2014

WISC-IV and Intellectual Disability: A Pilot Study on Hidden Floor Effects

Allyssa M. Lanza
Antioch University - New England

Running Head: WISC-IV AND INTELLECTUAL DISABILITY

WISC-IV and Intellectual Disability:

A Pilot Study on Hidden Floor Effects

by

Allyssa Lanza

M.S., Antioch University New England, 2012
B.A., University of New Hampshire, 2008

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Psychology in the Department of Clinical Psychology
at Antioch University New England, 2014

Keene, New Hampshire

Department of Clinical Psychology

DISSERTATION COMMITTEE PAGE

The undersigned have examined the dissertation entitled:

WISC-IV and Intellectual Disability:
A Pilot Study on Hidden Floor Effects

presented on July 14, 2014

by

Allyssa Lanza

Candidate for the degree of Doctor of Psychology
and hereby certify that it is accepted*.

Dissertation Committee Chairperson:
Gargi Roysircar, PhD

Dissertation Committee members:
E. Porter Eagan, PsyD
Gina Pasquale, PsyD

Accepted by the

Department of Clinical Psychology Chairperson
Kathi A. Borden, PhD

on **7/14/14**

* Signatures are on file with the Registrar's Office at Antioch University New England.

Dedication

For all individuals with intellectual disability, and their teachers.

Acknowledgments

This project would not have been possible without Gargi Roysircar. She has been my advisor, chair, motivator, and mentor. Her work ethic and dedication to her students is unparalleled. I would like to thank her for her faith in this project and in me. It has been a gift to be able to learn and work with Dr. Roysircar, and I am a better psychologist for her mentorship.

I would also like to thank my committee members, Gina Pasquale and Porter Eagan. Both have brought unique perspectives to the study, and were willing to help when few others were. Thank you for your creativity, patience, and willingness to participate.

Thank you also to those who helped obtain data for this study. I truly appreciate the time and footwork volunteered to help find eligible protocols, which at times seemed like the proverbial needles in the haystack. Without your help this study could not have come to fruition.

I would also like to thank all of the members of the community at Antioch who create an environment for such a project to exist. In particular, thank you to Vincent Pignatiello for our many scholarly and professional conversations. Thank you to the staff who foster a fantastic learning environment and who always encourage students to make the world better. Thank you to Liz Allyn, Catherine Peterson, Nancy Richard, and Joy Guerriero for being the glue that holds everything together. Thank you to my fellow students, who supported me through all of the trials of higher education. Thank you for sharing your creativity, spirit, and brilliance.

Finally, thank you to my family, both given and chosen. I would like to particularly thank Mum, Dad, Kurt, Auntie, Emmy, Sarah, Kate, Matt, Courtney, Katrin, Kaylee, TJ, and Adrienne. Your patience, willingness to learn, and unconditional love has given me the strength to be a better person and psychologist. Thank you.

Quote

“Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid.”

— Albert Einstein

Table of Contents

Abstract 1

Chapter 1 1

Statement of the Problem 1

Research Questions 8

Definition of Terms 8

Summary 13

Chapter 2: A Review of Relevant Literature 14

Wechsler Intelligence Tests 14

Wechsler Four-Factor Model 14

Standardization of the WISC-IV 15

Concerns about WISC-III and WAIS-III for Assessing LD 18

Distribution Problem and Floor Effects with Low IQ 20

Comparison of WISC-IV and WAIS-III with Low IQ 22

Evaluating Floor Effects in the UK Version of the WISC-IV 23

Ability of WAIS-III to Asses Variability of Clients 24

Accessibility Needs of Individuals with ID 26

Resources for People with ID 27

Significance of Study and Potential Stakeholders 31

Theoretical Framework 33

Summary 33

Chapter 3: Method 35

Whitaker and Gordon's (2012) Method of Assessment 35

Setting 38

Participants 38

Measures 41

Procedures 42

Research Hypotheses 43

Data Analyses 43

Summary 47

Chapter 4: Results 48

Tests of Major Hypothesis 48

Rescoring Protocols Using Adjusted Scores 48

Testing Differences in Index and FSIQ Scores 56

Summary 58

Chapter 5: Discussion 59

Limitations of the Research 59

Future Directions of Research 61

Reflections on the Study 63

References 66

Appendix A: IRB Approval 70

Appendix B: IRB Extension Request 71

Appendix C: IRB Extension Permission 72

List of Tables

Table 1: Raw to Scaled Score Conversion Guide for Digit Span, ages 7:8 to 7:11 37

Table 2: Whitaker and Wood's (2012) Raw to Scaled Score Conversion Guide for Digit Span, ages 7:8 to 7:11 38

Table 3: Ages and Location of Each Child 40

Table 4: WISC-IV Indexes and Core Subtests 42

Table 5: Formulae for Creating Adjusted Scaled Scores 45

Table 6: Child 1 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores 49

Table 7: Child 2 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores 50

Table 8: Child 3 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores 51

Table 9: Child 4 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores 52

Table 10: Child 5 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores 53

Table 11: Child 6 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores 54

Table 12: Child 7 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores 55

Table 13: Means and Standard Deviations of Index and FSIQ Scores 67

Table 14: Wechsler FSIQ Ranges at 90% and 95% Confidence Intervals and Adjusted FSIQ Scores 58

List of Figures

Figure 1. Predicted line based on Wechsler raw to scaled score chart

for Digit Span (DS), Ages 7:8 to 7:11.. 36

Abstract

This study is a pilot re-creation of research in the United Kingdom (UK) by Whitaker and Gordon (2012) that assesses for possible floor effects in the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV). Their study suggested that the Index and Full Scale IQ (FSIQ) scores of low IQ adolescents taking the WISC-IV (UK version) were significantly inflated because low raw scores were converted to scaled scores of 1. Whitaker and Gordon assessed for score inflation and resulting floor effects by creating an alternative scoring system based on the relationship between the lowest raw scores that convert to each WISC-IV scaled score. Since the WISC-IV is the most commonly used intelligence test in school settings in the United States (Riccio, Houston, & Harrison, 1998), similar findings were assessed in the US version of the WISC-IV by completing a pilot replication of the Whitaker and Gordon study. Additionally, I created my own adjusted scoring system that more modestly altered WISC-IV scaled scores. These scaled scores were created based on using the mean of the raw scores that could be converted to each WISC-IV scaled score. The study consisted of 7 de-identified protocols of New England students who obtained a FSIQ less than or equal to 70 and obtained at least one scaled score of 1. Results of the study, however, suggested that Index and FSIQ were not significantly affected by either Whitaker and Gordon's or my alternative scoring systems. The limitations of the study were the small sample size and related constricted demographics. I concluded that this area of IQ research on intellectual disability warrants investigations with large diverse populations.

Keywords: WISC-IV, intellectual disability, floor effects,
intelligence testing, intellectual assessment

WISC-IV and Intellectual Disability:

A Pilot Study on Hidden Floor Effects

Chapter 1

The Wechsler Intelligence Scale for Children, fourth edition (WISC-IV), is an intelligence test commonly used to assess for intellectual disorders (Riccio, Houston, & Harrison, 1998). The Diagnostic and Statistical Manual, fifth edition (DSM-5) encourages a well-rounded psychological assessment, including testing and evaluation of functional impairment, to diagnose intellectual disability (ID); (American Psychiatric Association, 2013). Since a formal diagnosis of ID is required to access disability services (Social Security Administration, 2014, April 3), the WISC-IV indirectly affects social, occupational, and financial governmental and social services. This chapter reviews research that raises concerns about the ability of the WISC-IV to assess ID, highlights the potential negative impact of a possibly invalid intelligence test for test-takers with ID, and outlines the present study.

Statement of the Problem

Clinical presentations of people with ID vary significantly, and their needs for educational, occupational, and social success are highly dependent on their unique strengths and weaknesses. Intellectual testing is commonly used to identify overall cognitive capacity, as well as the particular ways an individual learns and processes information. Information gathered from intelligence testing, such as that provided from the WISC-IV, informs what services are appropriate for children and adolescents with ID. Inflation of abilities leads to a disservice for children and adolescents requiring support and assistance and can have lasting consequences in their ability to access services as adults.

For over a century, intelligence testing has been used to measure a person’s ability to process information (Wechsler, 2003). Many intelligence tests have been revised to suit changing definitions of intelligence and re-normed to represent changes in the population’s cognitive abilities. Currently, intelligence tests are commonly used not only to assess overall intellectual ability, but to inform individualized education programs and aid in differential diagnosis.

The rationale for the study was based on problems identified in Wechsler tests and the significance of an ID diagnosis. There is ample research that older and UK editions of Wechsler intelligence tests have not accurately assessed low intellect (MacLean, McKenzie, Kidd, Murray, & Schwannauer, 2011; Whitaker, 2008; Whitaker, 2010; Whitaker & Wood, 2008). However, there is a lack of research about the US version of the WISC-IV. There is insufficient evidence from non-Wechsler funded research that supports the WISC-IV’s ability to assess low IQ. Due to the significance of an ID diagnosis and the prevalence of Wechsler tests in determining ID, it is imperative that research explore these concerns.

Research suggests that the specific abilities of people with ID vary, meaning that there are larger differences between strengths and weaknesses of people with ID than people with average intellect. MacLean et al. (2011) found that the WAIS-III index scores overgeneralize the large range of abilities represented by subtest scores. While the specific abilities of people with typically developing intellect often cluster around an index score value, one index score may not best represent the varied abilities of a person with ID. The authors argued that it was likely that other Wechsler intelligence tests also overgeneralize due to their use of indexes.

Wechsler tests have been criticized for insufficient norm samples for low IQ (Whitaker, 2008; Whitaker, 2010; Whitaker & Wood, 2008). Since norming samples are used to standardize the test and to create scores that meaningfully compare an individual's results to the general population, such an oversight poses many problems. In particular, an insufficient norming sample is unlikely to represent the abilities of the population in that range. A small sample selection increases the likelihood that anomalies of the group will be generalized to the entire population. A larger sample decreases the likelihood that the sample group will be inappropriately homogeneous or that less common traits will be assumed to be more prevalent in the population, or both. Although the norming sample for children with low IQ has been expanded, Wechsler notes problems with sample selection bias.

Wechsler (2003) reported that the increased norming sample for children with low IQ improved the WISC-IV's ability to assess variability within indexes; however, further verification is necessary. I was unable to find research that corroborates that variability in indexes has improved. Due to the large differences between the indexes, FSIQs of children with low cognitive abilities may overgeneralize their highly varied strengths and weaknesses.

Additionally, hidden floor effects have been hypothesized in Wechsler tests. Floor effects occur when a score, such as a FSIQ, cannot accurately measure below a particular value. Wechsler (2003) stated that the WISC-IV's floor is at a FSIQ of 40. Studies of UK versions of the WAIS-III, WAIS-IV, and WISC-III suggested floor effects may occur as high as FSIQ of 70 (Whitaker, 2008; Whitaker, 2010; Whitaker & Wood, 2008).

Whitaker and Wood (2008) posited that floor effects may occur at FSIQ of 70 for two reasons: (a) because low raw scores are scaled to scores of 1, including raw scores of zero, and (b) because the distribution of intellect in the population is assumed to be normal bell-shaped. Whitaker (2008, 2010) and Whitaker and Wood (2008) argued that the scaled score of 1 represents both children who perform extremely poorly and children who cannot perform the task at all. As a result, he hypothesized that the scaled score of 1 becomes meaningless because it does not differentiate between severe low ability and no ability to complete subtest tasks.

Whitaker’s (2008; 2010) point may be better understood by citing examples of scoring scenarios where scaled scores of 1 are obtained. The WISC-IV Administration and Scoring Manual (Wechsler, 2003) provides scoring tables to convert raw scores to scaled scores. A 7-year-old may get a scaled score of 1 on the Coding subtest for raw scores of zero through 7. The scaled score of 1 might indicate any number of weaknesses in an examinee, such as the child not understanding the task despite sample items, poor fine-motor coordination, or low processing speed. The supplementary processing speed subtest Cancellation may be used to replace Coding in scoring the index or FSIQ. However, raw scores of zero through 9 also are scaled to 1 on the Cancellation subtest. In sum, a scaled score of 1 encompasses a wide range of low ability, and there are few options within WISC-IV to measure the low ability precisely. Although an assessor may be able to distinguish the cause of poor performance through qualitative information, the objectivity of the measure is compromised.

It may not be as apparent for an assessor to determine why a child or adolescent performed poorly in the next scoring example. In this situation, two 16-year-old adolescents can obtain the same scaled score on Picture Concepts in two vastly different ways. One adolescent might not be able to meet the baseline to start at item 7, the standard start point for teens in this age range. He or she might struggle through the initial items, and receive a scaled score of 1. Another adolescent could meet baseline criteria at item 7 and score points up until item 11. This teen also receives a scaled score of 1. The supplementary for the Perceptual Reasoning index is Picture Completion should the clinician deem this subtest invalid. On Picture Completion, an examinee may receive raw scores between zero and 15, with no indication in the scaled score whether he or she was able to perform at the start point for 16-year-olds (item 10), or if the test-taker had to return to earlier items.

Whitaker and Wood’s research (2008) suggested that due to the high number of scaled scores of 1 received by people with FSIQ below 70, it is possible FSIQ scores are inflated. For children and adolescents with FSIQ scores in the 60s and 70s, a few lower points may be the deciding factor for a disability diagnosis or disability services. They hypothesized that more test-takers would get FSIQ scores below 70 if the Wechsler (2003) scoring system used a scoring method that did not scale extremely low raw scores to scaled scores of 1.

In addition, Whitaker (Whitaker, 2008; Whitaker, 2010; Whitaker & Wood, 2008) expressed concern that Wechsler tests assumed a normal distribution in a population’s intellect. Instead, Whitaker argued that intellect is likely bimodal. He cited the increase in disorders affiliated with low intellect (i.e. autism spectrum disorders). He suggested that an assumed normal curve may affect the standardization of the test. The Wechsler Technical and Interpretive Manual (*Manual*; Wechsler, 2003) does not include the distribution of FSIQ scores collected from their standardization sample. The *Manual* refers to collecting stratified samples based on age, sex, race, parent education level, and geographic location (pp. 20-21). Due to the limited psychometric information in the *Manual*, it is unclear if and how an assumed distribution affects low IQ scores on the WISC-IV.

Finally, the study was important due to the significance of an ID diagnosis and the role intelligence testing plays in ID assessment. The DSM-5 diagnostic criteria for ID include: deficits in general cognitive abilities, significant problems in functioning as a result of cognitive deficits, and onset during the developmental period (American Psychiatric Association, 2013). The DSM-5 encourages diagnosis not to be based entirely on intelligence testing, but instead on a thorough psychological evaluation in conjunction with testing. Consistent with mental retardation diagnosis in the DSM-IV, ID would be suggested by an intelligence quotient (IQ) score at least two standard deviations below the mean (at or below a FSIQ of 70 on the WISC-IV; Wechsler, 2003).

I argued that while diagnosis of intellectual disability will be improved when testing is used in conjunction with assessment; I had concerns about organizations that may rely only on intelligence testing scores for quick disbursement of services. Specifically, insurance companies and social services agencies are inundated with requests for services and may use test results to inform the services they will or will not provide. While competent assessors might integrate both qualitative and quantitative data into their assessment, it is likely that many providers will continue to use test scores as determinants for service disbursement. Thus, the WISC-IV must accurately measure low intellect not only to accurately inform assessors, but so that children and adolescents may access services provided by providers that primarily rely on test scores.

Data analyses of the present study measured the difference in index and FSIQ scores with the creation of a theoretical scaled score of zero, and the extent to which children and adolescents' scores were inflated so that they no longer met intelligence testing diagnostic criteria for ID. This study tested the limits to which scaled scores of 1 inflated index and FSIQ scores in WISC-IV protocols with FSIQ below 70 by creating two alternative scoring systems. The first alternative scoring system changed all raw scores of zero to theoretical scaled scores of zero to see if and to what extent raw scores of zero affected index and FSIQ. The second alternative scoring system changed all scaled scores of 1 to theoretical scaled scores of zero to measure if index and FSIQ scores significantly differed when all scaled scores of 1 were assumed to be inflated. By using these two alternative scoring systems, this study assessed if Whitaker's (2008, 2010) hypothesis about inflated scaled scores warrants further investigation.

The range of FSIQ below 70 was selected because individuals with these scores are at increased risk for being misdiagnosed. I hypothesized that children and adolescents with very low intellect were at less risk for being misdiagnosed due to the severity of their limitations and because of revisions to the DSM-5 (2013) that emphasized functional impairment. I argued that even slight inflation in FSIQ scores below 70 may lead some assessors and support service agencies to deny diagnosis of ID since functional impairment would be less overt than in further lower IQs. As a result, people in this IQ range are at increased risk to miss criteria for ID and subsequent services.

Grant Funding for the Study

The present study was awarded a $10,000 grant by the Social Security Disability Determination Small Grants Program. The grant was awarded for research that may inform how social security disability funds are distributed. Funding was not dependent on study results, and the grant program did not express any investment in a particular finding from the study.

Research Questions

The study answered the following questions:

1. Are hidden floor effects hypothesized in the UK version of the WISC-IV present in the US version?
2. Is there evidence that scaled scores of 1 significantly inflate index and FSIQ scores?
3. Is there a significant difference in the number of children and adolescents who might qualify for a diagnosis of ID when a scaled score of zero is utilized?

Definition of Terms

Intelligence. Wechsler (1939) defined intelligence as the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his or her environment. He said that intelligence is not defined by one single ability, but the composite of multiple abilities. For example, indexes of the WISC-IV break intelligence into four main factors, or indexes, described in this section under “index score.” Collectively, these four factors represent major domains that comprise effective cognitive abilities.

Mental retardation (MR). Prior to the DSM-5, the diagnosis of MR was given to people with significantly impacted cognitive abilities. The Diagnostic and Statistical Manual of Mental Disorders, fourth edition, text revision (DSM-IV-TR; American Psychiatric Association, 2000), had three criteria for MR. First, a person had significant deficits in intellectual abilities as evidenced by an IQ score of 70 or below on a standardized intelligence measure. The second criterion was significant impairments in adaptive functioning. This was determined by problems in two or more areas of “communication, self-care, home living, social/interpersonal skills, use of community resources, self-direction, functional academic skills, work, leisure, health, and safety” (American Psychiatric Association, 2000, p. 41). Finally, the third criterion was the onset of symptoms must occur before age 18.

Intellectual Disability. The DSM-5 diagnosis of intellectual disability (also called intellectual developmental disorder) replaced the DSM-IV diagnosis of MR. There are many similarities between the MR and ID diagnoses, but the key difference is that ID heavily focuses on functional impairment rather than intelligence testing. The three criteria for the diagnosis of ID are described below.

The first criterion is that a person experiences difficulty in general mental abilities related to “reasoning, problem-solving, planning, abstract thinking, judgment, academic learning and learning from experience confirmed by both clinical assessment and individualized, standardized intelligence testing” (American Psychiatric Association, 2013, 33). Intellectual deficits must be both objectively measured by intelligence tests, such as the WISC-IV, but also observed by trained clinical professionals. In other words, a FSIQ less than 70 without observed impairment would not suffice for diagnoses, nor would a person with a FSIQ in the normal range but with challenges in the domains listed above.

The second criterion is that the deficit in mental abilities must significantly affect performance in one or more aspects of daily life, like “communication, social participation, and independent living, across multiple environments, such as home, school, work, and community” (American Psychiatric Association, 2013, p. 33). This is slightly more descriptive than the criterion for MR, which did not specify any requirement for needed support.

The final required criterion is that onset must occur during the developmental period (American Psychiatric Association, 2013). This is different from the MR criterion that required that a person have symptoms present as a child or adolescent. The term “developmental period” allows flexibility in the observation of symptoms to young adulthood, where some individuals in the very mild 65-70 IQ range may display significant difficulties adjusting to independent living. For young adults who may have had their intellectual needs neglected as children and adolescents, this offers them opportunities to be accurately diagnosed retroactively and to be potentially provided support services for people with ID.

Severity of ID is classified as mild, moderate, severe, or profound. Criteria for each ID specifier are based on qualitative information and the individual's functioning (American Psychiatric Association, 2013). This is different from the DSM-IV's MR severity specifiers, which were based on actual or estimated IQ level (American Psychiatric Association, 2000). Additionally, the DSM-5 has created a separate diagnosis of unspecified intellectual disability for when a person over the age of 5 is unable to be assessed due to physiological or co-morbid disorders that impact assessment. This has replaced the DSM-IV's diagnosis of MR, unspecified.

Raw score. After the administration of each subtest on the WISC-IV, a raw score is calculated based on the scoring criteria in the manual (Wechsler, 2003). Often 0-2 points are awarded for each item on a subtest, based on the accuracy of the responses. The sum of the items provides a raw score for each subtest.

Scaled score. The scaling process translates the child's subtest raw score to a standard score that is meaningful when the child is compared to their same-aged peers. This is practical since the raw scores of young children are likely to much lower than older adolescents. The scaled scores range from 0-19 with a mean at 10. Once scaled scores are calculated, a child's performance on each subtest can be easily mapped as at above or below the average abilities of other children within that age group. These scores are calculated using a table in the WISC-IV scoring handbook or by scoring software, and the scaled scores were developed based on the test norming sample (Wechsler, 2003).

Index score. Indexes represent a person's relative intellectual strengths and weakness (Flanagan, & Kaufman, 2009). Each index is comprised of particular subtests, and index scores are calculated from the standard scores of subtests. The index scores are used to measure ability, such as Verbal Comprehension (VCI), Working Memory (WMI), Perceptual Reasoning (PRI), and Processing Speed (PSI). The VCI is calculated from the subtest scores on Similarities, Vocabulary, and Comprehension (Information and Word Reasoning are supplemental subtests). The PSI is calculated from Block Design, Picture Concepts, and Matrix Reasoning (Picture Completion is a supplementary subtest). The WMI is calculated from Letter-Number Sequencing and Digit Span (with Arithmetic is a supplementary subtest). The PSI is calculated from Coding and Symbol Search (Cancellation is a supplementary subtest). Indexes are thought to have more utility in assessing strengths and weaknesses because subtests are too specific and variable to reliably infer broad abilities.

Full scale IQ (FSIQ). The FSIQ is a numerical value that represents an individual's general intellectual ability. It is calculated from the index scores and does not represent any relative strengths or weaknesses. For many with typical intellectual development, this value suffices to generalize intellect because an individual's intellectual strengths and weaknesses tend not to differ greatly (Flanagan & Kaufman, 2009). In other words, a person with a typically developing intellect may have personal strengths and weaknesses, but generally his or her abilities measured on the indexes will not significantly differ. The FSIQ can still represent how the person generally performs. For people with ID, FSIQ is less useful because there tends to be greater differences between the abilities represented within an index score (MacLean et al., 2011). Additionally, differences between subtests are likely to be significant. By using scores that generalize multiple abilities, the significant differences in strengths and weaknesses are lost in an averaged value.

Floor effects. This term represents a phenomenon that occurs when a test is unable to measure below a particular value. A common result of floor effects is that an examinee obtains an inaccurate, higher score. Examples of how floor effects could occur include not having enough “easy” items on a subtest so that the examinee can meet an appropriate baseline, or when there are not enough easy items to describe the examinee's abilities to perform on the subtest. Hidden floor effects refer to floor effects that are not necessarily obvious to an examiner. An example of a hidden floor effect would be if the items required to meet baseline were significantly easier than the later items on a subtest. Although the examinee may be able to meet baseline criteria, the items of the test still measure beyond the abilities of the examinee.

Summary

Intelligence testing is a key component in the diagnosis, treatment, and support of people with ID. It is crucial that intelligence tests provide accurate clinical information about a person's overall intellectual ability, strengths, and weaknesses. Using measures that do not properly assess a child or adolescent's intellectual ability can lead to misdiagnoses, denial of appropriate interventions, and limited government benefits. Limitations of the WISC-IV must be researched in order for it to either be improved or not used for people with low cognitive ability. Next is a brief literature review on Wechsler intelligence tests, features of MR/ID, and floor effects seen in UK versions of the Wechsler measures.

Chapter 2: A Review of Relevant Literature

This chapter summarizes pertinent literature that informed the study's research. The purpose of the literature review was to inform the reader about key topics that are related to the study and to briefly review current understanding of the assessment of ID. Specifically, this chapter provides information about ID and how assessment is used in the diagnosis and treatment of people with this disorder. The WISC-IV is reviewed, and its normative scoring, assessment practice, and use with an ID population are presented. Finally, specific needs and accessibility options for people with ID is described.

Wechsler Intelligence Tests

Standardized tests are necessary to diagnose intellectual disorders in the DSM-5 (2013). Additionally, intelligence tests are frequently used to inform treatment because they may identify strengths and a weakness in a student's learning style. Thus, the assessment measures used to assess for ID must be studied and critiqued for their validity with the ID population.

Wechsler Four-Factor Model

The Wechsler intelligence tests were the primary measures for assessing intellectual ability for many years. Published in 1939 by David Wechsler, the Wechsler-Bellevue Intelligence Scale was the first of the Wechsler IQ tests (Wechsler, 2003). At the time, they were developed without theory, and Wechsler believed that tests gave insight into a client's personality. Since then, Wechsler-based assessment tests have been revised to incorporate a four factor model of understanding intelligence and are used internationally. The most common two Wechsler tests are described here.

The Wechsler Intelligence Scale for Children (WISC) has had four editions, with the most recent being completed in 2003. The second measure, for people aged 16 years to 90 years old, is the Wechsler Adult Intelligence Scale (WAIS). This tool has had four editions, with the most recent release in 2008. Both the WAIS-IV and WISC-IV are based on the Wechsler four-factor model. The four factors, or indexes, are Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed.

The Wechsler model theorizes that the indexes represent the four main domains of intelligence (Wechsler, 2003). The Verbal Comprehension Index (VCI) measures verbal knowledge and comprehension, and is often seen as a good predictor of scholastic achievement. The Perceptual Reasoning Index (PRI) measures fluid reasoning and perceptual and organizational skills. Fluid Reasoning is the ability to apply learned skills to novel or unfamiliar situations, which often utilizes perceptual and organizational skills. The Working Memory Index (WMI) measures short-term auditory memory, concentration, and attention. Finally, the Processing Speed Index (PSI) measures the speed that one processes nonverbal visual information.

Standardization of the WISC-IV

The WISC-IV is an intelligence test for children and adolescents aged 6 years to 16 years, 11 months (Wechsler, 2003). It purports to measure intellect from $40 \leq \text{FSIQ} \leq 160$. Norms for the test were developed in a five-stage process, beginning with Conceptual Development. The Pilot Stage focused on details related to the new Wechsler subtests, such as content, relevance of the items, subtests floor effects, and the order of the subtests. The National Tryout State used information from a stratified sample of 1,270 children. Stratification was based on information collected in U.S. Censuses (1998 and 2000), including age, sex, race, parent education level, and geographic region. Additional data were collected for special groups, including children with MR (*n* not noted in the *Manual*). The Standardization Phase used a stratified sample of 2,200 children, with 200 samples per age bracket. The Arithmetic subtest was based on a stratified subsample of 1,100 children, with 100 children per age group. Samples were identified using trained recruiters and independent examiners. Some children were excluded from the study, including those who were taking medication that might depress performance, such as antipsychotics and antidepressants. Approximately 5.7% of the norming sample was added to “accurately represent the population of children attending school” (Wechsler, 2003, p. 23). No further information was provided to clarify the demographics of the 5.7% or children attending school. Within each age bracket, samples were collected from a range of intellectual ability falling within a normal curve. In other words, more individuals with average intellectual ability (90<FSIQ<110) were sampled than individuals with extremely high or low intellectual ability. Validity for the WISC-IV was assessed by the sample’s scores to scores from the Children’s Memory Scale, Gifted Rating Scale, Baron EQ, Adapted Behavior Assessment System II and other Wechsler tests (i.e. the WISC-III, WAIS-III, Wechsler Preschool and Primary Scale of Intelligence III, Wechsler Abbreviated Scale of Intelligence, and Wechsler Individual Achievement Test). Finally, there was Final Assembly and Evaluation and Quality Assurance Procedures.

Wechsler (2003) reported that 120 children were used to create a standardization sample for FSIQ scores less than or equal to 70. This is approximately 5.5% of their 2,200 person norming sample. Wechsler cites that literature estimates 2.5% to 3.0% of the general population meets criteria for MR, and that 2.2% of children would test below a FSIQ of 70 if there was a normal distribution of intellect. Of 120 children used in this sample, 63 were in the mild severity group (60.5≤FSIQ≤73) and 57 children were in the moderate severity group (46.6≤FSIQ≤58.2). They reported less variability in the index scores in this sample (standard deviations of 9.1 to

11.6, depending for mild MR and 7.5 to 11.0 for moderated MR) than in the general sample (SD=15).

Despite this information, Whitaker (2008, 2010) posited that the sample was likely affected by selection bias. He reported children in the sample were relatively high-functioning and that hospitalized children were excluded from the study. Additionally, I argue that excluding children on medication likely affected the sample due to the high prevalence of children with MR/ID on medication. Depending on the abilities and functioning of children in the WISC-IV sample, it is possible that their sample was not highly representative of children and adolescents with MR/ID.

I also posit that Wechsler is minimizing the variability of index scores within each sample. Although the standard deviations for each index may be less than the general population, Wechsler (2003) found that 16.7% of children with FSIQ less than 79 points had PRI scores 15 or more points higher than VCI scores, and 10.2% of children in this range had VCI scores 15 or more points higher than their PRI scores. Thus, at least 26.9% of children with FSIQ scores below 79 have index scores that differ by 15 or more points. This suggests to me that variance is likely not occurring within each index, but between the indexes. No information on the prevalence of invalid FSIQ scores was reported in the manual.

All Wechsler intelligence tests are based on a similar scoring system. The assessment consists of a number of individual tasks that fall under subtests. Each subtest has a specific mode of administration, which is described in the administration manual. The examinee is explained the task of the subtest, and in some instances, practice items are given to help the examinee orient himself or herself to the task. A start point is predetermined based on the age of the examinee, unless the examiner is modifying the administration to accommodate the examinee’s need. Many of the subtests require a minimum baseline achievement for the subtest to be valid.

The examinee receives points for correct and partially correct responses, as specified by the administration manual. The sum of these points is the raw score for a subtest. The raw score is then compared to scores of a normative age or grade equivalent sample, and a scaled score is created. The scaled scores of a subtest can range from 0 to 19, with a mean of 10. The scaled scores allow for the scores of the examinee to quickly be understood as being at, above or below the peer average of 10.

Subtests are then grouped into one of four indexes (Wechsler, 2003). The WISC-IV Verbal Comprehension Index includes the subtests similarities, vocabulary, comprehension, word reasoning, and information (supplemental). The Working Memory Index is made of the subtests digit span, letter-numbering sequencing, and arithmetic (supplemental). The Perceptual Reasoning Index is comprised of subtests block design, matrix reasoning, visual puzzles, and picture completion (supplemental). Finally, the Processing Speed Index is made of subtests symbol search, coding, and cancellation (supplemental). The FSIQ is calculated from the Index scores.

Concerns about WISC-III and WAIS-III for Assessing LD

Whitaker (2008) discussed three concerns about using the UK versions of the WISC-III and WAIS-III with people who have learning disabilities. The first concern was about significant differences in the difficulty level of some test items with regard to low test scores. The second was about floor effects greater than what are acknowledged in the administration manuals. The third concern was that the WISC-IV, UK version, norming system was based on the inaccurate assumptions that low-end IQ scores falls within a normal distribution. While Whitaker also discussed norming problems with the UK version, this review will focus on areas of concern applicable to US populations. Although this study focused on only one of his concerns, all three will be reviewed.

Whitaker's (2008) first concern about UK Wechsler tests is that the degree of difficulty on the practice portion of the subtests exceeds the degree of difficulty of the actual test items. In other words, understanding the directions to subtests often requires more ability than was needed to perform the task of the subtest. Although some subtests have demonstration items within the instructions, many instructions are entirely verbal and, therefore, abstract. In order for someone to excel on those subtests with only verbal instructions, he or she must first be able to comprehend the complex, and often lengthy, instructions. Thus, the person is not just being assessed for the task of the subtest, but also on the ability to attend to and understand the directions. This is problematic because if a child is unable to meet baseline criteria for a subtest because of the difficulty level of the instructions, it appears that he or she is unable to complete the task for which the subtest is assessing.

Whitaker's (2008) second concern was that the floor effect acknowledged for both the UK versions of the WAIS-III and WISC-III was much higher than described for people with learning disabilities. The reported limits of the UK WAIS-III and WISC-III were IQs of 45 and 40, respectively. Whitaker reported that because of the ways that raw scores were scaled, people with a “true” IQ in the 30s were instead receiving FSIQ scores in the 40s. He hypothesized that this was because raw scores of zero are scaled to 1 on all subtests (Wechsler, 2003). For individuals who score zero due to challenges of ID and not because of age, this creates a floor effect. Whitaker argued that it is not practical that a raw score of zero receives a scaled score of 1 because a raw score of zero could indicate no ability whatsoever. He argued that a scaled score of zero, which does not exist on the UK or US versions of the WISC-IV (Wechsler, 2003), should be created to differentiate between an individual who obtained zero points on a subscale and someone who obtained very few points (Whitaker, 2008). Since the scaled score of 1 may represent people who could not perform on the subtest, he posited that the lowest scaled score that could be interpreted with confidence would be 2.

Finally, Whitaker (2008) said that because the UK and US WAIS-III and WISC-IV norms are based on the theoretical normal distribution, low-end scores are not given an accurate percentile rank. In a normal distribution, it is relatively easy to identify a percentile rank because the population is balanced on either side of the mean (Anastasi & Urbina, 1997). When a curve is skewed, the population is not evenly plotted on either side of the mean. Instead, more or less of the population may be above or below the mean. Wechsler models base their percentile ranks on an assumed normal distribution (mean/median/mode FSIQ = 100). Whitaker wrote that low IQ scores did not follow a normal distribution because the intellect distribution is bimodal with a second mode estimated in the low intellect range. As a result, a person estimated in a low percentile rank may actually be at a higher percentile rank because there are more people with lower IQs than estimated. Therefore, placing low IQ scores in an ideal distribution provides them inaccurate percentile ranks. Thus, the IQ distribution should be represented as a bimodal distribution, and not as a normal distribution. As a result of using an assumed inappropriate normal distribution, people with ID are placed at much lower percentile ranks than if the test had been normed assuming a theoretical bimodal distribution.

Distribution Problems and Floor Effects with Low IQ

Whitaker and Wood (2008) elaborated on Whitaker’s (2005) points about the floor effects and distribution of scaled scores on the UK and US version WISC-III and WAIS-III. They discussed that both UK version WISC-III and WAIS-III manuals (Wechsler, 1991; Wechsler,

1997) state that a FSIQ should not be calculated unless the client has raw scores above 0 on a minimum of three Verbal and three Performance subtests. The authors argued that the stipulation was not sufficient. They argued that the UK WISC-IV scoring system should be changed so that raw scores of zero can create reliable and valid FSIQs.

Whitaker and Wood (2008) collected data from 49 UK version WAIS-III and 50 WISC-III assessment protocols to analyze for floor effects and scoring problems resulting from assuming a normal distribution of scaled scores. The study analyzed the floor effects for tests that received FSIQs in the 50s, 60s, and 70s. They measured the influence of scaled scores of 1 despite raw scores of 0, and how scaled scores of 1 may influence the three groups of IQ scores. They also measured how the distribution of scaled scores indicated hidden floor effects.

Whitaker and Wood (2008) administered and reviewed UK versions of the WISC-III and WAIS-IV. The average age of adults administered the WAIS-III was 40 years 4 months, and the average age of children administered the WISC-III was 11 years 9 months. No information was provided on the samples' racial or cultural diversity. Scores fell within 50-59, 60-69, and 70 plus. Thirteen clients given the WISC-III had FSIQs less than 50, and 1 client on the WAIS-III scored less than 50. Scores of 40-49 were not included in the analysis.

There was no significant difference in the mean FSIQs between the UK versions of the WAIS-III and the WISC-III. On the WISC-III, for scores in the 40s, 50s, and 60s, there were more scaled scores of 1 than any other scaled score. For FSIQs in the 70s, there were more scaled scores of 5 than of 1. Taking all FSIQ scores together, the distribution was bimodal (peaks for 1 and 5) and scaled scores of 1 were the second most frequent scaled score. Overall, the study found that the UK WISC-III had a relatively large number of scaled scores of 1 for FSIQs less than 60. To a lesser extent, the WAIS-III also had a high number of scaled scores of 1 for FSIQs less than 60. For Full Scale IQ scores above 60, the WISC-III showed more scale scores of 1 than would have been expected by chance alone.

Whitaker and Wood (2008) wrote that it was unclear why there was a difference between the number of scaled scores of 1 between the two measures. Since the groups did not have significant difference in the distribution of mean IQs, it was unlikely that the anomaly was due to a sampling error. One explanation was that the UK WISC-III may have had harder criteria to achieve a scale score of 2 than the UK version of the WAIS-III. The authors looked at the requirements for a 16-year-old (an age at which a client could take either test) to get a 2 on both measures. They found that the raw score required for a scale score of 2 on the WISC-III were much higher than the raw score needed on the WAIS-III. It is likely that these occurrences were seen on both measures because the measures were created using the same four factor model for scoring.

One reason that Whitaker and Wood (2008) attributed to the floor effects was the relatively small sample of people with low IQs used to norm the UK version of the WAIS-III and WISC-III. Each test was normed with approximately 200 people for each age range. This meant that there were only five people with IQ of 70 or scaled score of 4, and no people below an IQ score 58 or scaled score 2 (Whitaker & Wood, 2008). In other words, there were too few subjects to effectively norm scaled scores of 2 and 3. The result was an inaccurate assumed normative curve and a floor effect much higher than reported by the WISC-III and WAIS-IV on the measures.

Comparison of WISC-IV and WAIS-III for Low IQ Scores

Whitaker (2008) compared the abilities of UK versions of the WISC-III and WISC-IV to assess low IQ. Whitaker expressed the same three concerns about the UK WISC-IV for assessing children with low IQs as he did for the UK WISC-III in his previous article (Whitaker, 2005). In addition, he presented evidence that the UK WISC-IV may give lower IQ scores than the UK WAIS-III when assessing for low IQ. This was similar to the findings of Whitaker and Wood (2008) that did assessment with the UK WISC-III and WAIS-III.

Whitaker (2008) discussed that the WISC-IV had slightly improved from the WISC-III and the WAIS-III, but that he still had concerns about its ability to assess low IQ. He repeated that it is still uncertain if children with low IQ are able to understand the subtest directions to the extent necessary for completing the subtest tasks. It is unclear if the floor effects of the WISC-III found by Whitaker and Wood (2008) can be found on the WISC-IV. Finally, as discussed previously, the current percentile ranks of the WISC-IV are based on a normative population’s distribution that does not represent the actual distribution of low IQ scores.

Whitaker (2008) suggested options to address these problems. Similar to his previous writing (Whitaker, 2005), he suggested changing scaled scores so that a raw score of 0 does not ever result in a scaled score of 1. Finally, he advocated for increasing the sample size for norming individuals with ID so that a more accurate distribution can be created.

Evaluating Floor Effects in the UK Version of the WISC-IV

Whitaker and Gordon (2012) researched Whitaker's (Whitaker, 2008; Whitaker, 2010; Whitaker & Wood, 2008) hypothesis that scaled scores of 1 create a hidden floor effect in the UK version of the WISC-IV. To measure their hypothesis, they created adjusted scaled scores for raw scores that otherwise would have been scaled to 1. From the scoring tables found in the *Manual* (2003), they extrapolated the algorithm the Wechsler tests used to distribute scaled scores less than or equal to 10. Whitaker and Gordon then applied the algorithm to very low raw scores, creating adjusted scores of 0 and below. They calculated indexes and FSIQ using the adjusted scores to see if, and to what extent, they varied.

Whitaker and Wood (2008)found that 45 out of 66 raw scores that would have been scaled to 1 qualified for a lower scaled score using their method. Furthermore, nine out of 17 subjects had a reduced FSIQ after their scores had been adjusted. Of these nine scores, four were within six points of the original FSIQ, and five had a greater reduction. These change in these scores created a significantly different score distribution, and Whitaker and Wood posited that the change in score distribution suggested of a floor effect in the UK version of the WISC-IV.

Ability of WAIS-III to Assess Variability of Clients

MacLean et al. (2011) did a study measuring the invariance of assessment of ID with the US version of the WAIS-III to see if the WAIS-III index scores accurately represented the abilities of people with ID. They acknowledged that for people with IQs in the normative range of 90<FSIQ<110, the WAIS four-factor model can accurately assess intelligence because the abilities measured in each index were similar enough to be represented by one index score. The authors reported that people with low IQ often had varying and unpredictable strengths and weaknesses between subtests within each index. The formulation of an index score for people with low IQ (IQ of 70 or below) did not accurately represent the abilities of each person. In other words, a person with very similar scale scores on subtests within an index could potentially receive the same index score as someone with highly varying subtest scaled scores.

MacLean et al. (2011) reviewed 404 US WAIS-III tests from an intellectual disability service. The files were divided based on level of impairment, with one sample containing 140 tests with Full Scale IQs of 55-69, and the other sample containing 264 assessment with scores less than 55. The assessment files included information on gender, age, subtest scores, and intervention and support requirements.

A confirmatory factor analysis was performed on subtest scores using both 11 and 13 subtest scores to a hypothesized normative population curve. The object assembly subtest was not used in the analysis because it does not contribute to the Full Scale IQ or index scores. The four factor model was tested by comparing the variability in the scores of the subtests with their respective index.

Results of a Kolmogorov-Smirnov test, used to assess the probability of distribution, suggested that all the subtest distributions from the Wechsler tests differed from a normal distribution for both samples. To adjust for the difference from the mean expected by the scores in the samples, a ROBUST option scaled statistic was used to correct the sampling distribution so that it was closer to the mean of the normal distribution and to evaluate the goodness-of-fit for the model. Bentler (as cited in MacLean et al., 2011) explained that the ROBUST statistic is best when using non-normative data. Both the comparative fit index and root mean square error of approximation were used to assess goodness-of-fit, and neither found that the hypothesized distribution had a good fit for either 11 or 13 subtests.

The results of the study (MacLean et al., 2011) suggested that the four factor model used in the US WAIS-III is not appropriate for assessing individuals with ID because it assumes a normative population distribution. This assumption causes the varying strengths and weaknesses of people with ID to be overgeneralized within the index scores.

MacLean et al. (2011) hypothesized three reasons for why they did not find the model to be a good fit. The reasons were a statistical flaw caused by floor effects evidenced by a positive skew in their collected data; that the data were Full Scale IQ scores that were based on flawed index scores; and that there is a real difference in abilities with people with low IQ that is underrepresented in index scores. It is unclear which of the above three reasons was causing a poor fit between the hypothesized distribution and the collected sample distribution of scores from ID subjects.

Accessibility Needs of Individuals with ID

Yalon-Chamovitz (2009) is an occupational therapist who wrote about the accessibility needs of people with ID in order to build a conceptualization for treatment. She argued that legislative rights for individuals with ID allowed them to have the maximum “independence, privacy, and dignity.” Accessibility, defined as the availability of services and resources to the greatest number of people possible, is a large component of the needs of people with ID. Yalon-Chamovitz described how accessibility is emphasized in the rights of people with physical and sensory disabilities, but is not as present in the rights of people with ID. She highlights four main areas of need for accessibility for individuals with ID.

The first domain of need is *pace*, or the rate at which people function. She sites many studies that show people with ID have slower reaction times and processing speeds in many settings and for many tasks. Historically, people with ID were expected to adjust over time and develop faster pace, but the shift from a medical model of conceptualization to a social model placed focus on accommodation rather than adaptation.

The second area of need for accommodations for people with ID is an appropriate *complexity* level of communication. Complexity can refer to many parts or areas needing attention, or it can refer to need for a high level of understanding or problem-solving process. Yalon-Chamovitz (2009) gave an example of poor accessibility as when someone speaks louder to a person with ID rather than speaking in simpler terms. She wrote that complex communication limits accessibility to information for people with ID, and that simple language can greatly improve the ability for a person with ID to thrive. Thus, if intelligence tests, such as the WISC-IV, do not use simple instructions, these tests may not be accessible for people with ID. If tests are not accessible for people with ID, then the test results likely do not represent a client's full abilities.

The third accessibility issue Yalon-Chamovitz (2009) discussed was *literacy*. She cited that people with ID have significantly lower literacy skills than the general population (Kirsh, Jungeblut, Jenkins, & Kolstad, 1993), meaning that people with ID miss a great deal in a society driven by literacy. Solutions to illiteracy issues include simple language, pictorial communication, and alternative options, such as available audio. Therefore, if an intelligence test did not have accommodations for people with limited literacy, it may be presumed that the test would not be accessible for those people.

Yalon-Chamovitz's (2009) fourth and final accessibility need for people with ID was an *elimination of stigma*. She wrote that the needs of people with ID are largely not met because of stigma. When lawmakers, providers, and laypersons hold stigma, accessibility to services is often denied. Thus, stigma around ID needs to be removed in order to gain rights for people with ID and have these rights enforced appropriately.

Resources for People with ID

Recent changes in legislation have provided people with ID access to rights that facilitate many of the needs that Yalon-Chamovitz (2009) discussed. Legislation, such as the Americans with Disabilities Act (ADA; 1990) and the ADA Amendments Act of 2008 (2008), provided rights to people with ID, such as equal pay and access to public entities. The Acts were passed to promote equality and diminish discrimination against individuals with physical and mental disabilities. Subsequently, the Acts have allowed individuals with ID access to specialized education and social services, such as financial disability benefits.

To be covered by the Acts, a person must fall under one of these criteria: “(A) physical or mental impairment that substantially limits one or more major life activities of such individual; (B) a record of such an impairment; or (C) being regarded as having such an impairment” (ADA, 2008). When compared to the diagnostic criteria for both MR and ID, as defined above, it appears that all people who qualify for MR (and likely ID) will also be eligible for the ADA. Unfortunately, the ADA allots services based on an individual level of need, and thus does not address the larger societal problems related to expected pace, complexity of communication, literacy, and stigma.

Additionally, as assistance is provided on an individual basis, services are not necessarily standardized. This has potentially large implications for individuals with ID who do not have advocates to ensure that all of their needs are appropriately managed. Instead, the satisfactory level of services is subjective to those giving and obtaining services, and potentially below the level of services required for the person with ID to achieve.

This is highlighted in the process for people to get supplemental security income (SSI) through the Social Security Administration (SSA). Per their definition, a disability is defined as the “inability to engage in any substantial gainful activity (SGA) by reason of any medically determinable physical or mental impairment(s)....” (Social Security Administration, 2014, April 4). These impairments must impact the person or child for longer than one year, and meet requirements determined in a sequential evaluation process. The sequential evaluation process for children and adolescents (aged 22 or less) includes a “review of the child's current work activity (if any), the severity of his or her impairment(s), and an assessment of whether his or her impairment(s) results in marked and severe functional limitations” (Social Security Administration, 2014, April 3).

In order to meet SSI criteria for Intellectual Disability, one of six possible requirements must be observed in the sequential evaluation process. Requirement A is for children aged 1 to 3, and is supported by functioning no more than two-thirds of the child's chronological age. Requirement B is for children who are grossly dependent on others, and whom would be inappropriate to assess with standardized intellectual tests due to limited functioning. Requirement C is a verbal, performance, or FSIQ of 59 or less. Requirement D is a verbal, performance, or FSIQ of 60 through 70 and another physical or mental impairment creating significant deficits in functioning. Requirement E is a verbal, performance, or FSIQ of 60 to 70 and resulting in documented impairment in age-appropriate social functioning, personal functioning, or difficulties in maintaining concentration, persistence, or pace. Finally, requirement F is marked impairment in age-appropriate cognitive or communicative function, and another documented physical or mental impairment causing additional challenges.

In sum, three out of 6 requirements are directly linked with intellectual testing results (requirements C, D, and E). One other requirement (requirement A) is for younger children that may not have access to valid intelligence testing in their age range, and another requirement (requirement B) is for children with whom intellectual testing of any level would be too challenging because of the child's impairments. Thus, five out of six options for individuals to qualify for SSI are based, at least in part, on intelligence testing results.

Significance of the Study and Potential Stakeholders

The WAIS-IV is used to measure intellectual ability because it is considered an invariant test of intellect between the FSIQ range of 40 to 160 (Wechsler, 2003). If any hidden floor effects were found in the WISC-IV, it would suggest that the test does vary when assessing lower tiers of intellect. Variant intelligence tests could have ramifications for individuals with ID, their families, schools, and organizations that fund and support people with ID.

People with ID. People with ID are the largest stakeholders in this study. People with ID are reliant on intelligence tests for diagnosis and access to many support services (Yalon-Chamovitz, 2009). These support services may include, but are not limited to, access to individualized education, occupational supports, assisted housing, and subsidized income. Misdiagnoses or inflation of an individual's abilities could lead to denial of services for his or her lifetime. Since the WISC-IV is the most commonly used assessment measure to diagnose ID (Riccio et al., 1998), it is imperative that it be an accurate test of low intellectual ability.

People whose FSIQ is around the intellectual testing cutoff range for ID diagnosis (FSIQ between 60 to 80) may particularly benefit from findings of this study. While the DSM-5 urges assessors to integrate qualitative data about functioning into their assessments of ID, there is no guarantee that all assessors can comprehensively assess for ID. The abilities of people within the ID range vary significantly (Yalon-Chamovitz, 2009). For example, it is not uncommon for a person with mild MR/ID to maintain relationships, a low-paying job, or independent living. This person may easily be overlooked for an ID diagnosis based on his or her basic level of functioning. Additionally, a person with average verbal abilities may be overlooked for a diagnosis of ID due to a less apparent processing speed or working memory deficit. It is crucial that intelligence tests, which are intended to make assessment objective and uniform, measure intellect accurately. The author argues that intellectual testing must be accurate so that people with mild ID are not denied access to support, social services, or rights such as the American’s with Disabilities Act of 2008 (ADA) and Individuals with Disabilities Education Improvement Act (IDEIA; 2004),

Families and Caregivers. Families and caregivers of people with ID are also affected by the results of intelligence testing. Caregivers of children and adolescents with ID are often in need of in-home assistance; social workers to organize care; and financial support for medications, travel expenses, and supplemental care for other children in the family. In addition to access to these resources, improper diagnosis may limit caregiver's access to subsidized familial supports, such as ID education, parenting seminars, and support groups (Yalon-Chamovitz, 2009).

Schools. Schools are one of the largest providers of intellectual assessment for children and adolescents. They often diagnose intellectual delays and provide initial interventions to help children learn and build skills. Legislative acts, such as the ADA and the IDEIA, mandate schools to provide adequate education for students. The results of intelligence testing often inform the level of care or service that a school must provide because test scores are considered objective tests of ability. Adapted educational services are often financially taxing on schools because schools must pay for alternative programming. It is crucial for schools to use appropriate tests so that they may provide the most ethical and financially viable services.

Social Services Agencies. Similar to schools, many public social service agencies base access to service on diagnosis and assessment results. A person must have evidence that he or she is disabled in order to have access to costly and in-demand services. If the WISC-IV is not an appropriate assessment for ID, services may not be distributed appropriately. The ramifications for this include denial of needed services, higher expenses for agencies, and strains on limited resources.

Theoretical Framework

In the past 20 years, the field of assessment has shifted to a Cattell-Horn Carroll (CHC) model of assessment conceptualization (McGrew & Wendling, 2010). Instead of using indexes and full-scale IQ (FSIQ) scores, the CHC model measures over 70 narrow abilities grouped into broad abilities (Newton & McGrew, 2010). Many modern IQ tests were developed or adapted to this model, including the Woodcock-Johnson Tests of Cognitive Abilities and the revised versions of Wechsler Adult Intelligence Scales. Unfortunately, as McGrew and Wendling highlight, much of the praise received for the CHC model is based on the highly integrated Woodcock-Johnson and does not necessarily translate well to the Wechsler four factor scales.

Additionally, a new model for diagnosing ID has emerged out of the Individuals with Disabilities Education Act (IDEA, 1990) that focuses on a student's Response to Intervention (RTI; Newton & McGrew, 2010). This three-tiered intervention model identifies students with learning challenges and providing schools a structured format to adapt curriculum to student needs. Supporters of the model posit that it provides superior identification of learning challenges, as well as a cost-effective method of supporting students with alternative educational needs (Dombrowski, Kamphaus, & Reynolds, 2004). Since RTI uses its own techniques to identify and define learning disabilities, there is a debate about the usefulness of intelligence testing with the RTI model. Some theorists suggest that the use of intelligence testing and a CHC model can complement the RTI model when designing interventions (Restori, Gresham, & Cook, 2008). Thus, although there are dissenting opinions about intelligence testing in modern assessment theory, determining the accuracy of the intelligence scales can still be useful.

The study is further informed by Yalon-Chamovitz's (2009) description of the accessibility needs of people with ID. Accessibility is defined as the availability of resources to the greatest possible number of people. Yalon-Chamovitz suggests that people with ID need accommodations for successful daily living. Specifically, people with ID need simple language, lower literacy expectations, a slower pace to complete tasks, and an elimination of stigma. These accessibility needs are considered when assessing the WISC-IV's ability to accurately measure lower tiers of intellect.

Summary

The needs of people with ID have been made clear by Yalon-Chamovitz (2009) and supported by the ADA and IDEA (ADA, 2009; IDEA 2007). Even with changes in the DSM-5, intellectual assessment is a required component for individuals to be appropriately diagnosed and eligible for the described services. Additionally, as there is a change in models from a CHC model to a RTI model, it is crucial that the strengths and weaknesses of people with ID be understood fully in order for them to receive appropriate interventions.

Evidence against the appropriateness of the Wechsler four-factor model has been found in multiple studies in the United Kingdom that posit the four-factor model is inappropriate not only for individuals with ID (MacLean et. al., 2011; Whitaker, 2008; Whitaker, 2010) but potentially even for individuals with specific learning disabilities (Whitaker, 2005). Unfortunately, the evidence presented in these studies has not prevented (or may not be known to) American school systems, assessment agencies, and private practices from using Wechsler four-factor assessment tools to determine IQ and cognitive strengths and weaknesses of people with ID. Continued studies must be performed to see if the WAIS-IV has the same variance as the WAIS-III and to promote education on findings amongst assessors.

Research has consistently supported Whitaker’s (2005) concerns about assessing people with low IQ with the WISC-III and WAIS-III. MacLean et al. (2011) found that the four-factor model used in the WAIS-III was not a good fit for people with low IQs. Since norming practices, floor thresholds, and models used have not changed from the WAIS-III to the WAIS-IV, it is likely that similar problems will be seen when using the new edition.

Research has consistently shown that floor effects can be seen on multiple UK versions of the Wechsler intelligence tests, including the UK version of the WISC-IV (Whitaker & Gordon, 2012). The problems cited when assessing individuals with ID seem to be correlated more with Wechsler's process of scaling extremely low raw scores to 1 than with other variables. In order to see if these findings are unique to the UK version of the WISC-IV, Whitaker and Gordon's study (2012) was replicated using the US version of the WISC-IV.

Chapter 3: Method

I evaluated for floor effects in the WISC-IV when assessing children and adolescents with FSIQs below 70. Historical data were collected from de-identified schools in New Hampshire and Massachusetts. I used the adjusted scoring system proposed by Whitaker and Gordon (2012) to assess if the WISC-IV scoring system inflates the index and FSIQ scores of children and adolescents with intellectual disability (ID). Results of the study will be shared with the Social Security Determination Small Grants Program, in accordance with our grant agreement. This chapter describes the methodology for the study.

Method of Assessment

Whitaker and Gordon (2012) began creating their adjusted scoring system in England by using the data available in the scoring charts of *Wechsler Intelligence Scale for Children–Fourth Edition: Administration and Scoring Manual* (2003). Since the WISC-IV does not provide the equation they used to determine how raw scores would be converted to scaled scores, Whitaker and Gordon found an algorithm by plotting the mathematical relationship between the raw scores and scaled scores less than 10. Only scaled scores less than 10 were included because they felt the mathematical relationship would be simpler when only low scores were used. They observed that raw score to scaled score relationship did not continue as expected with very low raw scores, and that instead it stopped abruptly at the scaled score of one. They hypothesized that this represented the suspected floor effect. Figure 1 is a visual representation I created of one of the graphs using the Wechsler raw to scaled score tables.

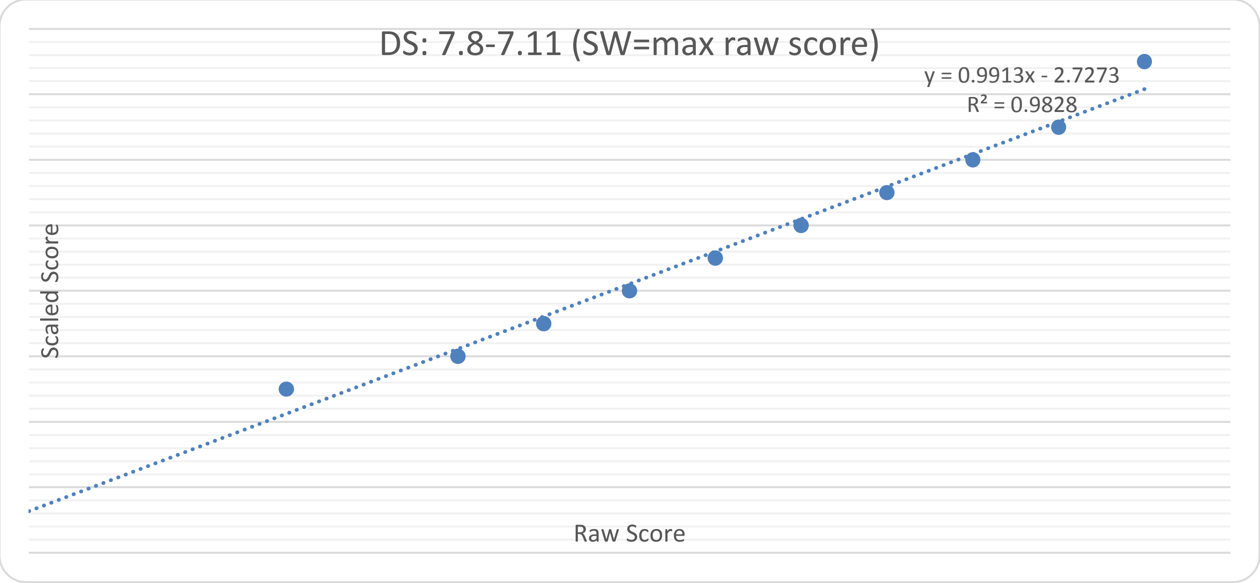


Figure 1. Predicted line based on Wechsler raw to scaled score chart for Digit Span (DS), Ages 7:8 to 7:11

Whitaker and Gordon (2012) created adjusted scaled scores by allowing the relationship between raw scores and scaled scores to continue below a scaled score of 1. For example, there is a linear mathematical relationship between raw scores and scaled scores on the subtest Digit Span for children aged 7 years, 8 months to 7 years, 11 months. Figure 1 shows that a straight line and a linear equation best fit the points provided by the WISC-IV conversion charts. I created Table 1 to show the 1:1 linear relationship between raw and scaled scores until it reaches low raw scores.

Table 1

Raw to Scaled Score Conversion Guide for Digit Span, ages 7:8 to 7:11

Raw Scores	Scaled Scores
0	1
1	1
2	1
3	1
4	2
5	2
6	3
7	4
8	5
9	6

Whitaker and Wood (2012) continued the mathematical relationship found via the scores in the WISC-IV *Manual* (2003) to the very low raw scores. Where the WISC-IV does not allow scaled scores to go below 1, their adjusted scores did not have a lower limit. Table 2 shows how the scores were adjusted for Digit Span, ages 7 years, 8 months to 7 years, 11 months.

Table 2

Whitaker and Gordon's (2012) Raw to Scaled Score Conversion Guide for Digit Span, ages 7:8 to 7:11

Raw Scores	Scaled Scores
0	-1
1	0
2	0
3	1
4	2
5	2
6	3
7	4
8	5
9	6

Whitaker and Wood (2012) found that index and FSIQ scores were significantly lower when using the adjusted scores. Since the UK version of the WISC-IV is different than the U.S. version, I investigated whether similar results could be seen in the U.S. version of the WISC-IV. To do so, I replicated the aforementioned study and created another method of adjusting scores (see Data Analyses section).

Setting. The data were collected from assorted schools located in New Hampshire and Massachusetts. One set of data was collected from a southwestern New Hampshire school district that serves students from over 200 square miles, 14 school buildings, and approximately 4,200 students. The other set of data was collected from a suburban Massachusetts school for students with emotional, behavioral, and developmental disorders.

Participants. The sample consisted of 7 students who scored a FSIQ below 70 on the WISC-IV. Students were identified by their schools as needing intelligence testing for numerous reasons including, but not limited to, admission, overall poor academic achievement, concerns about particular areas of learning, or unexplained conduct problems. Participants ranged from the ages of 7 years, 8 months to 15 years, 10 months at the time of testing. I did not know gender, race, and socioeconomic status. Table 3 lists the ages and locations of each participant.

Table 3

Ages and Location of Each Child

ID	Age	Location
Child 1	15 years, 10 months	New Hampshire
Child 2	11 years, 8 months	New Hampshire
Child 3	7 years, 8 months	Massachusetts
Child 4	9 years, 4 months	Massachusetts
Child 5	12 years, 7 months	Massachusetts
Child 6	8 years, 8 months	Massachusetts
Child 7	8 years, 9 months	Massachusetts

Measures. Students were administered the Wechsler's Intelligence Scale for Children, fourth edition (WISC-IV; Wechsler, 2003). Participants had been administered the 10 core WISC-IV subtests as specified in the administration manual (Wechsler, 2003). Original scoring was done according Wechsler's protocol. Table 4 lists each of the subtests under its respective index.

Table 4

WISC-IV Indexes and Core Subtests

Verbal Comprehension Index
Vocabulary
Similarities
Comprehension
Perceptual Reasoning Index
Block Design
Picture Concepts
Matrix Reasoning
Working Memory Index
Digit Span
Letter-Number Sequencing
Processing Speed Index
Coding
Symbol Search

Procedures. Per APA's ethical codes (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), each school gave permission to access archived data of children and adolescents without parental consent because the data were owned by the schools for administrative reasons (See Appendix A for IRB approval). By giving permission to access their test data, the schools received the findings of the study.

In order to protect the identities of the students whose tests were used in the study, all WISC-IV protocols were de-identified prior to coming into my possession. Physical protocols were linked with electronic data using randomly assigned numbers. Raw, scaled, index, and FSIQ scores were recorded from the WISC-IV protocols into a password-protected Microsoft Excel file. The hard copies of de-identified assessment protocols were stored in a locked filing cabinet at my residence.

Research Hypotheses

The research hypotheses for the study were as follows:

1. Hidden floor effects observed in the UK versions of Wechsler intelligence tests will be seen in the current study's New England sample.

2. Index and FSIQ scores will be significantly lower using adjusted scores than Wechsler scores.

3. A significant number of students who did not meet diagnostic testing criteria for an ID diagnosis will now meet criteria.

Data Analyses

The methods of analysis were informed by the research of Whitaker and Gordon (2012), and were performed with the support of Wright State University's Statistical Consulting Center. Two sets of adjusted scores were created for subtests receiving a scaled score of 1. The first set was made using the exact method of Whitaker and Gordon (2012). The second set (referred heretofore as Lanza's adjusted scores) was informed by Whitaker and Gordon, but was modified.

Whitaker and Gordon's (2012) adjusted scores were created by finding the mathematical relationship between the WISC-IV (U.S. version) raw scores to scaled scores for each subtest receiving a scaled score of 1. It is important to note that many subtests have a range of raw scores that convert to the same scaled score. In this instance, Whitaker and Gordon took the highest score of the raw score range to graph (personal communication, August 20, 2013). For example, if raw scores 0 to 3 received a scaled score of 1, Whitaker and Gordon (2012) plotted the relationship between a raw score of 3 and a scaled score of 1.

I modified their method and used the mean of the raw scores in the range when calculating the mathematical relationship between raw scores and scaled scores. I considered that the highest raw score might not best represent the scaled score. Instead, I used the average raw score to correlate with each scaled score in order to represent the mean raw score value of each scaled score. For example, if raw scores 0 to 3 received a scaled score of 1, I plotted the relationship between a raw score of 2 (the integer closest to the mean of raw scores 0, 1, 2, and 3) and a scaled score of 1.

The raw and scaled scores were graphed using Microsoft Excel for both Whitaker and Gordon, and Lanza's adjusted scores. Using regression analysis, Microsoft Excel identified the slope of the line that best fit the plotted points. This provided an algorithm for where the line would continue should it be allowed to continue past a scaled score of 1. Index and FSIQ scores were then calculated using the new scaled scores according to Wechsler scoring system and norms tables. Table 5 lists the formulae found for each subtest requiring an adjusted score.

Table 5

Formulae for Creating Adjusted Scaled Scores

Subtest	Age Group	Algorithm for Whitaker and Gordon's Adjusted Scores	Algorithm for Lanza's Adjusted Scores
Vocabulary	11:8-11:11	$y = 0.3636x - 3.5542$ $R^2 = 0.9983$	$y = 0.3701x - 3.4377$ $R^2 = 0.997$
Comprehension	9:4-9:7	$y = 0.5687x - 1.7222$ $R^2 = 0.996$	$y = 0.5824x - 1.6928$ $R^2 = 0.9972$
Comprehension	11:8-11:11	$y = 0.5539x - 3.5285$ $R^2 = 0.997$	$y = 0.5558x - 3.3653$ $R^2 = 0.9988$
Matrix Reasoning	11:8-11:11	$y = 0.0021x^2 + 0.4762x - 1.969$ $R^2 = 0.9981$	$y = 0.0015x^2 + 0.5012x - 1.9635$ $R^2 = 0.9988$
Digit Span	7:8- 7:11	$y = 0.9913x - 2.7273$ $R^2 = 0.9828$	$y = 0.9727x - 2.5217$ $R^2 = 0.9832$
Digit Span	11:8-11:11	$y = x - 6$ $R^2 = 1$	$y = x - 6$ $R^2 = 1$
Letter-Number Sequencing	11:8-11:11	$y = 0.0321x^2 + 0.0688x - 0.346$ $R^2 = 0.9967$	$y = 0.027x^2 + 0.1737x - 0.7363$ $R^2 = 0.9966$
Coding	8:8- 8:11	$y = 0.2941x - 1.4103$ $R^2 = 0.9962$	$y = 0.2945x - 1.0813$ $R^2 = 0.9985$
Coding	11:8-11:11	$y = 0.2687x - 4.3091$ $R^2 = 0.9984$	$y = 0.0031x^2 + 0.0345x + 0.1833$ $R^2 = 0.9948$

Subtest	Age Group	Algorithm for Whitaker and Gordon’s Adjusted Scores	Algorithm for Lanza’s Adjusted Scores
Coding	15:8-15:10	$y = 0.2248x - 6.0339$ $R^2 = 0.9988$	$y = 0.2315x - 6.0166$ $R^2 = 0.9968$
Symbol Search	11:8-11:11	$y = -0.004x^2 + 0.6109x - 3.342$ $R^2 = 0.9966$	$y = -0.0068x^2 + 0.7129x - 3.96$ $R^2 = 0.9966$
Symbol Search	12:4-12:7	$y = -0.0029x^2 + 0.577x - 3.9768$ $R^2 = 0.999$	$y = -0.0044x^2 + 0.6383x - 4.3093$ $R^2 = 0.9995$
Symbol Search	15:8-15:10	$y = -0.0044x^2 + 0.6334x - 6.9781$ $R^2 = 0.9986$	$y = -0.0059x^2 + 0.716x - 7.7435$ $R^2 = 0.9993$

Adjusted scaled scores were created for subtests on which a student obtained a scaled score of 0. Adjusted scaled scores for other subtests were not calculated since they would not be used in the study sample. Although the Lanza adjusted score equations were all different from the Whitaker and Gordon adjusted score equations, it is important to note that this did not necessarily create differences in the two adjusted scoring systems. This is discussed further in the Results chapter.

Adjusted scores were then added together to create the Sum of Scaled Scores for each index and for the FSIQ. Occasionally, new index scores could not be calculated using the Wechsler scoring charts because they were too low. When this happened, new index scores were created using a very similar process as the creation of adjusted scores. The mathematical relationship between the sums of the scaled scores was plotted with the index scores to produce an equation. The equation was then used to predict what the index score would be if lower scaled scores had been available in the original Wechsler scoring.

Once new FSIQ and index scores were created for each sample, the Wechsler scores and adjusted scores were compared using two-tailed paired-sample *t*-tests. In order to determine differences between the Wechsler and both the Lanza adjusted scores and Whitaker and Gordon (2012) adjusted scores, the data were analyzed using Microsoft 10 Excel and Statistical Package for the Social Sciences, version 20 (SPSS 20). Both Whitaker and Gordon's and Lanza’s adjusted scores were run as “post” scores to see if both score adjustment procedures created significant difference in the FSIQ and index scores.

Summary

The study evaluated for a hidden floor effect in the U.S. version of the WISC-IV. Evidence supports that the Wechsler four-factor model poses problems to assessing the IQ of children and adolescents with ID, although no study exists specifically that evaluates the U.S. version of the WISC-IV with regard to suggested problems. Findings from the study can be used to promote more appropriate assessment in organizations that diagnose and allocate services to children and adolescents with ID. As a result, more appropriate services can be provided to children, their families, and schools.

Chapter 4: Results

Analyses for the study were performed on 7, WISC-IV protocols of students. Adjusted scores were created using the same process as Whitaker and Gordon (2012). Additional consultation was sought from direct communication with Simon Whitaker to better understand the methodology of Whitaker and Gordon. Analyses for the study were done with the support of Wright State University's Statistical Consulting Center.

Tests of Major Hypotheses

Rescoring Protocols Using Adjusted Scores

Each protocol was rescored using Whitaker and Gordon Adjusted Scores and Lanza Adjusted Scores. Out of the seven total protocols used in this study, a total of four had changes to their index scores using an adjusted scoring system. Two had had changes in index scores using Lanza’s adjusted scores, and four had changes using Whitaker and Gordon’s (2012) adjusted scores. One of the protocols had no difference in scores with either method. A total of three protocols had changes to their FSIQ using an alternative scoring method. The Wechsler, Whitaker and Gordon adjusted, and Lanza adjusted scores are listed in the tables 6 through 12 below.

Table 6

Child 1 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores

Subtest/Index	Raw Scores	Wechsler Scores	Whitaker & Gordon Adjusted Scores	Lanza Adjusted Scores
VCI		81	81	81
Vocabulary	28	4	4	4
Similarities	24	8	8	8
Comprehension	27	8	8	8
PRI		75	75	75
Block Design	30	6	6	6
Picture Concepts	18	8	8	8
Matrix Reasoning	15	4	4	4
WMI		68	68	68
Digit Span	11	3	3	3
Letter-Number Sequencing	15	6	6	6
PSI		50	45*	45*
Coding	23	1	-1*	-1*
Symbol Search	14	1	-1*	-1*
FSIQ		63	60*	60*

Note. * Highlights a difference in score

Table 7

Child 2 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores

Subtest/Index	Raw Scores	Wechsler Scores	Whitaker & Gordon Adjusted Scores	Lanza Adjusted Scores
VCI		53	47 *	53
Vocabulary	11	1	0*	1
Similarities	7	4	4	4
Comprehension	7	1	0*	1
PRI		57	57	57
Block Design	18	5	5	5
Picture Concepts	10	3	3	3
Matrix Reasoning	6	1	1	1
WMI		50	42*	40*
Digit Span	4	1	-2*	-2*
Letter-Number Sequencing	0	1	0*	-1*
PSI		50	42*	40*
Coding	16	1	0*	2*
Symbol Search	6	1	0*	0*
FSIQ		43	40*	40*

Note. * Highlights a difference in score

Table 8

Child 3 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores

Subtest/Index	Raw Scores	Wechsler Scores	Whitaker & Gordon Adjusted Scores	Lanza Adjusted Scores
VCI		75	75	75
Vocabulary	13	5	5	5
Similarities	10	8	8	8
Comprehension	7	4	4	4
PRI		67	67	67
Block Design	3	3	3	3
Picture Concepts	8	6	6	6
Matrix Reasoning	7	5	5	5
WMI		65	65	65
Digit Span	3	1	1	1
Letter-Number Sequencing	8	7	7	7
PSI		75	75	75
Coding	30	5	5	5
Symbol Search	18	6	6	6
FSIQ		64	64	64

Note. * Highlights a difference in score

Table 9

Child 4 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores

Subtest/Index	Raw Scores	Wechsler Scores	Whitaker & Gordon Adjusted Scores	Lanza Adjusted Scores
VCI		75	75	75
Vocabulary	26	8	8	8
Similarities	15	8	8	8
Comprehension	5	1	1	1
PRI		69	69	69
Block Design	10	5	5	5
Picture Concepts	12	6	6	6
Matrix Reasoning	8	4	4	4
WMI		77	77	77
Digit Span	11	7	7	7
Letter-Number Sequencing	9	5	5	5
PSI		65	65	65
Coding	20	4	4	4
Symbol Search	7	3	3	3
FSIQ		65	65	65

Note. * Highlights a difference in score

Table 10

Child 5 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores

Subtest/Index	Raw Scores	Wechsler Scores	Whitaker & Gordon Adjusted Scores	Lanza Adjusted Scores
VCI		83	83	83
Vocabulary	31	7	7	7
Similarities	19	8	8	8
Comprehension	18	6	6	6
PRI		77	77	77
Block Design	22	6	6	6
Picture Concepts	15	7	7	7
Matrix Reasoning	17	6	6	6
WMI		86	86	86
Digit Span	16	9	9	9
Letter-Number Sequencing	13	6	6	6
PSI		53	50*	53
Coding	24	2	2	2
Symbol Search	8	1	0*	1
FSIQ		67	64*	67

Note. * Highlights a difference in score

Table 11

Child 6 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores

Subtest/Index	Raw Scores	Wechsler Scores	Whitaker & Gordon Adjusted Scores	Lanza Adjusted Scores
VCI		71	71	71
Vocabulary	13	4	4	4
Similarities	8	6	6	6
Comprehension	22	9	9	9
PRI		88	88	88
Block Design	22	9	9	9
Picture Concepts	13	8	8	8
Matrix Reasoning	12	7	7	7
WMI		77	77	77
Digit Span	11	7	7	7
Letter-Number Sequencing	8	5	5	5
PSI		62	59*	59*
Coding	4	1	0*	0*
Symbol Search	9	5	5	5
FSIQ		70	69*	69*

Note. * Highlights a difference in score

Table 12

Child 7 Wechsler, Whitaker and Gordon Adjusted, and Lanza Adjusted Subtest Scores

Subtest/Index	Raw Scores	Wechsler Scores	Whitaker & Gordon Adjusted Scores	Lanza Adjusted Scores
VCI		67	67	67
Vocabulary	15	5	5	5
Similarities	9	6	6	6
Comprehension	5	2	2	2
PRI		79	79	79
Block Design	14	7	7	7
Picture Concepts	10	6	6	6
Matrix Reasoning	13	7	7	7
WMI		68	68	68
Digit Span	9	5	5	5
Letter-Number Sequencing	6	4	4	4
PSI		68	68	68
Coding	7	1	1	1
Symbol Search	7	7	7	7
FSIQ		64	64	64

Note. * Highlights a difference in score

Testing Differences in Index and FSIQ Scores

Ten paired *t*-tests were used to determine if the Whitaker and Gordon (2012) or Lanza adjusted index and FSIQ scores were significant differently from the Wechsler index and FSIQ scores. There were no significant differences noted between indexes or FSIQ between the Wechsler scores or the Whitaker and Gordon (2012) Adjusted or Lanza Adjusted. Table 13 shows the results of the *t*-tests.

Table 13

Means and Standard Deviations of Index and FSIQ Scores

Index	Wechsler	Whitaker & Gordon (2012) Adjusted	Lanza Adjusted
VCI *Mean (SD)*	72.1 (9.32)	71.3 (14.53)	72.1 (9.32)
PRI *Mean (SD)*	73.1 (9.91)	73.1 (9.91)	73.1 (9.91)
WMI *Mean (SD)*	70.1 (11.45)	69 (13.93)	68.7 (14.58)
PSI *Mean (SD)*	60.4 (9.71)	58.1 (11.84)	59.3 (10.69)
FSIQ *Mean (SD)*	62.3 (8.83)	60.9 (9.56)	61.3 (9.79)

Based on the paired-samples *t*-tests, neither Whitaker and Gordon's (2012) method of rescaling nor the Lanza method of scoring created significant changes to the sample's FSIQ. Furthermore, it should be noted that the adjusted FSIQ scores that were affected by the alternative scoring system still fell within the Wechsler predicted FSIQ range with 95% confidence. Only the protocol of Child 3 had Whitaker and Gordon (2012) and Lanza adjusted FSIQ scores outside of the FSIQ range expected with 90% confidence. Table 14 shows the FSIQ ranges for each protocol to highlight how the adjusted scores do, or do not, fall within the expected ranges.

Table 14

Wechsler FSIQ Ranges at 90% and 95% Confidence Intervals and Adjusted FSIQ Scores

Protocol	Wechsler FSIQ	FSIQ Range (90% CI)	FSIQ Range (95% CI)	Whitaker & Gordon (2012) FSIQ	Lanza FSIQ
Child 1	63	60–68	59–69	60	60
Child 2	43	41–49	40–50	40	40
Child 3	64	61–69	60–70	64	64
Child 4	65	62–70	61–71	65	65
Child 5	67	61–72	63–73	64	67
Child 6	70	67–75	66–76	69	69
Child 7	64	61–69	60–70	64	64

Summary

Both Whitaker and Gordon’s (2012) and Lanza’s adjusted scores affected the subtest scores of the protocols used. Five paired-sample *t*-tests were conducted to determine differences between the Wechsler scores and Lanza adjusted scores, and five paired-sample *t*-tests were conducted to determine differences between Wechsler scores and Whitaker and Gordon scores. There was no significant difference found in the index scores or FSIQs of either adjusted scoring system.

Chapter 5: Discussion

Results of the study suggest no floor effects were present in the sample of US WISC-IV protocols. This conclusion is based on the lack of finding that FSIQ and index scores changed when very low raw scores were weighted lower with two adjusted scoring systems. Although Index and FSIQ were lowered using both Whitaker and Gordon's (2012) and Lanza's adjusted scoring system, this difference was minimal and is better understood as a result of their adjusted scoring method.

Results of the study must be interpreted cautiously for many reasons. The first reason is that the study had limited data. Not only was the sample less than half of Whitaker and Gordon's (2012) sample, the data sources were limited to two school locations. A small sample may inflate or miss any findings found in a larger sample. Second, any claims that the WISC-IV is not a valid measure of cognitive abilities for low intellect may negatively impact individuals with ID. I speculate that individuals diagnosed with ID may need to be reassessed with other measures, which can be costly for schools and families, as well as emotionally taxing for the individual. Furthermore, individuals previously diagnosed with ID may be required to reapply for social services that based eligibility on cognitive assessment. Therefore, it is important for readers of this study to be familiar with its limitations.

Limitations of the Research

The first limitation to the study is that Whitaker and Gordon's hypotheses are challenging to investigate in a research study. Many of Whitaker's arguments would require hundreds of protocols to have adequate sample sizes. For example, an international study would need to be conducted to see if his hypothesis that intellect is bimodal could be supported. Another example would be collecting a new norming sample for the WISC-IV to test the hypothesis that the existing sample is inadequate. Since these studies do not exist at this time, smaller studies such as this can only allude to support the hypothesis that there is a floor effect, which this study, however, could not.

Another potential limitation of the study is its sampling method. Protocols for the study were provided by individuals from two New England schools. The sample of students represented an extremely narrow part of the U.S. population. Based on demographics of the surrounding areas of the schools, students were likely to be Caucasian and from middle to lower socioeconomic status. If this study were to be expanded or replicated, it would be useful to gather a larger number of protocols from various sources, socioeconomic status, race, and regions within the United States. All results of the study would be found only in students with similar demographics. The study’s results should not be assumed to be present in other populations.

Another sampling challenge is that it is common for students with severe ID to be administered only parts of the WISC-IV. This prevents the students from being overworked or challenged with unreasonable tasks. This study's sample is not likely to be representative of an ID population; however, it may still inform the utility of the WISC-IV to measure low IQ.

Finally, the sample collected for the study did not control for variables that have been controlled for in the WISC-IV norming sample. Specifically, the study did not determine if the child or adolescent was on pharmacological medication at the time of testing. I argue that while some medication might influence WISC-IV performance, such as creating sedating effects or increasing a child's ability to focus, children in this sample were more representative of the population of children and adolescents with ID.

Future Directions for Research

Research on intellectual disabilities would benefit from an expanded version of the present study to see if and to what extent other floor effects are observed using both Whitaker and Gordon's (2012) and Lanza's adjusted scoring systems. Both methods reduced Index and FSIQ scores, and I posit that her original research hypothesis of floor effects observed in index scores of people with low IQ may be found in a larger sample. However, due to the significant results observed in the study by Whitaker and Gordon, it is important to understand if their findings are unique to the UK version of the test or because of their larger sample. A new, larger sample would need to be matched with a WISC-IV sample of ID scores before paired sample t-tests could be done on index and FSIQ scores. It would be helpful if the publishers of the WISC-IV would release their normative data.

Furthermore, I hope that efforts in replicating this study may be applied to the Wechsler Intelligence Scale for Children, Fifth Edition, which is planned to be released in Fall 2014 (Pearson Education, Inc., 2014, June 12). I anticipate that a larger sample may be collected if data collections sites participate from the release date of the new measure. I also believes that it will be important to measure for possible floor effects in the newest version due to the basic scoring problems explained earlier.

I hypothesize that there may other ways of addressing hidden floor effects other than an adjusted scoring system. One method I believe is viable is modifying the validity criteria for indexes. In the current WISC-IV *Manual*, a FSIQ can only be calculated if a client receives raw scores of 0 on three Verbal Comprehension and three Perceptual Reasoning subtests. The Working Memory and Processing Speed Index scores can only be calculated if one or more subtests has a raw score greater than 0. I posit that there could be increased research on this area to see if all raw scores of 0 could lead to invalid subtest and index scores. Additionally, it may be helpful if other very low raw scores, especially for adolescent age norms, may invalidate subtests.

Finally, there may be other methods of improving the WISC-IV's validity in assessing children with ID. The WISC-IV scoring system measures for significant differences between subtests that may invalidate the index score. Similar to when a FSIQ is invalid due to highly varied index scores, sometimes subtests scores differ so greatly that they are not best represented by one index score. The WISC-IV has tables available to calculate critical values for these differences. I propose that it may be helpful to expand research on these values to see if children with low IQ may benefit from a different set of critical values. I believe that this may help expand understanding of intra-individual differences that are hallmark of children with ID.

To evaluate the reliability and rarity of the difference between an individual's subtest scores, measurement error and reliability of the scores need to be taken into account. Payne and Jones (1957; Florio & Ley, 2006) suggested evaluating such differences by standardizing each subtest, using Z-scores, then adjusting by an appropriate standard error. To evaluate whether the observed difference is reliable, that is, the difference is not due to measurement error or chance, the following difference measure is used:

$$Z_{diff} = \frac{Z_x - Z_y}{\sqrt{2-\left(r_{xx} - r_{yy}\right)}}, \qquad (1)$$

where Z_x and Z_y are the individual's Z-scores for the two subtests and r_{xx} and r_{yy} are the reliabilities (Cronbach's alpha) of the two subtests (Florio & Ley, 2006; Payne & Jones, 1957). The two-tailed probability associated with this difference score is the chance that the observed difference is due to measurement error. Participants will be identified whose difference in subtest scores is not due to measurement or random error.

It is important to note that a reliable difference does not indicate the difference is rare. Another formula suggested by Payne and Jones (1957) assesses the abnormality or rarity of the observed difference using the correlation between the two subtests:

$$Z_{diff} = \frac{Z_x - Z_y}{\sqrt{\left(2 - 2r_{xy}\right)}}, \qquad (2)$$

where r_{xy} is the correlation between the two subtests. The two-tailed probability associated with the difference score indicates the percent of the population who would have a difference this large. Participants whose difference score on two subtests is at or above the critical difference score show significant difference in subtest scores at the individual level. These participants will need a special intervention that will address particular cognitive deficits.

Reflections on Conducting the Study

I am grateful to have completed a pilot study as the preliminary research on an alternative scoring system on the U.S. version of the WISC-IV. I believe that the study better outlined procedures on how to use Whitaker and Gordon’s (2012) adjusted scoring system, as well as explored the benefits of Lanza’s adjusted score method. Should the study be replicated on matched samples, it is anticipated that the scoring procedures outlined in the current pilot study would streamline the replication process.

I am also grateful for the individuals and schools that supported this study. I am particularly grateful to those who spent many hours educating meon how to obtain data within educational settings, as well as sharing contacts with other supporters with whom I could network. I would like to thank those who volunteered their time to locate and de-identify protocols used for this research.

Finally, I would like to express her gratitude to those who create and work on intelligence test measures. This study focused on one concern in a largely efficient and effective test. Although I believe the concern is very important to investigate, I also understand that the Wechsler system has offered meaningful information to individuals with ID and their caregivers for decades.

I believe that it is important to question all assessment tools, especially the most revered. Although I believe that clinicians and researchers are well-intentioned, I recognize that they are also representative of the *zeitgeist*. Just as political history is blemished with discrimination and oppression, early psychology often pathologized marginalized groups (Canady, 1943; Guthrie, 1998; Herrnstein & Murray, 1994). There clearly has been effort to improve our tests and interventions (Rosenthal & Jacobson, 1968) to meet the needs of underserved groups; however, there is work to be done.

Assessments are valued for their neutrality and norms, but often they are neither socially or culturally neutral nor normed adequately in diverse populations. Instead, group of various races, ethnicities, and socioeconomic statuses are viewed as “specialty populations,” among whom only samples of convenience are tested. An intelligence measure must have special norms for people with low IQ, including adequate representation and sample size.

I encourage clinicians to utilize their clinical expertise in assessment and emphasize it in their reports. I hope that clinicians will not only include their behavioral observations, but use them regularly to challenge assessment measures. I believe that assessment report-writing is more likely to be accepted when clinicians readily include observations and clinical opinions that support, critique, as well as oppose particular measures. I believe that clinicians trained to embrace diversity do provide culturally-competent interventions and are active in social advocacy that will quickly outpace tests that are inadequate for certain groups. I also believe that inclusion of clinical expertise and qualitative information in reports may lead to a rich data-pool from which improved measures may develop.

It is my hope that this pilot study may act as a stepping stone providing direction to assessors to critically think about tests and their interpretation. In particular, I hope that assessors will be mindful in interpreting scaled scores of 1 on the WISC-IV, and that they will be critical-minded about very low raw scores being scaled to 1. I hopes that assessors might regularly use corroborating measures when there are numerous scaled scores of 1 on a WISC-IV protocol, and that they might consider including a discussion paragraph in their reports about the possibility of floor effects on the test. Finally, I hope that assessors embrace the spirit of the DSM-5 Intellectual Disability diagnosis, and focus their assessments on describing how an individual is likely to function given their low intellectual abilities. I believe that this will allow children and adolescents with ID to receive access to services they require.

Finally, I am hopeful that the study may motivate readers to critically think about the utility and limitations of even the best and most widely used measures. The Wechsler intelligence tests have ample research citing how they may be used as helpful interventions with children and adolescents. I also feel that it is important for all consumers of test data to be aware of how test design, norming samples, and hidden floor or ceiling effects may bias results. I believe that with continued research on the limits of commonly used measures, testing and assessment practice will become stronger and serve better children and adolescents with intellectual disability.

References

ADA Amendments Act of 2008, Pub. L. 110–325, 122 Stat. §3553 (2008).

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

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (Revised 4th ed.)*. Washington, DC: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5 (5th ed.)*. Washington, DC: Author.

Americans with Disabilities Act of 1990, Pub. L. No. 101-336 (1990).

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Canady, H. G. (1943). The problem of equating the environment of Negro-White groups for intelligence testing in comparative studies. *Journal of Social Psychology, 17,* 3-15.

Dombrowski, S. C., Kamphaus, R. W., & Reynolds, C. R. (2004). After the Demise of the Discrepancy: Proposed Learning Disabilities Diagnostic Criteria. *Professional Psychology: Research and Practice, 35*(4), 364-372. doi:10.1037/0735-7028.35.4.364

Flanagan, D. P., & Kaufman, A. S. (2009). *Essentials of WISC-IV assessment (2nd ed.).* Hoboken, NJ: John Wiley & Sons.

Florio, T., & Ley, P. (2006). Assessment of differences. Retrieved from psychassessment.com.au

Guthrie, R. V. (1998). *Even the rat was white: A historical view of psychology.* Boston: Allyn & Bacon.

Herrnstein, R.J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in*

American life. New York, NY: Free Press.

Individuals with Disabilities Education Act (IDEA) Amendments of 1997, Pub. L. No. 105-17.

Individuals with Disabilities Education Improvement Act (IDEIA) Pub. L. 108-446 (2004).

Kirsh. I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the findings of the National Adult Literacy Survey.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

MacLean, H., McKenzie, K., Kidd, G., Murray, A. L., & Schwannauer, M. (2011). Measurement invariance in the assessment of people with an intellectual disability. *Research in Developmental Disabilities*, *32*(3), 1081-1085. doi:10.1016/j.ridd.2011.01.022

McGrew, K. S., & Wendling, B. J. (2010). Cattell-Horn-Carroll Cognitive-Achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools, 47*(7), 651-675.

Newton, J. H., & McGrew, K. S. (2010). Introduction to the special issue: Current research in Cattell–Horn–Carroll–based assessment. *Psychology in the Schools, 47*(7), 621-634.

Payne, R. W., & Jones, H. G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology, 18*, 115-121.

Pearson Education, Inc. (2014, June 12). http://www.pearsonclinical.com/education/products/100000771/wechsler-intelligence-scale-for-childrensupsupfifth-edition--wisc-v.html

Restori, A. F., Gresham, F. M., & Cook, C. R. (2008). "Old Habits Die Hard:" Past and current issues pertaining to response-to-intervention. *California School Psychologist*, 1367-78.

Riccio, C. A., Houston, F., & Harrison, P. L. (1998). Assessment practices for children with severe mental retardation. *Journal Of Psychoeducational Assessment*, *16*(4), 292-301.

doi:10.1177/073428299801600401

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom.* New York: Holt, Rinehart, & Winston.

Social Security Administration. (2014, April 3). http://www.ssa.gov/disability/professionals/bluebook

Social Security Administration. (2014, April 2). http://www.ssa.gov/disability/professionals/bluebook/112.00-MentalDisorders-Childhood.htm#112_05

Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore, MD US: Williams & Wilkins Co. doi:10.1037/10020-000

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children – Fourth Edition: Administration and Scoring Manual*. San Antonio,   TX: Pearson.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale-Third Edition*. London: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children – Fourth Edition: Administration and Scoring Manual*. San Antonio,   TX: Pearson.

Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale–Fourth Edition*. San Antonio, TX: Pearson.

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

Whitaker, S. (2008). WISC-IV and low IQ: Review and comparison with the WAIS-III. *Educational Psychology in Practice, 24*(2), 129-137. doi:10.1080/02667360802019180

Whitaker, S. (2010). Error in the estimation of intellectual ability in the low range using the

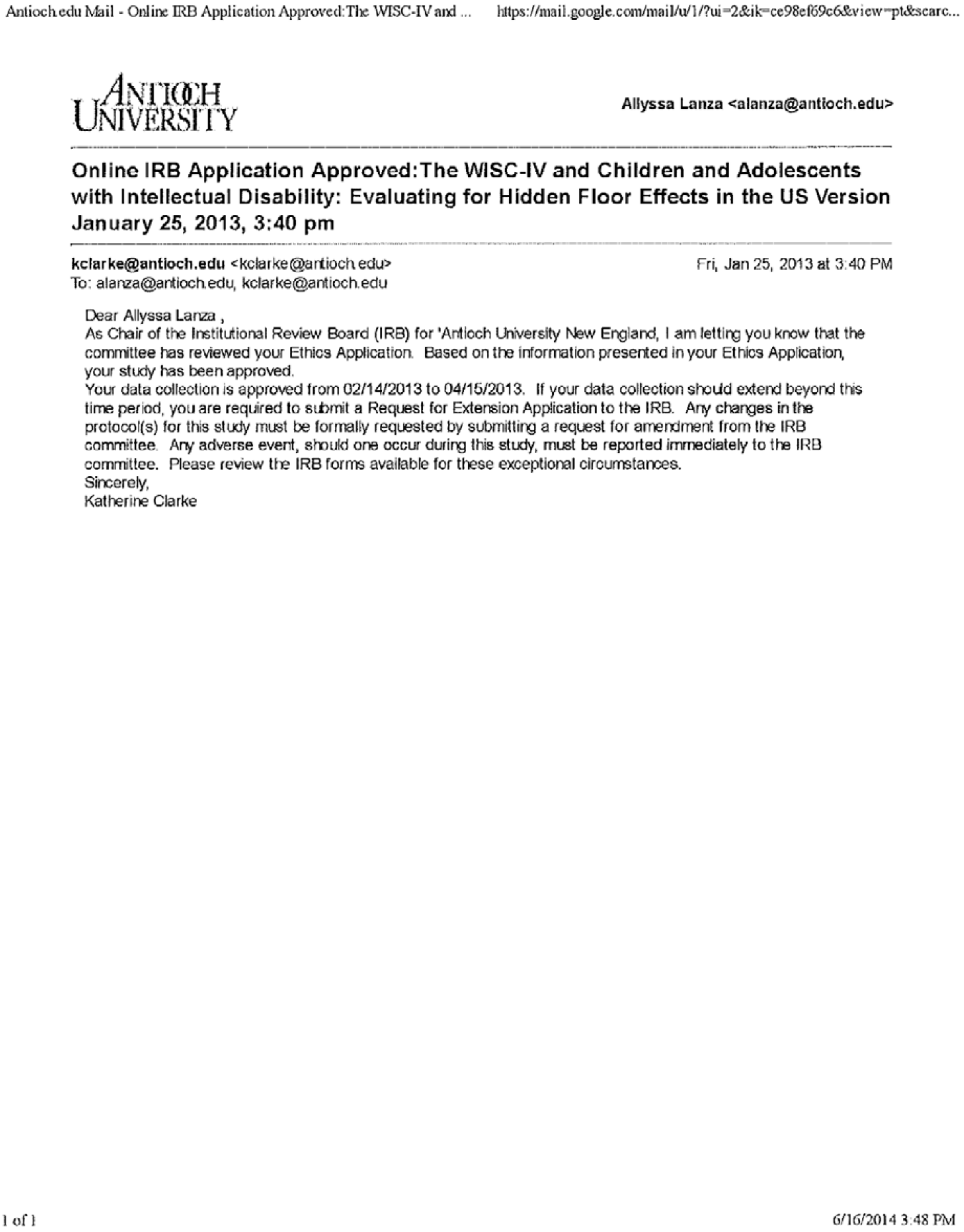
WISC-IV and WAIS-III. *Personality & Individual Differences, 48*(5), 517-521. doi:10.1016/j.paid.2009.11.017

Whitaker, S., & Gordon, S. (2012). Floor effects on the WISC-IV. *International Journal of Developmental Disabilities*, *58*(1), 1-9. doi: 10.1179/2047387711Y.0000000012

Whitaker, S., & Wood, C. (2008). The distribution of scaled scores and possible floor effects on the WISC-III and WAIS-III. *Journal of Applied Research in Intellectual Disabilities, 21*(2), 136-141. doi:10.1111/j.1468-3148.2007.00378.x

Yalon-Chamovitz, S. (2009). Invisible access needs of people with intellectual disabilities: A conceptual model of practice. *Intellectual & Developmental Disabilities, 47*(5), 395-400

Appendix A: IRB Approval



Antioch.edu Mail - Online IRB Application Approved:The WISC-IV and ... https://mail.google.com/mail/u/1/?ui=2&ik=ce98ef69c6&view=pt&searc...

ANTIOCH UNIVERSITY

Allyssa Lanza <alanza@antioch.edu>

Online IRB Application Approved:The WISC-IV and Children and Adolescents with Intellectual Disability: Evaluating for Hidden Floor Effects in the US Version January 25, 2013, 3:40 pm

kclarke@antioch.edu <kclarke@antioch.edu> Fri, Jan 25, 2013 at 3:40 PM
To: alanza@antioch.edu, kclarke@antioch.edu

Dear Allyssa Lanza ,
As Chair of the Institutional Review Board (IRB) for 'Antioch University New England, I am letting you know that the committee has reviewed your Ethics Application. Based on the information presented in your Ethics Application, your study has been approved.
Your data collection is approved from 02/14/2013 to 04/15/2013. If your data collection should extend beyond this time period, you are required to submit a Request for Extension Application to the IRB. Any changes in the protocol(s) for this study must be formally requested by submitting a request for amendment from the IRB committee. Any adverse event, should one occur during this study, must be reported immediately to the IRB committee. Please review the IRB forms available for these exceptional circumstances.
Sincerely,
Katherine Clarke

1 of 1 6/16/2014 3:48 PM

Appendix B: IRB Extension Request

ANTIOCH UNIVERSITY IRB

Investigator's name: Allyssa Lanza

Project Title: The WISC-IV and Children and Adolescents with Intellectual Disability: Evaluating for Hidden Floor Effects in the US Version

The proposed revisions listed below are submitted for approval for the above referenced, approved research project.

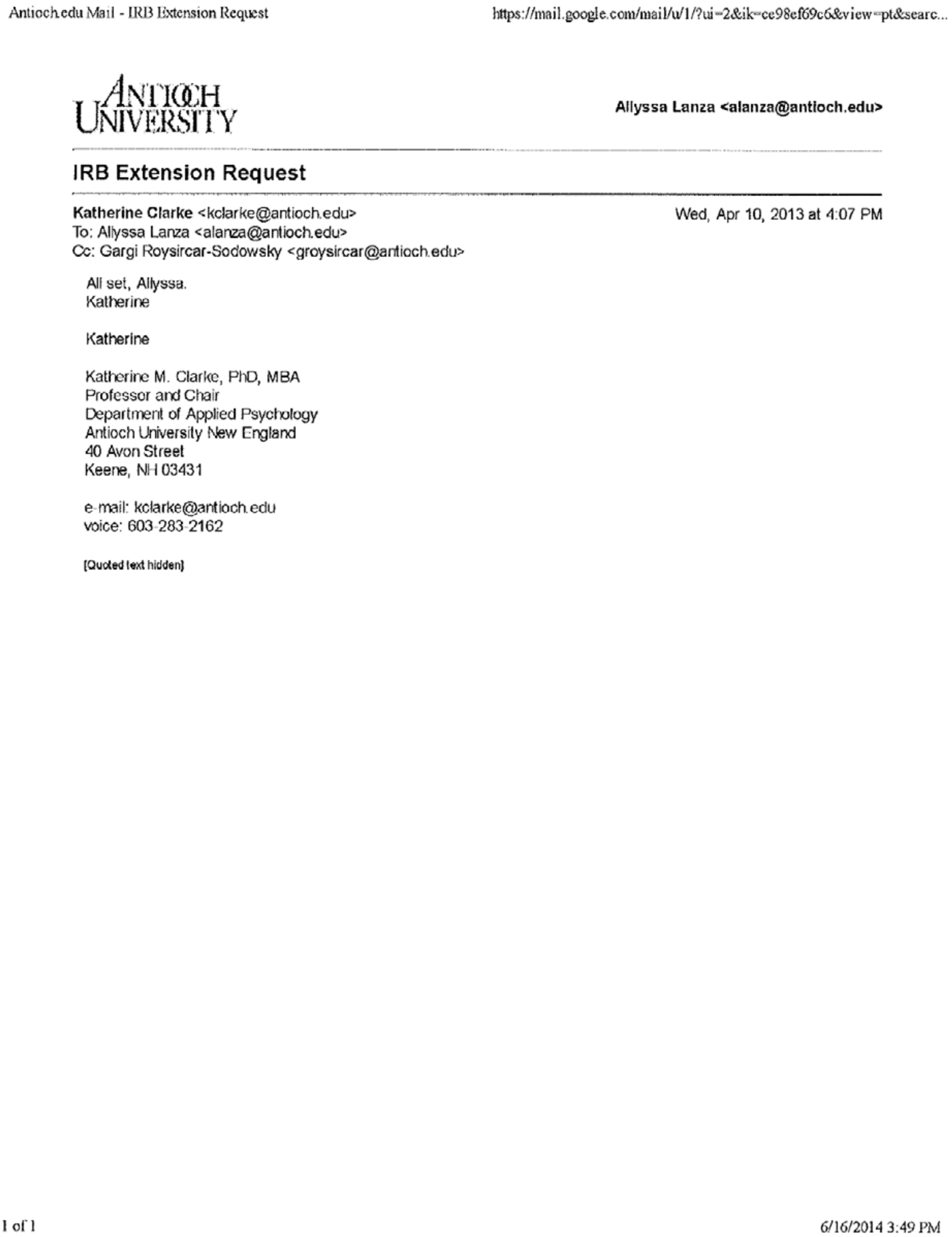
The general purpose of this study is to: Evaluate for hidden floor effects in the US version of the Wechsler Intelligence Scale for Children, fourth edition (WISC-IV)

<u>Originally Approved</u>	<u>Proposed Change</u>
1. Data collection will occur up until April 14, 2013	1. Data collection will occur up until April 14, 2014

The risks:benefits ratio will change in the following ways:

It is not anticipated that the risk/benefit ratio will be affected by these changes.

Appendix C: IRB Extension Permission



Antioch.edu Mail - IRB Extension Request https://mail.google.com/mail/u/1/?ui=2&ik=ce98ef69c6&view=pt&searc...

ANTIOCH UNIVERSITY

Allyssa Lanza <alanza@antioch.edu>

IRB Extension Request

Katherine Clarke <kclarke@antioch.edu> Wed, Apr 10, 2013 at 4:07 PM
To: Allyssa Lanza <alanza@antioch.edu>
Cc: Gargi Roysircar-Sodowsky <groysircar@antioch.edu>

All set, Allyssa.
Katherine

Katherine

Katherine M. Clarke, PhD, MBA
Professor and Chair
Department of Applied Psychology
Antioch University New England
40 Avon Street
Keene, NH 03431

e-mail: kclarke@antioch.edu
voice: 603-283-2162

[Quoted text hidden]

1 of 1 6/16/2014 3:49 PM