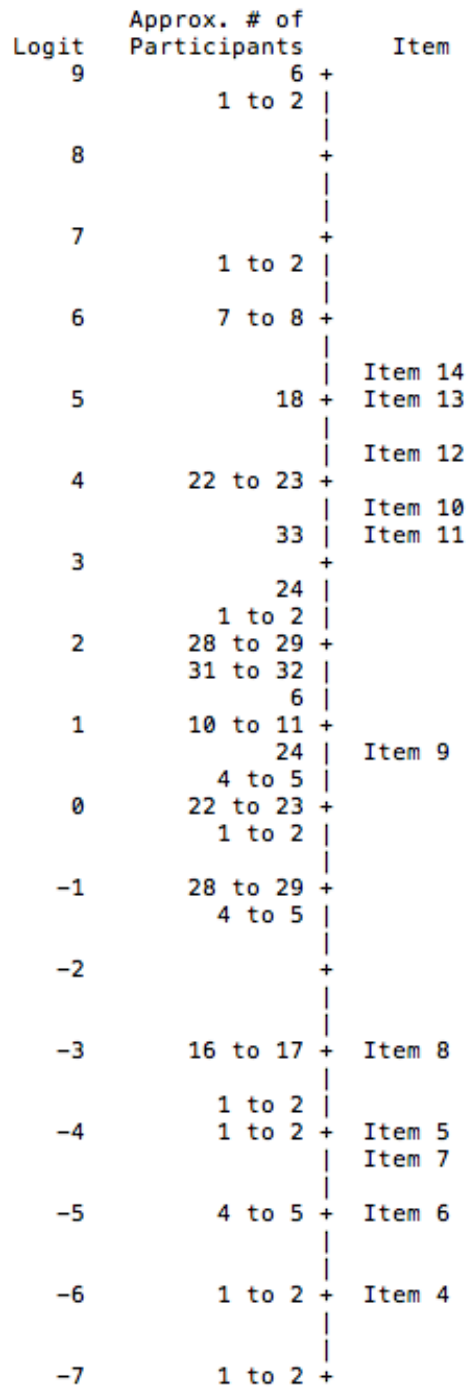


Figure 1. Item difficulty and person ability levels are displayed in logits, which appear on the extreme left. The number of participants with a certain ability level appears to the left of the vertical line. Items appear to the right of the line. Participants' locations correspond to their ability level while items' locations correspond with their difficulty level.

### ***Item Discrimination***

Item discrimination refers to an item's ability to differentiate individuals of varying ability levels. It is indicative of the relationship between the item and the underlying construct being measured (Furr & Bacharach, 2013). Ideally, in Rasch modeling, all item discrimination values are assumed to be equal to one (Bond & Fox, 2007). That is, all items are assumed to differentiate individuals equally, given that items' level of difficulty. Because item discrimination values are assumed to be equal to one, they are not calculated in Rasch modeling. However, it is empirically unlikely that all item discrimination values will actually equal one (Linacre, 2014b). Instead, guidelines suggest that all item discrimination values should be between 0.5 and 1.5 (Linacre, 2009). An item with a discrimination outside these values indicates that the item does not distinguish between individuals with high and low ability levels as would be expected given that item's level of difficulty (Linacre, 2014b). To evaluate how well item discrimination values conform to Rasch modeling assumptions, item discrimination was estimated outside the Rasch model using Winsteps software.

The results revealed that discrimination values ranged from .75 to 1.32. All values fell within the guidelines of .5 to 1.5, indicating that items discriminated approximately as would be expected in the model.

### ***Item Difficulty***

Item difficulty was examined using the partial credit model in order to assess the assertion of the hypothesis, which stated that the items that comprise the Block Design do not proceed in order of difficulty. Similar to multiple regression, where the estimate method is ordinary least squares, the Rasch model uses joint maximum likelihood

estimation, where it maximizes model to data fit. Item difficulty is estimated via a Rasch model using joint maximum likelihood estimation, which optimizes the model to data fit through an iterative calculus based approach. Item difficulty represents how difficult it is to answer an item correctly (or partially correctly), is estimated in logits and exists on a true interval scale (Bond & Fox, 2007).

The results (Table 8) revealed that item difficulty values ranged from -6.13 to 5.34 and not all items proceeded in order of difficulty. Items six, seven and 11 were disordered. Item six had a difficulty of -5.23 while item five had a difficulty of -4.32. Item seven (-4.37) was easier than items five (-4.32) and six (-5.23). Finally, item 11 had a difficulty of 3.28 while item 10 had a difficulty of 3.44.

Table 8  
*Block Design Item Difficulty*

Item	Difficulty
4	-6.13
5	-4.32
<b>6</b>	<b>-5.23</b>
<b>7</b>	<b>-4.37</b>
8	-3.28
9	2.17
10	3.44
<b>11</b>	<b>3.28</b>
12	4.33
13	4.77
14	5.34

*Note.* Disordered items are in boldface.

It is possible that items one through three of the Block Design subtest were disordered. However, as previously discussed, there were too few responses provided to these items and these items were excluded from analyses.

The disordered items occur towards the beginning and end of the Block Design subtest. Because two of the disordered items are consecutive it is possible that individuals would meet the discontinue criteria before having the chance to earn points on an item that was easier than those they had already been administered.

The results revealed that items 6, 7 and 11 of the Block Design subtest are disordered. These results provide support for the hypothesis, which asserted that the items of the Block Design subtest would be disordered. However, the remaining items were consistent. As a result, the hypothesis for the Block Design subtest was partially accepted.

### *Similarities*

Items one through three occur before the start point and therefore were not administered to all participants. Because of this, there were only a few responses provided to items one and two. The number of responses was so few that standard dichotomous Rasch model was unable to produce statistics for these items (Linacre, 1997a). Thus, items one and two were excluded from analyses. This left a total of 16 items available for analyses in the Similarities subtest.

Examining the responses to the items of the Similarities subtest also revealed that items three, four and five had few people who were entirely unable to answer the questions (0) or provide a partially correct response to (1). The number of responses that received scores of 0 or 1 were so few that it was inadvisable to include these items

without first combining the 0 and 1 response categories (Bond & Fox, 2007).

***Ordered Andrich Thresholds, Observed Average and Sample Expectations***

One of the first steps in using a Rasch or partial credit model is to explore items' Andrich thresholds. Andrich thresholds are calculated by creating probability curves for each response option against the range of possible trait levels. Andrich thresholds are the points at which the probability curves for adjacent categories overlap. In order for Andrich thresholds to be calculated, participants must provide several responses to each response category; otherwise, the measurement model is unable to produce stable item calibrations (Bond & Fox, 2007).

Andrich thresholds refer to the ability level needed for participants to have a 50% chance of picking or generating one response versus picking or generating an adjacent response (Embretson & Reise, 2000). For example, in a rating scale where 0 is equivalent to an incorrect answer, 1 is equivalent to a partially correct answer and 2 is equivalent to a correct answer, thresholds exist between 0 and 1 and 1 and 2. Multiple Andrich thresholds only exist for items with more than two possible responses (e.g., incorrect, partially correct, correct), referred to as polytymous rating scales.

The standard dichotomous Rasch model and the partial credit model assume that items have ordered Andrich thresholds. Ordered Andrich thresholds ensure that the probability of obtaining a higher score on an item requires a higher trait level. For example, Andrich thresholds are ordered when the ability level needed for a participant to have a 50% chance of generating or picking a response option worth fewer points (e.g., a 1 point "partially correct" response) is less than the ability level needed for participants to have a 50% chance of generating or picking a response option worth more points (e.g., a

2 point “correct” response). Disordered thresholds indicate that the rating scale is not performing in the way the Rasch or partial credit model expect (Bond & Fox, 2007).

When disordered thresholds exist, it is necessary to adjust the rating scale so that it aligns with model expectations. Disordered thresholds are indicative of a problem with the way in which individuals receive credit on an item. For example, disordered thresholds can occur when there is a mix of individuals with high trait levels being more likely to obtain a lower score.

Adjusting disordered thresholds often requires combining response categories (Embretson & Reise, 2000). When the threshold between lower categories (e.g., between 0 and 1) is greater than the threshold between higher response categories (e.g., between 1 and 2), these response categories may be combined. The combined responses are then assigned the lower point value.

Because the Similarities subtest contains three possible response categories, it is a polytymous rating scale and therefore the Andrich thresholds must be examined. As previously discussed, items three, four and five had few people who provided responses that fell in the 0 and 1 categories. The 0 and 1 categories were combined, making the rating scale for these items dichotomous. As a result, there were not multiple Andrich thresholds for items three through five.

After combining response categories for items three through five, Andrich thresholds for the remaining items were examined. The results displayed in Table 9 revealed that items 8, 10, 11, 13 and 16 had disordered Andrich thresholds, suggesting that the response categories did not perform as the model would expect.

Table 9  
*Similarities Items' Andrich Thresholds*

Item	Andrich Threshold	
	Between 0 & 1	Between 1 & 2
6	-.15	.15
7	-1.10	1.10
<b>8</b>	<b>.99</b>	<b>-.99</b>
9	-.47	.47
<b>10</b>	<b>.55</b>	<b>-.55</b>
<b>11</b>	<b>.22</b>	<b>-.22</b>
12	-.40	.40
<b>13</b>	<b>.37</b>	<b>-.37</b>
14	-1.32	1.32
15	-.30	.30
<b>16</b>	<b>.12</b>	<b>-.12</b>
17	-1.17	1.17
18	-.60	.60

*Note.* Disordered thresholds are in boldface.

To eliminate the disordered thresholds, categories 0 and 1 were combined for items 8, 10, 11, 13, and 16. The resulting categories were 0 (the combination of all 0 and 1 responses) and 2, making the rating scale for these items dichotomous. As a result, there are not multiple Andrich thresholds for these items.

In the same way that items are intercorrelated in multiple regression analyses, the partial credit model assumes that the items that comprise a task are intercorrelated. As a result, combining response categories for some items will change the Andrich thresholds



of the other items. Therefore, after combining response categories, all data was reanalyzed. The results displayed in Table 10 revealed that the Andrich thresholds for all included items remained ordered.

Table 10  
*Similarities Andrich Thresholds After Combining Items' Response Categories*

Item	Andrich Threshold	
	Between 0 & 1	Between 1 & 2
6	-.23	.23
7	-1.29	1.29
9	-.60	.60
12	-.50	.50
14	-1.43	1.43
15	-.38	.38
17	-1.26	1.26
18	-.68	.68

Next, both observed averages and sample expectations were examined. Observed averages are the average ability levels required for participants in this sample to generate or pick a response in a certain category. Sample expectations are estimates of the average ability level needed for participants to make a response in a certain response category (Bond & Fox, 2007). The partial credit model assumes that observed averages and sample expectations increase as answer value increases (Embretson & Reise, 2000).

Table 11 shows the observed and sample expectations for the items of the Similarities subtest, all of which are all ordered.

Table 11

*Similarities Average & Expected Ability Levels After Combining Response Categories*

Item	Observed Average			Sample Expect		
	0	1	2	0	1	2
3	-1.49	.25	-	-3.11	.62	-
4	-.35	1.26	-	-1.01	1.28	-
5	-.66	1.25	-	-1.11	1.25	-
6	-1.30	-.58	1.39	-2.01	-.41	1.39
7	-2.01	.28	1.58	-1.59	.14	1.61
8	-.09	1.61	-	-.01	1.59	-
9	-.71	.60	1.79	-.81	.63	1.80
10	.18	1.67	-	.21	1.66	-
11	.55	1.81	-	.55	1.81	-
12	-.14	1.05	1.97	-.05	.97	1.99
13	.74	1.90	-	.74	1.90	-
14	.35	1.32	2.28	.36	1.28	2.36
15	.62	1.47	2.23	.62	1.37	2.32
16	1.02	2.54	-	1.14	2.28	-
17	.86	2.00	2.83	.97	1.84	3.01
18	1.24	1.96	2.92	1.19	2.00	3.06

***Item Fit***

Next, item fit was examined. Item outfit mean square equals the sum of the standardized residuals squared divided by the number of subjects, where the standardized residual equals the residual divided by the square root of the variance. It approximates a

chi square distribution (20014e). Mathematically, this is represented as  $\text{Infit} = \frac{\sum [(\text{residual}^2 / \text{model variance}) * \text{model variance}]}{\sum (\text{model variance})} = \text{average} [(\text{standardized residuals})^2 * \text{model variance}] = \text{model variance-weighted mean-square}$ .  
 Outfit is the chi square divided by degrees of freedom and is mathematically represented as  $\text{outfit} = \frac{\sum (\text{residual}^2 / \text{model variance})}{(\text{count of residuals})} = \text{average} [(\text{standardized residuals})^2] = \text{chi-square/degrees of freedom} = \text{mean-square}$ .

Infit mean square infit and outfit values quantify the extent to which items conform to measurement model expectations (Bond & Fox, 2007). Items with both mean square infit and outfit values greater than 1.3 indicate response patterns that are “too haphazard” (Bond & Fox, 2007, p. 240). When both values are above 1.3, stable item statistics cannot be calculated because participants responded in extremely unexpected ways to these items (Bond & Fox, 2007).

Results revealed that mean square infit values ranged from .80 to 2.46. Mean square outfit values ranged from .76 to 2.83. Item 3 was found to be misfitting, as both the mean square infit and outfit values were greater than 1.3. Therefore, before further analyzing the data, item 3 was removed. This yielded a total of 15 items available for the rest of the analyses.

In the same way that items are intercorrelated in multiple regression analyses, the partial credit model assumes that the items that comprise a task are intercorrelated. As a result, removing items will change the Andrich thresholds of the other items. Therefore, after removing item three, categories, all data was reanalyzed. Because item three was removed, new infit and outfit mean square statistics were calculated and are displayed in Table 12.

Table 12  
*Similarities Item Fit Statistics with Misfitting Item Removed*

Item	Infit Mean Square	Outfit Mean Square
4	1.08	1.21
5	.98	1.24
6	1.06	1.50
7	1.02	.93
8	.96	.90
9	1.04	.99
10	.97	.95
11	.97	1.16
12	.96	1.00
13	1.00	1.02
14	1.04	1.04
15	1.06	1.12
16	.80	.73
17	.93	.92
18	1.11	1.12

After removing item three, results revealed that mean square infit values ranged from .80 to 1.08. Mean square outfit values ranged from .73 to 1.50. No item had both infit and outfit values above the suggested cutoff of 1.3. Although item six had a high outfit value of 1.50, the infit value of 1.06 was within the acceptable range, indicating that this item was not substantially misfitting.

***Person Fit***

Person fit was evaluated next, to determine how well participants' responses aligned with expectations of the standard dichotomous Rasch or partial credit models' expectations. As with misfitting items, unstandardized mean square infit and outfit values are used to identify misfitting participants. These values are calculated in the same manner as item fit values, described above (Bond & Fox, 2007).

When both infit and outfit mean square values are greater than 1.3, participants are considered misfitting, indicating that they responded in unexpected ways (Bond & Fox, 2007). When a large number of participants are misfitting, employing a standard dichotomous Rasch or partial credit model is not useful because the model is unable to predict participant performance.

Results revealed that mean square person infit values ranged from .30 to 3.15 ( $M = .99, SD = .46$ ) and outfit mean square values ranged from .12 to 9.90 ( $M = .99, SD = 1.16$ ). Results also showed that 11% of participants ( $n = 33$ ) responded in unexpected ways.

***Dimensionality***

When using a standard dichotomous Rasch or partial credit model, it is important to assess the dimensionality of a measure. Dimensionality refers to how many latent constructs a test measures. The standard dichotomous Rasch model and partial credit model assume that tests measure only one underlying construct (Bond & Fox, 2007; Embretson & Reise, 2007), which is called unidimensionality.

First, an examination of the total raw explained variance was used to gauge the dimensionality of the Similarities subtest. Raw explained variance is an unstandardized

value that quantifies how much of participants' performance can be attributed to their ability and item difficulty. In the standard dichotomous Rasch model and the partial credit model, the raw explained variance is calculated in the same manner as would be done during an exploratory factor analysis. The two components entered into this analysis are participants' ability levels and item difficulty.

The results displayed in Table 13 revealed that the raw variance explained by person ability and item difficulty was 52.8%. Guidelines indicate that this can be interpreted as "good" unidimensionality (Fisher, 2007).

Table 13  
*Dimensionality of the Similarities Subtest*

	Raw Variance	Percentage
Total Variance Explained	16.7	52.8%
Variance Explained by Persons	6.5	20.7%
Variance Explained by Items	10.1	32.1%
Total Unexplained Variance	15.0	47.2%
Unexplained Variance Accounted for by 1 <sup>st</sup> Factor	1.5	9.7%
Unexplained Variance Accounted for by 2 <sup>nd</sup> Factor	1.4	9.3%
Unexplained Variance Accounted for by 3 <sup>rd</sup> Factor	1.3	8.7%
Unexplained Variance Accounted for by 4 <sup>th</sup> Factor	1.3	8.6%
Unexplained Variance Accounted for by 5 <sup>th</sup> Factor	1.3	8.5%

The next step in assessing dimensionality is to look at the unexplained variance to determine whether there is another factor, in addition to item difficulty and person ability, that may be systematically affecting task performance (Linacre, 2014a). When another factor is systematically and significantly affecting task performance, this is indicative of

multidimensionality (Bond & Fox, 2007).

To determine whether another factor can explain the unexplained, or residual, variance, a principal component analysis is performed on the residual variance. The principal component analysis identifies groups of items that account for portions of the residual variance. These groups of items are referred to as factors. The first factor is the one that explains the greatest portion of the residual variance. While there may be a number of these factors, Winsteps software includes up to five factors. Fewer factors are included if there are no items comprising additional factors (Linacre, personal communication, June 2, 2014).

Because the first factor explains the most residual variance, examining the residual variance requires comparing the unexplained variance accounted for by the first factor with the total residual variance. Results revealed that the unexplained variance accounted for by the first factor was 1.9 while total raw residual variance was 15. This means that 10% [ $1.5/15$ ] of the total unexplained variance was accounted for by the first factor. This is below the suggested cutoff of 15% (Linacre, 2013), demonstrating the unidimensionality of the Similarities subtest.

Next, the raw unexplained variance accounted for by all factors was examined. Guidelines suggest that when the raw values of all factors are less than 2, a measure is unidimensional (Linacre, 2014a). Results revealed that the raw unexplained variance of factors one through five ranged from 1.3 to 1.5, indicating that Similarities is unidimensional.

### ***Person Ability***

Person ability levels were examined next. Person ability is estimated via a Rasch

or partial credit model using joint maximum likelihood estimation, which optimizes the model to data fit through an iterative calculus based approach. Similar to multiple regression, where the estimate method is ordinary least squares, the Rasch model and partial credit model uses joint maximum likelihood estimation, where it maximizes model to data fit.

Person ability is the estimate of an amount of attribute a person has, for example, average verbal abstraction abilities. It is reported on the logit scale, where a lower value indicates a lower level of a trait, such as verbal abilities, and a higher value represents a higher level of a trait (Furr & Bacharach, 2008). A wide range of person ability helps ensure the accurate calculation of item statistics, such as item difficulty (Embretson & Reise, 2000).

Results revealed that ability levels in this sample ranged from -5.98 to 4.63 ( $M = 1.14$ ,  $SD = 1.46$ ), with fewer individuals having extreme ability levels. As will be discussed later, the left side of the item to person map (Figure 2) provides a graphical distribution of person abilities.

### ***Person Reliability***

Person reliability was examined next. Person reliability is analogous to classical test theory reliability as would be typically estimated using Cronbach's alpha. It can be interpreted in a manner similar to which test score reliability would be interpreted. Person reliability values above .50 are generally acceptable (Linacre, 2014c). Log-likelihood chi square values can also be examined to explore reliability. Significant chi square values represent significant departures from actual performance and model estimates (Linacre, 2014d).



The person reliability was .77, suggesting that the range of ability in the sample was sufficient, the items were sufficiently difficult, there were a suitable number of response options and the length of the task was appropriate (Linacre 2014b). However, the significant chi-square value of 4464.09,  $p = .01$  produced by the standard dichotomous Rasch or partial credit model indicates a departure of the data from Rasch modeling expectations (Embretson & Reise, 2000).

### ***Item Reliability***

Item reliability, which refers to the reproducibility of raw item scores, was examined next (Linacre, 1997a). Item reliability values fall between 0 and 1, with higher scores indicating greater reliability (Bond & Fox, 2007). The item reliability of .99 revealed that the sample was sufficiently large, there was an adequate range of item difficulty and the item statistics produced from these analyses are highly likely to be replicated (Linacre, 2014b).

### ***Item Coverage***

Item coverage was examined next. Item coverage is the extent to which items span the entire continuum of participants' ability levels. Item difficulty and person ability levels are reported in logits. By using a logit scale, it has been argued that what often are ordinal level raw scores are converted to interval level scores, whereby both individual person ability levels and item difficulties can be placed on the same continuum, as shown in the item to person map (Figure 2) (Embretson & Reise, 2000). Person abilities appear to the left of the graph and item difficulties appear on the right.

Upon examination of the item to person map displayed in Figure 2, it is evident that there is a range of items to differentiate people with low ability levels to above

average ability levels (-4 to just under 3 logits). However, it is worth noting there is a lack of coverage for individuals with very below average (e.g. -5 logits) and very above average (e.g. 4 to 5 logits) ability levels.

Figure 2. *Similarities Item Coverage*

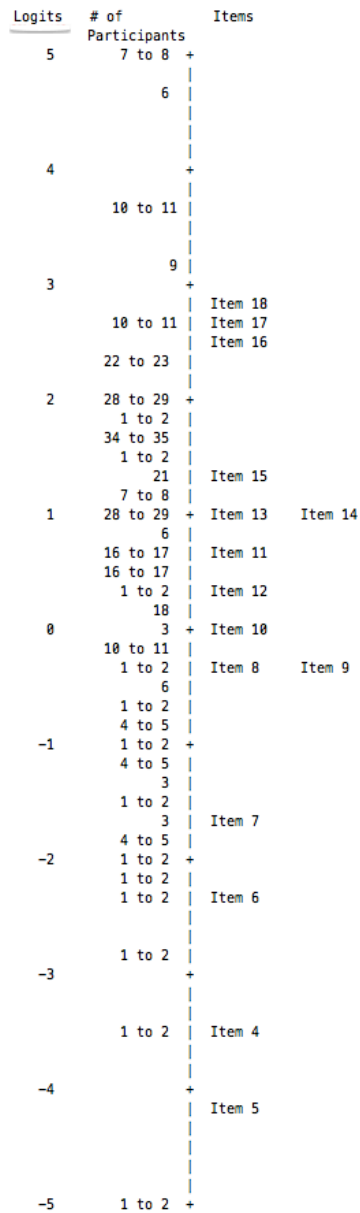


Figure 2. Item difficulty and person ability levels are displayed in logits, which appear on the extreme left. The number of participants with a certain ability level appears to the left of the vertical line. Items appear to the right of the line. Participants' locations correspond to their ability level while items' locations correspond with their difficulty level.

### ***Item Discrimination***

Item discrimination refers to an item's ability to differentiate individuals of varying ability levels. It is indicative of the relationship between the item and the underlying construct being measured (Furr & Bacharach, 2013).

Ideally, in Rasch modeling, all item discrimination values are assumed to be equal to one (Bond & Fox, 2007). That is, all items are assumed to differentiate individuals equally, given that items' level of difficulty. Because item discrimination values are assumed to be equal to one, they are not calculated in Rasch modeling. However, it is empirically unlikely that all item discrimination values will actually equal one (Linacre, 2014b). Instead, guidelines suggest that all item discrimination values should be between 0.5 and 1.5 (Linacre, 2009). An item with a discrimination outside these values indicates that the item does not distinguish between individuals with high and low ability levels as would be expected given that item's level of difficulty (Linacre, 2014b). To evaluate how well item discrimination values conform to Rasch modeling assumptions, item discrimination was estimated outside the Rasch model using Winsteps software.

The results revealed that discrimination values ranged from .88 to 1.33. All values fell within the guidelines of .5 to 1.5, indicating that items discriminated approximately as would be expected in a Rasch model.

### ***Item Difficulty***

Item difficulty was examined using a partial credit model in order to assess the assertion of the hypothesis, which stated that the items that comprise the Similarities subtest do not proceed in order of difficulty. Similar to multiple regression, where the estimate method is ordinary least squares, the Rasch model uses joint maximum

likelihood estimation, where it maximizes model to data fit. Item difficulty is estimated via a Rasch model using joint maximum likelihood estimation, which optimizes the model to data fit through an iterative calculus based approach. Item difficulty represents how difficult it is to answer an item correctly (or partially correctly), is estimated in logits and exists on a true interval scale, in which the intervals between difficulty levels have a consistent value (Bond & Fox, 2007).

The results displayed in Table 14 revealed that item difficulty values ranged from -4.16 to 2.78. Results also revealed that not all items proceeded in order of difficulty. Items five, 12 and 14 were disordered. All three of these items were easier than the items that immediately preceded them. Item five had a difficulty of -4.16 while item four had a difficulty of -3.53. Item 12 had a difficulty of .39, which made it easier than item 11, which had a difficulty of .69. Finally, item 14 had a difficulty of 1.00 while item 13 had a difficulty of 1.05. These results provide support for the hypothesis, which asserted that the items that comprise the Similarities subtest do not proceed in order of difficulty.

It is possible that items one through three of the Similarities subtest were disordered. However, as previously discussed, because items one and two occur before the start point, there were too few responses provided to these items. As a result, the partial credit model could not produce calculations of item difficulty. Therefore, items one and two were excluded. Additionally, item three was removed because it was misfitting. As a result, the partial credit model could not make stable calculations of item difficulty because response patterns to this item were extremely unexpected. Therefore, item three was excluded because any conclusions drawn based upon the difficulty value of these items would not be reliable and would possibly be erroneous.

Because two of the disordered items, item twelve and fourteen, occur towards the end of the subtest, it is possible that individuals would meet the discontinue criteria before reaching these items. If this happened, individuals would not have the chance to earn points on items that were easier than those they had already been administered.

Table 14

*Similarities Item Difficulty*

Item	Difficulty
4	-3.53
<b>5</b>	<b>-4.16</b>
6	-2.36
7	-1.65
8	-.40
9	-.30
10	-.01
11	.69
<b>12</b>	<b>.39</b>
13	1.05
<b>14</b>	<b>1.00</b>
15	1.33
16	2.46
17	2.71
18	2.78

*Note.* Disordered items are in boldface.

The results revealed that items five, 12 and 14 of the Similarities subtest are disordered, as they are easier than the items that immediately precede them. However, the

remaining items proceeded in order of difficulty. As a result, it was concluded that the results provided some support for the hypothesis.

### ***Vocabulary***

Items one through four occur before the start point of the Vocabulary subtest. Because of this, there were a limited number of responses provided to items one and two. As a result, the partial credit model was unable to produce item and person statistics for items one and two (Linacre, 1997a). As a result, items one and two were excluded from analyses, producing a total of 28 items available for analyses.

Examining the responses to the items of the Vocabulary subtest also revealed that item four had few people who provided responses that fell in the 0 and 1 response categories. The number of responses to these categories were so few that it was inadvisable to include these items without first combining the 0 and 1 response categories (Bond & Fox, 2007).

### ***Ordered Andrich Thresholds, Observed Average and Sample Expectations***

Because the Vocabulary subtest contains three possible response categories, it is a polytomous rating scale and therefore the Andrich thresholds were examined. However, items three and four had dichotomous rating scales, either because the original rating scale was dichotomous or because response categories were combined. As a result, there are not multiple Andrich thresholds for items three and four.

After combining response categories for item four, Andrich thresholds for the remaining response categories were examined. The results displayed in Table 15 showed that items 5, 7, 11, 17, 19, 20, 22 and 23 had ordered thresholds. Items 6, 8-10, 12-16, 18, 21, and 24-30 had disordered thresholds. This suggests that the response categories did

not discriminate as they were expected to according to the Rasch model.

To address these deviations from expectations, for items 6, 8-10, 12-16, 18, 21, and 24-30, all 0 and 1 responses were combined by recoding responses. All data were then reanalyzed.

Results (Table 15) revealed that the Andrich thresholds remained ordered for all items. These results indicate that the rating scale for these items were consistent with the expectations of the partial credit model.

Table 15  
*Vocabulary Items' Andrich Thresholds After Combining Select Items' Response Categories*

Item	Andrich Threshold	
	Between 0 & 1	Between 1 & 2
5	-.01	.01
7	-1.31	1.31
11	-1.19	1.19
17	-.28	.28
19	-.81	.81
20	-.30	.30
22	-.74	.74
23	-.68	.68

The results displayed in Table 16 shows the observed averages and sample expectations for the Vocabulary subtest. Results revealed that the response categories for all items of the Vocabulary subtest had ordered observed averages and sample expectations.

Table 16  
*Vocabulary Observed Averages and Sample Expectations*

Item	Observed Average			Sample Expect		
	0	1	2	0	1	2
3	-4.09	-	-1.03	-3.59	-	-1.14
4	-2.93	-	-.82	-3.31	-	-.80
5	-2.71	-.64	.89	-2.95	-1.19	.90
6	-1.11	-	.97	-1.09	-	.97
7	-.79	-.28	1.00	-2.57	-.80	1.08
8	-.81	-	7.14	-.98	-	5.02
9	-1.18	-	.97	-1.11	-	.96
10	-.51	-	1.17	-.59	-	1.19
11	-1.46	-.16	1.47	-1.45	-.06	1.42
12	-.86	-	1.18	-.57	-	1.13
13	-.64	-	1.17	-.50	-	1.14
14	-.19	-	1.48	-.09	-	1.43
15	-.19	-	1.36	-.13	-	1.33
16	-.10	-	1.32	-.11	-	1.32
17	-.68	.33	1.44	-.67	.23	1.47
18	.32	-	1.86	.39	-	1.78
19	-.28	.92	1.72	-.22	.70	1.87
20	-.13	.72	1.58	-.29	.56	1.69
21	.01	-	1.67	.29	-	1.54
22	-.47	.99	1.64	-.19	.66	1.79
23	-.36	.65	1.70	-.26	.58	1.72
24	.75	-	2.28	.83	-	2.11
25	.79	-	2.36	.87	-	2.16
26	.94	-	2.39	.98	-	2.28
27	1.19	-	2.40	1.20	-	2.38
28	1.33	-	4.80	1.14	-	5.42
29	1.39	-	7.14	1.42	-	5.67
30	1.77	-	2.77	1.72	-	2.99

### *Item Fit*

As shown in Table 17, infit values ranged from .70 to 1.26 while outfit values ranged from .32 to 9.90, indicating that no items were misfitting.



Table 17  
*Vocabulary Item Fit Statistics*

Item	Infit Mean Square	Outfit Mean Square
3	.70	.32
4	1.10	.94
5	1.15	1.05
6	1.04	.64
7	1.26	9.90
8	1.02	1.25
9	.88	2.18
10	1.05	1.02
11	.93	.95
12	.85	.66
13	.93	.81
14	.91	.88
15	.96	.88
16	1.00	1.11
17	1.04	.99
18	.92	.86
19	1.12	1.25
20	1.27	1.19
21	.79	.65
22	1.08	1.12
23	.97	1.07
24	.88	.77
25	.86	.79
26	.93	.85
27	.99	.95
28	1.17	1.73
29	.97	.89
30	1.09	1.48

### *Person Fit*

Results revealed that mean square person infit values ranged from .32 to 9.14 ( $M = 1.18$ ,  $SD = 1.27$ ) and outfit mean square values ranged from .10 to 9.90 ( $M = 1.22$ ,  $SD = 1.71$ ). Results also showed that 6.3% of participants ( $n = 19$ ) responded in unexpected ways. However, this did not prevent stable estimates of item difficulty and person ability

from being made.

### *Dimensionality*

Results (Table 18) revealed that the total raw variance explained by both person ability and items difficulty was 54%. Guidelines indicate that this can be interpreted as “good” unidimensionality (Fisher, 2007).

Table 18  
*Dimensionality of the Vocabulary Subtest*

	Raw Variance	Percentage
Total Variance Explained	32.1	54.0%
Variance Explained by Persons	12.3	20.7%
Variance Explained by Items	19.8	33.3%
Total Unexplained Variance	28.0	46.0%
Unexplained Variance Accounted for by 1 <sup>st</sup> Factor	1.8	6.3%
Unexplained Variance Accounted for by 2 <sup>nd</sup> Factor	1.6	5.9%
Unexplained Variance Accounted for by 3 <sup>rd</sup> Factor	1.6	5.6%
Unexplained Variance Accounted for by 4 <sup>th</sup> Factor	1.5	5.5%
Unexplained Variance accounted for by 5 <sup>th</sup> Factor	1.4	5.0%

Results revealed that the unexplained variance accounted for by the first factor was 1.8 while the total raw residual variance was 28.0. This means that 6.4% [1.8/28.0] of the variance was accounted for by the first factor. This is below the suggested cutoff of 25% (Reckase, 1979), suggesting unidimensionality.

Finally, the raw unexplained variance accounted for by all factors was examined. Guidelines suggest that when the raw values of all factors are less than 2, a measure is unidimensional (Linacre, 2014a). Results revealed that the raw unexplained variance of

factors one through five ranged from 1.4 to 1.8, indicating that Vocabulary is unidimensional.

### ***Person Ability***

Results revealed that ability levels in this sample ranged from -4.40 to 10.13 ( $M = .96, SD = 2.01$ ). With respect to variability, the first quartile was -.11, the second quartile was .83 the third quartile was 1.84 and the inter-quartile range (IQR) was 1.73 [1.81 - .11]. These results indicate that there was a large range of ability represented in the sample.

### ***Person Reliability***

The person reliability was .86 suggesting that the range of ability in the sample was sufficient, the items were sufficiently difficult, there were a suitable number of response options and the length of the task was appropriate (Linacre 2014b). Additionally, the non-significant chi-square value of 6367.70,  $p = .95$  indicates a good fit of the persons and the items to the Rasch model (Embretson & Reise, 2000).

### ***Item Reliability***

The item reliability of .99 for revealed that the sample was sufficiently large, there was an adequate range of item difficulty and the item statistics produced from these analyses are highly likely to be replicated if given to a sample with a similar range of ability levels (Linacre, 2014b).

### ***Item Coverage***

The results displayed in Figure 3 revealed that there was adequate coverage for people with very below average to above average ability levels (-5 to 5 logits) and one notable gap in item coverage. This gap existed for participants with very above average

ability levels (6 to 9 logits).

Figure 3. *Vocabulary Item Coverage*

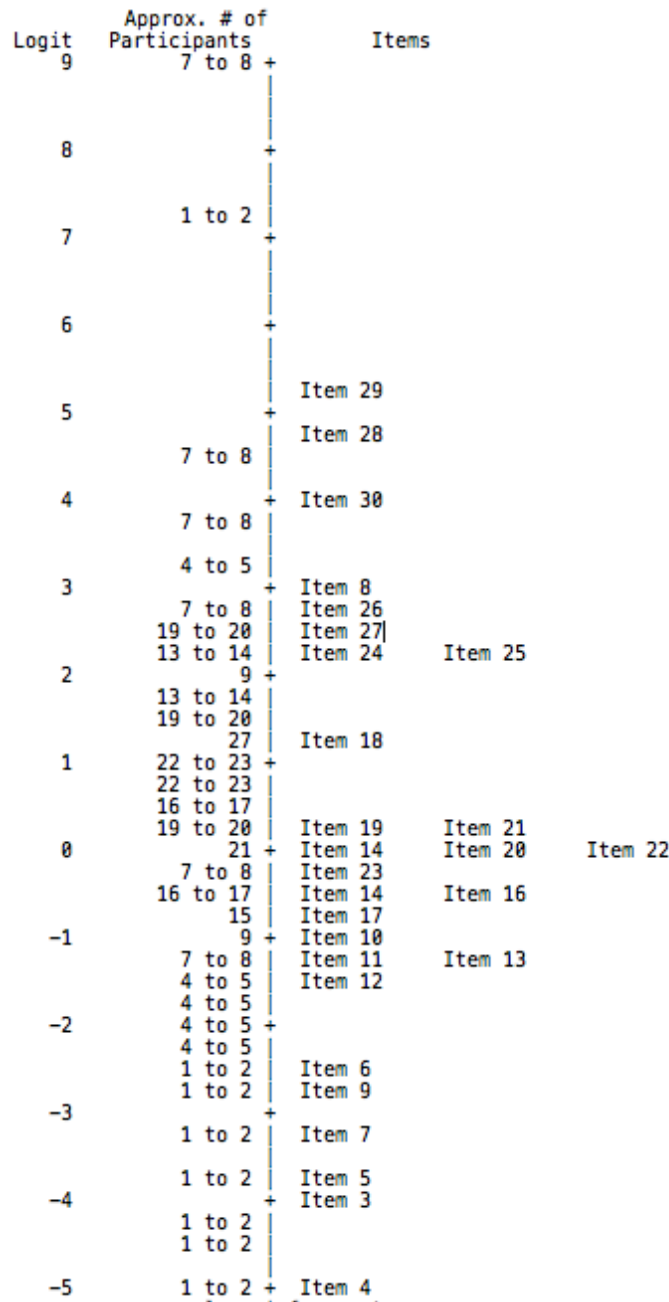


Figure 3. Item difficulty and person ability levels are displayed in logits, which appear on the extreme left. The number of participants with a certain ability level appears to the left of the vertical line. Items appear to the right of the line. Participants' locations correspond to their ability level while items' locations correspond with their difficulty level.

***Item Discrimination***

To evaluate how well item discrimination values conform to standard dichotomous Rasch model assumptions, item discrimination was estimated outside the Rasch model using Winsteps software. The results revealed that discrimination values ranged from .70 to 1.36. All values fell within the guidelines of .5 to 1.5, indicating that items discriminated approximately as would be expected in a Rasch model.

***Item Difficulty***

Item difficulty was examined using a partial credit model in order to assess the assertion of the hypothesis.

The results displayed in Table 19 revealed that item difficulty values ranged from -4.89 to 5.21. Results also revealed that 15 of the 28 items examined did not proceed in order of difficulty. Items four, six, 10, 12, 13, 14, 16, 17, 19, 20, 21, 22, 23, 27 and 30 were disordered. Item four was found to be easier than item three. Item six was harder than items seven, eight and nine. Item 10 was easier than item nine. Items 12 and 13 were easier than item 11. Item 14 was harder than items 15, 16 and 17. Items 16 and 17 were easier than items 14 and 15. Item 19 was easier than item 18. Item 20 was easier than items 14, 15, 16, 17, 18 and 19. Item 21 was easier than items 18 and 19. Item 22 was easier than items 18, 19, 20, and 21. Item 23 was easier than items 18, 19 and 22. Item 27 was easier than item 26. Finally, item 30 was easier than items 28 and 29.

It is possible that items one and two were disordered. However, as previously discussed, because items one and two occur before the start point, there were too few responses provided to these items. Therefore, these items were excluded.

Table 19

*Vocabulary Item Difficulty*

Item	Difficulty
3	-3.92
<b>4</b>	<b>-4.89</b>
5	-3.63
6	-2.60
<b>7</b>	<b>-3.26</b>
<b>8</b>	<b>-3.06</b>
<b>9</b>	<b>-2.66</b>
10	-1.03
11	-1.33
<b>12</b>	<b>-1.43</b>
<b>13</b>	<b>-1.34</b>
14	.02
<b>15</b>	<b>-.50</b>
<b>16</b>	<b>-.62</b>
<b>17</b>	<b>-.73</b>
18	1.23
<b>19</b>	<b>.24</b>
<b>20</b>	<b>-.07</b>
<b>21</b>	<b>.14</b>
<b>22</b>	<b>-.01</b>
<b>23</b>	<b>-.20</b>
24	2.20
25	2.34
26	2.63
<b>27</b>	<b>2.41</b>
28	4.84
29	5.21
<b>30</b>	<b>3.90</b>

*Note.* Disordered items are in boldface.

Because of the large amount of disordered items and because as many as five consecutive items are disordered, it is possible that an individual would meet the discontinue criteria without having the chance to earn points on items that were easier than those they had already been administered. The results revealed that 15 of the 28 examined items were disordered. These results provide support for the hypothesis, which asserted that the items of the Vocabulary subtest would be disordered. As a result, the hypothesis for the Vocabulary subtest was accepted.

### ***Digits Forward***

Examining the responses to the items of the Digits Forward task revealed that items one, two and three had few people who provided responses that fell in the 0 and 1 categories. The number of responses to these categories were so few that it was inadvisable to include these items without first combining the 0 and 1 response categories (Embretson & Reise, 2000).

### ***Ordered Andrich Thresholds, Observed Average and Sample Expectations***

One of the first steps in using a Rasch or partial credit model is to explore items' Andrich thresholds. Andrich thresholds are calculated by creating probability curves for each response option against the range of possible trait levels. Andrich thresholds are the points at which the probability curves for adjacent categories overlap. In order for Andrich thresholds to be calculated, participants must provide several responses to each response category; otherwise, the measurement model is unable to produce stable item calibrations (Bond & Fox, 2007).

Andrich thresholds refer to the ability level needed for participants to have a 50% chance of picking or generating one response versus picking or generating an adjacent

response (Embretson & Reise, 2000). For example, in a rating scale where 0 is equivalent to an incorrect answer, 1 is equivalent to a partially correct answer and 2 is equivalent to a correct answer, thresholds exist between 0 and 1 and 1 and 2. Multiple Andrich thresholds only exist for items with more than two possible responses (e.g., incorrect, partially correct, correct), referred to as polytymous rating scales.

The standard dichotomous Rasch model and the partial credit model assume that items have ordered Andrich thresholds. Ordered Andrich thresholds ensure that the probability of obtaining a higher score on an item requires a higher trait level. For example, Andrich thresholds are ordered when the ability level needed for a participant to have a 50% chance of generating or picking a response option worth fewer points (e.g., a 1 point “partially correct” response) is less than the ability level needed for participants to have a 50% chance of generating or picking a response option worth more points (e.g., a 2 point “correct” response).

Disordered thresholds indicate that the rating scale is not performing in the way the Rasch or partial credit model expect (Bond & Fox, 2007). When disordered thresholds exist, it is necessary to adjust the rating scale so that it aligns with model expectations. Disordered thresholds are indicative of a problem with the way in which individuals receive credit on an item. For example, disordered thresholds can occur when there is a mix of individuals with high trait levels being more likely to obtain a lower score.

Adjusting disordered thresholds often requires combining response categories (Embretson & Reise, 2000). When the threshold between lower categories (e.g., between 0 and 1) is greater than the threshold between higher response categories (e.g., between 1



and 2), these response categories may be combined. The combined responses are then assigned the lower point value.

Because the Digits Forward task contains three possible response categories, it is a polytymous rating scale and therefore the Andrich thresholds must be examined. However, because the 0 and 1 response categories were combined for items one through three, due to too few participants providing scores that fell into these response categories, the rating scales for these items were dichotomous. As a result, items one through three do not have multiple Andrich thresholds. After combining response categories for items one through three, Andrich thresholds for the remaining items were examined. Results (Table 20) revealed that all items had ordered thresholds.

Table 20  
*Digits Forward Items' Andrich Thresholds*

Item	Andrich Threshold	
	Between 0 & 1	Between 1 & 2
4	-1.81	1.81
5	-1.42	1.42
6	-1.17	1.17
7	-1.35	1.35
8	-.58	.58

Next, both observed averages and sample expectations for each item were examined. Observed averages are the average ability levels required for participants in this sample to generate or pick a response in a certain category. Sample expectations are estimates of the average ability level needed for participants to make a response in a certain category (e.g., correct or incorrect) (Bond & Fox, 2007).

The standard dichotomous Rasch model and the partial credit model assume that observed averages and sample expectations increase as category, or answer, value increases (Embretson & Reise, 2000). This means that both the observed averages and sample expectations for wrong responses (i.e., those worth 0 point) should be lower than the observed averages and sample expectations for partially correct responses (i.e., those worth 1 point), which should be lower than the observed averages and sample expectations for correct answers (i.e., those worth 2 points).

Table 21 shows the observed and sample expectations for each item. An examination of these values reveals that they are ordered for all items.

Table 21  
*Average & Expected Ability Levels for the Digits Forward Task*

Item	Observed Average			Sample Expect		
	0	1	2	0	1	2
1	7.10	-	2.24	-6.53	-	2.28
2	-7.20	-	2.33	-6.24	-	2.32
3	-3.60	-	2.66	-4.28	-	2.70
4	-4.93	-.40	3.65	-4.40	-.78	3.71
5	-.66	2.35	5.34	-.48	2.05	5.45
6	1.62	4.12	6.52	1.78	3.91	6.58
7	3.96	6.63	8.63	4.15	6.30	8.51
8	6.44	8.83	9.45	6.61	8.27	9.30

***Item Fit***

Next, item fit was examined. Item outfit mean square equals the sum of the standardized residuals squared divided by the number of subjects, where the standardized

residual equals the residual divided by the square root of the variance. It approximates a chi square distribution (Linacre, 2014e). Mathematically, this is represented as  $\text{Infit} = \frac{\sum [(residual^2 / model\ variance) * model\ variance]}{\sum (model\ variance)} = \text{average} [(standardized\ residuals)^2 * model\ variance] = \text{model\ variance-weighted\ mean-square}$ . Outfit is the chi square divided by degrees of freedom and is mathematically represented as  $\text{outfit} = \frac{\sum (residual^2 / model\ variance)}{(\text{count\ of\ residuals})} = \text{average} [(standardized\ residuals)^2] = \text{chi-square/degrees\ of\ freedom} = \text{mean-square}$ .

Infit mean square infit and outfit values quantify the extent to which items conform to measurement model expectations (Bond & Fox, 2007). Items with both mean square infit and outfit values greater than 1.3 indicate response patterns that are “too haphazard” (Bond & Fox, 2007, p. 240). When both values are above 1.3, stable item statistics cannot be calculated because participants responded in extremely unexpected ways to these items (Bond & Fox, 2007).

Results revealed that mean square infit values ranged from .69 to 1.45 while mean square outfit values ranged from .03 to 9.90. Item one was found to be misfitting (infit = 1.45, outfit = 9.90). Therefore, before further analyzing the data, item one was removed, producing a total of seven items available for further analyses.

The standard dichotomous Rasch model and partial credit model assume that the items that comprise a task are intercorrelated. Thus, removing one item may change the infit and outfit mean square values of the other items. Because item one was removed, new infit and outfit mean square statistics were calculated. Results revealed that mean square infit values ranged from .64 to 1.44. Mean square outfit values ranged from .03 to 9.90. Item three was found to be misfitting (infit = 1.44, outfit = 9.90). Therefore, before

further analyzing the data, item three was removed, producing a total of six items available for analysis.

After removing item three, new infit and outfit mean square statistics were calculated. The results displayed in Table 22 and revealed that mean square infit values ranged from .62 to 1.13 while mean square outfit values ranged from .40 to 9.90. No items were substantially misfitting, as no item had both mean square infit and outfit values above the suggested cutoff of 1.3 (Bond & Fox, 2007).

Table 22  
*Digits Forward Item Fit Statistics with Misfitting Items Removed*

Item	Infit Mean Square	Outfit Mean Square
2	1.11	9.90
4	1.13	9.90
5	.89	3.95
6	.85	1.33
7	.70	.58
8	.62	.40

### ***Person Fit***

Person fit was evaluated next, to determine how well participants' responses aligned with expectations of the standard dichotomous Rasch or partial credit models' expectations. As with misfitting items, unstandardized mean square infit and outfit values are used to identify misfitting participants. These values are calculated in the same manner as item fit values, described above (Bond & Fox, 2007).

When both infit and outfit mean square values are greater than 1.3, participants are considered misfitting, indicating that they responded in unexpected ways (Bond &

Fox, 2007). When a large number of participants are misfitting, employing a standard dichotomous Rasch or partial credit model is not useful because the model is unable to predict participant performance.

Mean square infit values ranged from 0 to 4.86 ( $M = .76$ ,  $SD = .79$ ) and outfit mean square values ranged from 0 to 9.90 ( $M = .73$ ,  $SD = 1.56$ ). Results revealed that 8.6% of participants ( $n = 26$ ) responded in unexpected ways, as was indicated by mean square infit and outfit values greater than 1.3. This did not prevent stable estimates of item difficulty and person ability from being made.

### ***Dimensionality***

When using a standard dichotomous Rasch or partial credit model, dimensionality must be assessed. Dimensionality refers to how many latent constructs a test measures. Both the standard dichotomous and partial credit models assume that tests measure only one underlying construct (Bond & Fox, 2007), which is called unidimensionality.

First, an examination of the total raw explained variance was used to gauge the dimensionality of the Digits Forward task. Raw explained variance is an unstandardized value that quantifies how much of participants' performance can be attributed to their ability and item difficulty. The raw explained variance is calculated in the same manner as would be done during an exploratory factor analysis. The two components entered into this analysis are participants' ability levels and item difficulty.

The results displayed in Table 23 revealed that the raw variance explained by person ability and item difficulty was 77.3%. Guidelines indicate that this can be interpreted as "good" unidimensionality (Fisher, 2007).

The next step in assessing dimensionality is to look at the unexplained variance to

determine whether there is another factor, in addition to item difficulty and person ability, that may be systematically affecting task performance (Linacre, 2014a). When another factor is systematically and significantly affecting task performance, this is indicative of multidimensionality (Bond & Fox, 2007).

To determine whether another factor can explain the unexplained, or residual, variance, a principal component analysis is performed on the residual variance. This identifies groups of items, referred to as factors, which account for portions of the residual variance. The first factor explains the greatest portion of the residual variance.

Table 23  
*Dimensionality of the Digits Forward Task*

	Raw Variance	Percentage
Total Variance Explained	24.6	77.3%
Variance Explained by Persons	11.2	35.1%
Variance Explained by Items	13.4	42.1%
Total Unexplained Variance	6.0	22.7%
Unexplained Variance Accounted for by 1 <sup>st</sup> Factor	1.4	24.0%
Unexplained Variance Accounted for by 2 <sup>nd</sup> Factor	1.3	22.0%
Unexplained Variance Accounted for by 3 <sup>rd</sup> Factor	1.1	19.1%
Unexplained Variance Accounted for by 4 <sup>th</sup> Factor	1.0	17.0%
Unexplained Variance accounted for by 5 <sup>th</sup> Factor	0.9	14.8%

Because the first factor explains the most residual variance, examining the residual variance requires comparing the unexplained variance accounted for by the first factor with the total residual variance. Results revealed that the unexplained variance accounted for by the first factor was 1.4 while total raw residual variance was 6.0. This

means that 23% [1.4/6] of the total unexplained variance was accounted for by the first factor. This is above the suggested cutoff of 15% (Linacre, 2013), which suggests that the Digits Forward task may be multidimensional.

Finally, the raw unexplained variance accounted for by all factors was examined. Guidelines suggest that when the raw values of all factors are less than 2, a measure is unidimensional (Linacre, 2014a). Results revealed that the raw unexplained variance of factors one through five ranged from 0.9 to 1.9, indicating that Digit Sequencing is unidimensional.

### ***Person Ability***

Person ability levels were examined next. Person ability is estimated via a Rasch or partial credit model using joint maximum likelihood estimation, which optimizes the model to data fit through an iterative calculus based approach. Similar to multiple regression, where the estimate method is ordinary least squares, the Rasch model and partial credit model uses joint maximum likelihood estimation, where it maximizes model to data fit.

Person ability is the estimate of an amount of attribute a person has, for example, average verbal abstraction abilities. It is reported on the logit scale, where a lower value indicates a lower level of a trait, such as verbal abilities, and a higher value represents a higher level of a trait (Furr & Bacharach, 2008). A wide range of person ability helps ensure the accurate calculation of item statistics, such as item difficulty (Embretson & Reise, 2000).

Results revealed that ability levels in this sample ranged from -12.94 to 10.39 ( $M = .81, SD = 4.55$ ), with fewer individuals having extreme ability levels. With respect to

variability, the first quartile was -1.11, the second quartile was 1.99, the third quartile was 3.34, and the inter-quartile range (IQR) was 2.23 [3.34- 1.11]. These results indicate that there was a large range of ability represented in the sample.

### ***Person Reliability***

Person reliability was examined next. Person reliability is analogous to classical test theory reliability as would be typically estimated using Cronbach's alpha. It can be interpreted in a manner similar to which test score reliability would be interpreted. Log-likelihood chi square values can also be examined to explore reliability. Significant chi square values represent significant departures from actual performance and model estimates (Linacre, 2014d).

The person reliability was .85, suggesting that the range of ability in the sample was sufficient, the items were sufficiently difficult, there were a suitable number of response options and the length of the task was appropriate (Linacre 2014b). Additionally, the non-significant chi-square value of 894.74,  $p = .89$  indicates a good fit of the persons and the items to the Rasch model (Embretson & Reise, 2000).

### ***Item Reliability***

Item reliability was examined next. The item reliability of 1.00 revealed that the sample was sufficiently large, there was an adequate range of item difficulty and the item statistics produced from these analyses are highly likely to be replicated (Linacre, 2014b).

### ***Item Coverage***

Item coverage was examined next. Item coverage is the extent to which items span the entire continuum of participants' ability levels.

Figure 4 revealed that there was adequate coverage for individuals with above



average ability levels to those with very above average ability levels (1 to 8 logits).

However, there were no items to differentiate individuals whose ability levels were very

below average, (-14 to -6), below average to average (-4 to -1), and above average (3 to 6)

ability levels.

Figure 4. *Digits Forward Item Coverage*

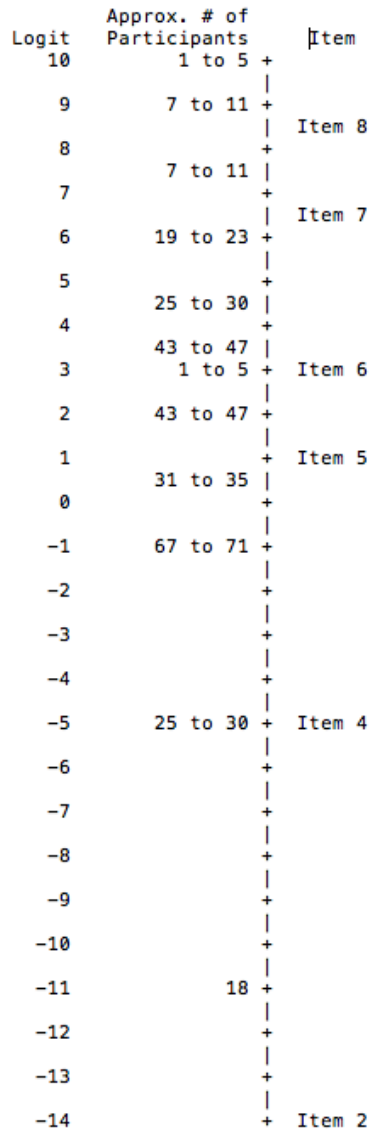


Figure 4. Item difficulty and person ability levels are displayed in logits, which appear on the extreme left. The number of participants with a certain ability level appears to the left of the vertical line. Items appear to the right of the line. Participants' locations correspond to their ability level while items' locations correspond with their difficulty level.

### ***Item Discrimination***

Item discrimination refers to an item's ability to differentiate individuals of varying ability levels. It is indicative of the relationship between the item and the underlying construct being measured (Furr & Bacharach, 2013).

Ideally, in the standard dichotomous Rasch model, all item discrimination values are assumed to be equal to one (Bond & Fox, 2007). That is, all items are assumed to differentiate individuals equally, given that items' level of difficulty. Because item discrimination values are assumed to be equal to one, they are not calculated in the standard dichotomous Rasch model. However, it is empirically unlikely that all item discrimination values will actually equal one (Linacre, 2014b). Instead, guidelines suggest that all item discrimination values should be between 0.5 and 1.5 (Linacre, 2009). An item with a discrimination outside these values indicates that the item does not distinguish between individuals with high and low ability levels as would be expected given that item's level of difficulty (Linacre, 2014b). To evaluate how well item discrimination values conform to Rasch modeling assumptions, item discrimination was estimated outside the Rasch model using Winsteps software.

The results revealed that discrimination values ranged from -1.15 to 1.35. Item two had a discrimination value (-1.15) that fell outside the suggested guidelines of 0.5 to 1.50. This indicates that item two is less able to discriminate between individuals with high and low ability levels than would be expected given its level of difficulty. All other items had acceptable discrimination values.

### ***Item Difficulty***

The hypothesis for Digits Forward asserted that the items that comprise this

subtest would not proceed in order of difficulty.

Item difficulty was examined using a partial credit model in order to assess the assertion of the hypothesis, which stated that the items that comprise the Similarities subtest do not proceed in order of difficulty. Similar to multiple regression, where the estimate method is ordinary least squares, the Rasch model uses joint maximum likelihood estimation, where it maximizes model to data fit. Item difficulty is estimated via a Rasch model using joint maximum likelihood estimation, which optimizes the model to data fit through an iterative calculus based approach. Item difficulty represents how difficult it is to answer an item correctly (or partially correctly), is estimated in logits and exists on a true interval scale, in which the intervals between difficulty levels have a consistent value (Bond & Fox, 2007).

The results displayed in Table 24 reveal that the difficulty values for the items that comprise the Digits Forward task range from -13.80 (item 2) to 8.44 (item 8). The items that were included in analyses proceed in order of difficulty. Items proceed from very easy (difficulty of -13.80) to very hard (difficulty of 8.44) without any disordering.

Table 24  
*Digits Forward Item Difficulty*

Item	Difficulty
2	-13.80
4	-4.78
5	.83
6	3.04
7	6.28
8	8.44

It is possible that the first item of the Digits Forward task is disordered. However, as previously discussed, because item one was misfitting. The standard dichotomous Rasch model could not make stable calculations of item difficulty because response patterns to this item were extremely unexpected. Therefore, item one was excluded because any conclusions drawn based upon the difficulty value of item one would not be reliable and would possibly be erroneous.

The results revealed that the items of the Digits Forward task proceed from easy to hard without any disordering. These results do not provide support for the hypothesis, which asserted that the items of the Digits Forward task would be disordered. As a result, the hypothesis for the Digits Forward task was rejected.

### ***Digits Backward***

Examining the responses to the items of the Digits Backward task revealed that items one and two had few people who provided responses that fell in the 0 and 1 categories. The number of responses to these categories were so few that it was inadvisable to include these items without first combining the 0 and 1 response categories (Bond & Fox, 2007). This increased the frequency of responses and allowed for stable item statistics to be calculated.

### ***Ordered Andrich Thresholds, Observed Average and Sample Expectations***

Because the Digits Backward task contains three possible response categories, it is a polytomous rating scale and therefore the Andrich thresholds must be examined. However, because the 0 and 1 response categories were combined for items one and two, due to too few participants providing answers that fell into these response categories, the rating scales for these items were dichotomous. As a result, items one and two do not

have multiple Andrich thresholds. After combining response categories for items one and two, Andrich thresholds for the remaining items were examined. Table 25 shows that all items had ordered Andrich thresholds.

Table 25  
*Digits Backward Items' Andrich Thresholds*

Item	Andrich Threshold	
	Between 0 & 1	Between 1 & 2
3	-.95	.95
4	-1.04	1.04
5	-1.39	1.39
6	-1.22	1.22
7	-.88	.88
8	-.58	.58

Table 26 shows the observed and sample expectations are ordered for all items.

Table 26  
*Average & Expected Ability Levels for the Digits Backward Task*

Item	Observed Average			Sample Expect		
	0	1	2	0	1	2
1	-.11	-	.40	-4.47	-	.51
2	-2.90	-	.46	-4.50	-	.49
3	-5.41	-2.20	1.06	-4.73	-2.89	1.13
4	-3.41	-.97	2.17	-2.87	-1.19	2.15
5	-1.20	1.20	4.16	-.90	.83	4.29
6	.92	3.16	5.76	.97	3.03	5.82
7	2.87	5.44	6.98	2.95	5.31	6.69
8	5.48	7.05	7.22	.67	.59	1.25

***Item Fit***

Results revealed that mean square infit values ranged from .66 to 1.42 while mean square outfit values ranged from .50 to 9.90. Item one was found to be misfitting (infit = 1.42, outfit = 9.90). Therefore, before further analyzing the data, item one was removed, producing a total of seven items available for further analyses.

Because item one was removed, new infit and outfit mean square statistics were calculated. Results revealed that mean square infit values ranged from .67 to 1.60. Mean square outfit values ranged from .51 to 9.90. Item two was found to be misfitting (infit = 1.60, outfit = 9.90). Therefore, before further analyzing the data, item two was removed, producing a total of six items available for analysis.

After removing item two, new infit and outfit mean square statistics were again calculated. Results revealed that mean square infit values ranged from .67 to 1.51 while mean square outfit values ranged from .51 to 3.56. Item three was found to be misfitting (infit = 1.51, outfit = 3.56). Therefore, before further analyzing the data, item three was removed, producing a total of five items available for analysis.

After removing item three, new infit and outfit mean square statistics were calculated. Table 27 shows that no items were substantially misfitting.

Table 27  
*Digits Backward Item Fit Statistics with Misfitting Items Removed*

Item	Infit Mean Square	Outfit Mean Square
4	1.30	1.63
5	.84	1.87
6	.88	.87
7	.66	.51
8	.86	.60

***Person Fit***

Mean square infit values ranged from .70 to 4.89 ( $M = .93$ ,  $SD = .77$ ) and outfit mean square values ranged from .06 to 9.90 ( $M = 1.00$ ,  $SD = 1.44$ ). Results revealed that 8% of participants ( $n = 24$ ) responded in unexpected ways. However, This did not prevent stable estimates of item difficulty and person ability from being made.

***Dimensionality***

The results displayed in Table 28 revealed that the raw variance explained by person ability and item difficulty was 65.8%, indicating “good” unidimensionality (Fisher, 2007).

The unexplained variance accounted for by the first factor was 1.5 while total raw residual variance was 5.0. This means that 30% [ $1.5/5$ ] of the total unexplained variance was accounted for by the first factor. This is above the suggested cutoff of 15% (Linacre, 2013), which suggests that the Digits Backward task may be multidimensional.

Table 28  
*Dimensionality of the Digits Backward Task*

	Raw Variance	Percentage
Total Variance Explained	11.2	65.8%
Variance Explained by Persons	6.0	35.3%
Variance Explained by Items	5.2	30.5%
Total Unexplained Variance	5.0	34.2%
Unexplained Variance Accounted for by 1 <sup>st</sup> Factor	1.5	30.4%
Unexplained Variance Accounted for by 2 <sup>nd</sup> Factor	1.5	29.8%
Unexplained Variance Accounted for by 3 <sup>rd</sup> Factor	1.1	22.3%
Unexplained Variance Accounted for by 4 <sup>th</sup> Factor	0.8	15.5%
Unexplained Variance accounted for by 5 <sup>th</sup> Factor	0.1	2.1%

Results revealed that the raw unexplained variance of factors one through five ranged from 0.1 to 1.5, indicating that Digits Backward is unidimensional.

### ***Person Ability***

Results revealed that person ability levels ranged from -6.49 to 6.55 ( $M = -2.05$ ,  $SD = 2.97$ ), with fewer individuals having extreme ability levels. With respect to variability, the first quartile was -4.75, the second quartile was -2.79, the third quartile was 0.0, and the inter-quartile range (IQR) was -4.75 [0 – 4.75].

### ***Person Reliability***

The person reliability was .78. However, the significant chi-square value of 693.11,  $p = .01$  indicated a departure from expectations (Embretson & Reise, 2000).

### ***Item Reliability***

The item reliability of 1.00 was within the acceptable range (Linacre, 2014b).

### ***Item Coverage***

Item coverage was examined next. Item coverage is the extent to which items span the entire continuum of participants' ability levels. Item difficulty and participant ability levels are plotted along a true interval scale in logits. This means that the distance between each point on the scale is the same and allows for a direct comparison between ability level and difficulty level.

Figure 5 revealed that there was adequate coverage for a majority of individuals. However, there were several distinct gaps in coverage. There were no items to differentiate individuals with very below average (-4 logits), below average (-2 to 0), slightly above average (1 to 2), above average, (3 to 4) and very above average (5 to 6) ability levels.



Figure 5. *Digits Backward Item Coverage*

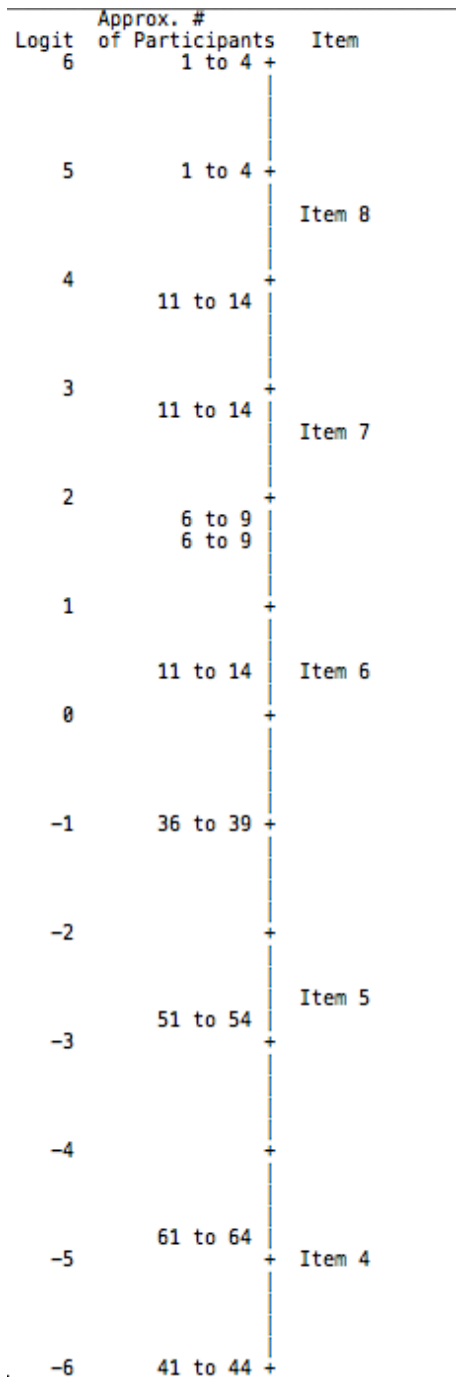


Figure 5. Item difficulty and person ability levels are displayed in logits, which appear on the extreme left. The number of participants with a certain ability level appears to the left of the vertical line. Items appear to the right of the line. Participants' locations correspond to their ability level while items' locations correspond with their difficulty level.

***Item Discrimination***

To evaluate how well item discrimination values conform to standard dichotomous Rasch model assumptions, item discrimination was estimated outside the Rasch model. The results revealed that discrimination values ranged from .45 to 1.37. Item four had a discrimination value (.45) that fell outside the suggested guidelines of 0.5 to 1.50. This indicates that item four is less able to discriminate between individuals with high and low ability levels than would be expected given its level of difficulty.

***Item Difficulty***

Item difficulty was examined using the partial credit model in order to assess the assertion of the hypothesis, which stated that the items that comprise the Digits Backward task do not proceed in order of difficulty.

The results displayed in Table 29 reveal that the difficulty values for the items that comprise the Digits Backward task range from -5.02 (item four) to 4.68 (item eight). Results also revealed that the items proceed in order of difficulty. Items proceed from easy (difficulty of -5.02) to hard (difficulty of 4.68) without any disordering. These results do not provide support for the hypothesis that asserted that the items of the Digits Backward task would be disordered.

Table 29

*Digits Backward Item Difficulty*

Item	Difficulty
4	-5.02
5	-2.61
6	.36
7	2.58
8	4.68

It is possible that items one through three of the Digits Backward task were disordered. However, as previously discussed, because items one through three were misfitting, the partial credit model could not make stable calculations of item difficulty because response patterns to these items were extremely unexpected. Therefore, items one through three were excluded because any conclusions drawn based upon the difficulty value of these items would not be reliable and would possibly be erroneous.

The results revealed that the items of the Digits Backward task proceed from easy to hard without any disordering. These results do not provide support for the hypothesis, which asserted that the items of the Digits Backward task would be disordered. As a result, the hypothesis for the Digits Backward task was rejected.

### ***Digit Sequencing***

Examining the responses to the items of the Digit Sequencing task revealed that items one and two had few people who provided responses that fell in the 0 and 1 categories. The number of responses to these categories were so few that it was inadvisable to include these items without first combining the 0 and 1 response categories (Bond & Fox, 2007).

### ***Ordered Andrich Thresholds, Observed Average and Sample Expectations***

Because the Digit Sequencing task contains three possible response categories, it is a polytymous rating scale and therefore the Andrich thresholds must be examined. However, because the 0 and 1 response categories were combined for items one and two, due to too few participants providing scores that fell into these response categories, the rating scales for these items were dichotomous. As a result, items one and two do not have multiple Andrich thresholds.

After combining response categories for items one and two, Andrich thresholds for the remaining items were examined. Table 30 shows all items had ordered thresholds.

Table 30  
*Digit Sequencing Items' Andrich Thresholds*

Item	Andrich Threshold	
	Between 0 & 1	Between 1 & 2
3	-1.64	1.64
4	-2.24	2.24
5	-1.42	1.42
6	-1.41	1.41
7	-1.50	1.50
8	-.94	.94

Table 31 shows that all observed and sample expectations are ordered.

Table 31  
*Average & Expected Ability Levels for the Digit Sequencing Task*

Item	Observed Average			Sample Expect		
	0	1	2	0	1	2
1	-4.86	-	1.15	-5.16	-	1.15
2	-3.10	-	1.16	-5.05	-	1.19
3	-6.30	-2.71	1.67	-5.52	-3.51	1.75
4	-4.00	-.33	3.19	-3.58	-.56	3.26
5	-1.77	1.24	3.82	-1.49	1.03	3.86
6	1.41	3.98	6.30	1.51	3.72	6.47
7	3.44	5.95	8.15	3.62	5.69	7.77
8	5.79	7.04	9.18	5.79	7.20	8.55

***Item Fit***

Table 32 shows that no Digit Sequencing items were substantially misfitting.

Table 32  
*Digit Sequencing Item Fit Statistics*

Item	Infit Mean Square	Outfit Mean Square
1	.94	.90
2	1.48	1.16
3	1.04	9.90
4	.98	1.19
5	.79	1.34
6	.84	.93
7	.65	.52
8	.86	.82

***Person Fit***

Mean square infit values ranged from 0 to 5.97 ( $M = .87, SD = .85$ ) and outfit mean square values ranged from 0 to 9.90 ( $M = .83, SD = 1.76$ ). Results revealed that 9% of participants ( $n = 27$ ) responded in unexpected ways, as was indicated by mean square infit and outfit values greater than 1.3. This did not prevent stable estimates of item difficulty and person ability from being made.

***Dimensionality***

First, an examination of the total raw explained variance was used to gauge the dimensionality of the Digit Sequencing task. The results displayed in Table 33 revealed that the raw variance explained by person ability and item difficulty was 77.5%.

Guidelines indicate that this can be interpreted as “good” unidimensionality (Fisher,

2007).

Table 33  
*Dimensionality of the Digit Sequencing Task*

	Raw Variance	Percentage
Total Variance Explained	32.6	77.5%
Variance Explained by Persons	11.8	28.1%
Variance Explained by Items	20.8	49.4%
Total Unexplained Variance	8.0	22.5%
Unexplained Variance Accounted for by 1 <sup>st</sup> Factor	1.5	19.0%
Unexplained Variance Accounted for by 2 <sup>nd</sup> Factor	1.4	17.6%
Unexplained Variance Accounted for by 3 <sup>rd</sup> Factor	1.2	15.5%
Unexplained Variance Accounted for by 4 <sup>th</sup> Factor	1.1	14.0%
Unexplained Variance accounted for by 5 <sup>th</sup> Factor	1.0	12.2%

Results revealed that the unexplained variance accounted for by the first factor was 1.5 while total raw residual variance was 8.0. This means that 18.75% [1.5/8] of the total unexplained variance was accounted for by the first factor. This is above the suggested cutoff of 15% (Linacre, 2013), which suggests that the Digit Sequencing task may be multidimensional.

Finally, the raw unexplained variance accounted for by all factors was examined. Guidelines suggest that when the raw values of all factors are less than 2, a measure is unidimensional (Linacre, 2014a). Results revealed that the raw unexplained variance of factors one through five ranged from 1.0 to 1.5, indicating that Digit Sequencing is unidimensional.

***Person Ability***

Results revealed that ability levels in this sample ranged from -9.18 to 9.18 ( $M = .97, SD = 3.37$ ), with fewer individuals having extreme ability levels. With respect to variability, the first quartile was -1.81, the second quartile was 1.43, the third quartile was 2.92, and the inter-quartile range (IQR) was 1.11 [2.92 – 1.81]. These results indicate that there was a large range of ability represented in the sample.

***Person Reliability***

The person reliability was .86, suggesting that the range of ability in the sample was sufficient, the items were sufficiently difficult, there were a suitable number of response options and the length of the task was appropriate (Linacre 2014b). Additionally, the non-significant chi-square value of 1122.96,  $p > .99$  indicates a good fit of the persons and the items to the Rasch model (Embretson & Reise, 2000).

***Item Reliability***

The item reliability of 1.00 revealed that the sample was sufficiently large, there was an adequate range of item difficulty and the item statistics produced from these analyses are highly likely to be replicated (Linacre, 2014b).

***Item Coverage***

Item coverage was examined next. Item coverage is the extent to which items span the entire continuum of participants' ability levels. Item difficulty and participant ability levels are plotted along a true interval scale in logits. This means that the distance between each point on the scale is the same and allows for a direct comparison between ability level and difficulty level.

Figure 6 revealed that there were no items to differentiate individuals whose

ability levels were slightly below average (-4 to -2 logits), slightly above average (1 to 4) and extremely above average (9 to 10).

Figure 6. *Digit Sequencing Item Coverage*

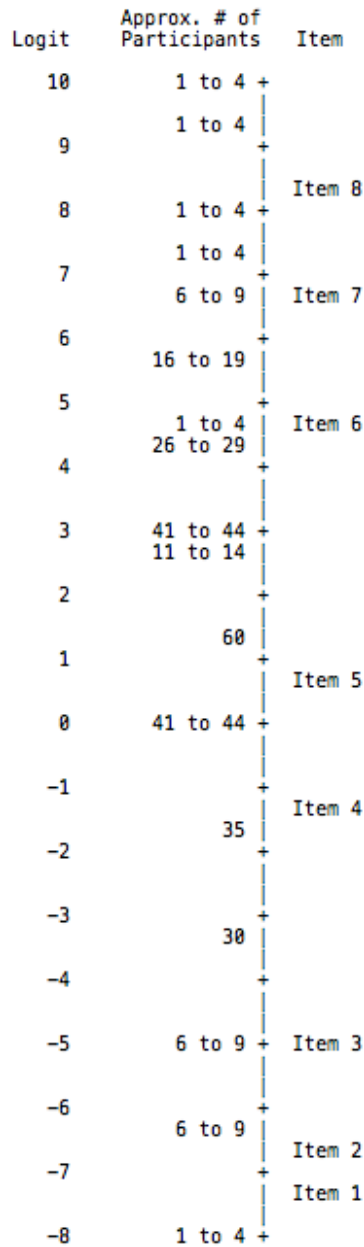


Figure 6. Item difficulty and person ability levels are displayed in logits, which appear on the extreme left. The number of participants with a certain ability level appears to the left of the vertical line. Items appear to the right of the line. Participants' locations correspond to their ability level while items' locations correspond with their difficulty level.



***Item Discrimination***

To evaluate how well item discrimination values conform to standard dichotomous Rasch model assumptions, item discrimination was estimated outside the Rasch model using Winsteps software. The results revealed that discrimination values ranged from .61 to 1.41. All items had discrimination values that were within the suggested guidelines of 0.5 to 1.50 (Linacre, 2014b).

***Item Difficulty***

Item difficulty was examined using the partial credit model in order to assess the assertion of the hypothesis, which stated that the items that comprise the Digit Sequencing task do not proceed in order of difficulty.

The results displayed in Table 34 reveal that the difficulty ranges from -13.80 to 8.39 and that disordering occurred. Item three was disordered and was easier than both items one and two. All other items proceeded in order of difficulty.

Table 34

***Digit Sequencing Item Difficulty***

Item	Difficulty
1	-7.37
2	-6.74
<b>3</b>	<b>-13.80</b>
4	-1.30
5	.65
6	4.69
7	6.76
8	8.39

*Note.* Disordered item is in boldface.

The results revealed that item three of the Digit Sequencing task is disordered, as it is easier than items one and two.

### ***Hypothesis 2***

The second hypothesis was that the items of the Matrix Reasoning, Visual Puzzles, Information and Arithmetic subtests of the WAIS-IV do not proceed in order. In order to analyze this hypothesis, performance on these subtests was examined using a standard dichotomous Rasch model because it is appropriate for items with dichotomous answer options. For these subtests, the probability of a participant receiving a score of 1 on an item is mathematically represented as  $P(X_{is} = 1 | \theta_s, \beta_i) = \frac{e^{(\theta_s - \beta_i)}}{1 + e^{(\theta_s - \beta_i)}}$  where  $X_{is}$  represents an individual's response,  $\theta_s$  stands for an individual's ability level,  $\beta_i$  refers to the difficulty of the item and  $e$  is the symbol for the base of the natural logarithm (Embretson & Reise, 2000).

### ***Matrix Reasoning***

Because item one occurs before the start point, too few responses were provided. As a result, the standard dichotomous Rasch model was unable to produce reliable statistics for this item. Consequently, item one was excluded from analyses. This produced a total of 25 items available for the rest of analyses.

### ***Ordered Andrich Thresholds, Observed Average and Sample Expectations***

Because Matrix Reasoning has a dichotomous rating scale, there are not multiple Andrich thresholds. Table 35 shows the observed and sample expectations for the items of the Matrix Reasoning subtest. These results revealed that values were ordered for all items.

Table 35  
*Matrix Reasoning Observed Averages and Sample Expectations*

Item	Observed Average		Sample Expect	
	0	1	0	1
2	-3.02	-.66	-4.53	-.41
3	-5.41	.15	-4.44	.05
4	-2.88	1.95	-2.33	1.95
5	-1.68	1.99	-1.63	1.99
6	.75	2.0	-.57	2.05
7	.01	2.08	-.33	2.11
8	-.14	2.25	.05	2.22
9	.21	2.12	-.16	2.16
10	.08	2.31	.30	2.27
11	.80	2.31	.65	2.34
12	.21	2.36	.56	2.30
13	.42	2.41	.90	2.34
14	1.36	2.34	1.12	2.38
15	1.44	2.45	1.36	2.47
16	1.25	2.50	1.38	2.47
17	1.61	2.61	1.60	2.62
18	1.52	2.64	1.65	2.59
19	1.86	2.60	1.77	2.64
20	1.85	2.68	1.86	2.68
21	2.03	2.82	2.05	2.81
22	2.23	2.64	2.01	2.75
23	2.10	2.91	2.13	2.89
24	2.17	3.08	2.25	2.99
25	2.45	3.10	2.41	3.23
26	2.59	3.58	2.64	3.40

*Item Fit*

Table 36 revealed that mean square infit values ranged from .28 to 2.46 while

mean square outfit values ranged from .08 to 2.44. None of the items were misfitting.

Table 36  
*Matrix Reasoning Item Fit Statistics*

Item	Infit MNSQ	Outfit MNSQ
2	2.46	1.03
3	.28	.08
4	.90	.16
5	.83	1.61
6	1.27	2.24
7	1.14	1.12
8	.87	.88
9	1.17	1.13
10	.88	.71
11	1.05	1.52
12	.83	.66
13	.79	.54
14	1.09	1.60
15	1.05	1.05
16	.92	.93
17	1.01	.99
18	.92	.83
19	1.06	1.07
20	.99	1.02
21	.99	.96
22	1.14	1.37
23	.97	.97
24	.91	.88
25	1.06	1.18
26	.91	.81

***Person Fit***

Mean square infit values ranged from .45 to 2.02 ( $M = .97$ ,  $SD = .31$ ) and outfit mean square values ranged from .13 to 9.90 ( $M = .99$ ,  $SD = 1.21$ ). Results revealed that 11.33% of participants ( $n = 34$ ) responded in unexpected ways, as was indicated by mean square infit and outfit values greater than 1.3. This did not prevent stable estimates of item difficulty and person ability from being made.

***Dimensionality***

Results (Table 37) revealed that the total raw variance explained by person ability and items difficulty was 35.1%, which is indicative of “poor” dimensionality (Fisher, 2007). These results suggest that the Matrix Reasoning subtest is likely measuring more than one underlying construct.

Table 37  
*Dimensionality of the Matrix Reasoning Subtest*

	Raw Variance	Percentage
Total Variance Explained	13.9	35.1%
Variance Explained by Persons	5.4	13.8%
Variance Explained by Items	8.4	21.4%
Total Unexplained Variance	25.0	64.9%
Unexplained Variance Accounted for by 1 <sup>st</sup> Factor	1.9	7.5%
Unexplained Variance Accounted for by 2 <sup>nd</sup> Factor	1.8	7.2%
Unexplained Variance Accounted for by 3 <sup>rd</sup> Factor	1.6	5.6%
Unexplained Variance Accounted for by 4 <sup>th</sup> Factor	1.5	6.1%
Unexplained Variance Accounted for by 5 <sup>th</sup> Factor	1.4	5.8%

Results revealed that the unexplained variance in the first factor was 1.9 while the

total raw unexplained variance was 25. This means that 7.6% [1.9/25] of the variance was accounted for by the first factor. This is below the suggested cutoff of 25% (Reckase, 1979), suggesting unidimensionality.

Finally, dimensionality was assessed by examining the raw unexplained variance in each factor. Guidelines suggest that when these values are less than 2, a measure is unidimensional (Linacre, 2014a). Results revealed that the raw value for factors one through five ranged from 1.4 to 1.9, suggesting unidimensionality. Two of the three measures of dimensionality suggest that the Matrix Reasoning subtest is unidimensional.

### ***Person Ability***

Person ability levels were examined next. Results revealed that ability levels in this sample ranged from -5.41 to 6.33 ( $M = 1.64$ ,  $SD = 1.86$ ) With respect to variability, the first quartile was .92, the second quartile was 2.17, the third quartile was 2.94, and the inter-quartile range (IQR) was 1.25 [2.17 - .92].

### ***Person Reliability***

The person reliability of .82 for Matrix Reasoning subtest suggests that the range of ability in the sample was adequately large, the items were sufficiently difficult, there were a suitable number of response options and the length of the task was appropriate (Linacre 2014b). Additionally, the non-significant chi-square value of 4496.10,  $p > .99$  indicates a good fit of the data to the model (Embretson & Reise, 2000).

### ***Item Reliability***

The item reliability of .96 for the Matrix Reasoning subtest revealed that there was an adequate range of item difficulty, the same was sufficiently large, and the item statistics produced from these analyses are highly likely to be replicated.

**Item Coverage**

Figure 7 reveals that there is adequate coverage for individuals with ability levels ranging from below average to above average (-3 to 4 logits). However, there are no items with difficulty levels match to participants with very below (-4 to -5) and very above average (5) ability levels.

Figure 7. *Matrix Reasoning Item Coverage*

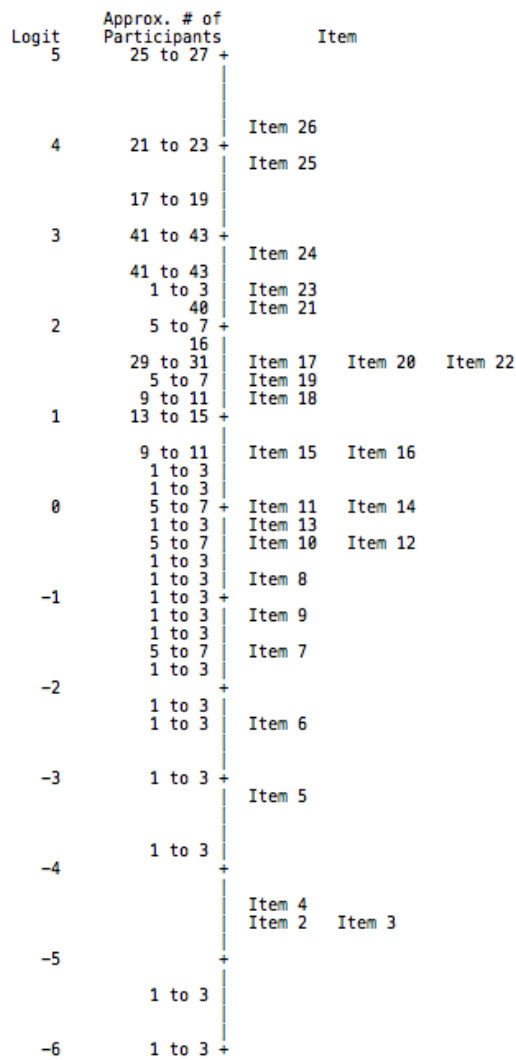


Figure 7. Item difficulty and person ability levels are displayed in logits, which appear on the extreme left. The number of participants with a certain ability level appears to the left of the vertical line. Items appear to the right of the line. Participants' locations correspond to their ability level while items' locations correspond with their difficulty level.

***Item Discrimination***

To evaluate how well item discrimination values conform to standard dichotomous Rasch model assumptions, item discrimination was estimated outside the Rasch model using Winsteps software. The results revealed that discrimination values ranged from 0.00 to 1.50. Item two had a discrimination value (.00) that fell outside suggested guidelines. This indicates that item two is less able to discriminate between individuals with high and low ability levels than would be expected given its level of difficulty.

***Item Difficulty***

Item difficulty was examined using a standard dichotomous Rasch model in order to assess the assertion of the hypothesis, which stated that the items that comprise the Matrix Reasoning subtest do not proceed in order of difficulty.

Item difficulty ranged from -4.65 to 4.18. The results displayed in Table 38 revealed that there were a number of disordered items. Item 3 was easier than item 2. Item 9 was easier than item 8. Item 11 was easier than item 12. Item 16 was easier than item 15. Item 17 was harder than items 18, 19 and 20. Finally, item 22 was easier than item 21. These results provide support for the hypothesis, which asserted that the items of Matrix Reasoning do not proceed in order of difficulty.

It is possible that item one of the Matrix Reasoning was disordered. However, as previously discussed, there were too few responses provided to this item. As a result, Rasch modeling could not produce calculations of item difficulty. Therefore, item one was excluded.



Table 38  
*Matrix Reasoning Item Difficulty*

Item	Item Difficulty
2	-4.58
<b>3</b>	<b>-4.65</b>
4	-4.45
5	-3.30
6	-2.45
7	-1.65
8	-.71
9	-1.21
10	-.36
11	-.02
<b>12</b>	<b>-.31</b>
<b>13</b>	<b>-.29</b>
14	.07
15	.61
<b>16</b>	<b>.57</b>
<b>17</b>	<b>1.56</b>
18	1.20
19	1.40
20	1.53
21	2.20
<b>22</b>	<b>1.67</b>
23	2.38
24	2.72
25	3.87
26	4.18

*Note.* Disordered items are in boldface.

Because the disordered items occur throughout the Matrix Reasoning subtest, it is

possible that an individual would meet the discontinue criteria before reaching these items. If this happened, individuals would not have the chance to earn points on items that were easier than those they had already been administered.

The results revealed that items 3, 12, 13, 16, 17 and 22 of the Matrix Reasoning subtest are disordered. These results provide support for the hypothesis, which asserted that the items of the Matrix Reasoning subtest would be disordered. However, the remainder of the items was ordered according to difficulty. As a result, the hypothesis for the Matrix Reasoning subtest was partially accepted.

### ***Visual Puzzles***

Items one through four of the Visual Puzzles subtest occur before the start point. Because of the limited number of responses provided to these items, the standard dichotomous Rasch model was unable to produce item and person statistics (Linacre, 1997a). As a result, items one through four were excluded from analyses. This produced a total of 22 items.

### ***Ordered Andrich Thresholds, Observed Average and Sample Expectations***

Because the rating scale for Visual Puzzles is dichotomous, there are not multiple Andrich thresholds and thus the first examination involved observed averages and sample expectations.

An examination of the results displayed in Table 39 revealed that observed averages and sample expectations are ordered for all items of the Visual Puzzles subtest, indicating that receiving a higher score was associated with having a higher person ability level.

Table 39

*Visual Puzzles Observed Averages and Sample Expectations*

Item	<u>Observed Average</u>		<u>Sample Expect</u>	
	<u>0</u>	<u>1</u>	<u>0</u>	<u>1</u>
5	-.29	.44	-6.41	.46
6	-4.14	.47	-6.07	.48
7	-3.59	.62	-4.26	.65
8	-1.71	.99	-2.04	1.08
9	-1.86	1.15	-1.81	1.13
10	-1.64	1.10	-1.52	1.08
11	-.77	1.21	-.83	1.23
12	-.20	1.44	-.34	1.51
13	-.58	1.39	-.45	1.35
14	.15	1.87	.19	1.82
15	.32	1.94	.40	1.87
16	.48	2.12	.64	1.97
17	.87	2.04	.91	2.02
18	1.41	2.24	1.31	2.35
19	1.71	2.16	1.50	2.43
20	1.29	2.58	1.50	2.39
21	1.42	2.66	1.63	2.49
22	1.84	2.64	1.85	2.64
23	2.11	2.97	2.17	2.86
24	2.41	2.82	2.34	3.01
25	2.37	3.16	2.42	3.05
26	2.65	3.22	2.65	3.23

***Item Fit***

Results revealed that mean square infit values ranged from .79 to 1.50 and mean square outfit values ranged from .72 to 9.90. Item 5 (infit = 1.50, outfit = 9.90) was misfitting and was therefore removed, yielding a total of 21 items.

Because item three was removed, new infit and outfit mean square statistics were calculated. Results (Table 40) revealed that no items were misfitting.

Table 40  
*Visual Puzzles Item Fit Statistics with Misfitting Item Removed*

Item	Infit Mean Square	Outfit Mean Square
6	1.26	9.90
7	1.04	5.45
8	1.21	2.24
9	.98	.98
10	.90	.87
11	1.02	1.14
12	1.12	1.16
13	.88	.89
14	.95	.88
15	.90	.91
16	.83	.76
17	.97	.92
18	1.11	1.15
19	1.25	1.36
20	.79	.72
21	.79	.76
22	.99	.99
23	.92	.87
24	1.12	1.17
25	.93	.88
26	1.00	1.05

### *Person Fit*

Person fit was examined next, to determine if participants responded in a manner consistent with model expectations. Mean square infit values ranged from .14 to 2.64 (*M*

= .92,  $SD = .37$ ) and outfit mean square values ranged from .08 to 9.90 ( $M = .97$ ,  $SD = 1.36$ ). Results revealed that 9% of participants ( $n = 28$ ) responded in unexpected ways, as was indicated by mean square infit and outfit values greater than 1.3. However, this did not prevent stable estimates of item difficulty and person ability from being made.

### *Dimensionality*

The results displayed in Table 41 revealed that the total variance explained was 39.2%. Guidelines suggest that anything below 50% is indicative of “poor” dimensionality (Fisher, 2007). These results suggest that the Visual Puzzles subtest is likely measuring more than one underlying construct.

Results revealed that the unexplained variance in the first factor was 1.8 while the total raw unexplained variance was 21.0. This means that 8.6% [ $1.8/21$ ] of the variance was accounted for by the first factor. This is below the suggested cutoff of 15% (Linacre, 2013), indicating unidimensionality.

Table 41  
*Dimensionality of the Visual Puzzles Subtest*

	Raw Variance	Percentage
Total Raw Variance explained by Measures	14.3	39.2%
Variance explained by Persons	6.2	17.1%
Variance explained by Items	8.0	22.1%
Total Unexplained Variance	21.0	60.8%
Unexplained Variance Accounted for by 1 <sup>st</sup> Factor	1.8	8.5%
Unexplained Accounted for by 2 <sup>nd</sup> Factor	1.7	8.0%
Unexplained Variance Accounted for by 3 <sup>rd</sup> Factor	1.6	7.5%
Unexplained Variance Accounted for by 4 <sup>th</sup> Factor	1.3	6.3%

Finally, dimensionality was assessed by examining residual variance explained by each factor. Results revealed that the raw value for factors one through four ranged from 1.3 to 1.8, suggesting unidimensionality.

### ***Person Ability***

Person ability is the estimate of an amount of attribute a person has, for example, average verbal abstraction abilities. Results revealed that ability levels in this sample ranged from -9.92 to 4.72 ( $M = -.02$ ,  $SD = 2.20$ ). With respect to variability, the first quartile was -1.07, the second quartile was .12, the third quartile was 1.48, and the inter-quartile range (IQR) was .41 [1.48 – 1.07].

### ***Person Reliability***

Person reliability, which is equivalent to test reliability in classical test theory, was examined next. The person reliability of .83 suggests that the range of ability in the sample was adequately large, items were sufficiently difficult, there were a suitable number of response options and the length of the task was appropriate (Linacre 2014b). The non-significant chi-square value of 3564.75,  $p > .99$  also indicates a good fit of the persons and the items to the model (Embretson & Reise, 2000).

### ***Item Reliability***

The item reliability of .99 revealed that there was an adequate range of item difficulty, the same was sufficiently large, and the item statistics produced from these analyses are highly likely to be replicated.

### ***Item Coverage***

Figure 8 reveals that were few items with difficulty levels match to participants with very below average (-9 to -6 logits), below average (-3 to -1) and very high above

average (4 to 5) ability levels.

Figure 8. *Visual Puzzles Item Difficulty*

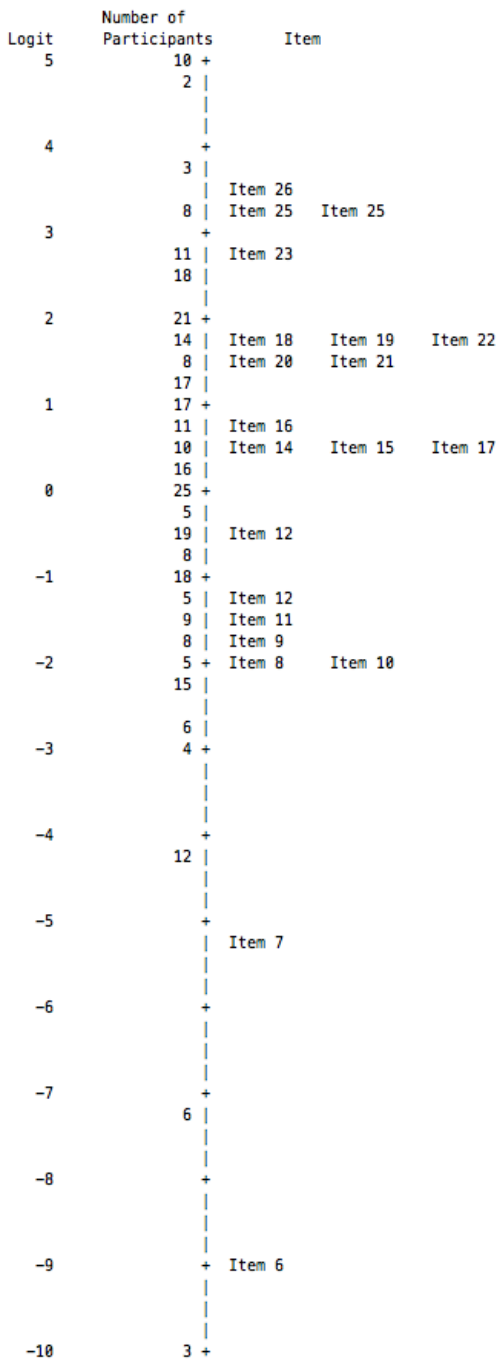


Figure 8. Item difficulty and person ability levels are displayed in logits, which appear on the extreme left. The number of participants with a certain ability level appears to the left of the vertical line. Items appear to the right of the line. Participants' locations correspond to their ability level while items' locations correspond with their difficulty level.

***Item Discrimination***

To evaluate how well item discrimination values conform to standard dichotomous Rasch model assumptions, item discrimination was estimated outside the Rasch model using Winsteps software. Results revealed that item discrimination values ranged from .29 to 1.64. Most values fell within the guidelines of .5 to 1.5, indicating that items discriminated approximately as would be expected in a Rasch model (Linacre, 2009). However, items 20 (discrimination = 1.64) and 21 (discrimination = 1.62) had high discrimination values, indicating that they are better able to differentiate individuals with high and low ability levels than would be expected given their difficulty. Additionally, items six (discrimination = .31) and 19 (discrimination = .29) had a low discrimination value, indicating that they are less able to discriminate between individuals with high and low ability levels as would be expected given their difficulty.

***Item Difficulty***

Item difficulty was examined using a standard dichotomous Rasch model in order to explore the hypothesis, which asserted that the items that comprise the Visual Puzzles subtest do not proceed in order of difficulty.

The results displayed in Table 42 reveal that item difficulty ranged from -9.01 to 3.48. The results displayed in Table 42 also revealed that not all items proceeded in order of difficulty. Items 10, 13, 17, 20, 21, 22 and 25 were disordered. Items 10, 13, and 25 were easier than the items that immediately preceded them. Item 17 was easier than items 14 through 16. Items 20 and 21 were easier than items 18 and 19. Item 22 was easier than item 19. These results provide support for the first hypothesis, which asserted that the items that comprise the Visual Puzzles subtest do not proceed in order of difficulty.



Because the disordered items occur throughout the Visual Puzzles subtest and because up to three consecutive items are disordered, it is possible that individuals would meet the discontinue criteria before having the chance to earn points on items that were easier than those they had already been administered

Table 42  
*Visual Puzzles Item Difficulty*

Item	Item Difficulty
6	-9.01
7	-5.19
8	-2.03
9	-1.84
<b>10</b>	<b>-2.08</b>
11	-1.47
12	-.42
<b>13</b>	<b>-1.15</b>
14	.57
15	.60
16	.85
<b>17</b>	<b>.39</b>
18	1.72
19	1.84
<b>20</b>	<b>1.45</b>
<b>21</b>	<b>1.50</b>
<b>22</b>	<b>1.76</b>
23	2.65
24	3.20
<b>25</b>	<b>3.17</b>
26	3.48

*Note.* Disordered items are in boldface.

It is possible that items one through five of the Visual Puzzles subtest were disordered. However, as previously discussed, there were too few responses provided to items one through four. As a result, the standard dichotomous Rasch model could not produce calculations of item difficulty. Therefore, items one through four were excluded. Additionally, item five was removed because it was misfitting. As a result, Rasch modeling could not make stable calculations of item difficulty because response patterns to this item were extremely unexpected. Therefore, item five was excluded because any conclusions drawn based upon the difficulty value of these items would not be reliable and would possibly be erroneous.

The results revealed that items 10, 13, 17, 20, 21, 22 and 25 of the Visual Puzzles subtest are disordered. These results provide support for the hypothesis, which asserted that the items of the Visual Puzzles subtest would be disordered. However, the other items were ordered according to difficulty. As a result, the hypothesis for the Visual Puzzles subtest was partially accepted.

### ***Information***

Information items one and two occur before the start point. The standard dichotomous Rasch model was unable to produce item and person statistics from the limited number of responses provided to these items (Linacre, 1997a). As a result, items one and two were excluded from analyses. This produced a total of 24 items available for analyses.

Examining the responses to the items of the Information subtest also revealed that item, three had too few people who provided responses that fell in the 0 response category. These items were excluded, producing a total of 23 items available for analyses.

***Ordered Andrich Thresholds, Observed Average and Sample Expectations***

Because the items of the Information subtest contain dichotomous rating scales, the items do not have multiple Andrich thresholds. An examination of the observed averages and sample expectations revealed that these values were ordered (Table 43).

Table 43  
*Information Observed Averages and Sample Expectations*

Item	Observed Average		Sample Expect	
	0	1	0	1
4	-3.58	.27	-3.64	.27
5	-2.5	.48	-2.68	.50
6	-2.30	.57	-2.56	.62
7	-2.39	.65	-2.45	.66
8	-2.62	.68	-2.26	.63
9	-2.08	.73	-1.79	.69
10	-.92	.92	-1.13	.99
11	-.97	1.17	-.87	1.12
12	-.96	1.05	-.89	1.03
13	-.45	1.07	-.66	1.14
14	-.57	1.30	-.41	1.24
15	-.37	1.38	-.21	1.32
16	.08	1.62	.16	1.56
17	.65	1.80	.60	1.87
18	.73	1.88	.72	1.90
19	1.12	1.76	.92	1.99
20	.94	2.32	1.08	2.14
21	1.26	1.93	1.06	2.06
22	1.38	2.78	1.48	2.51
23	1.53	2.69	1.62	2.52
24	1.80	2.61	1.76	2.77
25	2.13	2.90	2.16	2.86
26	2.30	3.36	2.36	3.16

***Item Fit***

Item fit was examined next. Table 44 revealed that none of the items were misfitting, as no items had mean square infit and outfit values that exceeded 1.3.

Table 44

***Item Fit Statistics for Information***

Item	Infit Mean square	Outfit Mean Square
4	1.08	.33
5	1.25	.79
6	1.11	1.44
7	1.01	1.31
8	.82	.44
9	.81	.69
10	1.18	1.10
11	.93	.79
12	.95	.93
13	1.12	1.25
14	.88	.78
15	.88	.80
16	.91	.98
17	1.07	1.11
18	1.01	1.09
19	1.23	1.35
20	.83	.76
21	1.17	1.25
22	.83	.69
23	.87	.83
24	1.09	1.12
25	.97	.92
26	.85	.92

***Person Fit***

Person fit was examined next, to determine if participants responded in a manner consistent with model expectations. Mean square infit values ranged from .34 to 2.41 ( $M = .96$ ,  $SD = .35$ ) and outfit mean square values ranged from .17 to 7.04 ( $M = .90$ ,  $SD = .72$ ). Results revealed that 13.3% of participants ( $n = 40$ ) responded in unexpected ways, as was indicated by mean square infit and outfit values greater than 1.3. This did not prevent stable estimates of item difficulty and person ability from being made.

***Dimensionality***

Dimensionality was examined next. Dimensionality refers to how many latent constructs a test measures. The standard dichotomous Rasch model assumes that tests measure only one underlying construct (Bond & Fox, 2007), which is called unidimensionality.

Results (Table 45) revealed that the total raw variance explained by both person ability and items difficulty was 41.3%. Guidelines suggest that anything below 50% is indicative of “poor” dimensionality (Fisher, 2007). These results suggest that the Information subtest is likely measuring more than one underlying construct.

The unexplained variance accounted for by the first factor was 2.1 while the total raw residual variance was 23.0. This means that 10.95% [ $2.1/23.0$ ] of the variance was accounted for by the first factor. This is below the suggested cutoff of 25% (Reckase, 1979), suggesting unidimensionality.

Finally, results revealed that the raw value for factors one through five ranged from 1.3 to 2.1, suggesting that the Information subtest is multidimensional. Two of three examinations of the dimensionality suggest that the Information subtest is

multidimensional, suggesting that it is measuring more than one construct.

Table 45  
*Dimensionality of the Information Subtest*

	Raw Variance	Percentage
Total Variance Explained	16.4	41.3%
Variance Explained by Persons	6.8	17.1%
Variance Explained by Items	9.6	24.2%
Total Unexplained Variance	23.0	9.3%
Unexplained Variance Accounted for by 1 <sup>st</sup> Factor	21.1	9.3%
Unexplained Variance Accounted for by 2 <sup>nd</sup> Factor	1.7	7.3%
Unexplained Variance Accounted for by 3 <sup>rd</sup> Factor	1.6	7.1%
Unexplained Variance Accounted for by 4 <sup>th</sup> Factor	1.4	6.2%
Unexplained Variance accounted for by 5 <sup>th</sup> Factor	1.3	5.6%

### ***Person Ability***

Person ability is the estimate of an amount of attribute a person has, for example, average verbal abstraction abilities. A wide range of person ability helps ensure the accurate calculation of item statistics, such as item difficulty (Embretson & Reise, 2000).

Results revealed that ability levels in this sample ranged from -6.53 to 4.92 ( $M = .04$ ,  $SD = 2.04$ ). With respect to variability, the first quartile was -1.31, the second quartile was .23, the third quartile was 1.51, and the inter-quartile range (IQR) was .02 [1.51 - 1.31]. These results indicate that there was an adequate range of ability represented in the sample.

### ***Person Reliability***

Person reliability, which is equivalent to test reliability in classical test theory,

was examined next. The person reliability of .82 for Information suggests that the range of ability in the sample was adequately large, the items were sufficiently difficult, there were a suitable number of response options and the length of the task was appropriate (Linacre 2014b). Additionally, the non-significant chi-square value of 383.83,  $p > .99$  indicates a good fit of the persons and the items to the Rasch model (Embretson & Reise, 2000).

### ***Item Reliability***

Next, item reliability, which is dependent upon an adequate range of item difficulty levels as well as an adequate sample size (Linacre, 2014b), was examined. Item reliability values fall between 0 and 1, with higher scores indicating greater reliability (Bond & Fox, 2007).

The item reliability of .99 for revealed that the sample was sufficiently large, there was an adequate range of item difficulty and the item statistics produced from these analyses are highly likely to be replicated if given to a sample with a similar range of ability levels (Linacre, 2014b).

### ***Item Coverage***

Item coverage was examined next. Items and persons are then plotted vertically on a true interval scale (logits). Figure 9 revealed that there was adequate coverage for individuals with ability levels that were just above average to very above average 1 to 4 logits). Item difficulty levels were fairly well matched to participants' ability levels. However, there were several gaps in coverage. There were no items to differentiate individuals whose ability levels were very below average (between -5 and -4), slightly below average (-2), average (0) and extremely above average (5).

Figure 9. *Information Item Coverage*

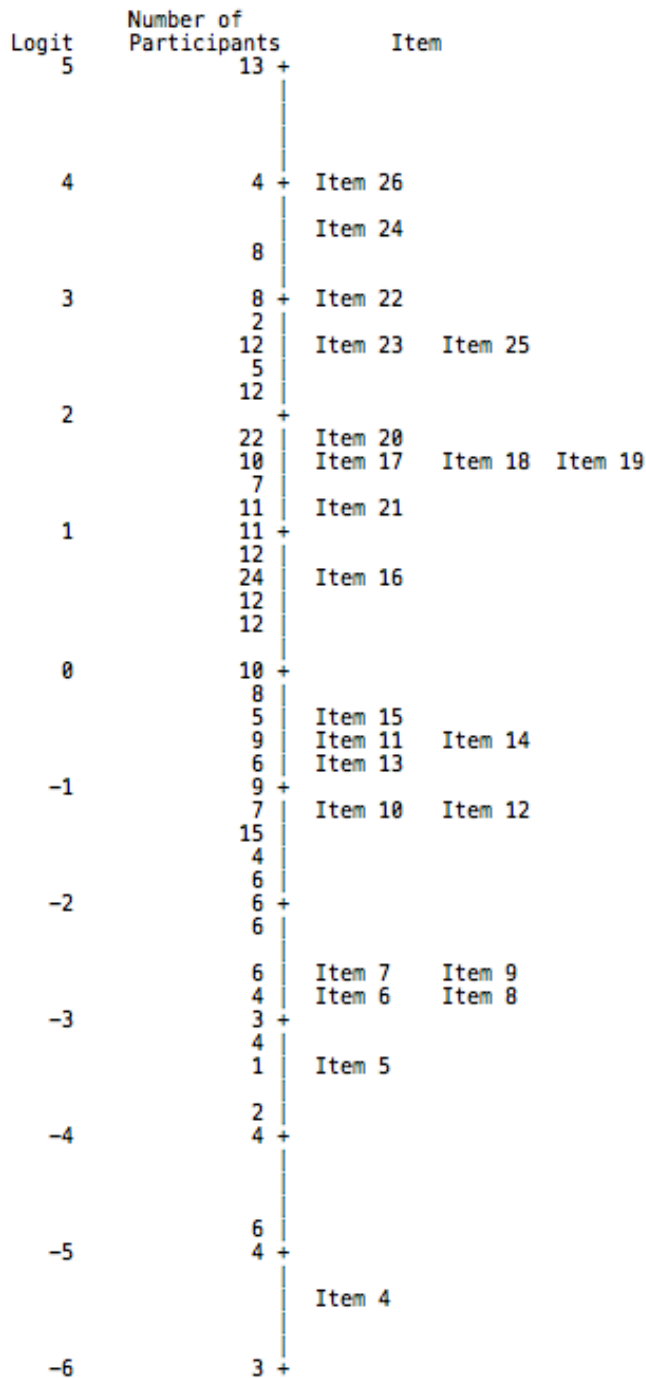


Figure 9. Item difficulty and person ability levels are displayed in logits, which appear on the extreme left. The number of participants with a certain ability level appears to the left of the vertical line. Items appear to the right of the line. Participants' locations correspond to their ability level while items' locations correspond with their difficulty level.



***Item Discrimination***

To evaluate how well item discrimination values conform to standard dichotomous Rasch model assumptions, item discrimination was estimated outside the Rasch model using Winsteps software. The results revealed that discrimination values ranged from .35 to 1.44. All values fell within the guidelines of .5 to 1.5, indicating that items discriminated approximately as would be expected in a Rasch model. Item 21 had a discrimination value (.35) that fell outside suggested guidelines. This indicates that item 21 is less able to discriminate between individuals with high and low ability levels than would be expected given its level of difficulty.

***Item Difficulty***

The hypothesis for Information asserted that the items that comprise this subtest would not proceed in order of difficulty. Item difficulty was examined using a standard dichotomous Rasch model in order to assess the assertion of the hypothesis.

Item difficulty ranged from -5.83 to 3.98. The results displayed in Table 46 reveal that there were a number of disordered items. Items 8 and 9 were both easier than item 7. Item 8 was also easier than item 6. Item 13 was easier than item 12. Item 17 was harder than items 18 and 19. Item 21 was easier than items 17 through 20. Item 23 was easier than item 22. Finally, item 25 was easier than item 24.

Because the disordered items occur throughout the subtest and as many as three consecutive items are disordered in the Information subtest, it is possible that individuals would meet the discontinue criteria before reaching these items. If this happened, individuals would not have the chance to earn points on items that were easier than those they had already been administered.

It is possible that items one through three of the Information subtest were disordered. However, as previously discussed, because items one through three had to be excluded.

Table 46  
*Item Difficulty of the Information Subtest*

Item	Item Difficulty
4	-5.38
5	-3.31
6	-2.72
7	-2.50
<b>8</b>	<b>-2.77</b>
<b>9</b>	<b>-2.54</b>
10	-1.19
11	-.68
<b>12</b>	<b>-1.14</b>
<b>13</b>	<b>-.87</b>
14	-.53
15	-.39
16	.53
<b>17</b>	<b>1.62</b>
18	1.50
19	1.56
20	1.86
<b>21</b>	<b>1.13</b>
22	2.95
<b>23</b>	<b>2.62</b>
24	3.56
<b>25</b>	<b>2.70</b>
26	3.98

*Note.* Disordered items are in boldface.

The results revealed that eight items the Information subtest are disordered.

*Arithmetic*

Items one through five were excluded from analyses because there were too few responses provided to these items.

*Ordered Andrich Thresholds, Observed Average and Sample Expectations*

Because the rating scale is dichotomous, there are not multiple Andrich thresholds.

Table 47 shows that the observed and sample expectations were ordered for each item.

Table 47

*Arithmetic Observed Averages and Sample Expectations*

Item	Observed Average		Sample Expect	
	<u>0</u>	<u>1</u>	<u>0</u>	<u>1</u>
6	-1.81	.30	-5.71	.33
7	-4.93	.34	-5.40	.34
8	-4.19	.60	-4.21	.61
9	-2.00	1.26	-2.22	1.35
10	-1.99	1.35	-2.08	1.39
11	-1.20	1.75	-1.16	1.72
12	-.54	1.77	-.52	1.76
13	.01	1.96	-.03	1.98
14	.45	2.16	.50	2.13
15	.80	2.49	.86	2.41
16	1.31	2.73	1.33	2.69
17	1.51	2.85	1.59	2.74
18	2.04	3.06	2.03	3.07
19	2.22	3.27	2.28	3.17
20	2.48	3.34	2.52	3.28
21	2.78	3.86	2.85	3.53
22	3.08	3.53	3.06	3.71

*Item Fit*

Mean square infit values ranged from .25 to 1.18 while mean square outfit values ranged from .12 to 3.07. Item 6 was substantially misfitting (infit = 1.51, outfit = 3.07). Therefore, before further analyzing the data, item six was removed, yielding a total of 16 items available for the rest of the analyses. New fit statistics (Table 48) revealed that no items were misfitting.

Table 48

*Arithmetic Item Fit Statistics with Misfitting Item Removed*

Item	Infit Mean Square	Outfit Mean Square
7	1.05	2.49
8	1.04	1.33
9	1.18	1.95
10	1.10	1.04
11	.95	.89
12	.96	.96
13	1.01	1.22
14	.94	.96
15	.91	.87
16	.95	.95
17	.89	.85
18	1.03	.92
19	.91	.92
20	.95	.93
21	.86	.68
22	1.06	1.03

***Person Fit***

Results revealed that 12% of participants ( $n = 37$ ) responded in unexpected ways, as was indicated by mean square infit and outfit values greater than 1.3. However, this did not prevent stable estimates of item difficulty and person ability from being made.

***Dimensionality***

First, an examination of the total raw explained variance was used to gauge the dimensionality of the Arithmetic subtest. The results displayed in Table 49 revealed that the total raw variance explained by the measures was 47.7%. This means that just over 47% of the variance in performance on the Arithmetic subtest was explained by participant's ability levels and item difficulty. Guidelines suggest that anything below 50% is indicative of "poor" dimensionality (Fisher, 2007). These results suggest that the Arithmetic subtest is likely measuring more than one underlying construct.

Table 49  
*Dimensionality of the Arithmetic Subtest*

	Raw Variance	Percentage
Total Variance Explained	15.1	47.7%
Variance Explained by Persons	7.1	22.3%
Variance explained by Items	8.0	25.4%
Total Unexplained Variance	16.0	52.3%
Unexplained Variance Accounted for by 1 <sup>st</sup> Factor	1.6	10.1%
Unexplained Variance Accounted for by 2 <sup>nd</sup> Factor	1.6	10.0%
Unexplained Variance Accounted for by 3 <sup>rd</sup> Factor	1.5	9.4%
Unexplained Variance Accounted for by 4 <sup>th</sup> Factor	1.4	8.9%
Unexplained Variance Accounted for by 5 <sup>th</sup> Factor	1.2	7.5%

Dimensionality was also assessed by examining the residual variance explained by the first factor. Results revealed that the explained variance by the first factor was 1.6 while the total raw residual variance was 16.0. This means that 10% (1.6/16) of the variance was accounted for by the first factor. This is below the suggested cutoff of 15% (Linacre, 2013), indicating unidimensionality.

Finally, dimensionality was assessed by examining the residual variance explained by each factor. Guidelines suggest that when the raw values of all factors are less than 2, a measure is unidimensional (Linacre, 2014a). Results revealed that the raw value for factors one through five ranged from 1.2 to 1.6, suggesting that the Arithmetic subtest is unidimensional.

### ***Person Ability***

Results revealed that ability levels in this sample ranged from -9.35 to 6.40 ( $M = -.57$ ,  $SD = 2.56$ ), with fewer individuals having extreme ability levels. With respect to variability, the first quartile was -1.82, the second quartile was -.05, the third quartile was 1.25, and the inter-quartile range (IQR) was .57 [1.25 – 1.82].

### ***Person Reliability***

The person reliability of .82 suggests that the range of ability in the sample was adequately large, the items were sufficiently difficult, there were a suitable number of response options and the length of the task was appropriate (Linacre 2014b). Additionally, the non-significant chi-square value of 2431.53,  $p > .99$  indicates a good fit of the persons and the items to the model (Embretson & Reise, 2000).

### ***Item Reliability***

The item reliability of .99 revealed that the sample was sufficiently large, there

was an adequate range of item difficulty and the item statistics produced from these analyses are highly likely to be replicated.

**Item Coverage**

Figure 10 revealed that there were few items well suited for participants with very below (-8 to -6 logits), below average (-5 to -3), and very above average ability levels (5).

Figure 10. *Item Coverage for Arithmetic*

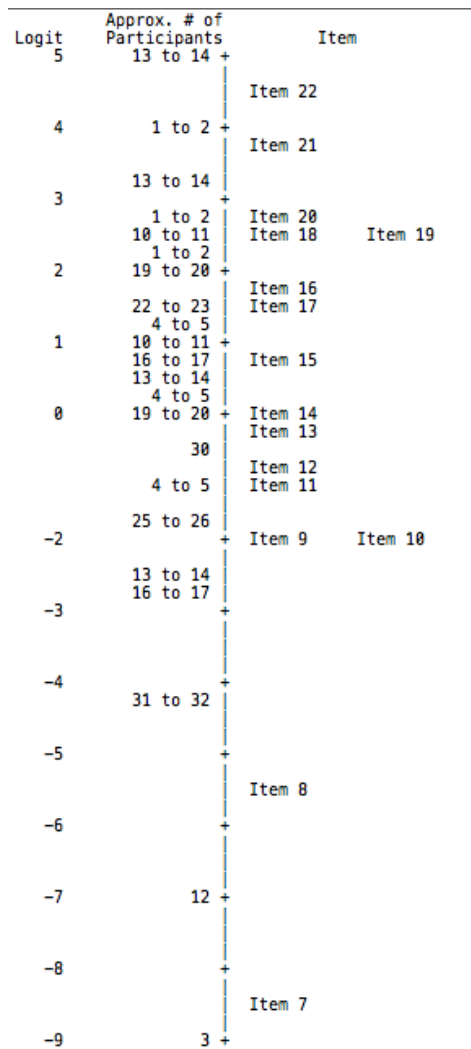


Figure 10. Item difficulty and person ability levels are displayed in logits, which appear on the extreme left. The number of participants with a certain ability level appears to the left of the vertical line. Items appear to the right of the line. Participants' locations correspond to their ability level while items' locations correspond with their difficulty level.

### *Item Discrimination*

To evaluate how well item discrimination values conform to standard dichotomous Rasch model assumptions, item discrimination was estimated outside the Rasch model using Winsteps software. Results revealed that item discrimination values ranged from .76 to 1.25. All values fell within the guidelines of .5 to 1.5, indicating that items discriminated approximately as would be expected in a Rasch model (Linacre, 2009).

### *Item Difficulty*

Item difficulty was examined using the standard dichotomous Rasch model in order to assess the assertion of the hypothesis that stated that the items that comprise the Arithmetic subtest do not proceed in order of difficulty. Similar to multiple regression, where the estimate method is ordinary least squares, the Rasch model uses joint maximum likelihood estimation, where it maximizes model to data fit. Item difficulty is estimated via a Rasch model using joint maximum likelihood estimation, which optimizes the model to data fit through an iterative calculus based approach. Item difficulty represents how difficult it is to answer an item correctly (or partially correctly), is estimated in logits and exists on a true interval scale, in which the intervals between difficulty levels have a consistent value (Bond & Fox, 2007).

The results displayed in Table 50 reveal that item difficulty ranged from -8.41 to 4.61. Results also revealed that not all items proceeded in order of difficulty. Item 17 (difficulty = 1.49) was easier than item 17 (difficulty = 1.65). Item 19 (difficulty = 2.38) was easier than item 18 (difficulty = 2.48). These results provide support for the hypothesis, which asserted that the items that comprise the Arithmetic subtest do not



proceed in order of difficulty.

It is possible that items one through six of the Arithmetic subtest were disordered.

However, as previously discussed, these items were excluded from analyses.

Table 50  
*Arithmetic Item Difficulty*

Item	Item Difficulty
7	-8.41
8	-5.56
9	-2.08
10	-1.95
11	-.89
12	-.88
13	-.26
14	.08
15	.83
16	1.65
<b>17</b>	<b>1.49</b>
18	2.48
<b>19</b>	<b>2.38</b>
20	2.63
21	3.87
22	4.61

*Note.* Disordered items are in boldface.

These results also reveal that an individual who responded incorrectly to items 16, 17 and 18 would meet the discontinue criteria before being administered item 19. This

would result in an individual not having the chance to earn points on item 19, which is easier than item 18.

The results revealed that items 17 and 19 of the Arithmetic subtest are disordered, as they are easier than the items that immediately precede them. These results provide support for the hypothesis, which asserted that the items of the Arithmetic subtest would be disordered. However, the remaining items were consistent. As a result, the hypothesis for the Arithmetic subtest was partially accepted.

## CHAPTER V

### Discussion

The purpose of this study was to use a standard dichotomous Rasch model or partial credit model to examine the difficulty of the items of the core verbal comprehension, perceptual reasoning and working memory indices of the WAIS-IV. It was hypothesized that not all the items of the WAIS-IV subtests would proceed in order of difficulty. In addition to analyzing item difficulty, Andrich thresholds, observed averages, sample expectations, item fit, person fit, dimensionality, person ability, person reliability, item reliability, item coverage and item discrimination were also examined to assess how well the items of the WAIS-IV conformed to Rasch modeling expectations.

In regards to item difficulty, results revealed that WAIS-IV subtests fall into one of three categories: optimally ordered, near optimally ordered and sub-optimally ordered. Optimally ordered subtests, Digits Forward and Digits Backward, had no disordered items. Near optimally ordered subtests were those with one to three disordered items and included Digit Sequencing, Arithmetic, Similarities and Block Design. Sub-optimally ordered subtests consisted of Matrix Reasoning, Visual Puzzles, Information and Vocabulary, with the number of disordered items ranging from six to 16. Disordering may have significant implications for ceiling and basal rules, which affect test scores and total administration time. Implications of the effects on ceiling and basal rules are discussed below.

#### *Ceiling Rules*

Ceiling rules were designed to shorten overall test administration time by stopping administration of a subtest when items were deemed too difficult for the

participant. In order for this rule not to detract from the constructive validity of the WAIS-IV subtests, items must proceed in order of difficulty. When items are disordered, individuals may not have the opportunity to answer questions that they might have answered correctly. This can artificially reduce individuals' overall raw scores.

Because this study produced item difficulty values, another implication should also be considered. Increases in difficulty between consecutive items should be taken into consideration when evaluating the effectiveness of ceiling rules. If item difficulty values increase minimally, it may make sense to have a higher ceiling rule. Individuals may be able to answer items that only slightly increase in difficulty for a variety of reasons, including gaining a better understanding of the task, creating or modifying problem-solving strategies, having more comfort or familiarity with item stimuli or changes in self-perception of ability.

### ***Optimally Ordered Subtests***

Results of this study lend empirical support to the use of discontinue rules on the Digits Forward and Backward tasks as they currently exist. The items of the Digits Forward and Backwards task proceed in order of difficulty. So when individuals discontinue, it is because they have incorrectly answered two consecutive trials of an item. Because the items continue to get more difficult, it is unlikely that individuals would correctly answer items after the discontinue rule. This means that subtest administration is stopped when individuals meet their true ceiling ability. This increases the likelihood that individuals' scores on these tasks reflect their true ability.

***Digits Forward.*** Because all items of the Digits Forward task proceeded in order of difficulty, it is considered an optimally ordered subtest. This finding can be used to

support the construct validity of the Digits Forward task. Because items proceed in order of difficulty, the discontinue rule does stop administration until individuals have received all items of which they have a significant likelihood of answering correctly. This means that individuals' scores are likely accurate gauges of their real ability.

Because the discontinue rule for the Digits Forward task is failing both trials of an item, looking at changes in item difficulty between consecutive items does not have practical implications for the discontinue rule of this subtest.

***Digits Backward.*** The Digits Backward task is considered optimally ordered because all items proceeded in order of difficulty. As a result, the discontinue rule does not stop administration until individuals have received all items of which they have a significant likelihood of answering correctly. This increases the likelihood that Digits Backwards is accurately measuring individuals' performance.

As with Digits Forward, because the discontinue rule for the Digits Backward task is failing both trials of an item, looking at changes in item difficulty between consecutive items does not have practical implications for the discontinue rule of this subtest.

### ***Near Optimally Ordered Subtests***

There was one disordered item in Digit Sequencing, two disordered items in Arithmetic and three disordered items each in Block Design and Similarities, making these subtest nearly optimally ordered. This means that when individuals meet the discontinue rule, depending on at what item they reach it, they may not have the chance to answer items that are easier than ones they have already been administered. This may be an artifact that erroneously reduces individuals' subtest raw scores. This not only

suggests the need to reorder the items of these subtests and/or modify the discontinue criteria, but calls into question the construct validity of the subtest, since individuals' raw scores may not be accurate reflections of their ability levels.

There are two reasons why, despite the disordering of items, measurement inaccuracy is likely to be low near optimally ordered subtests. First, there are few disordered items, meaning that individuals would be unlikely to lose more than two or three raw score points. This is unlikely to reduce overall index scores by more than one to two points. While this reduction may be important in certain circumstances, such as determining a learning disability or competency, in many other circumstances, the reduction may have no impact on overall test interpretation. Second, because several of the disordered items occur in the beginning of the subtests, individuals may be less likely to meet the discontinue rule early.

The actual impact of disordered items on individuals' raw score was examined by determining the modal discontinue item for each subtest, determining how many easier items occurred after this point, and calculating the reduction in raw points this would have caused most people. Implications for individual subtests are discussed below.

***Digit Sequencing.*** Based on the modal discontinue item of item 6 and the fact that item four is the only disordered item, most individuals will not have their raw scores artificially reduced due to disordering of items. Most individuals are administered all items that they are likely to have a chance of answering correctly.

Because the discontinue rule for the Digit Sequencing task is failing both trials of an item, looking at changes in item difficulty between consecutive items does not have practical implications for the discontinue rule of this subtest.

**Arithmetic.** The impact of disordering on individuals' raw scores for the Arithmetic subtest is likely to be minimal. At item 15, most participants received scores of 0, meaning that item 18 was the last item administered to most people. As item 19 is easier than item 18, the majority of individuals who discontinue at item 18 would not have the opportunity to answer this item and may have their raw score artificially reduced by one point. No other items after item 18 are disordered. Depending upon individuals' ages, it is possible that a reduction of one raw score point has no impact on individuals' scaled scores. Thus, while there are two disordered items in the Arithmetic subtest, the disordering is unlikely to have a significant impact on individuals' scores.

Because the changes in difficulty between consecutive items are minimal, it likely makes sense to keep the ceiling rule of the Arithmetic subtest at three. Three is the highest discontinue rule and is used across most subtests.

**Block Design.** Because most of the disordered items occur early in the Block Design subtest, it is unlikely that disordering would have a significant impact on individuals' raw scores. For example, Block Design items would proceed in order of difficulty if arranged as follows: 4, 6, 7, 5, 8, 9, etc. Most participants in this study answered item four correctly. Therefore, even if participants incorrectly answered item five, they would be unlikely to meet the discontinue rule and would still have the opportunity to answer two easier items. Indeed the modal discontinue item was item 10, which occurs well after 2/3 of the disordered items. However, item 11 was found to be easier than item 10. Therefore, the overall impact of disordering on most individuals' raw scores is likely to be a reduction of one point. Depending upon individuals' ages, it is possible that a reduction of one raw score point has no impact on individuals' scaled

scores. Thus, while there are three disordered items in the Block Design subtest, the disordering is unlikely to have a significant impact on individuals' scores.

Because the changes in difficulty between consecutive items are minimal, it may make sense to increase the discontinue rule from two to three. This may allow individuals the opportunity to answer some items that are only two tenths of a point harder than an item they had just missed. Especially because the stimuli of the Block Design subtest are so variably, individuals may be able to answer items that are slightly harder because of the way they perceive the stimuli.

***Similarities.*** Most individuals do not start to consistently receive scores of 0 until item 16 of the Similarities subtest. As all the disordered items occur before item 16, disordered items have no effect on most individuals' raw scores.

Because the changes in difficulty between consecutive items are minimal, it likely makes sense to keep the ceiling rule of the Similarities subtest at three. Three is the highest discontinue rule and is used across most subtests. Especially because individuals can receive partial credit and not meet the discontinue rule, keeping the ceiling rule at three items is likely adequate for most individuals.

### ***Sub-optimally Ordered Subtests***

The effect of disordered items on ceiling rules is likely to have the biggest impact on subtests that were sub-optimally ordered. Matrix Reasoning had six disordered items, Visual Puzzles had seven disordered items, Information had eight disordered items and Vocabulary had 16 disordered items. This disordering may seriously impact the construct validity of these subtests by not permitting individuals to answer items that were easier than ones they had already been administered. This may suggest that reordering is



necessary to improve the psychometric properties of some WAIS-IV subtests.

Clinicians may consider testing the limits on subtests that are sub-optimally ordered. While this will not affect individuals' scores, it may provide more accurate information about individuals' true abilities. Based on the item difficulty values produced by this study, clinicians could also test the limits only with those items they believe individuals would have answered, based on their item difficulty value.

As with near optimally ordered subtests, the actual impact of disordered items on individuals' raw score was examined by determining the modal discontinue item for each subtest, determining how many easier items occurred after this point, and calculating the reduction in raw points this would have caused most people. Implications for individual subtests are discussed below.

***Matrix Reasoning.*** While six disordered items might be thought to have a major impact on most individuals' raw scores, results reveal that this is not the case. Because the modal discontinue item was the last item of the subtest, item 26, most individuals are given the opportunity to answer all items. This means that most individuals' raw scores are not going to be artificially reduced because they will be given the opportunity to answer items that were easier than ones they had already been administered.

Because the changes in difficulty between consecutive items are minimal, it likely makes sense to keep the ceiling rule of the Matrix Reasoning subtest at three.

***Visual Puzzles.*** Even though seven items are disordered, the manner in which they are disordered means that only six of the Visual Puzzles items would maintain their current numerical designation if reordered according to difficulty. However, given the fact that most people do not meet the discontinue criteria until item 26, which is the last

item of the subtest, the disordering is unlikely to significantly reduce most individuals' raw scores.

As with Matrix Reasoning, because the changes in difficulty between consecutive items are minimal, it likely makes sense to keep the ceiling rule of the Visual Puzzles subtest at three.

**Information.** Item 20 was the item at which most people discontinue the Information subtest. As item 21 is easier than item 20, this means that most people will not have the opportunity to answer one item that was easier than items they had already been administered. While several other disordered items occur after this, they are all harder than item 20. This means that individuals' overall raw scores are unlikely to be reduced by more than one raw score point due to disordering, indicating a minimal impact of disordered items.

As with the other sub-optimally ordered subtests, because the changes in difficulty between consecutive items are minimal, it likely makes sense to keep the ceiling rule of the Information subtest at three.

**Vocabulary.** While some of the disordered items occur early in the Vocabulary subtest, many others occur later, where individuals may be more likely to meet the discontinue rule before being administered items they may have been able to answer. If items of the Vocabulary subtest were reordered to proceed in difficulty, only five of the items would retain their ordinal rank. The amount of disordering may suggest that individuals may not be administered items that they may be able to answer. However, as most people discontinue at item 27, and there is only one disordered item that occurs after item 27, it is unlikely that disordering significantly affects individuals' raw scores.

Similar to the other sub-optimally ordered subtests, because the changes in difficulty between consecutive items are minimal, it likely makes sense to keep the ceiling rule of the Vocabulary subtest at three.

### ***Basal Rules***

Basal rules were designed to decrease overall test administration time by not administering items that are too easy for most individuals. Because basal rules do not stop test administration, they are less likely to affect test scores than ceiling rules. The major implication of improperly ordered items on basal rules is increasing administration time by having to administer items that are too easy.

Because this study produced item difficulty and person ability values, other implications should also be considered. Comparing item difficulty level with the modal person ability level may be a good way to determine a start point that has a difficulty value commensurate with most individuals' ability levels. In addition, when there are drastic increases in item difficulty between consecutive items that occur early in the subtest, it may prove useful to add basal items with only minor changes in item difficulty.

### ***Optimally Ordered Subtests***

Because both optimally ordered subtests, Digits Forward and Digits Backward, start with item one, there are no basal rules for these subtests. Therefore, the main implication of this study's results on the basal rules is to consider a more appropriate start point, based on the most commonly occurring person ability for each subtest.

***Digits Forward.*** Findings from this study suggest that, while properly ordered, items 1, 2 and 4 of the Digits Forward task are very easy for most participants. Almost all individuals correctly answered both trials of items 1 and 3 while item 2 had a difficulty

level of greater than -13 logits and item 4 had a difficulty of greater than -4 logits, while the modal person ability was -1.11. This may suggest increasing the start point to item five, since most individuals are correctly able to answer items 1 through 4.

If the start point was changed to item five, a basal rule would need to be created so that the performance of individuals whose ability levels were at the lowest end of the range could still be accurately measured. The lowest person ability level for Digits Forward was -12.94, which was roughly equivalent with the item difficulty of item 2 (-13.80). In order to administer item 2 to individuals who incorrectly answered item 5, the basal rule would have to be at least three items. Changing the start point and the basal rule may help prevent a majority individuals from being administered items that are too easy for them while also ensuring that even individuals with the lowest ability levels are administered items commensurate with their ability levels.

*Digits Backward.* Similar to the findings from Digits Forward, results for the Digits Backward task suggest that the start item be changed. The modal person ability was 0.0, which was most closely associated with the difficulty level of item six (.36). This suggests that item six may be a more appropriate start point for most individuals, as items one through four were too easy for most individuals.

Changing the start point would require creating a basal rule. As the lowest person ability level was -6.49, it may make sense to have individuals reverse to item 3. While the difficulty value of item three could not be calculated it was because most individuals correctly answered this item. Creating a basal rule of three would ensure that individuals who incorrectly answered item six, the new proposed start point, would reverse far enough to receive an item commensurate with their ability levels.

Looking at the changes in difficulty between consecutive items suggests that they are fairly evenly spaced and do not, therefore, indicate the need to add items to more accurately capture individuals' true ability levels.

### *Near Optimally Ordered Subtests*

Because some of the disordered items in the near optimally ordered subtests occur early, this suggests that it is necessary to reorder items so that the first item administered is the easiest item included in standard administration. This may prevent administering reversal items to some individuals when it is not necessary.

***Digit Sequencing.*** If items were reordered according to difficulty, item 4 would become the first item and therefore the start point for this subtest, according to standard administration procedures. However, because the modal person ability was 1.43, having an item with a difficulty value of -13.80 as the start point would likely mean that this item was too easy for most individuals. The modal person ability level is numerically closest to the difficulty values of item six (.65), it may make sense to change the start point to item six.

If the start point was changed to item six, the lowest person ability level for this subtest could be used to establish a basal rule. The lowest person ability was -9.18, which is closest in value to the difficulty level of item one. However, if items were reordered according to difficulty, item one would become item two. In order for individuals who incorrectly answered item six to be administered this item, there would need to be a basal rule of four.

When optimally ordered, the change in difficulty between the first item of Digit Sequencing and the second item would be greater than six points. Adding an item with a

difficulty between values of perhaps three points greater than the first item may improve the ability of this subtest to accurately gauge the performance of individuals with low ability levels.

*Arithmetic.* Because none of the early items of the Arithmetic subtest are disordered, the disordering will not affect the basal rules. Based on the modal person ability level of -4.24, it may make sense to change the starting point to item 8, which has a difficulty level of -5.56. This would prevent most individuals being administered six items that are too easy for them.

If the start point were changed to item eight, the basal rule would need to be reevaluated. The lowest person ability level of -9.35, which is numerically closest to item seven, which has an item difficulty of -8.41. However, items 1 through six were excluded from analyses and it is therefore possible that item 6 may be closer in difficulty level to the lowest person ability level. If the starting item was changed to item eight, in order to insure that individuals with the lowest ability level receive items that are targeted to their ability level, the basal rule could be decreased from three to two.

As with most of the other subtests that are near optimally ordered, the increases in item difficulty between consecutive items do not drastically increase. Therefore, these results do not suggest a need to add more items in order to accurately measure individuals' performance on this subtest.

*Block Design.* The impact of disordered items on the basal rule of the Block Design subtest was mixed. Because item 5 is harder than items 6 and 7, individuals may answer five incorrectly and therefore not meet the basal rule of two and may be administered reversal items unnecessarily. This would increase overall administration

time and would indicate the need to reorder items in order to ensure the basal rules are functioning as they were intended to. However, because item 5 has a difficulty value of -4.32 and the modal person ability level was 3.20, it is unlikely that most individuals would fail to meet the basal rule. Based on the modal person ability level, results suggest that most individuals would be able to answer items up to number 11 (difficulty = 3.28) with relative ease, suggesting that this may be a more appropriate start point than item 5.

The lowest person ability level was -7.48, which was numerically closest to the difficulty of item 4 (difficulty = -6.13). However, as this item is more difficult than the lowest person ability level, it may make sense to have individuals reverse to item three. The difficulty of item three could not be analyzed in this study because nearly all participants correctly answered it, suggesting that even those with the lowest ability level could correctly answer this item. The basal rule would need to be eight to insure that individuals who answered item 11 wrong would be administered an item that could accurately gauge their ability level.

Because the items proceed without major jumps in difficulty levels between consecutive items, results do not suggest a need to add more items in order to accurately measure individuals' performance on this subtest.

*Similarities.* While several items of the Similarities subtest were disordered, results suggest that this will not prevent most people from meeting the basal rule. While the starting item, item four, is harder than item five, the modal person ability level was 1.61 and the item difficulty of item four was -3.53, making it unlikely that individuals would be unable to answer enough items to meet the basal rule of two.

Based on the modal person ability and item difficulty levels, many individuals are

administered items that are too easy for them. With a difficulty level of 1.05, item 13 would likely be a good starting point. If the start point was changed to item six, the lowest person ability level for this subtest could be used to establish a basal rule. The lowest person ability was -5.98, which is numerically closest in value to item five (difficulty = -4.16). If items were reordered according to difficulty, item five would be switched with item four, which would mean that a basal rule of two would need to be established in order to insure that individuals were administered at least one item commensurate with their ability level. Since the basal rule for the Block Design subtest is already two, this would not result in any changes to the reversal rule as it currently exists.

Because the items proceed without major jumps in difficulty levels between consecutive items, results do not suggest a need to add more items in order to accurately measure individuals' performance on this subtest.

### ***Sub-optimally ordered subtests***

As with near optimally ordered subtests, because some of the disordered items in the sub-optimally ordered subtests occur early, this suggests that it is necessary to reorder items so that the first item administered is the easiest item included in standard administration. This may prevent administering reversal items to some individuals when it is not necessary.

***Matrix Reasoning.*** Because the disordered items of the Matrix Reasoning subtest occur before the start point or later in the subtest, they do not affect basal rules. The modal person ability of 2.94 suggests that the item 24 (difficulty = 2.72) may be a more appropriate start point than the current start point of item four, which has a difficulty -4.45, making it too easy for most individuals.



As the person lowest ability level is  $-5.41$ , it may make sense to increase the basal rule if the start point is changed, so that even individuals with the lowest ability levels would have the opportunity to answer items commensurate with their ability level. The suggested start point, based on the most frequently occurring person ability, is item 24. However, if reordered according to difficulty, item 24 would become item 23. As item 23 is the suggested start point and the item difficulty most commensurate with the lowest person ability level is item 3, this would require drastically increasing the basal rule.

Because the items proceed without major jumps in difficulty levels between consecutive items, results do not suggest a need to add more items in order to accurately measure individuals' performance on this subtest.

**Visual Puzzles.** Because the disordered items of the Visual Puzzles subtest occur before the start point or later in the subtest, they do not affect basal rules. As the modal person ability for the Visual Puzzles subtest was  $-.50$ , it may make sense to change the start point to item 12 (difficulty =  $-.42$ ) so that the start point is commensurate with most individuals' ability levels.

As the person lowest ability level is  $-9.92$ , it may make sense to increase the basal rule if the start point is changed, so that even individuals with the lowest ability levels would have the opportunity to answer items commensurate with their ability level. This would require drastically increasing the basal rule, as item 6 (difficulty =  $-9.01$ ) is numerically closest to the lowest person ability. However, because item six is slightly harder than the lowest ability level, it may make sense to make item five the start point. Item five was excluded from analyses and so the difficulty value of this item could not be calculated.

Because the items in the beginning of the test proceed without major jumps in difficulty levels between consecutive items, results do not suggest a need to add more items in order to accurately measure individuals' performance on this subtest.

**Information.** The disordered items of the Information subtest do not occur at the very beginning of the subtest and therefore do not affect the basal rules. Based on the modal person ability level of 0.0, it may make sense to change the starting point to item 15, which has a difficulty level of -.39. This would prevent most individuals being administered 11 items that are too easy for them.

If the start point were changed to item 11, the basal rule would need to be reevaluated. The lowest person ability level of -6.53, which is numerically closest to item four, which has an item difficulty of -5.38. However, items one through three were excluded from analyses and it is therefore possible that item three may be closer in difficulty level to the lowest person ability level. If the starting item was changed to item 11, in order to insure that individuals with the lowest ability level receive items that are targeted to their ability level, the basal rule would need to be increased from two to seven.

The increases in item difficulty between consecutive items do not drastically increase. Therefore, these results do not suggest a need to add more items in order to accurately measure individuals' performance on this subtest.

**Vocabulary.** While one early item of the Vocabulary subtest was disordered, results suggest that this will not prevent most people from meeting the basal rule. While the starting item, item three, is harder than item four, the modal person ability level was 1.30 and the item difficulty of item three was -3.92, making it unlikely that individuals would be unable to answer enough items to meet the basal rule of two.

Based on the modal person ability level of 1.30, it may make sense to change the starting point to item 18, which has a difficulty level of 1.23. This would prevent most individuals being administered 15 items that are too easy for them.

If the start point were changed to item 18, the basal rule would need to be reevaluated. The lowest person ability level of -4.40, which is numerically closest to item four, which has an item difficulty of -4.89. In order to insure that individuals with the lowest ability level receive items that are targeted to their ability level, the basal rule would need to be increased from two to 15.

The increases in item difficulty between consecutive items do not drastically increase. Therefore, these results do not suggest a need to add more items in order to accurately measure individuals' performance on this subtest.

### *Limitations*

Several sample characteristics contributed to the limitations of this study. This study had limited ability to analyze items that occur early in subtests because the sample largely represented individuals with average intellectual functioning who met basal criteria and were not administered reversal items. Additionally, the average education level of 14 years reduced the likelihood that individuals would need to be administered items before the start point. A majority of the sample had at least one psychological diagnosis and just over 25% had two or more diagnoses, which is likely not reflective of the overall population and may limit the generalizability of the results. Lastly, up to 13% of participants were substantially misfitting and it is unclear how this affected results.

There were also several limitations related to statistical analyses. Several items had discrimination values that were outside the range of suggested values. However,

Rasch modeling is a one-parameter model and item discrimination is not a factor in determining item difficulty. A two-parameter model would have included item discrimination values in item difficulty calculations. A comparison between the difficulty estimates produced by one-parameter and two parameter-models would have helped gauge the impact of item discrimination on item difficulty. Rasch modeling does not take into consideration differential item functioning, which occurs when individuals who belong to a certain group (i.e., ethnicity, gender) do not have the same chance of correctly answering an item as do individuals with the same ability level that belong to another group. Examining differential item functioning in combination with sample demographics could have allowed for a more in depth understanding of why certain results were obtained.

#### *Future Studies*

Future studies should attempt to replicate this study's results, improve upon the limitations of this study, seek to explain several findings and ascertain the effects of item ordering on participant performance. First, results of this study must be replicated. While results revealed good item and person reliability for most subtests, there were several subtests for which the model fit was less than ideal. Reproducing the results of this study would lend support to the reliability of the findings.

Second, future studies should break standard test administration by ignoring basal and ceiling rules to collect data on more items. Having all individuals answer all items decreases the likelihood that the statistical model will be unable to produce item statistics due to too few responses. This would allow for item difficulty values to be calculated for more basal items than were examined in this study. Having all individuals answer all

items would also allow for a comparison between individuals' scores using standard scoring guidelines, where items after the discontinue rule has been met are not included in individuals' scores, and scores where all correctly answered items contribute to individuals' scores.

In order to help clarify why some items are more or less difficult than would be expected given their ordinal rank, future studies should include a qualitative aspect, where, after each subtest, individuals are asked describe what they found easy and difficult about each item. Finally, future research should examine the effects of item ordering on participant performance. While this study revealed that only minimal reductions in index scores likely result from the prematurely stopping test administration, it is not known if disordering has other impacts on performance, perhaps by increasing or decreasing an individual's confidence.

### *Conclusion*

Two major implications of the result of this study were considered: the impact on individuals' scores and the impact on overall test administration time. While the number of disordered items ranged from 0 to 16, the overall impact on raw scores is deemed minimal. Because of where the disordered items occur in the subtest, most individuals are administered all the items that they would be expected to answer correctly. A one-point reduction in any one subtest is unlikely to significantly affect overall index scores, which are the scores most commonly interpreted in the WAIS-IV. However, if an individual received a one-point reduction across all subtests, this may have a more noticeable impact on index scores. In cases where individuals discontinue before having a chance to answer items that are easier, clinicians may consider testing the limits. While this would have no

impact on raw scores, it may provide clinicians with a better understanding of individuals' true abilities. Based on the findings of this study, clinicians may consider administering only certain items in order to test the limits, based on the items' difficulty value.

This study found that the start point for most subtests is too easy for most individuals. For some subtests, most individuals may be administered more than 10 items that are too easy for them. Other than increasing overall administration time, it is not clear what impact, of any, this has. However, it does suggest the need to reevaluate current start items so that they are the true basal for most people.

Based on the results of this study, it is suggested that the next version of the WAIS using Rasch modeling to determine item difficulty. There are several limitations in the way item difficulty on the WAIS-IV is calculated. First, the way item difficulty is calculated means that item difficulty is highly sample dependent. For example, in CTT the standard error of measurement is a mathematical representation of expected changes in scores due to error and is an important indicator of reliability. Standard error is consistent among scores for the same population and changes when the population changes. The use of *p*-values, or the proportion of people correctly answering an item, to determine item difficulty (Embretson & Reise, 2000) means that a sample of individuals with above average intelligence will produce lower difficulty values than a sample of individuals with below average intelligence (Hambleton & Jones, 1993). Additionally, while the overall order of the items is adequate for different samples, the distance between more difficult items may be greater for a sample with lower abilities than it is for a sample with above average abilities (Embretson and Reise, 2000).

Proponents of standard dichotomous Rasch model argue that it has distinct advantages above both CTT-based methods as well as other IRT models (Bond & Fox, 2007; Embretson & Reise, 2000; Furr & Bacharach, 2013; Hambleton & Jones, 1993) because of the principle of monotonicity, also referred to as specific objectivity, the principle of additivity or double cancellation, which “establishes that two parameters are additively related to a third variable” (Embretson & Reise, 2000, p. 148). In other words, because of the principle of monotonicity, in Rasch modeling, probability of correctly answering an item is the additive function of individuals’ ability, or trait level, and the item’s degree of difficulty. As ability increases, so does an individual’s probability of answering that item. Because only item difficulty and person ability affect an individual’s chance of correctly answering an item, inter-individual comparisons can be made even if individuals did not receive identical items or items of the same difficulty level. This is why Rasch modeling is referred to as a *test-free measurement*.

It is not uncommon for test developers who create tests within the framework of CTT to assume that scores exist upon an interval scale because they are normally distributed, or forced to be normally distributed (Embretson & Reise, 2000). The standard scores for many tests are normalized using either percentile matching or nonlinear transformations, both of which change the distance between scores, potentially changing the scale from interval to ordinal. Even scores that arise from a normal distribution (e.g., Full Scale IQ) cannot be assumed to be on an interval level scale without empirical evidence that the scores have interval scale properties.

Using a standard dichotomous Rasch or partial credit model would result in more accurate estimates of item difficulty, which would allow for all items to be ordered in

accordance of the assertions of the WAIS-IV, which posit that all items proceed in order of difficulty. Additionally, item difficulty is less likely to change when the sample of individuals tested changes, making item difficulty values generalizable to different populations. Using item response theory to determine item difficulty would likely result in the reordering of items and additional research is needed to determine what impact, if any, this reordering would have on individuals' performance.

Using item response theory to determine item difficulty would require making sure that all subtests conform to Rasch modeling expectations as well as the use of different software. Results of this study suggest that all subtests conform to Rasch modeling expectations and therefore do indicate that it would be appropriate for the next version of the WAIS to use item response theory to determine item difficulty.



### References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.
- Baghaei, P. (2008). Local dependency and Rasch measure. *Rasch measurement transactions*, 21(3), 1105-1106. Retrieved from <http://www.rasch.org/rmt/rmt213b.htm>
- Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV): what does the WAIS-IV measure? *Psychol Assess*, 22(1), 121-130. doi: 10.1037/a0017767
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model : fundamental measurement in the human sciences* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.
- Bornstein, R. A., McLeod, J., McClung, E., & Hutchison, B. (1983). Item difficulty and content bias on the WAIS-R Information subtest. *Canadian Journal of Behavioral Science*, 15(1), 27-34. doi.org/10.1037/h0080672
- Climie, E. A., & Rostad, K. R. (2011). Wechsler Adult Intelligence Scale (4th ed.). *Test Reviews*, 29(6), 581-586. doi: 10.1177/0734282911408707
- Costello, R. M., & Connolly, S. G. (2005). Item difficulty scaling for WAIS-III picture arrangement. *Journal of clinical psychology*, 61(6), 781-786. doi: 10.1002/jclp.20059

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*.

New York: Cengage learning.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function*

*System (D-KEFS)*. San Antonio, TX: The Psychological Corporation.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New

York: Psychology Group.

Fisher, Jr., W.P. (2007). Rasch Rating Scale instrument quality criteria. *Rasch*

*Measurement Transactions*, 21(1), 1095. Retrieved from

<http://www.rasch.org/rmt/rmt211m.htm>

Frank, G. (1983). *The Wechsler enterprise : an assessment of the development, structure,*

*and use of the Wechsler tests of intelligence* (1st ed.). Oxford ; New York:

Pergamon Press.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An Introduction* (Second ed.).

Los Angeles: SAGE Publications.

Hambleton, R.K. and Jones, R.W. (1993). An NCME instructional module on

Comparison of classical test theory and item response theory and their

applications to test development. *Instructional Topics in Educational*

*Measurement*. Retrieved from [http://ncme.org/linkservid/66968080-1320-](http://ncme.org/linkservid/66968080-1320-5CAE6E4E546A2E4FA9E1/showMeta/0/)

[5CAE6E4E546A2E4FA9E1/showMeta/0/](http://ncme.org/linkservid/66968080-1320-5CAE6E4E546A2E4FA9E1/showMeta/0/)

Heath, J. A., & Leathem, J. M. (1998). Order of item difficulty on the WAIS-R picture

arrangement subtest: data from a traumatically brain-injured sample. *Perceptual*

*and motor skills*, 87(1), 243-250. doi: 10.2466/pms.1998.87.1.243

Lezak, M. D., & Lezak, M. D. (2004). *Neuropsychological assessment* (4th ed.). Oxford; New York: Oxford University Press.

Linacre, J.M. (1997). KR-20 / Cronbach Alpha or Rasch Person Reliability: Which tells the “truth’ ?” *Rasch Measurement Transactions*, 11(3), 580-581. Retrieved from <http://www.rasch.org/rmt/rmt1131.htm>

Linacre, J.M. (1997, June). Guidelines for Rating Scales. Retrieved from <http://www.rasch.org/rn2.htm>.

Linacre, J.M. (2009, October). 713. Guideline for Item Discrimination. Message posted to <http://www.rasch.org/forum2009.htm>.

Linacre, J.M. (2014a). Dimensionality: Contrasts and variances. Retrieved from <http://www.winsteps.com/winman/index.htm?principalcomponents.htm> In

Linacre, J.M. (2014). A user’s guide to Winsteps Ministeps Rasch-Model Computer Programs Program Manual 3.81.0. Published online: John M Linacre. Retrieved from <http://www.winsteps.com/winman/copyright.htm>

Linacre, J.M. (2014b). Discrimination = report item discrimination = no. Retrieved from <http://www.winsteps.com/winman/discrim.htm>. In Linacre, J.M. (2014). A user’s guide to Winsteps Ministeps Rasch-Model Computer Programs Program Manual 3.81.0. Published online: John M Linacre. Retrieved from <http://www.winsteps.com/winman/copyright.htm>

Linacre, J.M. (2014c). Reliability and separation of measures. Retrieved from <http://www.winsteps.com/winman/reliability.htm>. In Linacre, J.M. (2014). A user’s guide to Winsteps Ministeps Rasch-Model Computer Programs Program

Manual 3.81.0. Published online: John M Linacre. Retrieved from

<http://www.winsteps.com/winman/copyright.htm>

Linacre, J.M. (2014d). Global fit statistics. Retrieved from

<http://www.winsteps.com/winman/index.htm?diagnosingmisfit.htm> In Linacre,

J.M. (2014). A user's guide to Winsteps Ministeps Rasch-Model Computer

Programs Program Manual 3.81.0. Published online: John M Linacre. Retrieved

from <http://www.winsteps.com/winman/copyright.htm>

Linacre, J.M. (2014e). Misfit diagnosis: Infit outfit mean-square standardized. Retrieved

from <http://www.winsteps.com/winman/diagnosingmisfit.htm> In Linacre, J.M.

(2014). A user's guide to Winsteps Ministeps Rasch-Model Computer Programs

Program Manual 3.81.0. Published online: John M Linacre. Retrieved from

<http://www.winsteps.com/winman/copyright.htm>

Linacre, J.M. (2014, June 2) [Personal Communication]

McNemar, Q. (1956). Wechsler Adult Intelligence Scale. *Psychological Test Reviews*,

159. doi.org/10.1037/h0038765

Myers, J.L., Well, A.D., Lorch, Jr., R.F. (2010). *Research design and statistical analysis*,

(3<sup>rd</sup> ed.). New York: Routledge.

Norman, R. P., & Wilensky, H. (1961). Item difficulty of the WAIS information subtest

for a chronic schizophrenic sample. *Journal of clinical psychology*, 17, 56-57.

doi: 10.1002/1097-4679(196101)17:1<56::AID-JCLP2270170120>3.0.CO;2-E

Randolph, C. (1998). *Repeatable battery for the assessment of neuropsychological status*.

San Antonio, TX: The Psychological Corporation.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests.

Chicago: University Press. doi: 10.1177/014662168100500413

Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Education Statistics*, 4: 207-230. doi:

10.3102/10769986004003207

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition*. Itasca, IL: Riverside Publishing.

Ryan, J. J., & Lopez, S. J. (1999). Order of item difficulty on picture arrangement: extending the discussion to the WAIS-III. *Perceptual and motor skills*, 88(3 Pt 1), 1053-1056. doi: 10.2466/pms.1999.88.3.1053

Sattler, J. M., & Ryan, J. J. (2009). *Assessment with the WAIS-IV*. San Diego: Jerome M. Sattler, Publisher, Inc.

Tennant, A. (2004). Disordered thresholds: An example from the Functional Independence Measure. *Rasch Measurement Transactions*, 17(4), 945-948.

Retrieved from: <http://www.rasch.org/rmt/rmt174a.htm>. doi: 10.2340/165019770871

Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.

Tulsky, D. S., Saklofske, D. H., & Zhu, J. (2003). Revising a Standard: An Evaluation on the Origin and Development of the WAIS-III. In D.S. Tulsky, D.H. Saklofske, G.J. Chelune, R.K. Heaton, R.J. Ivnik, A. Prifitera, M.F. Ledbetter & R. Bornstein. *Clinical Interpretation of the WAIS-III and WMS-III*. USA: Academic Press.

Wechsler, D. (1939a). *The measurement of adult intelligence*. Baltimore: Williams and Wilkins. doi: org/10.1037/10020-000

- Wechsler, D. (1939b). *Wechsler-Bellevue intelligence scale*. New York: The Psychological Corporation.
- Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Baltimore: Williams and Wilkins. doi: 10.1037/10020-000
- Wechsler, D. (1946). *Wechsler-Bellevue Intelligence Scale, Form II*. New York, NY: The Psychological Corporation.
- Wechsler, D. (1955a). *WAIS Manual*. New York, NY: The Psychological Corporation.
- Wechsler, D. (1955b). *Wechsler Adult Intelligence Scale*. New York, NY: The Psychological Corporation.
- Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore: Williams and Wilkins. doi: org/10.1037/11167-000
- Wechsler, D. (1975). Intelligence defined and undefined: A relativistic appraisal. *American Psychologist*, 30, 135-139.
- Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1992). *Wechsler Individual Achievement Test*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1997a). *WAIS-III Administration and scoring manual*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1997b). *Wechsler Adult Intelligence Scale-Third Edition*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (2003). *Wechsler intelligence scale for children* (4th ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008a). *WAIS-IV Administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008b). *WAIS-IV Technical and interpretive manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008c). *Wechsler Adult Intelligence Scale (4th ed.)*. San Antonio TX: The Psychological Corporation.

Whipple, G. M. (1909). A range of information test. *Psychological review*, *16*, 347-351.  
doi: org/10.1037/h0073114

Zhu, J. J. (2013, July 10). [Personal communication].