ABSTRACT

KROUSE, HAILEY ELIZABETH. The Reliability and Validity of the WISC-IV with Deaf and Hard-of-Hearing Children. (Under the direction of Jeffery P. Braden.)

The present study examined the reliability and validity of the Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV) for use with Deaf and Hard-of-Hearing (D/HOH) children. The participants, who were psychologists working directly with D/HOH children, entered data on D/HOH children (n= 128) in encrypted Excel spreadsheets sent via email. Results revealed that 8 of the 10 WISC-IV subtests/composites assessed were significantly more reliable ($p < .05$) compared to the split-half internal consistency reliabilities reported in the *WISC-IV Technical and Interpretive Manual* (2003) for the normative sample. In addition, the mean Perceptual Reasoning Index ($M = 93.21$) and Verbal Comprehension Index ($M = 80.86$) for this sample were significantly lower ($p < .001$) than the population mean ($M = 100$). Although the mean Verbal Comprehension Index was not significantly lower than one standard deviation from the mean ($M = 85$) Interrelationships among the WISC-IV subtests for this sample were assessed through Pearson Product Moment correlations. Of the 44 correlations, 29 were significantly greater than zero (i.e., the 95% confidence interval did not contain zero). Overall, the results support the reliability of the WISC-IV for D/HOH children. However, the evidence for the validity of the WISC-IV with D/HOH children is inconclusive. Further research is needed to investigate the validity of the WISC-IV (e.g., convergence, test-criterion, factor structure) for use with D/HOH children.

The Reliability and Validity of the WISC-IV with Deaf and Hard-of-Hearing Children

by
Hailey Elizabeth Krouse

A thesis submitted to the Graduate Faculty of
North Carolina State University
In partial fulfillment of the
Requirements for the Degree of
Master of Science

Psychology

Raleigh, North Carolina

April 25, 2008

APPROVED BY:

_______________________________ _______________________________
Patricia W. Collins, Ph.D. John L. Nietfeld, Ph.D.

____
Jeffery P. Braden, Ph.D.
Chair of Advisory Committee

DEDICATION

To my parents, Charles and Anne Krouse, whose unconditional love for each other, optimistic outlook on life and unwavering support throughout the past 26 years have helped me keep life's trivialities in perspective and realize what is truly important in this world.

I love you!

BIOGRAPHY

Hailey Elizabeth Krouse was born on April 29, 1982 in Morehead City, NC. Growing up on the Crystal Coast, her childhood memories are filled with afternoon picnics at the ocean's edge, galloping through the sandy dunes and splashing in the salty waves. When Hailey was not quite 2 years old, her sister, Laura, was born. They instantly became the best of friends and unexplainably share a bond that grows stronger with every passing day. When Hailey was 6 years old, she began praying for a little brother. Less than a year later, Stephen was born, completing (in more ways than one) the Krouse family.

Hailey attended St. Egbert's Catholic School from kindergarten through fifth grade. This loving, challenging, Catholic education provided the solid foundation upon which she built the rest of her educational career. Following St. Egbert's, Hailey attended Morehead City Middle School and then entered West Carteret High School in 1996. Beginning her freshman year, Hailey was an active member of the West Carteret Marching Patriots Colorguard. It was there, on the marching band field, that she met and fell in love with her boyfriend of 11 years, Michael Ridgeway. In addition to being the captain of the colorguard, Hailey was also a member of the National Art Honor Society and the Vice President of the National Honor Society. Hailey graduated from West Carteret High School, in the top 10 of her class, in 2000.

In the fall of 2000, Hailey officially became a Tarheel attending the University of North Carolina at Chapel Hill. After thoughts of pursuing medicine and a career as nurse practitioner midwife, Hailey realized psychology was her true passion. Following in the footsteps of her mother and grandmother, Hailey majored in psychology. She worked with

Dr. J. Steven Reznick, a professor in developmental psychology, and completed an undergraduate honors thesis investigating the relationship between working memory and language in infants. In addition to psychology, Hailey also studied Spanish and spent a semester studying abroad in Mexico during her junior year. After being inducted in the Phi Beta Kappa Honors Society, Hailey graduated in the spring of 2004 with a Bachelor of Arts in psychology with highest honors.

During her senior year at UNC-Chapel Hill, Hailey spent several months investigating and applying to various doctoral level school psychology graduate programs in the southeast U.S. However, it did not take her long to realize that North Carolina State University was the right place for her. After countless hours of reading, writing, studying, collecting data and exploring the field of school psychology, Hailey will receive her Master's of Science in the summer of 2008. Some of her other accomplishments while at NCSU include, teaching statistics to undergraduate students, teaching study skills to middle school students, co-authoring a review for the Mental Measurements Yearbook with Dr. Jeff Braden, collaborating with Dr. John Begeny in the implementation and investigation of reading interventions with elementary school children, and presenting posters at local conferences. Currently, Hailey is pursuing her PhD at NCSU in school psychology.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Jeff Braden, whose patience, support and guidance made this accomplishment possible. His time and knowledge have been invaluable during this learning process. I am indebted to the psychologists from across the country who have so generously dedicated their time and energy collecting and entering the data for this project. Truly, without their help, this study would not have been possible. I would also like to thank Patsy Collins and John Nietfeld for their time, feedback and collaboration on this project. I would also like to thank Jason Allaire and Shevaun Neupert for their statistical guidance. Thank you to Ann Schulte for her general encouragement to get this project done! I would also like to extend a special thanks to my "research team," Lauren Joyce and Michelle Tayrose. I would not be here without their academic and emotional support.

I am forever grateful to my parents, who have always loved and believed in me and from a young age, helped me grow the wings I've needed to reach my goals. Special thanks for my sister, Laura, who has lived with and put up with me throughout graduate school. I know it has not been an easy feat! I would also like to thank my brother, Stephen, who teaches me to seize the moment and realize that everyday is a gift. A very special thanks to my grandmother, Verne Reintjes, who has always lead by example. I am proud to follow in her footsteps. Her emotional and financial support throughout college and graduate school has allowed me to concentrate on my school work and reach my goals.

I would also like to thank my best friends, Ginger Kelly and Jessica Robinson, who have always been there for me when I have needed them. I have always felt I could count on them for excellent advice and a relaxing cup of tea. And, finally, I want to thank my boyfriend, Michael Ridgeway, who has literally held me up and kept me going when no one else could. You're the only one in this world who truly gets me. I owe much of my success to you. Thank You!

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

Introduction

The Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV) (Wechsler, 2003) has recently replaced the Wechsler Intelligence Scale for Children-Third Edition, or WISC-III. The new edition brought with it several significant changes in the test materials, most notably the elimination of the Performance and Verbal Scale dichotomy, and the adoption of four Index scores. To assess and enhance the clinical utility of the WISC-IV, 16 special group validity studies were conducted concurrently with the scale's standardization. However, deaf and hard-of-hearing (D/HOH) children[1] were not included among these special assessments. Therefore, there are no data relative to the WISC-IV with D/HOH children (Wechsler, 2003). Because the Wechsler Scales for Children are often used as part of a diagnostic assessment for D/HOH children, and because the WISC-IV is replacing the WISC-III as the preferred cognitive test to be used with this population, its applicability to D/HOH children should be evaluated.

The main purpose of this study is to assess the reliability and validity of the WISC-IV with D/HOH children. To justify and provide a context for this goal, I will provide the definitions of terms unique to the D/HOH population, followed by the prevalence rate of hearing impairments among children. Next, the importance of reliability and validity will be discussed, followed by a description of the types of intelligence tests that are appropriate for use with this special group. After that, an overview of findings of the reliability and validity of the previous Wechsler Scales for the cognitive assessment of D/HOH children will be provided. The discussion then switches to a focus on the new WISC-IV, the changes made and the implications for D/HOH children. Finally, the main purpose of the current study is described followed by hypotheses, methods for data collection, and proposed analyses.

Definition of Terms and Prevalence Rates

Before beginning the discussion on intelligence testing with D/HOH children, it is first imperative to review terms unique to this special population. Specifically, hearing loss, modes of communication, and prevalence rates are explored to define the D/HOH population.

Hearing Loss

Degree of Hearing Impairment

The degree of hearing impairment or severity is defined using two dimensions, the frequency and the intensity. The frequency is described as the pitch or tone of the sound, which is measured in Hertz (Hz) or cycles per second. The intensity or loudness of the sound is measured in decibels (dB). Thus, a hearing impairment is a defined level of intensity needed for a person to perceive sound at a specific frequency (Sattler & Hardy-Braz, 2002). It is important to note that there is no standard agreement as to the definitive criteria for each category; however, the characteristics commonly used to describe the different degrees of hearing loss are presented on the next page in Table 1 (Northern & Downs, 2002).

Pure Tone Average

Pure Tone Average (PTA) is another way to describe the severity of a hearing loss. Most speech sounds are created using frequencies between 500 and 2,000 Hz. Therefore, when diagnosing hearing impairments, the major concern becomes what intensity is needed to hear frequencies in this speech range. An individual's PTA is the average intensity ratings (measured in dB) needed to hear frequencies of 500, 1,000, and 2,000 Hz (Tye-Murray, 2004).

Table 1

Degrees of Hearing Loss and Brief Descriptions of Typical Hearing Abilities

Average Hearing Level	Classification of Hearing Loss	Typical Hearing Abilities (500-2,000 Hz, or the range in which most speech sounds are produced)
20 - 40 dB	Mild	These children comprehend vowel sounds clearly but may not understand unvoiced consonant sounds (e.g., /t/, /p/).
41 – 70 dB	Moderate	These children have difficulty comprehending most speech sounds when there is background noise.
71 – 90 dB	Severe	These children only comprehend the loudest speech sounds. They cannot comprehend any speech sounds at normal conversational levels.
> 90 dB	Profound	These children usually do not comprehend speech or other sounds.

Type of Hearing Impairment

There are three categories of hearing impairments: (a) conductive impairments, (b) sensorineural impairments, and (c) mixed impairments. Conductive impairments imply that the problem transmitting sound lies in the outer or middle ear. Some examples of conductive impairments include: blockage of the auditory canal, fluid in the ear associated with chronic ear infections, rupture of the eardrum, and calcification of the bones in the middle ear. Sensorineural impairments result from damage in the cochlea (i.e., the organ that converts impulses in the fluid of the inner ear to neural impulses) or the auditory nerve (i.e., the nerve connecting the cochlea to the brain). The third category of hearing impairments is called mixed. This category describes losses that include both conductive and sensorineural impairments. Finally, when describing the nature of a hearing loss, it is also essential to explain whether the impairment is unilateral (i.e., present in one ear) or bilateral (i.e., present in both ears) (Northern & Downs, 2002).

Age of Onset

Medical descriptions of hearing loss often distinguish between congenital onset (i.e., a hearing impairment present at birth) and adventitious onset (i.e., a hearing impairment acquired after birth). However, psychologists typically distinguish between prelingual and postlingual onset of hearing loss (Braden, 2000). A prelingual hearing loss is a hearing impairment that occurs before the acquisition of oral language skills. A postlingual hearing loss is a hearing impairment that occurs after the acquisition of oral language skills (Tye-Murray, 2004). The age at which oral language skills are “acquired” is debated. Some psychologists suggest that 2 years of age should be the cutoff to define prelingual from postlingual onset, due to the fact that children at this age begin to utter one-word sentences.

However, other psychologists argue that 5 years of age should be the point of distinction because this is the age at which most children have acquired basic grammar, syntax, and coherent conversational skills (Braden, 1994).

Distinction Between Deaf and Hard-of-Hearing

The distinction between deaf and hard-of-hearing individuals is usually made based on the degree or the severity of the hearing impairment. Typically, hard-of-hearing individuals are those who have mild or moderate hearing loss (i.e., a hearing loss between 20 and 70 dB); whereas deaf individuals have severe or profound hearing loss (i.e., a hearing loss greater than 70 dB) (Tye-Murray, 2004). A child with a hearing loss can generally respond to auditory stimuli, including speech, whereas deafness prevents a child from understanding sound in all or most of its forms (National Dissemination Center for Children with Disabilities, 2004).

It is important, however, to raise the issue of medical versus functional definitions of deaf and hard-of-hearing. Many medical definitions of hearing loss revolve around numbers (e.g., PTA) and anatomical abnormalities. This is quite different than functional definitions, which emphasize one's ability to acquire and use oral/auditory language. There is a clear, but imperfect, relationship between medical and functional definitions of hearing loss. For example, a more severe hearing loss, as defined in medical terms, generally leads to a more restricted ability to understand and use speech. However, there are important exceptions to this relationship. Amplification, such as hearing aids, makes it possible to alter a person's functioning without changing his or her medical classification. In other words, a person who is medically classified as deaf may, with the help of hearing aids, no longer function like a deaf person (i.e., this person now uses speech as the primary mode of communication). In a functional sense, the distinction between deaf and hard-of-hearing is the ability to comprehend and produce oral/auditory speech (with or without amplification) (J.P. Braden, personal communication, September 11, 2006).

Complicating the issue even further is the fact that states use a wide variety of terminology to describe hearing loss. Some of these terms include, deaf, hard-of-hearing, fluctuating hearing loss, unilateral hearing loss, hearing handicapped, hearing disorder, aural handicapped, auditorily handicapped, and severely handicapped. The definitions proposed by The Individuals with Disabilities Education Improvement Act of 2004 (IDEIA-04) for deafness and hearing impairment do not use specific measurable degrees of hearing loss (e.g., in decibels) to determine the criteria for the two categories. One of the reasons why state definitions are inconsistent is the lack of a common, federal definition. States may use the federal definitions to determine eligibility for special education, or they can create their own criteria (as long as their standards meet the minimum requirements established in the federal regulations) (Bienenstock & Vernon, 1994).

Terminology and eligibility requirements vary across states. For example, in one state a child with a hearing loss of 70 dB might be classified as deaf, whereas in another state the child would be classified as hard-of-hearing. A student with a hearing loss of 35 dB might be eligible to receive services in one state, but not in a neighboring state. Some states classify children with unilateral hearing impairments as eligible for special education services, and other states do not. This inconsistent classification system makes it difficult to conduct valid research, gather meaningful demographic data, and communicate clearly with professionals from other states (Bienenstock & Vernon, 1994).

Modes of Communication

American Sign Language

American Sign Language (ASL) is a manual system of communication used by members of the Deaf Culture in the U.S. ASL is different from English in its grammar and syntax and is considered a distinct language (Tye-Murray, 2004). ASL does not have a written form (beyond a few coding systems used only in research).

Other Sign Languages

Another manual system of communication is known as manually coded English (MCE). MCE is comprised of signs that directly correspond to English words. The grammar and syntax rules of MCE are also synonymous with those of English. Typically, a person using MCE will sign and speak at the same time; this is known as simultaneous communication (Tye-Murray, 2004). MCE is used primarily in educational contexts to teach students with hearing impairments. In contrast, ASL is primarily used in social contexts between members of the Deaf community.

Aural/Oral Language

Aural/oral language is the same language used by persons with normal hearing. The child with a hearing impairment who successfully uses this mode of communication has "typical" expressive language skills, but may use speechreading (i.e., lipreading) to aid in receptive language. The main emphasis behind using this model is that instead of adapting the communication mode to fit the needs of the child, the child adapts to the communication norm (i.e., speech) (Tye-Murray, 2004).

Cued Speech

Cued Speech uses phonemically based hand gestures to supplement speechreading. Eight different handshapes are used to distinguish consonant sounds, and six hand locations around the face and neck are used to distinguish vowel sounds. Thus, the individual speaks while simultaneously cueing the message. The gestures help distinguish between similar visual speech patterns. Alone, the hand signals are uninterpretable; however, when coupled with lip movement and sound, speech recognition increases (Tye-Murray, 2004). Like MCE, Cued Speech is an effort to visually represent spoken English to D/HOH "listeners," and is used almost exclusively in educational settings.

Prevalence Rates

Because data on the U.S. deaf population have not been collected since 1971, only approximate estimations of prevalence rates are available. Gallaudet University states that one of the best current estimates of U.S. D/HOH population is that of Holt, Hotto, and Cole (1994). Holt et al. published estimates based on surveys conducted in 1990-1991 by the National Center for Health Statistics (NCHS) of the U.S. Department of Health and Human Services. Tables 2, 3, and 4 provide prevalence information created from these estimates.

Table 2

Percent Estimates of the Deaf or Hard-of-Hearing US Population Based on Age Groups

Age Group	Percent
3 – 17 years	1.8 %
18 – 34 years	3.4%
35 – 44 years	6.3%
45 – 54 years	10.3%
55 – 64 years	15.4%
65 years and older	29.1%
Total	8.6%

Table 3

Estimates of the US Population Who are Deaf and Hard-of-Hearing by Age Groups and Gender

Age Group	Male	Female *
3 – 17 years	541,000	427,000
18 – 44 years	3,018,000	1,672,000
45 - 64 years	3,946,000	1,963,000
65 years and older	4,497,000	4,232,000
Total	12,002,000	8,293,000

Note. * Due to rounding, the total in this column does not equal the sum of the numbers

Table 4

Percent Estimates of the Age at Onset of Hearing Loss in the US Deaf and Hard-of-Hearing Population

Age of Onset	Percent
Before 3 years	5.4%
3 – 18 years	14.2%
19 years and over	76.3%
Unknown	4.1%
Total	100.0%

According to the United States Department of Education (2003), during the 2001-2002 school year, more than 71,000 children between the ages of 6 and 21 received special education services under the category "hearing impairments," which includes deafness. This means that approximately 1.2% of the children currently receiving special education services are deaf or hard-of-hearing. However, this is almost certainly an underestimate of the total number of children with hearing loss. It is most likely the case that some D/HOH children are receiving special education services under another category, whereas others are not counted because they receive only regular education services (National Center on Birth Defects and Developmental Disabilities, 2004).

Summary

D/HOH children likely account for a little over 1% of the children receiving special education services in the United Sates. Although this is a relatively small percentage of students, the low numbers should not minimize the importance of understanding this complex impairment. Due to the fact that hearing loss is a multifaceted phenomenon and that D/HOH children, as a group, are extremely heterogeneous, it is important to observe and understand the terms used to describe the unique functioning of D/HOH children. When describing hearing loss it is important to note the child's degree of hearing loss, often discussed in terms of PTA, as well as his or her classification of severity (i.e., mild, moderate, severe and profound). Considering the type of hearing impairment (i.e., conductive, sensorineural, or mixed) as well as the age of onset (i.e., congential, prelingual, or postlingual) are also important for understanding the implications of the hearing loss on the child's functioning. In addition, it is critical to note that there are multiple distinctions between "deaf" and "hard-of-hearing" individuals (e.g., medical versus functional definitions) and that different states use different definitions and eligibility requirements. Finally, psychologists should understand the various modes of communication used among this special group (e.g., ASL, MCE, aural/oral language, and cued speech) and the differences between these methods of communication.

Assessing the Intelligence of Deaf and Hard-of-hearing Children

More than 90% of D/HOH children are born and raised in households in which the primary mode of communication is speech (Gallaudet Research Institute, 2002). As a consequence, children with prelingual hearing losses are denied the early and consistent access to language available to normal hearing children. The nonstandard exposure to language reduces opportunities for accessing and exploring the surrounding community and culture, which leads to an impoverished knowledge base on which to acquire and build new information. Hearing loss and the subsequent dearth of language exposure undoubtedly affects children's development of language, and may affect other cognitive abilities as well. As a result, many D/HOH children appear to have intellectual deficits. For example, many D/HOH children do not talk, and if they do, their speech is often delayed, unintelligible, and immature compared to normal hearing peers (Braden, 2005). It is not uncommon for D/HOH children to make unusual noises, express deficits in adaptive behavior related to communication skills, and use tantrums or other "less linguistically mature forms of behavior" to get what they want (Braden, 2005, p. 352-353). These behaviors often lead parents and others (e.g., teachers, physicians, psychologists) to believe that the child also has cognitive deficits.

Although behaviors associated with hearing loss may be misdiagnosed as indicators of cognitive difficulties, hearing loss can and does occur with cognitive deficits. For example, meningitis, neurological damage, and some syndromic genetic conditions can cause deafness as well as cognitive delays. Hearing loss and the subsequent nonstandard exposure to language may also affect intellectual development, especially language and auditory dependent cognitive processes, such as verbal comprehension, verbal reasoning, and auditory processing. Even though many D/HOH children may have normal development in some areas of functioning, psychologists must assess whether the child's behavioral, academic, and/or linguistic deficits are a consequence of the hearing impairment, an intellectual impairment, or a result of multiple factors. A valid and reliable appraisal of the intellectual abilities of D/HOH children is often a crucial component of this assessment process (Braden, 2005).

Psychometric Properties of Intelligence Tests

Psychological testing and assessment is one of the most important contributions made by behavioral science to society. The proper use of tests can help educators and psychologists make informed decisions about children that can lead to more equitable access to education. However, improper use of tests can cause serious and detrimental consequences to test takers and other parties affected by test-based decisions (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). The *Standards for educational and psychological testing* (AERA, APA, NCME, 1999) (the *Standards*) was published to provide all participants involved in the testing process (e.g., those who develop, publish, administer, take, and review the test, and those who use the test results for decision-making) a set of criteria to help evaluate tests, testing practices, and the effects of test use so that negative consequences of testing are minimized. "Test developers and those selecting and interpreting tests need adequate knowledge of psychometric principles such as validity and reliability"(AERA, APA, NCME, 1999, p. 2). The *Standards* have an entire chapter devoted to assessment of clients with disabilities (Chapter 10). Standard 12.13 states:

> Those who select tests and draw inferences from test scores should be familiar with the relevant evidence of validity and reliability…in situations in which the selection of tests may be problematic (e.g., verbal subtests with deaf clients), a brief description of the rationale for using or not using particular measures is advisable (p. 133).

Therefore, psychologists administering various intelligence tests to D/HOH examinees should be aware of the psychometric properties of the test, and understand how such properties might support or limit the use of the test with these particular clients.

Reliability

Reliability refers to consistency in measurement (AERA, APA, NCME, 1999). A test must be identified as reliable before it can be declared valid. In other words, reliability is a necessary, but not sufficient, precursor of validity. *The Standards* (AERA, APA, NCME, 1999) identifies three broad categories of reliability coefficients: (1) alternate-form, (2) test-retest, and (3) internal consistency. Alternate-form coefficients are calculated from scores obtained on independent administrations of parallel test forms. Test-retest, or stability, coefficients are derived from the independent administrations (i.e., two or more) of the same instrument to the same set of examinees. Finally, internal consistency coefficients indicate the homogeneity or interrelatedness of the test items. Internal consistency coefficients are calculated based on the relationship among scores of individual items or subtest items from a single test on a single administration.

There are several advantages of assessing the internal consistency reliability over the other two types of reliability measures. First, common intelligence tests, like the WISC-IV, do not provide alternate forms needed to evaluate alternate-form reliability. Second, assessing internal consistency only requires that the test be administered once. Stability coefficients can be problematic because retesting often results in an inflated correlation between the first and second scores due to examinee recall of initial responses or familiarity with the idiosyncratic features of the test (AERA, APA, NCME, 1999). Stability coefficients also confound consistency of the test over time with the stability of the trait being measured (e.g., one would expect fairly high stability for IQ, but not for mood). And third, Wechsler (2003) reports evidence of WISC-IV internal consistency reliability for 16 special group populations. Therefore, investigating the internal consistency reliability of the WISC-IV with D/HOH children would allow direct comparisons between the reliability for D/HOH examinees and examinees from other special groups reported in the WISC-IV manual.

There are three main methods that can be used to calculate the internal consistency reliability for tests of intelligence. One method is called split-half reliability. Split-half reliability correlates pairs of scores obtained from equal halves of the test (e.g., correlates odd responses with even responses). Due to the fact that these correlations are only based on half the test, they are usually an underestimate of the reliability of the whole test. Therefore, they may be adjusted using the Spearman-Brown formula. It should be noted here that this is the method used in the *WISC-IV Technical and Interpretive Manual* for reporting the internal consistency reliability of the WISC-IV with other groups of children (Wechsler, 2003).

A second method of calculating internal consistency is called Cronbach's (1951) coefficient alpha. Cronbach's coefficient alpha can be described as the average of all possible split-half coefficients, corrected by the Spearman-Brown formula. In other words, instead of just splitting the test once (as in split-half reliability), coefficient alpha is the mean correlation among all possible splittings of a test.

A third method of calculating internal consistency was developed by Kuder and Richardson (1937) and is referred to as the Kuder-Richardson formula 20, or KR-20. This reliability calculation is extremely similar to coefficient alpha; however, KR-20 is limited to cases in which each test item is scored either right or wrong (e.g., either 1 or 0), whereas coefficient alpha may be applied to items producing any number of ordinal scores (Gregory, 2004).

In most cases, these methods of calculating internal consistency reliability are suitable to use with tests of intelligence. However, it is important to note that internal consistency reliability (including split-half coefficients, coefficient alphas, and KR-20 coefficients) is not an appropriate estimate of reliability for "speed tests" (i.e., tests in which speed is the key factor). Speed tests typically comprise items of minimal difficulty. Given sufficient time, examinees are likely to complete all items correctly. The challenging component of a speed test is the time limit. Due to the fact that speed tests involve relatively simple tasks, usually all items completed are correct, and only items not attempted are incorrect. Therefore, calculating internal consistency reliability for speed tests will result in a spuriously high reliability coefficient. It is better to calculate the reliability of a speed test using either the test-retest method or the split-half reliability from two, separately timed half tests. In the latter option, the Spearman-Brown correction is needed (Gregory, 2004).

Validity

Validity is the most fundamental aspect in the development and evaluation of a test. Validity refers to the degree to which the intended meaning of test scores (i.e., the constructs and concepts the test purport to measure) is supported by theory and evidence. According to *The Standards* (AERA, APA, NCME, 1999), there are five sources of evidence that may be used to support validity: (1) test content, (2) response processes, (3) internal structure, (4) relations with other variables, and (5) consequences of testing. Evidence based on test content often includes logical analyses and experts' evaluation of the content of the measure (including items, tasks, formats, wording, and processes required of examines). In general, it addresses the extent to which the content of a measure represents a specified content domain. Evidence based on response processes examines the extent to which the responses required of the examinees fit the intended construct of interest. The evidence of response processes generally comes from analyses of individual responses. Evidence based on internal structure examines the extent to which the internal components of a test conform to the construct of interest. This is often assessed using factor analysis and differential item functioning (DIF). Evidence based on relations to other variables examines the relationship of test scores to variables external to the test. This type of validity evidence is most often assessed through group comparisons and correlational studies. Finally, evidence of consequences of testing pertains to anticipated and unanticipated consequences of a test. It is important to indicate whether specific consequences (whether positive or negative) are likely to be realized (Goodwin & Leech, 2003).

It is important to note that the 1999 edition of *The Standards* (AERA, APA, NCME) brought about many changes to the conceptualization of validity. The previous versions of *The Standards* (AERA, APA, NCME, 1985; APA, AERA, NCME, 1966) defined validity in terms of a tripartite model. In this view, validity was categorized into specific types: content, criterion-related, and construct validity (Goodwin & Leech, 2003). Due to these recent changes, validity research conducted prior to 1999 (as well as research conducted shortly after) does not "fit" with the current validity framework. This makes reviewing validity research slightly more challenging. In an attempt to incorporate the old view of validity in with the new, Goodwin and Leech (2003) discuss how research using the tripartite model of validity can be integrated within the newly distinguished sources of validity evidence. For example, research assessing DIF was considered construct validity evidence and is now considered evidence based on internal structure. When considering the validity of intelligence tests for use with D/HOH examinees, the most common sources of validity evidence are DIF, factor analyses, analyzing item interrelationships, correlational studies with other intelligence and achievement tests, and group comparison studies that assess the differences between means scores for D/HOH and normally hearing examinees (e.g., Braden, 1989; Braden, 1990; Braden, Kostrubala & Reed, 1994; Brill, 1962; Hirshoren, Hurley & Kavale, 1979; Kelly & Braden, 1990; Maller, 1996; 1997; Maller & Braden, 1993; Sisco & Anderson, 1978; Slate & Fawcett, 1995). DIF and factor analyses are now considered validity evidence based on internal structure. Correlational studies and group comparison studies are now considered validity evidence based on relations to other variables. Each of these sources of evidence is discussed in more depth below.

Evidence based on internal structure

Validity evidence based on internal structure examines the extent to which the internal components of a test match the construct of interest. DIF and factor analytic studies are two common methods for assessing the internal structure of a test (Goodwin & Leech, 2003). DIF, commonly identified as item bias, is a statistical procedure that determines if a specific subgroup (e.g., the D/HOH population) is more or less likely to answer a particular question correctly because it is easier or more difficult for that group. In other words, DIF assesses the probability of a correct response to an item between a deaf and hearing sample of equal ability (e.g., matched IQ as measured by a set of nonbiased items; Maller, 2003).

A second common method of assessing for internal structure is through factor analysis. A test's factor structure should be equivalent for deaf and hearing samples if the test measures the same construct in both groups. If relationships between subtests differ across groups, scores may have different meanings for the two groups (i.e., the subtests are not assessing the same construct in the two groups). There are two types of factor analysis, (1) exploratory (EFA) and (2) confirmatory (CFA). EFA is used when researchers do not have a priori theory regarding the structure of the test. CFA is used when researchers want to test a hypothesized theoretical model (Maller, 2003).

Finally, the internal structure of a test can be examined through an analysis of item interrelationships. One way to do this is to provide a table of the correlations among subtests within a test battery. This illustrates the pattern of relationships within the test and provides a measure of test homogeneity. If a test measures a single construct (e.g., working memory, processing speed, verbal IQ, nonverbal IQ), then its subtests will be homogeneous (i.e., internally consistent). Homogeneity is important in certifying the internal structure (i.e., validity) of a test (Gregory, 2004).

Evidence based on relations with other variables

Validity evidence based on relations to other variables is the most extensive of the five sources of evidence (Goodwin & Leech, 2003). According to Goodwin and Leech (2003), the "old" construct validity that assessed group comparisons of mean scores is included in the new validity framework under evidence based on relations to other variables.

Therefore, known-group comparison studies that test hypotheses about expected differences in average scores across different groups is one way to provide validity evidence. (The use of group comparisons of average scores will be discussed in more depth under the section entitled "Verbal versus Nonverbal Scales.") In addition to including group comparisons of mean scores, this type of validity evidence also includes correlational studies that assess relationships between (a) measures of the same construct, (b) measures of different constructs, and (c) some criterion the measure or test is expected to predict.

The relationship between a test score and another measure intended to assess the same construct (e.g., the relationship between WISC-IV scores and Woodcock-Johnson—Third Edition, Tests of Cognitive Abilities scores) provides convergent evidence. The relationship between a test score and another measure intended to assess a different construct (e.g., the relationship between WISC-IV scores and the Woodcock-Johnson—Third Edition, Tests of Achievement scores) provides discriminant evidence. Both of these relationships can provide evidence of a test's validity (AERA, APA, NCME, 1999).

The relationship between a test and some criterion that the test is expected to predict is referred to as a "test-criterion relationship." In a test-criterion relationship, the "test" is the measure of interest and the "criterion" is a socially valued external variable. In the case of an intelligence test, one common criterion is performance in school. There are two types of test-criterion relationships, (a) predictive, and (b) concurrent. A predictive relationship reveals how well a test can predict a criterion variable at a later point in time (e.g., a third grade IQ score predicting a high school Grade Point Average, GPA). A concurrent relationship reveals the association between the test and a criterion variable when the information was gathered at or near the same point in time (e.g., the relation between an IQ score and an achievement score acquired on the same day) (AERA, APA, NCME, 1999).

Summary

All individuals who give tests or interpret test scores need evidence of reliability and validity to understand the consistency and meaning of examinee scores. Therefore, psychologists who use intelligence tests with D/HOH children should be familiar with the psychometric properties of the test specific to the D/HOH population. Reliability refers to consistency in measurement. There are three distinct methods for calculating reliability (i.e., alternate-form, test-retest and internal consistency). The most common measure of reliability for the Wechsler Scales with D/HOH examinees has been internal consistency reliability. Split-half reliability coefficients, corrected by the Spearman-Brown formula, are reported in the WISC-IV Technical and Interpretive Manual for 16 groups of children. Therefore, to allow for direct comparison, this method will be the focus of this study. However, it is important to keep in mind that although this method of measuring internal consistency is suitable for most tests of intelligence, it is not appropriate to use for tests in which speed is a key factor.

Reliability is a necessary, but not sufficient, precursor of validity. Validity refers to the degree to which the intended meaning of test scores is supported by theory and evidence. There are five sources of evidence that may be used to support the validity of a test (1) test content, (2) response processes, (3) internal structure, (4) relations with other variables, and (5) consequences of testing. One common source of evidence for the validity of the Wechsler Scales comes from assessing the internal structure of the tests. This is generally accomplished through factor analyses, DIF, and analyzing item interrelationships (i.e., test homogeneity). Evidence for the validity of Wechsler Scales with D/HOH examinees is also provided through assessing its relationship with other variables. These "other" variables can include measures of the (a) same construct, (b) different construct, or (c) some criterion the test is expected to predict (identified as either concurrent or predictive relationships). This area of validity evidence also includes group comparison studies of average scores (e.g., differences in verbal and nonverbal scores for D/HOH children; discussed in more depth below).

Types of Intelligence Tests

Verbal versus Nonverbal Batteries

Before discussing verbal and nonverbal batteries of intelligence, it is first necessary to note the difference between the terms "test" and "scale." Too often, these terms are used interchangeably, when, in fact, a clear distinction should be made between the two. The term test simply refers to a series of questions or exercises that are given to an individual, or group of individuals, with the intent to measure a specific construct (e.g., intelligence). Psychometrists distinguish between a test and a scale by noting that scales arrange questions or items in a graduated series (e.g., items are arranged from easy to hard). Specifically, the Wechsler Scales (with the exception of the WISC-IV) distinguish between language loaded and language reduced collections of subtests, termed Verbal Scales and Performance Scales. A more detailed discussion of language loaded (i.e., verbal) and language reduced (i.e., nonverbal) test batteries, the Wechsler Verbal and Performance Scales, and their implications for assessing D/HOH is provided below.

The subtests of verbal and nonverbal batteries are based on different assumptions about the individual's previous life experiences and exposure to certain events. Verbal subtests require the use of language in all aspects of the testing (e.g., in the directions, administration, cognitive processes elicited, and responses). Therefore, the subtests within verbal batteries presume that the examinee has been exposed to the verbal content of the items as often as those individuals in the normative sample. With this assumption intact, the verbal score is thought to reflect the aptitude or intelligence of the individual. In contrast, the subtests of nonverbal batteries aim to minimize or eliminate the need for language. Nonverbal batteries have limited linguistic content, provide opportunities for the examinee to understand directions and cognitively process the information with little or no language, and respond nonverbally. Verbal batteries can be thought of as "language-loaded," and nonverbal batteries as "language-reduced" (Braden & Anathasiou, 2005). Many believe nonverbal batteries, or "language-reduced" batteries, to be more appropriate for D/HOH examinees due to the reduced reliance on exposure to and comprehension of spoken language (Braden, 1994; Braden & Athanasiou, 2005).

This distinction between verbal and nonverbal batteries is important when discussing the intellectual assessment of hearing-impaired individuals. For the majority of prelingually deaf children, exposure to language does not begin until after the deafness is diagnosed and interventions have begun, which can be months, even years, after the hearing loss occurs. Even after the diagnosis, the degree of language exposure provided within the child's family varies greatly, depending on the resources available, the skills and training needed to learn alternative modes of communication, and the motivation of the family and child. For

D/HOH children, the hearing impairment limits opportunities to observe and participate in oral communication. D/HOH children experience fewer, less frequent linguistic interchanges than their normal hearing peers. However, not only is the frequency of their language exposure greatly reduced, but they also experience less intense and complex interactions. That is, parents and others frequently reduce the complexity of their messages when speaking to D/HOH children. Whereas this approach increases the likelihood of a successful interchange, the limited frequency, reduced complexity, and lower intensity of language interactions cumulates over time, resulting in poor language development for these children. Therefore, it cannot be assumed that D/HOH examinees have been exposed to the verbal content of the items on a verbal battery as often as those individuals in the normative sample. This fact undermines the usefulness of verbal batteries to reflect D/HOH examinees' intellectual abilities (Braden, 1994).

Verbal IQs of D/HOH children

"There is uniform agreement that systematic deprivation of exposure to verbal, socially specific knowledge impairs performance on verbal scales independent of an individual's underlying aptitude" (Braden, 1994, p. 76). Many researchers have argued that D/HOH children's lack of consistent exposure to verbal, socially specific knowledge impairs performance on verbal batteries and, therefore, scores from language-loaded tests do not accurately reflect the individual's intellectual abilities. It is more likely the case that the depressed verbal IQ score reflects the fact that the deaf examinee has been denied the opportunity to acquire verbal and social knowledge than limited intellectual ability (Braden, 1994).

Few researchers have investigated the validity of the Wechsler Verbal Scales (VS) for use with D/HOH children. Researchers may be discouraged by the fact that Verbal IQ is attenuated by hearing loss and that most D/HOH children (especially deaf children) do not acquire and develop spoken English skills in the same manner as normal hearing children (Maller & Braden, 1993). D/HOH children tend to obtain Verbal IQ scores that fall one standard deviation below the mean score of hearing examinees (Braden, 1994; Maller & Braden, 1993). These children also repeatedly score lower on verbal subtests and composites in comparison to nonverbal or performance subtests and composites (Braden, 2005). The difference between verbal and nonverbal or performance scores is statistically reliable and clinically meaningful, with most studies reporting deficits of approximately 15 points, or about 1 standard deviation (Braden, 1994; Sullivan & Montoya, 1997). However, not only do D/HOH children tend to score lower on verbal tests of intelligence, but DIF has also revealed that verbal test items function differently for D/HOH examinees compared to examinees with normal hearing (Maller, 1996, 1997, 2003).

Because language-loaded intelligence tests confound language skills with intelligence, the use of verbal intelligence tests could result in inappropriate consequences for D/HOH individuals (Braden, 2000). For example, diagnosis and placement decisions may be unduly influenced by low verbal scores, which could have profound implications for the forensic, educational, and vocational services, opportunities, and outcomes experienced by D/HOH examinees. Because of the risk of inappropriate diagnosis, many experts strongly discourage the use of verbal tests of intelligence to assess the intellectual abilities of this special population (Braden, 1994; 2000).

However, some researchers (e.g., Sullivan & Montoya, 1997) insist that verbal cognitive batteries should be used to estimate the intelligence of children with hearing impairments. Sullivan and Montoya (1997) state:

> The historic taboo against the use of verbal intelligence tests with deaf and hard-of-hearing children needs to be reexamined for several reasons. The majority of them are educated in settings where they must compete with hearing peers in academic subjects that are language based (Allen, 1994). Verbal IQ is a better predictor than Performance IQ of reading and math achievement among deaf and hard-of-hearing children (Sullivan & Brookhouser, 1996; Sullivan & Burley, 1990). Finally, for deaf and hard-of-hearing children and youth to obtain higher paying jobs in adulthood, higher levels of numeracy, English literacy, and face-to-face communication skills with hearing peers are required (Allen, 1994; Schildroth, Rawlings, & Allen, 1991) (p. 320).

Although researchers debate the use of verbal measures of intelligence for assessing the cognitive abilities of D/HOH individuals, many researchers advocate the use of verbal batteries to predict academic achievement for this population. Research demonstrates that verbal scores are better predictors of academic achievement and occupational success in D/HOH individuals than nonverbal scores of intelligence (Kelly & Braden, 1990; Maller & Braden, 1993). Also, verbal scores can be useful measures of incidental learning and language acquisition, which can help estimate performance in educational, vocational and social contexts (Braden, 2005). It has been found that the Verbal subtests of the Wechsler

Adult Intelligence Scale-Revised (WAIS-R) are better at predicting literacy in deaf adults than the Performance subtests (Moores et al., 1997). It has also been suggested that the WAIS-R Verbal Scale is a better predictor of deaf students' success in college than the Performance Scale (Falberg, 1983). It may be that English language skills underlie both verbal intelligence and academic achievement (Maller & Braden, 1993). Some examiners also use verbal batteries in an attempt to identify D/HOH children with unusual strengths or weaknesses. Verbally gifted D/HOH children should be exposed to challenging educational experiences and D/HOH children with verbal learning disabilities should receive appropriate support services (Maller, 2003).

Nonverbal IQs of D/HOH children

Nonverbal measures of intelligence are recommended for assessing the cognitive abilities of D/HOH children (Bradley-Johnson & Evans, 1991; Sullivan & Vernon, 1979; Zieziula, 1982). Braden, Kostrubala, and Reed (1994) posit two reasons why experts recommend nonverbal intelligence batteries for use with this population: (1) nonverbal batteries reduce language demands and allow demonstrations to ensure task comprehension, and (2) nonverbal batteries yield similar IQs for D/HOH and normal hearing children (although the authors specifically question the validity of the second reason). Nonverbal intelligence batteries can be divided into two categories, (1) performance batteries, which usually require an individual to manipulate materials to solve problems in speeded conditions, and (2) motor-free nonverbal batteries, which typically require the individual to select a response from a set of options in untimed conditions (Braden, Kostrubala, & Reed, 1994). The most common performance battery used with D/HOH children is the Wechsler

Performance Scales (PS), and the most popular motor-free nonverbal battery used with this population is Raven's Progressive Matrices (Braden, 1994).

Although performance and motor-free batteries are both considered nonverbal measures of intelligence, there is evidence to suggest that motor-reduced nonverbal batteries do not yield the same results as performance batteries with D/HOH children. Braden (1994) conducted a meta-analysis, using 195 studies, to examine the IQ of D/HOH subjects tested with performance batteries. He found the grand mean (*M* = 99.95) and the average standard deviation (*SD* = 15.36) to be well within the average limits (i.e., *M* = 100, *SD* = 15). These findings are virtually identical to those of normal-hearing individuals. However, the results are slightly different using motor-free nonverbal batteries. In the same meta-analysis, Braden aggregated the results of 77 studies assessing intelligence in D/HOH individuals using motor-free nonverbal batteries and found the mean IQ score was significantly lower (*M* = 94.57) compared to the mean IQ for normal hearing individuals (*M* = 100). The average reported standard deviation (SD = 15.95) however, was not significantly different from normal hearing norms (*SD* = 15). This means that in reference to motor-free nonverbal intelligence batteries, the distribution of scores among D/HOH individuals is lower than the distribution of scores among normal-hearing individuals; however, the spread of the scores is similar in both groups. These findings reiterate the conclusion that performance and motor-free nonverbal measures of intelligence do not produce equivalent results (Braden, Kostrubala, & Reed, 1994).

Critical issues in nonverbal measures of intelligence. The use of nonverbal measures to assess intelligence raises several critical issues. First, it should be noted that very few tests/scales completely eliminate language from directions, content, and responses (McCallum, 2003). Therefore, despite the title of the test or how it "claims" to measure intelligence, only those tests that actually eliminate or drastically reduce the need for language in understanding, processing, and responding to test items are considered truly "nonverbal" (Braden & Athanasiou, 2005).

A second critical issue deals with the cognitive processes the examinee uses to respond to test items. Or, in other words, Does the test measure nonverbal intelligence, or does the test measure intelligence nonverbally? Some researchers (e.g., Rourke, 2000; Rouke et al., 2002) claim that nonverbal cognitive processes actually differ from verbal cognitive processes; whereas, other researchers (e.g., McCallum, Bracken, & Wasserman, 2001) argue that nonverbal tests measure general intelligence. Factor-analytic evidence does not support a verbal-nonverbal dichotomy in intellectual abilities (e.g., Horn & Noll, 1997). Over time, the argument that processes underlying intelligence are consistent, whether or not they are verbally mediated, has taken precedence. However, researchers do not have a common framework for resolving conflicting findings regarding verbal and nonverbal intellectual abilities (Braden & Athanasiou, 2005).

A third critical issue is whether nonverbal tests can adequately represent the intended construct of interest (e.g., general intelligence, or *g*). In other words, does the nonverbal measure of intelligence adequately capture the range of cognitive processes thought to compose "intelligence?" Messick (1995) discusses two factors that invalidate assessment results: (1) construct underrepresentation, and (2) construct irrelevant variance. Construct underrepresentation occurs when the test too narrowly samples the construct of interest. This is important in the discussion of nonverbal assessment of intelligence because reduction or omission of language-loaded tests (i.e., verbal tests) may reduce the construct of interest (i.e., general intellectual ability, or *g*). A plethora of research (e.g., Carroll, 2005; Horn & Blankson, 2005; McGrew, 2005) suggests that cognitive abilities can be organized into a hierarchical three-tiered model, with *g* or general intelligence on top, second-order factors in the middle (e.g., crystallized ability, fluid abilities, visualization, long-term retrieval) and a large number of very specific abilities on the bottom (e.g., vocabulary knowledge, the ability to process and discriminate speech sounds, speed of eye movements). This model is referred to as the Cattell-Horn-Carroll (CHC) theory of cognitive abilities. Two of the second-order factors, crystallized ability and literacy or reading/writing ability, are strongly related to language. Most nonverbal measures of intelligence exclude these second-order factors, therefore reducing the representation of abilities thought to be important in the estimation of *g*. In other words, many nonverbal tests of intelligence exclude important domains of functioning thought to be fundamental to the construct of interest (i.e., general intelligence) (Braden & Athanasiou, 2005; Ortiz & Dynda, 2005). Using only the performance subtests of a test battery to assess the intelligence of D/HOH individuals may lead to construct underrepresentation. Therefore, the concept of construct underrepresentation is cited as support for using verbal intelligence measures with D/HOH children (Braden & Hannah, 1998; Maller, 1996).

It should also be noted, however, that the new Standford-Binet Intelligence Scales, Fifth Edition (Roid, 2003) includes both verbal (i.e., language loaded) and nonverbal (i.e., language reduced) measures of five CHC abilities. These abilities include, fluid reasoning, knowledge, quantitative reasoning, visual processing, and working memory (Roid, 2003). Therefore, there may be opportunities to broaden construct representation without confounding language use. However, there has been abundant research showing that *g* is fairly accurately and easily estimated from even small samples of tests. In other words, *g* is found to be relatively robust, and does not appear to be highly susceptible to underrepresentation. Although g is found to be relatively robust, CHC second-order factors would be much more vulnerable to underrepresentation, especially if they were systematically omitted from a test battery due to language loading (Jensen, 1998).

According to Messick (1995), the second factor that can invalidate assessment results is construct-irrelevant variance. Construct-irrelevant variance occurs when the test includes factors extraneous to the construct of interest. These irrelevant factors either serve to increase or decrease the difficulty of the test for a person or a particular group of people. As previously discussed, it has been found that verbal items function differently for hearing and D/HOH individuals. Verbal test items are more difficult for D/HOH examinees when compared to normal hearing examinees. This is most likely due to the fact that many D/HOH individuals are denied the opportunity to acquire verbal and social knowledge as a direct consequence of their hearing loss. Therefore, in this case, low verbal IQs are likely the result of depressed language abilities, not depressed intellectual abilities (Braden & Hannah, 1998; Maller 1996). Therefore, use of verbal batteries with D/HOH children could elicit construct-irrelevant variance, and invalidate the assessment outcome.

The Wechsler Intelligence Scales

The majority of practitioners agree that the Wechsler Performance Scales (PS) are appropriate for measuring the cognitive functioning of D/HOH children (Braden, 2005; Braden & Hannah, 1998; Sullivan & Vernon, 1979). The Wechsler PS continue to be the most widely used assessment of intellectual functioning for D/HOH children (Braden, 2005; Braden & Hannah, 1998). Throughout its history, the Wechsler PS have played a critical role in the ability to rule out mental retardation as a cause of the linguistic, social or academic delays often found among D/HOH children (Maller & Braden, 1993).

In this discussion, the Wechsler PS refer to the Performance subtests of the Wechsler Scales for adults, children, and preschoolers. Wechsler developed three separate intelligence scales for examinees of different ages. The Wechsler Adult Intelligence Scale is designed to assess the cognitive abilities of individuals 16 years and older. The Wechsler Intelligence Scale for Children is designed to assess the cognitive abilities of children between the ages of 6 years, 0 months to 16 years, 11 months. And, the Wechsler Preschool and Primary Scale of Intelligence is appropriate for assessing the cognitive abilities of children aged 2 years, 6 months through 7 years, 3 months.

The original Wechsler intelligence scale, the Wechsler-Bellevue Intelligence Scale, was published in 1939. The Wechsler-Bellevue Intelligence Scale consisted of both verbal and performance scales and produced an overall composite score. This was the intelligence battery from which all other Wechsler intelligence scales were derived. All Wechsler Intelligence scales (with the exception of the WISC-IV) include a Performance Scale and a Verbal Scale. The scores derived from these tests include a Performance IQ (PIQ) score, a

Verbal IQ (VIQ) score, and an overall Full Scale IQ (FSIQ) score. In 1955, the Wechsler Adult Intelligence Scale (WAIS) was published and 26 years later this scale was revised (i.e., Wechsler Adult Intelligence Scale—Revised, WAIS-R). In 1997, the third edition of the Wechsler Adult Intelligence Scale (WAIS-III) was published and continues to be the most current revision of this scale (Wechsler, 1997).

In 1949, Wechsler published his first intelligence scale for children, the Wechsler Intelligence Scale for Children (WISC). The Wechsler Intelligence Scale for Children-Revised (WISC-R) was published in 1974, and the Wechsler Intelligence Scale for Children-Third Edition (WISC-III) was published 19 years later (Wechsler, 1993). The most recent edition of this scale, the WISC-IV, was published in 2003.

The Wechsler Preschool and Primary Scale of Intelligence (WPPSI) was published in 1967 and revised in 1989 (Weschsler Preschool and Primary Scale of Intelligence—Revised, WPPSI-R). The latest edition of this scale, Wechsler Preschool and Primary Scale of Intelligence—Third Edition (WPPSI-III) was published in 2002.

Throughout the course of these revisions, research examining the utility of these scales for the intellectual assessment of D/HOH individuals has been conducted (with the exception of the WIPPSI, WIPPSI-R, and WIPPSI-III). In fact, the Wechsler PS have generated more studies on D/HOH children than any other measure of intelligence (Braden, 1994) and the WISC-III manual, for the first time, included a clinical study evaluating the construct validity of the test with D/HOH students (see Wechsler, 1991, p. 216). The evidence of the reliability and validity of the Wechsler Intelligence Scales with D/HOH individuals is discussed below.

The Reliability of the Wechsler Scales with D/HOH Examinees

The reliability of the Wechsler Scales for use with D/HOH examinees has not been as widely studied as some of the other psychometric properties of the tests (e.g., validity). Only one study (i.e., Evans, 1966) has assessed the test-retest reliability or stability of the Wechsler PS with this special population. Upon first testing of the WISC Performance Scale, the mean score was 98 and for the second testing, the mean score was 97. These findings supported the reliability of the WISC Performance Scale with deaf children

Research assessing the internal consistency of the Wechsler Scales for D/HOH children and adults is also lacking (Braden & Hannah, 1998). Only three studies have assessed the internal consistency of the Wechsler Scales with samples of D/HOH children and adults. The internal consistency of the WISC Performance Scale was assessed using the split-half method and the Spearman-Brown correction for whole test reliability. The reliability coefficients were calculated across four age groups (i.e., 5-6 years (r = .94); 7-8 years (r = .95); 9 –10 years (r = .91); and 11 – 12 years (r = .88)) and all correlations were found significant ($p < .01$). These results compared favorably with the estimates in Wechsler's (1949) original standardization sample with hearing children (Evans, 1980).

Internal consistency reliability was assessed for four of the WISC-R Performance Scale subtests (i.e., Object Assembly, Block Design, Picture Arrangement, and Picture Completion) using Cronbach's alpha and KR-20 coefficient (when responses were dichotomously scored). The reliability coefficients for Object Assembly (r = .62), Block Design (r = .80) and Picture Completion (r = .76) did not differ significantly from those reported for the normative sample (r = .68, .86, .77, respectively). However, the reliability coefficient reported for Picture Arrangement for this sample of deaf children was notably higher (*r* = .84) than that of the standardization sample (r = .73) (Hirshoren, Hurley and Kavale, 1979).

The internal consistency of the WAIS-III was assessed for groups of deaf and hearing adults. The WAIS-III was translated into ASL for the deaf sample. Cronbach's alpha was calculated for 11 subtests (i.e., Vocabulary, Similarities, Arithmetic, Digit Span, Information, Comprehension, Picture Completion, Block Design, Matrix Reasoning, Picture Arrangement, and Object Assembly), the PIQ, VIQ, and the FSIQ for both groups. Cronbach's alpha coefficients did not differ significantly between the hearing and deaf groups on all 11 subtests, the PIQ, and the FSIQ. However, the reliabilities for the VIQ were significantly different between the two groups. The deaf sample's VIQ reliability coefficient was significantly higher (*r* = .94) than that of the hearing sample (*r* = .83; *p* <.05) (Kostrubala, 1998). (A comprehensive review of these studies is included in Appendix A).

Although limited, the reliability evidence for the previous versions of the Wechsler PS has generally supported the use of these scales with D/HOH examinees. However, given the substantial changes between the WISC-IV and previous versions of the Wechsler (see next section), research addressing the reliability of the WISC-IV with D/HOH children would help examiners understand how this new scale will function with D/HOH examinees.

The Validity of the Wechsler Scales with D/HOH Examinees

Throughout the history of the Wechsler Scales, numerous studies have investigated the validity for use with D/HOH examinees. The majority of these studies have assessed the evidence of validity based on relations to other variables. This includes convergent evidence, test-criterion evidence, and group comparisons of average scores. A smaller portion of studies has investigated validity evidence based on internal structure, specifically factor analytic studies, DIF, and item interrelationships. Due to the recent changes in the conceptualization of validity in the current edition of *The Standards* (AERA, APA, NCME, 1999), validity evidence based on response processes, test content, and consequences of testing have not been explored for the Wechsler Scales with D/HOH examinees.

Evidence Based on Relations to Other Variables

Convergent evidence. Five studies report convergent evidence for the validity of the Wechsler PS for use with D/HOH children. A strong positive relationship was found between the WISC Performance Scale and the nonverbal battery of the Lorge-Thorndike Intelligence Scale ($r = .77$) (Lavos, 1962). A significant concurrent relationship was established between the PIQ of the WISC-R and the Learning Quotient of the Hiskey-Nebraska Test of Learning Aptitude ($r = .89$; $p < .05$) supporting the use of the WISC-R Performance Scale with this special population (Hishoren, Hurley and Kavale, 1979). A significant relationship ($p < .05$) was documented between the PIQs of the WISC and WISC-R ($r = .85$) and between each of the corresponding subtests (Picture Arrangement $r = .62$; Block Design $r = .79$; Object Assembly $r = .73$; Coding $r = .71$) (Brooks & Riggs, 1980). Significant correlations ($p < .008$) were calculated between the Performance Scale subtests and PIQ of WISC-R and the WAIS-R (Picture Completion $r = .470$; Picture Arrangement $r = .690$; Block Design $r = .692$; Object Assembly $r = .552$; Coding/Digit Symbol $r = .692$; Performance IQ $r = .744$) (Braden & Paquin, 1985). And, a significant correlation was found between the WISC-R and the WISC-III PIQs ($r = .93$; $p < .01$) (Slate & Fawcett, 1995). All of these studies support the Wechsler PS for use with D/HOH examinees.

Test-criterion evidence. Six studies have found test-criterion relationship evidence for the validity of the Wechsler PS for use with D/HOH children. One study found nonsignificant, but relatively strong correlations between the Wechsler Scales (included both WISC and WIAS scores) and the Gray-Votaw-Rogers Test ($r = .55$) and between these Wechsler Scales and the Stanford Achievement Test (SAT; $r = .54$) (Brill, 1962). Although these results are not significant, they are comparable to those found for hearing children and, therefore, suggest that the WISC and WAIS are relatively strong predictors of achievement for deaf children. In another study, a significant correlation was documented between the WISC Performance Scale and the Certificate of Secondary Education ($r =.81$; $p < .01$) (Evans, 1980). Significant correlations ($p <.05$) were also observed between the WISC-R PIQ and five subtests of the SAT (Word Meaning $r = .34$; Paragraph Meaning $r =.31$; Vocabulary $r = .30$; Spelling $r = .29$; Arithmetic Concepts r = .24) and the SAT Average Grade score ($r = .35$) (Hirshoren, Hurley, & Kavale, 1979). Similarly, significant correlations ($p < .01$) were established between the WISC-R PIQ and five SAT-Hearing Impaired (SAT-HI) subtest scaled scores (Reading Comprehension $r = .32$; Spelling $r = .24$; Concept of Number $r = .46$; Math Calculation $r = .31$; and Math Applications $r = .41$) and between the WISC-R PIQ and five SAT-HI percentile ranks (Reading Comprehension $r = .39$; Spelling $r = .33$; Concept of Number $r = .57$; Math Calculation $r = .42$; and Math Applications $r = .52$) (Kelly & Braden, 1990). Significant, low to moderate correlations were calculated between the WISC-III PIQ and the SAT-HI subtests (Total Reading $r = .46$; Total

Language *r* = .54; Total Math *r* = .63; *p* < .01) and significant, moderate to high correlations were calculated between the WISC-III VIQ and the SAT-HI subtests (Total Reading *r* = .80; Total Language *r* = .85; Total Math *r* = .83; *p* < .01) (Maller & Braden, 1993). Likewise, significant correlations (*p*< .01) were found between the WISC-III PIQ and three subtests from the Wide Range Achievement Test—Revised (WRAT-R) (Reading *r* = .41; Spelling *r* = .48; Arithmetic *r* = .64) (Slate & Fawcett, 1995).

Two studies, however, did not find test-criterion relationship evidence to support the validity of the Wechsler PS for use with this population. The PIQs from the WISC and the WISC-R were each correlated with reading achievement as measured by the SAT-HI (*r* = .18 and .19, respectively) and found to be nonsignificant (Brooks & Riggs, 1980). Similarly, correlations were calculated between the WISC-R PIQ and seven SAT-HI subtest age-based percentiles (Word Reading *r* = .04; Reading Comprehension *r* = .33; Spelling *r* = .36; Total Reading *r* = .02; Arithmetic Concepts *r* = .20; Arithmetic Computation *r* = .05; Arithmetic Applications *r* = .10) and between the WISC-R PIQ and seven SAT-HI subtest grade equivalents (Word Reading *r* = .27; Reading Comprehension *r* = .05; Spelling *r* = .22; Total Reading *r* = .37; Arithmetic Concepts *r* = .14; Arithmetic Computation *r* = -.08; Arithmetic Applications *r* = .20) and all correlations were nonsignificant (Braden, 1989). However, it should be mentioned that failure to find significant relationships between the WISC-R PS and the SAT—HI in these two studies may have been a function of the limited sample size as well as the use of metrics of the SAT (i.e., the use of grade equivalents from SAT lead to low correlations with IQ) (Braden, Wollack, & Allen, 1995).

Group comparisons of average scores. Six studies, as well as one meta-analysis, have assessed validity evidence for the Wechsler Scales for D/HOH examinees by comparing average scores from this special population and comparing them to normally hearing examinees. One study assessed the WISC PIQ and found the mean (*M* = 98) and standard deviation (*SD* = 15.9) for a sample of deaf children did not significantly differ from those of Wechsler's original normative sample (Murphy, 1957; as cited in Evans, 1980).

Two studies assessed the mean and standard deviations of the WISC-R PIQ with deaf children. One study reported that the mean (*M* = 95.7) and standard deviation (*SD* = 17.55) were significantly different from those of the normative sample ($p < .01$). However, upon closer examination, the data revealed that deaf and hearing children performed similarly on four of the six Performance subtests (i.e., Picture Completion, Block Design, Object Assembly, and Mazes) (Sisco & Anderson, 1978). The second study found that deaf children showed greater variability (*SD* = 17.84) and obtained a lower mean score (*M* = 88.07) compared to the normative sample (Hirshoren, Hurley & Kavale, 1979).

Two studies assessed the mean scores and standard deviations of deaf children on the WISC-III. The first study found that the mean PIQ for a sample of deaf examinees was not significantly different from that of the normative sample (*M* = 105.83). However, the mean VIQ (*M* = 81.12) and FSIQ (*M* = 92.17) were significantly lower than those of the normative sample, but not statistically different from one standard deviation below the normative mean. For all three scores, PIQ, VIQ, and FSIQ, there was greater variability among the deaf examinees compared to the normative sample (*SD* = 20.75; 20.34; 19.83; respectively) (Maller & Braden, 1993). The second study found similar results, in that the mean and standard deviation of the PIQ for the deaf sample (*M* = 102.32; *SD* = 11.03) did not differ significantly from those of the hearing sample (*M* = 104.55; *SD* = 10.75) (Braden, Kostrubala, & Reed, 1994).

A study was also conducted that assessed the mean scores of the WAIS-III with a group of deaf adults (Kostrubala, 1998). The mean PIQ for the deaf sample (*M* = 103.21) was not significantly different than that of the hearing sample (*M* = 102.33). However, the mean VIQ was significantly lower for the deaf sample (*M* = 82.73) compared to the hearing sample (*M* = 101.53). Additional analyses compared means for each subtest between the deaf and hearing samples. The differences between the Performance subtest means were not significant, except that deaf adults had a higher Object Assembly subtest mean. All Verbal subtest means were significantly lower ($p < .05$) for the deaf sample compared to those of the hearing sample.

Braden (1990) conducted a meta-analysis of 21 studies that reported Wechsler Performance Scale subtest scores (including the WISC, WISC-R, WAIS, and the Wechsler-Bellevue Scales) for samples of deaf examinees. The results indicated that the mean scores for all subtests fell within the average range. However, the Coding/Digit Symbol subtest is distinctly lower than other subtests. The unweighted and weighted average scaled scores for the five Performance Scale subtests are the following: Picture Completion (9.40; 9.41), Picture Arrangement (9.36; 9.26), Block Design (9.87; 9.89), Object Assembly (9.96; 9.92), and Coding/Digit Span (8.77; 8.77). In general, the results revealed that, although deaf individuals had mean PIQs slightly below the means for normal hearing individuals, the scores were well within the average range.

Evidence Based on Internal Structure

Factor analyses. Four studies have examined the factorial similarity of the Wechsler Scales across deaf and hearing examinees. One study (Braden, 1984) investigated that factorial similarity of the WISC-R Performance Scale for deaf and hearing samples and found that one factor was extracted. The factor loadings were quite similar in samples of D/HOH and normal hearing children, indicating that the WISC-R Performance Scale measures that same underlying trait for both groups. A second study (Sullivan & Schulte, 1992) factor analyzed the WISC-R with D/HOH children. Instead of finding three factors for this sample, as is found for hearing children (i.e., Verbal Comprehension, Perceptual Organization, and Freedom from Distractibility), only two factors were extracted for this population (i.e., Language Comprehension and Visual-Spatial Organization). These factors clearly correspond to the Verbal and Performance Scales. The Freedom of Distractibility factor was not identified in the deaf sample. Similar results were found in a third study (Sullivan & Montoya, 1997) that assessed the factor structure of the WISC-III with D/HOH children. The same two factors were identified (i.e., Language Comprehension and Visual-Spatial Organization) and the Freedom of Distractibility and Processing Speed factors were not found. The final study (Slate & Fawcett, 1995) factor analyzed the Performance Scale of the WISC-III with D/HOH children. Results indicated that the WISC-III Performance Scale separates into two factors (i.e., Perceptual Organization and Processing Speed), supporting the validity of this scale for use with D/HOH children. In sum, the research suggests that the Wechsler Scales load onto two factors for the deaf population, verbal and performance. This contrasts with the three to four factors found for the hearing population. Factor analysis research supports the similarity of the Wechsler PS for the deaf and hearing populations (i.e., the same factors have been extracted for both groups).

Differential item functioning (DIF). Two studies have assessed the DIF of various subtests of the WISC-III for deaf children. Maller (1996) found that numerous items of the WISC-III Verbal subtests exhibit DIF with a sample of deaf children. She later (Maller, 1997) found similar results when looking at the WISC-III Verbal subtests with another sample of deaf children. In addition, the Picture Completion subtest of the Performance Scale also functioned differently for deaf children. These findings suggest that items in the Verbal Scale of the WISC-III, as well as the Picture Completion subtest, have somewhat different meanings for deaf and hearing children.

Item interrelationships. One study assessed test homogeneity by examining the relationships among Verbal and Performance subtests of the WAIS-III for a sample of deaf and hearing adults. The findings indicated that, for the deaf sample, the Verbal Scale subtest scores were more alike than their Performance Scale subtest scores. Verbal Scale subtest correlations for the deaf sample ranged from moderate to high (.60 - .89); whereas, the Performance Scale subtest correlations ranged from low to moderate (.01 - .72). The subtest correlations for the hearing sample differed from those of the deaf sample. The subtest correlations from both the Verbal and Performance Scales for the hearing group ranged from low to moderate (.15 - .72; .09 - .58, respectively). All subtest correlations in the Verbal Scale for the deaf sample were statistically significant ($p < .01$); however, only 8 of the 15 Verbal Scale subtest correlations were significant at this level for the hearing sample. For the Performance Scale subtest correlations, 7 out of the 21 correlations were significant ($p < .01$)

for the deaf sample and 10 out of the 21 correlations were significant ($p < .01$) for the hearing sample. Overall, these findings suggest that the Verbal Scale subtests of the WAIS-III were more homogeneous, especially for the deaf sample (Kostrubala, 1998).

Summary

Although the findings have been inconsistent, overall, the studies generally support the validity of the Wechsler PS for use with D/HOH examinees. These studies support the belief that the Wechsler VS are inappropriate to use for the cognitive assessment of D/HOH examinees. (For a more extensive review of the studies, please see Appendix B.)

The WISC-IV

Changes from the WISC-III to the WISC-IV

In 2003, the newest version of the WISC was introduced. This latest version, the Wechsler Scale of Intelligence for Children- Fourth Edition (WISC-IV) (Wechsler, 2003), included significant modifications to the content and structure of the scale. Until the publication of the WISC-IV, all Wechsler Scales shared a common core structure that produced a VIQ, a PIQ, and a FSIQ score. However, the developers of the WISC-IV abandoned the VIQ/PIQ score structure, and replaced it with a four-factor structure. This four-factor structure consists of the following four indexes: (a) the Verbal Comprehension Index (VCI), (b) the Perceptual Reasoning Index (PRI), (c) the Processing Speed Index (PSI), and (d) the Working Memory Index (WMI). This new structure has elevated the four index scores to the primary level of interpretation (Wechsler, 2003). The FSIQ score continues to be the general composite score for the entire scale (Prifitera, Weiss, Saklofske, & Rolfhus, 2005). The FSIQ was retained because is still so widely used in assessment and research

(Saklofske, Rolfhus, Prifitera, Zhu, & Weiss, 2005), and because it is the best broad indicator of *g*.

The VCI is composed of subtests intended to measure comprehension, reasoning, and conceptualization (Wechsler, 2003). The VCI composite was adapted to require less acquired knowledge and more emphasis on reasoning and comprehension than its predecessor, the VIQ. However, it must be noted that verbal reasoning always requires some degree of acquired knowledge. For example, on the Similarities subtest, the ability to state how two things are alike requires prior knowledge of each of the words. The PRI is composed of subtests that claim to measure perceptual reasoning and organization. This composite has undergone the most extensive changes from the WISC-III. Overall, this composite intends to invoke more nonverbal reasoning and less visual spatial skills than its predecessor, the PIQ (Weiss, Prifitera, & Saklofske, 2005). Three new subtests (i.e., Matrix Reasoning, Picture Concepts, and Word Reasoning) replaced three old subtests (i.e., Mazes, Picture Arrangement, and Object Assembly). The three new subtests are thought to provide a more accurate measure of fluid reasoning and better assess the ability to reason with novel (i.e., less crystallized knowledge) information, while also reducing the emphasis on speed. The WISC-IV developers recommend that the VCI and PRI be substituted for the VIQ and PIQ, respectively, in clinical interpretations and evaluations (Prifitera, Weiss, Saklofske, & Rolfhus, 2005).

The PSI is composed of subtests intended to measure the speed of graphomotor and mental processing (Wechsler, 2003). The authors of the WISC-IV claim that the subtests directly measure speed and accuracy, as well as the ability to scan and track simple visual information. In other words, performance on the PSI subtests reveals the rapidity with which an examinee can process simple or routine information without making errors (Weiss, Prifitera, Saklofske, 2005).

The WMI is composed of subtests that intend to measure concentration, attention and working memory which is thought to be more in line with contemporary theories of cognition and working memory (Wechsler, 2003). Working memory is the ability to manipulate information held temporarily in memory. It should be noted that the subtests the make up the WMI composite are intended to tap verbal working memory, and not visual-spatial working memory (Weiss, Prifitera, Saklofske, 2005).

The fundamental structure of the WISC-IV was not the only component altered in the new edition; subtests and individual items were also modified. Subtests are either identified as core or supplemental. The core subtests are administered when composite scores (i.e., Index scores) are desired. The supplemental subtests are used to gather more information about specific cognitive skills. The supplemental subtests provide additional clinical information and allow the examiner to perform additional analyses. Also, when needed and deemed clinically appropriate, supplemental subtests can be used as substitutes for core subtests within certain limits. Five new subtests were added to the WISC-IV battery: Picture Concepts, Letter-Number Sequencing, Matrix Reasoning, Cancellation (supplemental) and Word Reasoning (supplemental). The Picture Arrangement, Object Assembly, and Mazes subtests were deleted from the WISC-IV. The Information, Picture Completion and Arithmetic subtests, which were core tests in previous versions, are now supplemental subtests (Wechsler, 2003). Although ten subtests were retained from the WISC-III (i.e., Block Design, Similarities, Digit Span, Coding, Vocabulary, Comprehension, Symbol Search, Picture Completion, Information, and Arithmetic), their item content, administration, and scoring procedures were revised. New items were added to 9 of the 10 subtests (i.e., all subtests except Coding). Table 5 provides a list of all the core and supplemental subtests that comprise the WISC-IV Index scores.

Table 5

The Core and Supplemental Subtests that Comprise the WISC-IV Index Scores

VCI	PRI	PSI	WMI
Similarities	Block Design	Coding	Digit Span
Vocabulary	Picture Concepts*	Symbol Search	Letter Number Seq.*
Comprehension	Matrix Reasoning*	*Cancellation**	*Arithmetic*
Information	*Picture Completion*		
*Word Reasoning**			

Note. New subtests are marked with an asterisk; supplemental subtests are italicized.

The WISC-IV and D/HOH Children

How might changes in the content and structure of the WISC-IV affect the assessment of D/HOH children? The majority of researchers and practitioners agree that nonverbal tests of intelligence, and more specifically the PS of the previous versions of the Wechsler, are desirable measures to use when assessing the cognitive functioning of D/HOH individuals. The PIQ, found in the previous versions of the Wechsler, tapped skills such as processing speed, manual dexterity, visualization, and fluid reasoning. However, this newest version of the Wechsler eliminated the VIQ-PIQ dichotomy and, instead, recommends the use of the PRI for measuring nonverbal fluid reasoning. This composite is thought to be a purer measure of fluid reasoning because it is less influenced by visualization, manual dexterity, and processing speed. The subtests that now compose the WISC-IV PRI are quite different than those that composed the WISC-III PIQ and the WISC-III POI. Table 6 provides a comparison of tests used in the WISC-III and WISC-IV to estimate intelligence using language-reduced composites. In the WISC-IV, Picture Arrangement and Object Assembly were eliminated entirely; Picture Completion was reclassified as a supplemental subtest (to be used only in place of one of the core subtests), and Coding was reassigned to another Index score (i.e., PSI). The one core subtest the WISC-III and the WISC-IV still have in common is Block Design. However, this subtest was modified to decrease the influence of time bonuses. These changes are likely to influence the assessment and score interpretations of the intelligence of D/HOH children (Braden, 2005), particularly because prior research shows D/HOH examinees tend to perform differently (i.e., lower) on nonverbal reasoning tests that reduce or eliminate the role of manual dexterity relative to performance tests (Braden, 1994; Braden, Kostrubala, & Reed, 1994).

Table 6

The Subtests that Comprise the WISC-III Perceptual Organization Index (POI), the WISC-III Performance Scale (PS), and the WISC-IV Perceptual Reasoning Index (PRI)

WISC-III POI	WISC-III PS	WISC-IV PRI
Block Design	Block Design	Block Design
Picture Completion	Picture Completion	Picture Concepts*
Object Assembly	Object Assembly	Matrix Reasoning*
Picture Arrangement	Picture Arrangement	*Picture Completion*
	Coding	
	Symbol Search	
	Mazes	

Note. New subtests are marked with an asterisk; supplemental subtests are italicized.

Purpose of the Present Study

Main Purpose

The changes in the WISC-IV are more substantial than any previous revisions of the scale, taking the WISC-IV far beyond most standard revisions (Prifitera et al., 2005). The lack of overlap in terms of both content and structure with previous versions of the Wechsler Scales raises significant questions about the ability to generalize prior validation research to the new version for the assessment of D/HOH children (Braden, 2005). The PS of the Wechsler Intelligence Scales for Children (i.e., WISC, WISC-R, WISC-III) are the most popular and widely researched instruments for assessing intelligence in D/HOH children (Braden, 1994, 2005). Although the new WISC-IV structure is based on theory and supported by factor-analytic and clinical research (Wechsler, 2003), there are no data available in the WISC-IV materials (i.e., the administration manual, technical manual, or technical reports) providing information on reliability or validity of the WISC-IV with D/HOH examinees. The recommendations for administering the WISC-IV to D/HOH examinees and interpreting the results (see pp. 12-18) are reasonable, but are not data-based, and do not address the fundamental characteristics of the test with D/HOH examinees. “The ability to draw on prior research in guiding the use of the new version is limited, speculative, and subject to confirmation with research using the WISC-IV with deaf and hard-of-hearing children” (Braden, 2005, p. 357). Therefore, the first purpose of this study was to obtain information to help examiners understand how the WISC-IV functions with D/HOH examinees.

Secondary Objective

Although the main purpose was foremost and paramount, this study did present a secondary objective within the methodological procedure. The secondary objective dealt with involving practitioners in the validation process in an effort to help bridge the gap between research and practice.

According to the *Standards*, test validation is a shared responsibility of the test developer and the test user. Relevant standards are:

> 1.2 The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described (AERA, APA, NCME, 1999, p. 17).

> 11.2 When a test is to be used for a purpose for which little or no documentation is available, the user is responsible for obtaining evidence of the test's validity and reliability for this purpose (AERA, APA, NCME, p. 113) (also see Chapter 10 for standards related to individuals with disabilities).

Using a test in a manner that has not been empirically supported violates the *Standards*. Professional judgment is required to evaluate the extent to which existing validity evidence applies to new situations and to determine if new evidence is necessary. If practitioners desire to use a test in a way that has not been validated, they are responsible for collecting data to inform test use (AERA, APA, NCME, 1999). However, it is acknowledged that this is much easier said than done—and that alternatives to not using tests may be less attractive than using tests that are likely to be of value. The WISC-IV is currently being used to assess the cognitive abilities of D/HOH children; therefore, psychologists who use the WISC-IV with D/HOH examinees have an ethical obligation to collect data. Research is needed to explore the reliability and validity of the WISC-IV with D/HOH children. The secondary objective of the current study was to test the feasibility of a process for including practitioners in the data collection process so that they can fulfill their ethical obligations outlined by the *Standards* and supply the data needed to evaluate the reliability and validity of the WISC-IV with D/HOH examinees.

Hypotheses

The main goal of the present study was to provide data that would help to answer the following questions: How internally consistent is the WISC-IV with D/HOH examinees? How does the WISC-IV relate to other measures of intelligence and achievement used by practicing psychologists for this special group? And, how well do the subtests correlate with each other? No study could provide a definitive answer to these questions, but I tested six discrete hypotheses that had direct bearing on these questions.

Hypothesis One: Internal consistency reliability

I predicted that the internal consistency coefficients for the D/HOH sample would not differ significantly from the internal consistency reliability coefficients reported for the normative sample in the *WISC-IV Technical and Interpretive Manual* (which will be considered to reflect population values). This hypothesis was tested using a two-tailed test of equality, Feldt's method of reliability comparison between two groups (Feldt, 1969). This method was chosen because it is the only one that directly compares the reliabilities from two different populations. An alpha level of $p = .05$ for each contrast was used.

Hypothesis Two: Relationships to other variables

I predicted that the mean PRI for the D/HOH sample would not differ significantly from the mean PRI reported for the normative sample in the *WISC-IV Technical and Interpretive Manual*. This hypothesis was tested using a two-tailed *t*-test with an alpha level of $p = .05$.

Hypothesis Three: Relationships to other variables

I predicted the mean VCI for the D/HOH sample would be significantly lower than the mean VCI reported for the normative sample in the *WISC-IV Technical and Interpretive Manual*. This hypothesis was tested using a one-tailed *t*-test with an alpha level of $p = .05$

Hypothesis Four: Relationships to other variables

I predicted that the correlations between (a) WISC-IV composite scores (i.e., FSIQ, VCI, PRI, WMI, and PSI) and achievement test scores, and (b) WISC-IV composite scores and other intelligence test scores, would be positive and reliably greater than zero (i.e., $\rho > 0$), given an alpha level $p = .05$. This hypothesis was tested by calculating 95% confidence intervals. If the lower bound of the interval was greater than zero, I concluded the hypothesis was supported.

Hypothesis Five: Relationships to other variables

I predicted that the correlation found between the VCI and achievement test score(s) would be greater than the correlation between the PRI and the achievement test score(s) (i.e., $\rho_{VCIAch} > \rho_{PRIAch}$) at alpha level $p = .05$. A *t*-test was used to test for the difference between these two dependent correlations from the same sample.

Hypothesis Six: Subtest interrelationships

I predicted that the correlations among WISC-IV subtests would be positive and reliably greater than zero (i.e., $p > 0$), given an alpha level $p = .05$. This hypothesis was tested by calculating 95% confidence intervals. If the lower bound of the interval was greater than zero, I concluded the hypothesis was supported.

CHAPTER 2

Method

Participants and Examinees

This study drew a distinction between participants and examinees. The participants were psychologists who volunteered to provide data on the WISC-IV with D/HOH children. The examinees were those D/HOH children on which the WISC-IV data were collected. In other words, the participants (i.e., the psychologists) tested the examinees (i.e., the D/HOH children) with the WISC-IV and supplied these data to me (i.e., the researcher).

Recruitment

The participants were practicing school psychologists who were recruited via sign-up sheets at area conferences, postings on deafness/psychology-related Listservs, and websites (see Appendixes C, D, and E for examples). Interested psychologists provided their names and e-mail addresses (e-mail served as the primary mode of communication for the duration of the study). These school psychologists provided archival data on D/HOH examinees who were tested with the WISC-IV for purposes other than research (e.g., educational placement, clinical or vocational assessment). Participating psychologists were also offered the opportunity to purchase materials from Harcourt Assessment at 50% less than retail as an incentive for their participation. The discount applied to all testing materials the psychologists might use to provide data for this project (e.g., the WISC-IV, Wechsler Individual Achievement Test-Second Edition, WIAT-III).

Participants were directed to identify examinees for inclusion based on a set of criteria. The inclusion criteria were the following: (a) the examinee must be between 6 years, 0 months and 16 years, 11 months of age, (b) have a hearing loss (i.e., ranging from mild to profound) sufficiently significant to be identified as having a hearing disability, (c) have prelingual onset of hearing impairment (i.e., defined in this study as a hearing loss that occurs prior to the age of 5), (d) have a hearing impairment as their primary disability (if they have more than one), and (e) have been tested on the WISC-IV as part of a previous psychological evaluation (e.g., educational placement, clinical diagnosis). Participants were requested to exclude examinees not meeting all the selection criteria.

Procedure

Participants were sent a consent form (see Appendix F) via e-mail, which they printed, signed, and returned by either ground mail or facsimile. Due to the insecure nature of e-mail, this consent form also had a space designated for the participants to supply a password. This password was used to encrypt the files and aid in the maintenance of confidentiality. Once the consent form was received, the participants were sent two pre-formatted Microsoft Excel spreadsheets to complete. One spreadsheet was for the participant to enter non-identifying examinee descriptive data and WISC-IV individual item data (see Appendix G). The second spreadsheet was for the participant to enter achievement test and other intelligence test data (see Appendix H). Both spreadsheets were encrypted; thus, the participants were required to enter their password to open the files. The spreadsheets were unique to each participant, meaning that each spreadsheet contained subject identification numbers specific to the participant, and empty cells to insert data. No participant's data was shared with other participants. Once participants entered their information into the spreadsheet, they returned the encrypted file to me via e-mail. Participants were able to insert information on as many examinees as they had available to them.

Variables

Demographic Data

A variety of demographic information was collected about each examinee. Tables 7 and 8 display the demographic data gathered for the examinees.

Table 7

General Background and Hearing Loss Data Gathered for each D/HOH Child

Variables	Possible Values
Age	6 years, 0 months – 16 years, 11 months
Grade	Kindergarten – 12th grade
Gender	Male, Female
Ethnicity	White, Black, Hispanic, Asian/Pacific Islander, Native American, Other/Unknown
Additional disabilities	Any DSM-IV Diagnosis or IDEIA-04 disability
Primary Mode of Communication	ASL, Other Sign, Aural/Oral, Cued Speech, Other
Degree of Hearing Loss	Mild, Moderate, Severe, Profound, Unknown
Pure Tone Average	
Better ear with aid	All possible PTA ranges
Better ear without aid	All possible PTA ranges
Type of Hearing Loss	Sensorineural, Conductive, Mixed, Unknown
Age of Onset	Congenital, Prelingual, Postlingual, Unknown

Table 8

Setting and Test Administration Data Gathered for each D/HOH Child

Variables	Possible Values
Setting	Residential School for the deaf, Commuter School for the deaf, Program within Public School, Public School with Supplemental Services, Other
Reason for Referral	Initial Evaluation, Triennial Reevaluation, Change of Service/Diagnosis, Social Service Eligibility, Other
Communication Modality used in Testing	American Sign Language, Other Sign, Aural/Oral, Cued Speech, Other
Date of Testing	All possible dates

WISC-IV Data

WISC-IV individual item data were also gathered for each examinee. This included individual item data and scaled scores from each of the 15 subtests as well as standard scores for each of the five composites. Table 9 displays the WISC-IV data gathered for each examinee.

Table 9

WISC-IV Data Gathered for each D/HOH Child

Individual Item Data and Scaled Scores	Standard Scores
Block Design	Verbal Comprehension Index
Similarities	Perceptual Reasoning Index
Digit Span	Working Memory Index
Picture Concepts	Processing Speed Index
Coding	Full Scale IQ
Vocabulary	
Letter-Number Sequencing	
Matrix Reasoning	
Comprehension	
Symbol Search	
Picture Completion	
Cancellation	
Information	
Arithmetic	
Word Reasoning	

Note. Supplemental Subtests are italicized.

Achievement and Other Intelligence Test Data

Each participant also entered available achievement data under one or more of the following categories: (a) Reading, (b) Arithmetic, (c) Written Language, and/or (d) Other.

The participant was asked to supply the name of the test, the name of the subtest or cluster, the standard score (based on normal-hearing norms), and the date the test was administered. In addition, each participant entered other intelligence test data (if available) for the examinee. Table 10 provides the categories under which participants could enter other examinee test data.

Table 10

Additional Achievement and Intelligence Test Data Gathered

Standard Scores for Achievement Tests	Standard Scores for Intelligence Tests
Reading	Other Intelligence Tests
Arithmetic	
Written Language	
Other	

Predictions and Analyses

Sample Size

I aspired to have a sample size of at least 100 examinees. However, I decided not run analyses with a sample size less than 28 examinees for the following reason. I conducted a one-tailed power analysis for the first hypothesis using the mean of the average internal consistency coefficients reported in the WISC-IV manual for the normative sample (i.e., r = .88) as the population value and r = .70 as cutoff for the lowest acceptable value for internal reliability. Using a power of .80 and the alpha level of .05, my calculations revealed a needed sample size of 28. Power analyses were conducted for all other hypothesis and the calculations revealed a required sample size lower than 28. Therefore, a sample size of 28 became the minimum sample size on which I would run analyses.

Due to the methodological procedures of this study, the sample of examinees was considered a “convenience sample.” This study used sampling methods that drew upon existing frameworks (e.g., Listservs) rather than systematic, random sampling procedures. Therefore, the sample is probably a nonrandom group.

Missing Data

The participants were asked to leave all unadministered WISC-IV items blank. Therefore, when analyzing the individual item WISC-IV data, it was assumed that all items below the basal were “passed” and all items beyond the ceiling were “failed.” The data were recoded to reflect this assumption. For subtests in which no items were administered, the data were coded as missing and excluded from subsequent analyses.

Combining Data

Because I asked participants to enter data from multiple achievement tests and other intelligence tests, I planned to combine scores from different tests if they measured the same construct. For example, I planned to combine measures of reading comprehension (e.g., from the WJ-III and WIAT-II), but did not plan to combine measures of reading comprehension and word reading. To ensure adequate power for testing obtained correlations, I decided to conduct correlation analyses if there were at least 28 scores representing a specific construct (e.g., reading comprehension, nonverbal intelligence).

However, even combining data left me with less than 28 individuals with achievement/intelligence data in any given domain. Therefore, I will not report correlations between the WISC-IV and achievement tests and between the WISC-IV and other intelligence tests in this thesis (but see Appendices K - N for a description of obtained achievement test and other intelligence test data).

CHAPTER 3

Analyses and Results

Participants and Examinees

Ten participants from nine states across the nation (i.e., California, Arizona, Texas, Illinois, Pennsylvania, Maryland, Massachusetts, New York, and Virginia) provided data on 128 D/HOH children (i.e., examinees). All examinees met the inclusion criteria (see pages 53-54 for a review of inclusion criteria). The examinees' ages ranged from 6.75 years to 16.83 years (M = 11.84, SD = 2.70). The median age was 11.5 years. All examines were tested between October 23, 2003 and April 4, 2007. Demographic data were aggregated across age groups. Frequencies for gender, ethnicity, grade, current educational placement, reasons for referral/testing, primary mode of communication, testing mode of communication, degree of hearing impairment, type of hearing loss, age of onset, and comorbid diagnoses are reported in Table 11. Frequencies, means and standard deviations for age of onset (in years), degree of hearing loss (in dB), and PTA (in dB) with and without hearing aids for the better ear are reported in Table 12.

Table 11

Demographic Information

Demographic Categories	Sample Size (*n*)	Percentage (%)
Total *N*	128	100.0%
Gender		
Male	69	53.9%
Female	55	43.0%
Missing	4	3.1%
Ethnicity		
White	51	39.8%
Black	9	7.0%
Hispanic	45	35.2%
Asian/Pacific Islander	8	6.3%
Native American	3	2.3%
Multiracial	1	0.8%
Other	5	3.9%
Unknown	2	1.6%
Missing	4	3.1%

Table 11 (continued)

Demographic Information

Demographic Categories	Sample Size (*n*)	Percentage (%)
Grade		
First	7	5.5%
Second	9	7.0%
Third	13	10.2%
Fourth	17	13.3%
Fifth	14	10.9%
Sixth	12	9.4%
Seventh	16	12.5%
Eighth	12	9.4%
Ninth	14	10.9%
Tenth	5	3.9%
Eleventh	2	1.6%
Missing	7	5.5%
Current Educational Placement		
Residential School for the deaf	62	48.4%
Commuter School for the deaf	19	14.8%
Program w/in Public School	15	11.7%
Public School w/ supplemental services	19	14.8%
Private Catholic School	2	1.6%

Table 11 (continued)

Demographic Information

Demographic Categories	Sample Size (*n*)	Percentage (%)
Current Educational Placement (continued)		
Regular Education	3	2.3%
Other	4	3.1%
Missing	4	3.1%
Reason for Referral/Testing		
Initial Evaluation for placement	5	3.9%
Triennial Reevaluation	58	45.3%
Change of service/diagnosis	12	9.4%
Admission Review	11	8.6%
Other	42	32.8%
Primary Mode of Communication		
American Sign Language (ASL)	83	64.8%
Other Sign	5	3.9%
Oral/Aural	22	17.2%
ASL & Oral/Aural	6	4.7%
Total Communication	7	5.5%
ASL & Sign Supported English	1	0.8%
Missing	4	3.1%

Table 11 (continued)

Demographic Information

Demographic Categories	Sample Size (*n*)	Percentage (%)
Testing Mode of Communication		
American Sign Language (ASL)	58	45.3%
Oral/Aural	18	14.1%
ASL & Oral/Aural	7	5.5%
Total Communication	16	12.5%
Signed Spoken English	1	0.8%
Missing	28	21.9%
Degree of Hearing Impairment		
Mild (20-40 dB)	4	3.1%
Moderate (4-70 dB)	14	10.9%
Severe (71-95 dB)	25	19.5%
Profound (> 95 dB)	80	62.5%
Missing	5	3.9%
Type of Hearing Loss		
Conductive	1	0.8%
Sensorineural	113	88.3%
Mixed	3	2.3%
Unknown	7	5.5%
Missing	4	3.1%

Table 11 (continued)

Demographic Information

Demographic Categories	Sample Size (*n*)	Percentage (%)
Age at Onset		
Congenital (at birth)	65	50.8%
Prelingual (before 5 years)	36	28.1%
Unknown	22	17.2%
Missing	5	3.9%
Comorbid Diagnoses		
None	32	25.0%
Asberger's Syndrome	1	0.8%
Pervasive Developmental Disorder	1	0.8%
Cerebral Palsy	1	0.8%
Depression	1	0.8%
Emotional/Behavioral Disorder	5	3.9%
Other Health Impaired (OHI)	1	0.8%
Learning Disability (LD)-Reading	3	2.3%
LD-Reading & Math	1	0.8%
LD-Language	3	2.3%
LD	2	1.6%
Attention Deficit/Hyperactivity Disorder (ADHD)	9	7.0%

Table 11 (continued)

Demographic Information

Demographic Categories	Sample Size (*n*)	Percentage (%)
Comorbid Diagnoses (continued)		
ADHD-Combined Type	2	1.6%
ADHD- Inattentive Type	1	0.8%
ADHD- Not Otherwise Specified	1	0.8%
ADHD & LD	1	0.8%
ADHD & LD-Nonverbal	1	0.8%
ADHD-Combined Type & Expressive Language Disorder	1	0.8%
ADHD-Combined Type & Adjustment Disorder	1	0.8%
ADHD-Combined Type & Spina Bifida Oculta	1	0.8%
ADHD & Post-Traumatic Stress Disorder & Disruptive Behavior Disorder & Speech Impairment & Gender Identity Disorder	1	0.8%

Table 11 (continued)

Demographic Information

Demographic Categories	Sample Size (*n*)	Percentage (%)
Comorbid Diagnoses (continued)		
ADHD-Combined Type & LD-Reading, Math, Written Language & Adjustment Disorder without anxiety	1	0.8%
Missing	57	44.5%

Note. Percentages may not equal 100% due to rounding.

Table 12

Demographic Information

Demographic Categories	*n*	*M*	*SD*
Age of Onset (in years)	80	1.10	2.62
Degree of Hearing Loss (in dB)	72	90.61	24.65
PTA of better ear without hearing aids (in dB)	81	85.99	26.71
PTA of better ear with hearing aids (in dB)	27	28.30	23.44

Note. PTA = Pure Tone Average; dB = decibels.

Data Checking

When the data arrived from the participants, I checked the data to identify anomalies. The individual item scores were summed and compared against the total raw score reported for each subtest. If there was a discrepancy between the individual item tallies and the total raw score for the subtest, the participant who supplied the data was contacted, via email, and asked to double-check the entry. Any corrections made by the participant were corrected in the data spreadsheet. However, if the participant did not respond to the email and correct the entry error, the individual item data for that particular subtest (for that particular examinee) were eliminated. Table 13 (next page) provides the number of individual item data cases excluded due to entry errors, as well as the number of individual item data cases retained per subtest.

Although some cases were removed from the individual item data, the subtest scaled scores and composite standard scores for that particular examinee were retained as reported. For example, the Block Design individual item data may not have matched the reported subtest raw score (and therefore eliminated) for one examinee; however, the examinee's reported Block Design scaled score, PRI standard score, and FSIQ standard score (to which the Block Design score contributed) were retained. As a result, the sample sizes used to calculate internal consistency reliabilities for subtests and composites (which rely on individual item data) differ from the sample sizes used to calculate means and standard deviations for subtests and composites (which rely on subtest scaled scores and composite standard scores). (For a comparison of these sample sizes, please see Tables 13 and 14).

Table 13

The Number of Individual Item Data Eliminated and Retained for Each Subtest

Subtest	*n* Eliminated	*n* Retained
Core Subtests		
Block Design	13	106
Similarities	5	59
Digit Span	3	33
Picture Concepts	5	115
Coding	17	102
Vocabulary	7	50
Letter-Number Sequencing	1	26
Matrix Reasoning	12	108
Comprehension	1	55
Symbol Search	36	81
Supplemental Subtests		
Picture Completion	2	46
Cancellation	3	33
Information	0	19

Table 13 (continued)

The Number of Individual Item Data Eliminated and Retained for Each Subtest

Subtest	*n* Eliminated	*n* Retained
Supplemental Subtests (continued)		
Arithmetic	1	11
Word Reasoning	2	5
Total	**108**	**849**

Sample Size Restrictions

As previously stated, I decided not to conduct analyses with a sample size smaller than 28 (as determined by power analysis calculations). As a result of sample size restrictions, analyses were not conducted with the following WISC-IV subtests and composite scores: Letter-Number Sequencing (*n* = 27), Information (*n* = 17), Arithmetic (*n* = 10), Word Reasoning (*n* = 5), and FSIQ score (*n* = 18). Therefore, some results in the following sections are not reported (NR).

Descriptive Analyses

The mean and standard deviation of each of the WISC-IV Subtests and the Composite Scales are reported in Table 14.

Table 14

The Mean and Standard Deviation of the WISC-IV Subtests and Composite Scales

WISC-IV Subtests and Composite Scales	*n*	*M*	*SD*
Core Subtests			
Block Design	119	8.96	2.66
Similarities	64	7.75	3.56
Digit Span	36	6.83	3.03
Picture Concepts	120	8.68	3.29
Coding	119	7.92	2.89
Vocabulary	57	5.16	3.52
Letter-Number Sequencing	27	NR	NR
Matrix Reasoning	120	9.00	3.30
Comprehension	56	7.80	4.51
Symbol Search	117	8.77	3.16
Supplemental Subtests			
Picture Completion	48	8.50	3.52
Cancellation	36	8.14	3.50
Information	17	NR	NR
Arithmetic	10	NR	NR
Word Reasoning	5	NR	NR

Table 14 (continued)

The Mean and Standard Deviation of the WISC-IV Subtests and Composite Scales

WISC-IV Subtests and Composite Scales	*n*	*M*	*SD*
Composite Scales			
Verbal Comprehension Index	57	80.86	19.16
Perceptual Reasoning Index	119	93.21	15.98
Working Memory Index	29	85.83	14.08
Processing Speed Index	116	91.31	15.36
Full Scale IQ	18	NR	NR

Note. The Core and Supplemental subtests have a mean of 10 and a standard deviation of 3; The Composite scales have a mean of 100 and a standard deviation of 15; NR denotes "Not Reported" due to sample size restrictions.

Restriction of Range

The correlation coefficient will be spuriously low if it is based on a sample of homogeneous subjects for whom there is a restriction of range on the characteristic being measured. To ensure that subtest intercorrelation coefficients are not unduly influenced by an unusual range of scores, I conducted a one-tailed, one sample *F* test for equal variances assuming a population variance of 9 (i.e., $SD = 3$, $SD^2 = 9$) for WISC-IV subtest scaled scores and a population variance of 225 (i.e., $SD = 15$, $SD^2 = 225$) for WISC-IV composite standard scores ($p < .05$). The *F* test for equal variances was only calculated for those subtests/composite scores for which the variance was less than the population variance (i.e., cases in which restriction of range could be a problem).

I calculated an *F* test of equal variance for the Block Design subtest ($SD = 2.67$), Coding subtest ($SD = 2.89$) and WMI composite ($SD = 14.08$). The *F*-ratio was significant for the Block Design subtest, indicating a restriction of range, $F(2199, 118) = 1.27$, $p < .05$. The subtest intercorrelations including the Block Design subtest were corrected for the restriction of range (see Table 16). *F*-ratios for the Coding subtest, $F(2199, 118) = 1.08$, $p > .05$, and the WMI, $F(2199, 28) = 1.13$, $p > .05$, were not significant, indicating restriction of range is not a problem for this subtest and composite.

Originally I proposed that I would use the Pearson-Lawley correction formula for restriction of range; however, after investigating this further, I found that this specific formula is only appropriate to use with multivariate correlations. In its place, I used the more common restriction of range correction formula for bivariate correlations (Guilford, 1965). For a description of this formula please see Appendix O.

Reliability

Hypothesis One: Internal Consistency Reliability

I predicted the internal consistency coefficients for the D/HOH sample would not differ significantly from the internal consistency reliability coefficients reported for the normative sample in the *WISC-IV Technical and Interpretive Manual* (Wechsler, 2003), which I considered to reflect population values. I estimated internal consistency reliability for the WISC-IV subtests and composites using the split-half method (i.e., splitting the test by odd and even items) and applying the Spearman-Brown correction for whole (equal length) test reliability to replicate the same procedures used in Wechsler (2003). The internal consistency reliabilities for the composites (i.e., VCI and PRI) were calculated using only the core subtests, as was done for the normative sample (J. Zhu, personal communication, April 7, 2008). Internal consistency coefficients were not calculated for Coding, Symbol Search, and Cancellation subtests, nor for the PSI, because the split-half reliability method is not a proper estimate of reliability for tasks in which speed is a key factor. Because Wechsler (2003) used test-retest stability data to estimate reliabilities for these scores, and I did not have test-retest data, I did not calculate reliability coefficients and could not compare sample coefficients to population values. Also, because the sample size was below 28, I did not calculate nor report internal consistency coefficients for four subtests (i.e., Letter-Number Sequencing, Information, Arithmetic, and Word Reasoning), or the WMI and FSIQ indexes.

Reliability coefficients for the sample were compared to those from the normative sample using Feldt's (1969) method of reliability comparison between two groups. Table 15 reports the reliability coefficients for the WISC-IV normative sample, reliability coefficients for the D/HOH sample, and the results of Feldt's test of significance.

Table 15

Reliability Coefficients for the Normative and D/HOH Samples and Feldt's Test of Significance

Subtests/Indexes (*n*)	Normative Sample Average r_{xx}*	D/HOH Sample r_{xx}	Feldt's *f(W)*
Subtests			
Block Design (106)	.86	.89	1.27
Similarities (59)	.86	.96	3.50****
Digit Span (33)	.87	.92	1.63**
Picture Concepts (115)	.82	.88	1.50***
Vocabulary (50)	.89	.93	1.57**
Matrix Reasoning (108)	.89	.93	1.57***
Comprehension (55)	.81	.91	2.11****
Picture Completion (46)	.84	.88	1.33
Indexes			
Verbal Comprehension (37)	.94	.97	2.00***
Perceptual Reasoning (90)	.92	.94	1.33**

Note. * Average reliability coefficients were calculated with Fisher's z transformation
** $p < .05$; *** $p < .01$; **** $p < .001$

The hypothesis that the D/HOH sample would produce reliabilities similar to those of the normative sample was not supported. With the exception of the internal consistency reliability coefficients for the Block Design and Picture Completion subtests, all other internal consistency correlation coefficients were significantly higher than those reported in the *WISC-IV Technical and Interpretive Manual* for the normative sample.

Validity

Hypothesis Two: Relationships to Other Variables

I predicted the mean PRI for the D/HOH sample would not differ from the mean PRI reported for the normative sample in the *WISC-IV Technical and Interpretive Manual*. A two-tailed, one-sample *t*-test with an alpha level of $p = .05$ was used to test the difference between the sample and normative (i.e., population) means.

This hypothesis was not supported. The mean PRI for the D/HOH sample (M = 93.21, SD = 15.98) was lower than the population mean (M = 100, SD = 15; $t\ (118) = -4.64$, $p < .001$).

Hypothesis Three: Relationships to Other Variables

I predicted that the mean VCI for the D/HOH sample would be lower than the mean VCI reported for the normative sample in the *WISC-IV Technical and Interpretive Manual*. A one-tailed, one-sample *t*-test with an alpha level of $p = .05$ was used to test the difference between the sample and normative (i.e., population) means.

This hypothesis was supported. The mean VCI for the D/HOH sample (M = 80.86, SD = 19.19) was lower than the population mean (M = 100, SD = 15; $t\ (56) = -7.53$, $p < .001$).

Hypothesis Four: Relationships to Other Variables

I predicted that the correlations found between WISC-IV scores and reported achievement test score(s) would be positive and significantly greater than zero ($\rho > 0$) at alpha $p = .05$. However, due to sample size restrictions, I did not test this hypothesis.

Hypothesis Five: Relationships to Other Variables

I predicted that the correlation between the VCI and the achievement test score(s) would be greater than the correlation between the PRI and the achievement test score(s). However, due to sample size restrictions, I did not test this hypothesis.

Hypothesis Six: Subtest Interrelationships

I predicted that the correlations among WISC-IV subtests would be positive and reliably greater than zero ($\rho > 0$) at alpha $p = .05$. Pearson product-moment correlations were calculated between the WISC-IV subtest scaled scores. Next, I calculated a 95% confidence interval for each correlation. If the lower bounds of the confidence intervals were greater than zero, I would conclude that my prediction was supported. However, if the confidence intervals included zero, the hypothesis would not be supported.

Forty-four Pearson product-moment correlations were calculated among the WISC-IV subtests (i.e., subtest correlations with $n > 28$). Of the 44 sets of correlation confidence intervals, 15 contain zero. Therefore, this hypothesis is only partially supported. See Table 16 for the correlation matrix.

Table 16

Correlations (with 95% Confidence Intervals reported) among WISC-IV Subtests

Subtests		**PCn**	**MR**	***PCm***	**SIM**	**VC**	**CO**	**DS**	**CD**	***SS***	***CA***
BD	***r***	**.48**	**.55**	**.55**	**.42**	.07	**.28**	**.52**	**.23**	**.41**	.18
	r	.44	.51	.51	.38	.06	.26	.47	.21	.37	.16
	95% C. I.	.33-.61	.42-.67	.32-.72	.19-.61	-.20-.32	.01-.51	.22-.72	.06-.40	.24-.55	-.17-.49
	N	119	119	48	61	56	53	35	117	115	34
PCn	***r***	--	**.61**	**.58**	**.53**	**.45**	**.36**	**.42**	**.19**	**.27**	.03
	95% C. I.	--	.48-.71	.35-.74	.32-.69	.22-.63	.10-.57	.11-.66	.01-.36	.09-.43	-.31-.36
	N	--	120	48	62	57	54	36	117	115	35
MR	***r***		**--**	**.55**	**.45**	**.37**	.27	**.49**	**.38**	**.44**	.29
	95% C.I.		**--**	.32-.72	.23-.63	.12-.58	.00-.50	.19-.71	.21-.53	.28-.58	-.05-.57
	N		**--**	48	62	57	54	36	117	115	35

Table 16 (continued)

Correlations (with 95% Confidence Intervals reported) among WISC-IV Subtests

Subtests		**SIM**	**VC**	**CO**	**DS**	**CD**	***SS***	***CA***
PCm	***r***	NR	NR	NR	NR	**.31**	**.47**	.07
	95% C.I.	NR	NR	NR	NR	.03-.55	.21-.67	-.30-.42
	N	NR	NR	NR	NR	47	47	30
SIM	***r***	--	**.66**	**.70**	NR	.04	.18	NR
	95% C.I.	--	.48-.79	.54-.81	NR	-.22-.29	-.08-.42	NR
	N	--	56	56	NR	60	59	NR
VC	***r***		--	**.65**	NR	-.09	.18	NR
	95% C.I.		--	.46-.78	NR	-.35-.18	-.09-.43	NR
	N		--	52	NR	56	55	NR
CO	***r***			--	NR	.12	.26	NR
	95% C.I.			--	NR	-.16-.38	-.01-50	NR
	N			--	NR	53	52	NR

Table 16 (continued)

Correlations (with 95% Confidence Intervals reported) among WISC-IV Subtests

Subtests		**CD**	***SS***	***CA***
DS	***r***	.20	.21	NR
	95% C.I.	-.15-.50	-.14-.52	NR
	N	34	33	NR
CD	***r***	--	**.58**	.34
	95% C.I.	--	.45-.69	.00-.61
	N	--	117	34
SS	***r***		--	**.37**
	95% C.I.		--	.04-.63
	N		--	34

Note. C.I. = Confidence Interval; BD = Block Design; PCn = Picture Concepts; MR = Matrix Reasoning; PCm = Picture Completion; SIM = Similarities; VC = Vocabulary; CO = Comprehension; DS = Digit Span; CD = Coding; SS = Symbol Search; CA = Cancellation.
Supplemental subtests are italicized.
* indicates correlations corrected for restriction of range.
Significant correlations are in bold font.
NR denotes “Not Reported” due to sample size restrictions.

Additional analyses

I calculated Pearson product moment correlations between degree of hearing loss (in dB) and WISC-IV Index scores with $n > 28$. Correlations were conducted between degree of hearing loss (in dB) and PRI ($n = 59$, $r = -.012$, $p > .05$) and between degree of hearing loss (in dB) and PSI ($n = 58$, $r = .041$, $p > .05$). The results failed to reveal significant relationships between degree of hearing loss and these index scores for this sample.

I also ran a one-tailed, one sample *t*-test to test whether the VCI mean ($M = 80.86$) for this sample was significantly different from one standard deviation below the population mean ($M = 85$). The t-test results revealed a non-significant difference, $t\ (56) = -1.63$, $p > .05$.

Chapter 4

Discussion

The primary purpose of this study was to assess the reliability and validity of the WISC-IV with D/HOH children. In 2003, the WISC-IV replaced the WISC-III as the preferred cognitive assessment to use with D/HOH children. However, no data relative to the WISC-IV with this special population were reported in the *WISC-IV Technical and Interpretive Manual* (Wechsler, 2003). The changes made in this fourth edition (in terms of the content and structure) are more substantial than any previous revision of this scale. Therefore, the ability to generalize prior validation research to the new version for the assessment of D/HOH children is limited (Braden, 2005). Because psychologists are using the WISC-IV to make educational decisions for D/HOH children, psychologists need data to understand how this scale functions for this specific and unique population of children.

A secondary objective of this study was to test the feasibility of a method for including practitioners in the data collection process. According to the *Standards*, test validation is the shared responsibility of the test developer and the test user (AERA, APA, NCME, 1999, standard 11.2). Therefore, psychologists who use the WISC-IV with D/HOH children have an ethical obligation to collect data to understand the reliability and validity of this scale for this special population. As a result, the data collection method for this study relied on psychologists working with D/HOH children. I piloted an electronic data collection process (i.e., encrypted electronic data sets sent via e-mail) to determine whether this process could elicit WISC-IV data for D/HOH children from psychologists across the nation.

The Reliability of the WISC-IV with D/HOH Children

Internal Consistency Reliability

The results revealed that the WISC-IV is likely to be more reliable with D/HOH children compared to the children in the standardization sample. With the exception of the Block Design and Picture Completion subtests, all other subtests and indexes calculated were significantly more reliable than the split-half reliability coefficients reported in the *WISC-IV Technical and Interpretive Manual* (Wechsler, 2003) for the normative sample.

The internal consistency coefficients reported for the normative sample were calculated using the basal and ceiling subtest rules established in the *WISC-IV Technical and Interpretive Manual* (J. Zhu, personal communication, March 31, 2008). Therefore, this study utilized the same internal consistency calculation procedures executed during the standardization of this scale. In other words, for both this sample and the normative sample, all items below the basal were coded as correct and all items above the ceiling were coded as incorrect prior to calculating the internal consistency reliability coefficients for the WISC-IV subtest and composite scores. This data coding procedure minimizes *within-child* variability. Calculating internal consistency reliabilities based on all responses (i.e., those below the basal and above the ceiling) would likely yield data with greater *within-child* variance because some children would respond to items below the starting point (i.e., basal) incorrectly and items above the stopping point (i.e., ceiling) correctly. Minimizing variability below the basal and above the ceiling (by coding all items as either correct or incorrect) decreases *within-child* variability and, statistically, yields more reliable data.

Whereas smaller *within-child* variance yields more reliable scores, greater *between-child* variability yields more reliable scores. All other things being equal, the greater the variance between examinees, the greater the reliability of scores. Therefore, restriction of range can pose a problem when calculating reliability. Restriction of range (or restriction in variability of scores) can result in spuriously low reliabilities, and spuriously low correlations with other variables. Homogenous samples often do not have enough variability in their scores to reveal a linear relationship.

In my sample of D/HOH children, I looked and then tested for restriction of range. Restriction of range was only found in the Block Design subtest. Six of the eight subtests I examined (i.e., Similarities [*SD* = 3.56], Digit Span [*SD* = 3.03], Picture Concepts [*SD* = 3.29], Vocabulary [*SD* = 3.52], Matrix Reasoning [*SD* = 3.30], and Comprehension [*SD* = 4.51]) had greater standard deviations than the population standard deviation (*SD* = 3). In addition, both the PRI (*SD* = 15.98) and the VCI (*SD* = 19.16) had greater standard deviations than the population standard deviation (*SD* = 15). Therefore, greater *between-child* variance in this sample may have contributed to higher internal consistency reliability coefficients.

Another plausible explanation for the significantly greater internal consistency reliabilities found for this sample is that the WISC-IV is simply more reliable with this special population. D/HOH children may score more consistently on the WISC-IV compared to the children in the normative sample. If a sample has very high ability (consistently passing many items) or very low ability (consistently failing many items) the internal consistency reliability coefficients can be higher than those calculated for the normative sample. The internal consistency reliability coefficients reported in the *WISC-IV Technical and Interpretive Manual* for the 16 special groups (e.g., Intellectually Gifted, Mentally Retarded) are often higher than those reported for the normative sample due to this reason (J. Zhu, personal communication, March 31, 2008).

Kostrubala (1998) compared the internal consistency reliabilities using Cronbach's alpha between deaf (n = 30) and normally hearing (n = 30) samples for the WAIS-III. She found that the internal consistency reliability for the deaf sample's VIQ was significantly higher than the hearing sample's internal consistency reliability VIQ coefficient. These findings are consistent with those of the present study, although neither her study nor mine allows direct tests of the various accounts for higher reliabilities found within samples of D/HOH children and adults.

The Validity of the WISC-IV with D/HOH Children

Group Comparisons of Mean Scores

Perceptual Reasoning Index (PRI). Contrary to my prediction, the mean PRI for the D/HOH sample (M = 93.21) was significantly lower than the population mean PRI (M = 100; $p < .05$). This finding does not support the notion that research on previous versions of the Wechsler Scales, which frequently recommended PIQ as the best measure of intelligence for D/HOH examinees, can be generalized to the WISC-IV.

This conclusion is surprising given the general consensus that nonverbal measures of intelligence are appropriate for assessing the cognitive abilities of D/HOH children (Braden & Athanasiou, 2005). Nonverbal measures of intelligence are recommended for use with D/HOH individuals because they reduce confounds between oral language deficits and intellectual abilities (Braden, 1990; Braden, Kostrubala & Reed, 1994; Kostrubala, 1998; Maller & Braden, 1993; Murphy, 1957; as cited in Evans 1980). There are four plausible explanations for why the mean PRI for this sample was significantly lower than the mean PRI for the population.

First, the lower mean PRI for this sample compared to the mean population PRI may be due to sample characteristics. It may be that a replication study would find different results using a larger, more diverse sample. It is important to note that in my sample, 105 of the 128 children fell within the Severe to Profound range of hearing loss (i.e., a hearing loss > 71 dB). The mean level of hearing loss for this sample was 90.61 dB. Therefore, the vast majority of children in this sample would be classified as deaf (as opposed to heard-of-hearing) due to the fact that they have a hearing loss greater than 70 dB. In addition, 83 of the 128 children in this sample relied exclusively on ASL as their primary mode of communication (i.e., they did not use any type of oral/aural communication modality). It is plausible that a sample comprised of deaf individuals (as compared to a sample of hard-of-hearing individuals) may score lower on a test of intelligence, especially if the test confounds with language (see next point). However previous research has found no such relationship between level of deafness and scores on performance based tests of intelligence **(**Braden, 1990; 1994; 2000**).**

Second, language demands may differ between the WISC-III PIQ subtests and the WISC-IV PRI subtests. Although nonverbal tests attempt to reduce language, language is not entirely eliminated from the subtests. Often, test content, directions, and examples require the use of language. Most performance or motor-free nonverbal tests (including the WISC-

III and WISC-IV) are not completely nonverbal (Braden & Anathasiou, 2005). As a result, verbal abilities may be tapped within these “nonverbal” tasks, introducing construct irrelevant variance (Messick, 1995).

For the WISC-III, all PIQ subtests, with the exception of the Picture Completion subtest, begin with the examiner modeling the task. Although verbal directions supplement this visual display, task understanding may be attained simply through watching the examiner. This is quite different from the subtests of the WISC-IV PRI, which, with the exception of the Block Design subtest, do not begin with the examiner modeling the task. The PRI subtests rely more heavily than WISC-III subtests on language for task comprehension.

Third, changes made to the content of the WISC-IV PRI may change what is being measured relative to the WISC-III PIQ. Weiss, Prifitera and Saklofske (2005) claim that, compared to the other WISC-IV composites, the PRI has undergone the most extensive changes from the earlier WISC-III PIQ composite. In fact, only one core Performance Scale subtest (i.e., Block Design) from the WISC-III PIQ remains on the WISC-IV PRI. The new WISC-IV PRI intends to elicit more nonverbal reasoning and reduces visual spatial and motor skills, relative to its predecessor. For example, all of the WISC-III PIQ subtests were either timed (i.e., Picture Completion and Coding) or were timed *and* required the manipulation of objects (i.e., Block Design, Picture Arrangement and Object Assembly). Although the PRI still includes the Block Design subtest, a timed task that requires the manual manipulation of blocks, the other two core PRI subtests (i.e., Picture Concepts and Matrix Reasoning) are not timed and do not require the manual manipulation of objects (only pointing to correct responses). The developers of the WISC-IV claim that the new PRI provides a better measure of fluid reasoning than PIQ because it is less influenced by visualization, manual dexterity and processing speed (Braden, 2005).

Therefore, it is important to consider the differences between performance and motor-free nonverbal tests of intelligence for this population. Previous research has shown that D/HOH examinees tend to score lower on nonverbal reasoning tests that reduce or eliminate the role of manual dexterity relative to performance tests that require the manual manipulation of items for task completion (Braden, 1994; Braden, 2005; Braden, Kostrubala & Reed, 1994). It appears that with the subtest changes, the PRI may no longer function like a performance battery, but instead function more like a motor-free nonverbal battery. In fact, even the Block Design subtest, although still timed, was modified for the WISC-IV to decrease the influence of time bonuses (an important characteristic of performance tests).

Fourth, it is possible that PIQs have simply overestimated *g*, and that PRI (as a better measure of *g*) is a more accurate and, therefore, lower reflection of general intelligence in D/HOH children. In evaluating this explanation, one can consider how the WISC-IV was modified in relation to the CHC theory of cognitive abilities. The CHC theory suggests that cognitive abilities can be organized into a hierarchical three-tiered model, with *g*, or general intelligence, on top, ten second-order factors in the middle (i.e., Fluid Intelligence, Quantitative Knowledge, Crystallized Intelligence, Reading and Writing, Short-term Memory, Visual Processing, Auditory Processing, Long-term Storage and Retrieval, Processing Speed, and Decision speed/Reaction time), and 76 very specific abilities on the bottom (e.g., vocabulary knowledge, the ability to process and discriminate speech sounds, speed of eye movements). In the WISC-III, five of the seven PIQ subtests (i.e., Block Design, Object Assembly, Picture Arrangement, Picture Completion, and Mazes) loaded heavily onto the CHC second-order “Visual Processing” factor. The other two PIQ subtests (i.e., Coding and Symbol Search) loaded primarily onto the CHC second-order “Processing Speed” factor (Alfonso, Flanagan, & Radwan, 2005). One goal of the WISC-IV modifications was for the PRI subtests to provide a more pure measure of fluid reasoning that is less influenced by visualization, manual dexterity and processing speed (Braden, 2005). Therefore, it can be assumed that the WISC-IV PRI better represents the second-order “Fluid Intelligence” factor than did PIQ.

Within the CHC theory, Fluid Intelligence loads more heavily on *g* than the Visual Processing and Processing Speed factors. Therefore, one explanation for the lower mean PRI for this sample could be that the new PRI provides a better measure of *g* than did the WISC-III PIQ. This explanation implies that previous research using WISCs’ PIQ may have overestimated *g* for D/HOH children.

Verbal Comprehension Index (VCI). As expected, the mean VCI for the D/HOH sample (M = 80. 86) was significantly lower than the known mean VCI for the population (M = 100; $p < .05$). There is general agreement among psychologists and researchers that verbal measures of intelligence are inappropriate to use with D/HOH children as an index of cognitive ability. It is understood that depressed verbal intelligence scores for this special population may not reflect cognitive ability, but rather may reflect the lack of opportunity D/HOH children have to acquire verbal and social knowledge (Braden, 1994). D/HOH children obtain Verbal IQs approximately one standard deviation below the mean score for hearing examinees (i.e., *M* = 85) (Braden, 1994; 2005; Maller & Braden, 1993; Sullivan & Montoya, 1997). The findings of this study are consistent with previous research. The mean VCI for this sample (*M* = 80. 86) was not significantly different than one standard deviation from the population mean (*M* = 85). This finding supports the validity of the WISC-IV; however, underscores the general consensus that verbal tests of intelligence (now including the WISC-IV VCI) are not appropriate for use with D/HOH children.

Subtest Interrelationships

Based on the assumption that all the subtests of the WISC-IV should measure general intelligence (i.e., *g*), it is expected that they should positively correlate with one another. I hypothesized that the correlations among the WISC-IV subtests would be positive and reliably greater than zero. However, of the 44 correlations between the available WISC-IV subtests, 15 correlation confidence intervals included zero. Therefore, the validity of the WISC-IV for D/HOH children is only partially supported by these findings.

Investigating these non-significant correlations, it was found that 13 of the 15 correlations included one of the PSI subtests (i.e., Coding, Symbol Search or Cancellation). The other two non-significant correlations were between a PRI subtest and a VCI subtest (i.e., Block Design – Vocabulary; Matrix Reasoning – Comprehension). All subtests within the same index were significantly correlated. In other words, all non-significant correlations were *between-index* correlations.

These findings partially support the validity of this scale for use with D/HOH children. "It is expected that the subtests contributing to a specific scale (e.g., Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed) would have higher correlations with each other than with subtests comprising other scales" (Wechsler, 2003, p. 49). This general assumption is supported, given the fact that the only non-significant correlations are among subtests within different indexes.

As previously mentioned, the WISC-IV is based on the CHC theory of cognitive abilities. The PSI subtests of the WISC-IV fall into the second-order Processing Speed factor of the CHC theory. Although all second-order factors load onto *g*, some factors (e.g., Fluid Intelligence) load more heavily on *g* than other factors (e.g., Processing Speed). In relation to the other second-order factors, the Processing Speed factor is one of the least *g*-loading factors (Alfonso, Flanagan, & Radwan, 2005). In fact, Wechsler (2003) points out that the Cancellation subtest appears to be the least *g*-loaded subtest of the WISC-IV. Therefore, the fact that the WISC-IV PSI subtests load minimally onto *g* is one plausible explanation for their low (non-significant) correlations with other subtests for this sample of D/HOH children. However, it is important to note that the two other non-significant correlations are comprised of subtests (i.e., Block Design—Vocabulary, Matrix Reasoning –Comprehension) that claim to primarily tap the two most *g*-loaded, second-order factors (i.e., Fluid Intelligence and Crystallized Intelligence), yet they do not correlate significantly with each other.

Another explanation for the non-significant correlations between the WISC-IV PSI subtests and the other WISC-IV subtests may be that speeded subtests may tap specific neuropsychological deficits that are more prevalent in D/HOH populations. Braden (1990) conducted a meta-analysis of 21 studies and found that on Wechsler Performance Scales, deaf individuals consistently scored lower on the Coding/Digit Symbol subtests relative to the other Performance subtests. Braden suggested that low Coding/Digit Symbol scores could be the result of neurological impairments more common in the deaf population. Following this logic, the WISC-IV PSI subtests may correlate poorly with other WISC-IV subtests because they are attenuated by subtle neuropsychological deficits present in this D/HOH sample.

In sum, the subtest inter-correlations only partially support the validity of the WISC-IV for use with D/HOH children. All non-significant correlations include subtests from two different indexes, thus are *between-index* correlations. It is unclear whether non-significant subtest interrelationships are due to differences in the factor loadings of the subtest, neuropsychological deficits present in this sample of D/HOH children or some other unknown cause. Research investigating the factor structure of the WISC-IV for D/HOH children, understanding how and to what degree the subtests load onto the four composite scores, may help shed light onto the subtest interrelationships and provide evidence for the validity of the WISC-IV with this special population.

Incorporating Psychologists into the Data Collection Process

The second objective of this study was to incorporate psychologists into the data collection process. Although quantitative data examining this process were not collected, I acquired some insights into how test users can share responsibility with test developers to acquire information on how tests should be used with D/HOH children.

I attempted a new method of collecting data (i.e., data collection via the Internet). The Internet allows access to people (and potentially data) that overcomes geographic and time barriers. Previous research on this topic has used geographically contiguous convenience samples (e.g., data from students in one or two programs serving D/HOH children). However, I explored whether access to relevant electronic Listservs would connect me with individuals who self-identified as interested professionals. I found that collecting data via the Internet yields a more heterogenous sample because data are gathered from across the country. This aspect of the data collection process increases the external validity of the study, which helps in generalizing the findings to D/HOH children across the US.

In addition to improving the external validity of this study, this methodology may also provide benefits to the individuals collecting the data. First and foremost, it allows the participants (in this case, psychologists) to contribute to the research in their field. By providing data, participants are able to directly influence their work and the work of their colleagues. Second, by providing data, the psychologists in this study were able to fulfill an ethical obligation to investigate the psychometric properties of the instruments they use with D/HOH children. Furthermore, as is true of any study collecting archival data, this methodology grants the participants the freedom for where and when to enter data. Unlike other data collection methodologies that require a participant to show up at a specific time and place, this process has no such setting or temporal constraints. In general, it appears that collecting data via the Internet is a specific methodological process that can help bridge the gap between research and practice.

Although there are many benefits of this methodology, there are several important challenges to mention as well. First, security is important; all electronic documents containing data should be encrypted. Second, it is important that the instructions for data entry are explicit and easy to understand. Because I did not communicate face-to-face with participants, there were more opportunities for misunderstandings about data entry (e.g., how to code data when entering it into the Excel spreadsheet, 1= mild hearing loss, 2 = moderate hearing loss, 3= severe hearing loss, 4 = profound hearing loss). I provided detailed directions on the Excel spreadsheets, with general directions at the top and specific instructions for each data cell in easy-to-view, pop-up windows at the top of each data column, yet still had data entry errors.

Finally, this approach challenges researchers to ensure that legitimate participants are providing the data. Because I did not have face-to-face contact with the participants, falsifying participant information is plausible. There are a number of safeguards researchers can take to minimize intentional (falsification) and unintentional errors. First, participants should be recruited from reputable websites and Listservs (as was done for this study). Second, researchers can require participants provide specific information that qualifies them for study participation (e.g., American Psychological Association member identification number). And third, if applicable, researchers can ask participants to supply professional titles, addresses and telephone numbers so that the researcher can follow up if any suspicion arises. For this study, there was no doubt about the legitimacy of the participants' qualifications. All participants were recruited through professional Listservs and and/or used their professional email addresses for study communication and data transmission. Furthermore, 7 of the 10 participating psychologists completed a Harcourt Assessment "qualification form" (which required proof of professional training) for ordering standardized test materials as part of the incentive for participation.

Due to the fact that (a) a sufficient number of qualified psychologists were recruited through Listservs and interested in contributing data, (b) data were collected relatively quickly (i.e., the majority of data were collected within six months), (c) there were minimal problems with data entry (i.e., participants understood how to enter data), (d) Excel spreadsheets were easily sent and received between the participants and researcher, and (e) there were minimal data entry errors, I conclude that this study was largely successful at using the Internet data collection methodology. Although this methodology is inappropriate for many areas of study (e.g., research relying on observational data to study behaviors) it was an effective, economical and efficient data collection process for research on a specific test with a specific population. Although there are possible shortcomings to my methodology (as there are with any data collection process), employing safeguards can prevent many problems. Overall, the benefits of using this procedure for data collection (e.g., larger and more heterogeneous samples, flexible nature of data entry) outweigh the risks (e.g., lower standardized control, data entry errors, requiring password protected data).

Implications for Practice

The main goal of this study is to inform assessment practice with D/HOH students. Therefore, it is important to consider the implications of these findings for psychologists, educators and others who work with D/HOH children.

First, it appears that the WISC-IV is a reliable assessment to use with D/HOH children. Test reliability is a necessary but not sufficient condition for validity. However, it is important for practitioners to understand that the internal consistency reliability was not assessed for all WISC-IV subtests and indexes. Internal consistency reliability coefficients were not calculated for the Coding, Symbol Search and Cancellation subtests, and the PSI, because the split-half reliability method is not a proper estimate of reliability for tasks in which speed is key factor. In addition, internal consistency reliability coefficients were not calculated for Letter-Number Sequencing, Information, Arithmetic, and Word Reasoning subtests, and the WMI and FISQ, due to sample size restrictions. Therefore, estimates of reliability for these subtests and indexes are currently unknown. Practitioners should exercise caution when interpreting WISC-IV scores, because reliability has only been assessed for 8 of the 15 subtests and 2 of the 5 indexes.

Second, it is important for practitioners to understand that, according to this study, the WISC-IV PRI does not function like the previous WISCs' PIQ. The WISC-IV developers encouraged practitioners to substitute the WISC-IV PRI for the WISC-III PIQ in clinical interpretations and evaluations, arguing that the PRI should be equivalent to the PIQ. Although the WISC-IV PRI is marketed as being sufficiently similar to previous versions of the Wechsler Scales that prior research generalizes well to the new PRI, my data suggest that PRI functions more like a motor-free nonverbal assessment of intelligence than a performance test of intelligence. The PRI/PIQ difference has implications for practitioners because D/HOH children tend to score lower on motor-free nonverbal assessments (like the WISC-IV PRI) compared to performance assessments of intelligence (like the WISC-III PIQ). Therefore, practitioners should exercise caution when interpreting the PRI for D/HOH examinees, especially because the PRI may be used for educational placement decisions.

Practitioners should also consider that a low PRI may not just reflect a D/HOH child's intelligence, but is likely confounded by the types of tasks given and the verbal language required to understand the task directions. These confounding variables alone or in interaction with each other may cause a spuriously low score that does not purely reflect the construct of interest (i.e., *g*) for this population. It is generally agreed that nonverbal measures of intelligence are appropriate to use with D/HOH children; however, practitioners must still exercise caution when interpreting the WISC-IV PRI for this special population. And, of course, this score alone should never be used to make educational decisions, but should only be part of a multi-method, multi-informant evaluation.

In general, verbal tests of intelligence are not recommended for the cognitive assessment of D/HOH individuals (Braden, 1994; 2000). This recommendation holds for the WISC-IV VCI as well. The WISC-IV VCI should not be used as a measure of intelligence for D/HOH children. The WISC-IV VCI confounds language skills with intelligence and, therefore, could result in inappropriate consequences for D/HOH children if used as an assessment of cognitive ability.

Limitations

There are several limitations to this study. The most obvious limitation is sample size. Although the study exceeded its target sample size (n = 100) the sample did not provide enough data to conduct correlational analyses between WISC-IV composite scores and achievement test and other intelligence test scores. As a result, hypotheses four and five were not tested.

In considering the limited sample size, it is also important to understand how this impacts the significance testing of the subtest inter-correlations. Statistically, smaller sample sizes produce larger confidence intervals, which are more likely to include zero. Therefore, a limited sample size will increase the odds of getting a non-significant correlation. Although possible, the non-significant, inter-subtest correlations found in this study do not appear to be the result of low sample size. When examining the data in this study, all non-significant, inter-subtest correlations had a sample size greater than or equal to 30, and 8 of the 15 non-significant correlations had a sample size over 50.

Another limitation of this study is the sample characteristics. Although this sample is described as D/HOH, the majority of the children in this sample classify as severely or profoundly deaf (i.e., hearing loss > 70 dB). Less than 20% of the children in the sample had a hearing loss less than 70 dB. Extrapolating these findings to children with mild or moderate degrees of hearing loss is not recommended.

A third limitation to this study deals with the data collection process. Because psychologists entered the data, there was no way to check for entry errors against the original protocols. Consequently, it is possible that some entry errors went unnoticed by the participants and may have influenced study results. For example, 12.7% of the cases had individual item data that did not match the total raw score.

A fourth limitation deals with standardized control of the testing environment. In this study, psychologists were asked to enter archival WISC-IV scores. In other words, psychologists entered WISC-IV test data that were used as part of a previous psychological evaluation (e.g., educational placement, clinical diagnosis). Because the WISC-IV data was archival, I had no control over the testing environment and/or situation. I assumed that all WISC-IV standardization procedures for test administration and scoring were used appropriately, but I cannot guarantee these conditions were met. Characteristics of the testing environment (e.g., noise level, lighting, visual distractions) as well as examinee characteristics during the assessment (e.g., fatigue, difficulty attending to tasks, motivation) were not reported nor analyzed. Therefore, the possible influence of these characteristics on the test scores and interpretations are unknown.

These procedures are somewhat different than those used for the WISC-IV standardization data collection process. In the WISC-IV standardization data collection process, specific quality control procedures were employed, including systematically training qualified examiners in test administration, scoring and data entry, double-checking all protocols, using a specifically designed computer program to check for errors, and having trained recruiters find children who met specified criteria for inclusion in the standardization sample (Wechsler, 2003). However, a discussion of the various testing environments and child characteristics during testing is not provided in the *WISC-IV Technical and Interpretive Manual*.

Directions for Future Research

Future research should explore the reliability and validity of the WISC-IV with D/HOH children. First, research should assess the internal consistency reliability for Letter-Number Sequencing, Information, Arithmetic, and Word Reasoning subtests and the WMI and FSIQ. The internal consistency reliabilities for these subtests and indexes were unable to be evaluated in this study because many cases in my sample lacked these data. In addition, stability data (i.e., test-retest) for the Coding, Symbol Search and Cancellation subtests and the PSI are needed because internal consistency is an inappropriate measure of reliability for tasks in which speed is a key factor. However, assessing the test-retest stability for all

WISC-IV subtests would provide researchers and practitioners with more information regarding the stability of this measure for D/HOH children.

Research is also needed to establish the validity of the WISC-IV with D/HOH children. Research exploring how the PRI functions with different, larger samples of D/HOH children would help clarify whether the findings of this study are anomalous or representative (i.e., whether my findings are due to sampling and procedural issues, or actually reflect values for D/HOH populations). Understanding the generalizabiliy of my research findings is especially important because the WISC-IV PRI is currently being used to represent the cognitive abilities of D/HOH children. In addition, this study was unable (due to sample size restrictions) to examine the validity of the WISC-IV for this special population by correlating it with achievement and other intelligence test measures. Future research is needed to investigate the validity evidence of the WISC-IV with D/HOH by assessing its relation with other variables (e.g., convergent evidence, test-criterion evidence).

Furthermore, studies investigating the factor structure of the WISC-IV with this special population (e.g., exploratory and confirmatory factor analysis) would also provide evidence for the validity of this assessment with D/HOH children. This type of research could help determine, if in fact, the PRI is more *g*-loaded than the previous WISCs' PIQ.

Conclusions

My results largely support the reliability of the WISC-IV for use with D/HOH children. The internal consistency reliabilities calculated for this sample of D/HOH children were the same or greater than those reported in the *WISC-IV Technical and Interpretive Manual* for the normative sample. In addition, the results of this study, consistent with previous studies, show that D/HOH children score significantly lower on verbal measures of intelligence. Scores from verbal tests of intelligence are confounded with language abilities and should not be used to assess the cognitive abilities of this special population.

This sample of D/HOH children had a mean PRI (*M* = 93.21) that fell significantly below the population mean of 100. This is inconsistent with the previous research showing the Wechsler Performance Scales function similarly with D/HOH and hearing children (Braden, 1984; Braden, 1990; Braden, Kostrubala, & Reed, 1994; Maller & Braden, 1993; Murphy, 1957; as cited in Evans, 1980). More research is needed to investigate the validity of the WISC-IV PRI subtests.

In addition, the validity evidence of the WISC-IV for D/HOH children is somewhat weakened by the fact that 15 of the 44 inter-subtest correlations were non-significant. It appears for this sample, the PSI subtests do not correlate strongly with other WISC-IV subtests.

This is the first study to explore the reliability and validity of the WISC-IV with D/HOH children. The results of this study generally support the reliability of this measure for D/HOH children; however, the validity of the WISC-IV with this special population is not yet established. Future research is needed to investigate this topic more thoroughly. In the mean time, practitioners working with D/HOH children should understand how these findings impact assessment selection and interpretation and are strongly encouraged to conduct multi-method, multi-informant evaluations in which the WISC-IV is only one component.

References

Alfonso, V.C., Flanagan, D. P., & Radwan S. (2005). The impact of the Cattell-Horn-Carroll Theory on test development and interpretation of cognitive and academic abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues, second edition* (pp. 185- 202). New York: Guilford Press.

Allen, T. E. (1994). *Who are the deaf and hard-of-hearing student leaving high school and entering postsecondary education?* (pp. 1 –16). U.S. Office of Special Education and Rehabilitative Services: Pelavin Research Institute.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1966). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Braden, J. P. (1984). The factorial similarity of the WISC-R performance scale in deaf and hearing samples. *Journal of Personality and Individual Differences, 5*, 405-409.

Braden, J. P. (1989). The criterion-related validity of the WISC-R performance scale and other nonverbal IQ tests for deaf children. *American Annals of the Deaf, 134*, 329-332.

Braden, J. P. (1990). Do deaf persons have a characteristic psychometric profile on the Wechsler Performance Scales? *Journal of Psychoeducational Assessment, 8*, 518-526.

Braden, J.P. (1994). *Deafness, Deprivation, and IQ*. New York: Plenum Press.

Braden, J. P. (2000). Deafness and hearing loss. In A. E. Kazdin (Ed.) *Encyclopedia of psychology* (vol. 2) (pp. 441-444). Washington, DC/Oxford: American Psychological Association/Oxford University Press.

Braden, J.P. (2005). Hard-of-hearing and deaf clients: Using the WISC-IV with clients who are hard-of-hearing or deaf. In A. Prifitera, D.H. Saklofske, L.G. Weiss (Eds.), *WISC-IV clinical use and interpretation* (pp. 351-376). Burlington, MA: Elsevier Academic Press.

Braden, J. P., & Anthanasiou, M. S. (2005). A comparative review of nonverbal measures of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues, second edition* (pp. 557-577). New York: Guilford Press.

Braden, J. P., & Hannah, J. M. (1998). Assessment of hearing impaired and deaf children with the WISC-III. In D. Saklofske & A. Prifiter (Eds.), *Use of the WISC-III in clinical practice* (pp. 175-201). Boston, MA: Houghton Mifflin.

Braden, J. P., Kostrubala, C.E., & Reed. J. (1994). Why do deaf children score differently on performance vs. motor-reduced nonverbal intelligence tests? *Journal of Psychoeducational Assessment*, 12, 357-363.

Braden, J.P., & Paquin, M.M. (1985). A comparison of the WISC-R and WAIS-R performance scales in deaf adolescents. *Journal of Psychoeducational Assessment, 3,* 285-290.

Braden, J.P., Wollack, J. A., & Allen, T. E. (1995). A reply to Kishor: Choosing the right metric. *Journal of Psychoeducational Assessment, 13*, 250-265.

Bradley-Johnson, S., & Evans, L. D. (1991). *Psychoeducational assessment of hearing-impaired students: Infancy through high school*. Austin, TX: Pro-ed.

Brill, R. G. (1962). The relationship of Wechsler IQs to academic achievement among deaf students. *Exceptional Children*, 325-321.

Brooks, C. R., & Riggs, S. T. (1980). WISC-R, WISC, and reading achievement relationships among hearing impaired children attending public schools. *Volta Review, 82*, 96-102.

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues, second edition* (pp. 69-76). New York: Guilford Press.

Council for Exceptional Children. (2005). Retrieved October 15, 2005, from http://www.cec.sped.org/

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

Evans, L. (1966). A comparative study of the Wechsler Intelligence Scale for Children (Performance) and Raven's Progressive Matrices with deaf children. *Teacher of the Deaf, 64*, 76-82.

Evans, L. (1980). WISC Performance Scale and Coloured Progressive Matrices with deaf children. *Journal of Educational Psychology, 50*, 216-222.

Fagan, K.T., & Wise, P.S. (2000). *School psychology: Past, present, and future* (2nd ed). New York: Longman.

Falberg, R. M. (1983). Psychological assessment of the verbal functioning of postsecondary educational program applicants: Improving predictive validity. In D. Watson, G. Anderson, P. Marut, S. Ouellette, & N. Ford (Eds.), *Vocational evaluation of hearing-impaired persons: Research and practice.* Warm Springs, AR: Univeristy of Arkansas.

Feldt, L. (1969). A test of the hypothesis that Cronbach's Alpha or Kuder-Richardson Coefficient Twenty is the same for two tests. *Psychometrika, 34,* 363-373.

Gallaudet Research Institute (2002). 2001-2002 regional and national survey. Washington, DC: Author. Available from http://gri.gallaudet.edu/Demographics/.

Goodwin, L.D., & Leech, N.L. (2003). The meaning of validity in the new Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*, 181-191.

Gregory, R. J. (2004). *Psychological testing: History, principles, and applications* (4th ed.). Boston, MA: Pearson Education Group, Inc.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education*. New York: McGraw Hill.

Hirshoren, A., Hurley, O. L., & Kavale, K. (1979). Psychometric characteristics of the WISC-R Performance Scale with deaf children. *Journal of Speech and Hearing Disorders, 44*, 73-79.

Holt, J., Hotto, S., & Cole, K. (1994). Demographic aspects of hearing impairment: Questions and Answers (3rd ed.). Gallaudet University. Available from http://gri.gallaudet.edu/Demographics/factsheet.html

Horn, J. L., & Blankson (2005). Foundations for better understanding of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (2nd ed., pp. 41-68). New York: Guilford Press.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Theoretical perspectives. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53-91). New York: Guilford Press.

Jenson, A.R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger Publishers.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometricka, 2*, 151-160.

Kelly, M., & Braden, J. P. (1990). Criterion-related validity of the WISC-R Performance Scale with the Stanford Achievement Test-Hearing Impaired Edition. *Journal of School Psychology, 28,* 147-151.

Kostrubala, C. (1998). Translation of the WAIS-III from English to ASL. Unpublished master's thesis. University of Wisconsin, Madison.

Lavos, G. (1962). W.I.S.C. psychometric patterns among deaf children. *The Volta Review*, 547-552.

Maller, S. J. (1996). WISC-III verbal item invariance across samples of deaf and hearing children of similar measured ability. *Journal of Psychoeducational Assessment, 14*, 152-165.

Maller, S. J. (1997). Deafness and WISC-III item difficulty: Invariance and fit. *Journal of School Psychology, 35*, 299-314.

Maller, S. J. (2003). Intellectual assessment of deaf people. In M. Marschark & P.E. Spencer (Eds.), *Oxford handbook of deaf studies, language and education* (pp. 451-463). New York, NY: Oxford University Press.

Maller, S.J., & Braden, J.P. (1993). The construct and criterion-related validity of the WISC-III with deaf adolescents. *Journal of Psychoeducational Assessment, Advances in Psychoeducational Assessment, Wechsler Intelligence Scale for Children, Third Edition [Monograph], 11*, 105-113.

Maltby, M T., & Knight, P. (2000). *Audiology: An introduction for teachers and other professionals*. London: David Fulton Publishers.

Marschark, M. (2003). Cognitive functioning in deaf adults and children. In M. Marschark & P.E. Spencer (Eds.), *Oxford handbook of deaf studies, language and education* (pp. 464-477). New York, NY: Oxford University Press.

Marschark, M., Lang, H.G., & Albertini, J.A., (2002). *Educating deaf students: From research to practice*. New York: Oxford University Press.

McCallum, R. S. (2003). Context for nonverbal assessment of intelligence and related abilities. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 3-22). New York: Kluwer Academic/Plenum.

McCallum, R. S., Bracken, B. A., & Wasserman, J. D. (2001). *Essentials of nonverbal assessment*. New York: Wiley.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilites. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues, second edition* (pp. 136-181). New York: Guilford Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Moores, D., Kluwin, T., Johnson, R., Cox, P., Blennerhassett, L., Kelly, L., Sweet, C., & Fields, L. (1987). *Factors predictive of literacy in deaf adolescents of deaf parents: Factors predictive of literacy in deaf adolescents in total communication programs*

(Contract No. NO1-NS-4-2365). Washington: National Institute of Neurological and Communicative Disorders and Stroke.

Murphy, L. J. (1957). Tests of abilities and attainments. In Ewing, A. W. G. (Eds.), *Educational Guidance and the Deaf Child*. Manchester University Press.

National Center on Birth Defects and Developmental Disabilities (2004). *Hearing loss*. Retrieved October 15, 2005 from http://www.cdc.gov/ncbddd/dd/ddhi.htm

National Dissemination Center for Children with Disabilities (2004). *Deafness & hearing loss: Disability fact sheet*. Retrieved April 8, 2006, from http://www.nichcy.org/pubs/factshe/fs3.pdf

Northern, J. L., & Downs, M. P. (2002). *Hearing in children* (5th ed). Baltimore, MD: Lippincott Williams & Wilkins.

Ortiz, S. O., & Dynda, A. M. (2005). Use of intelligence tests with culturally and linguistically diverse populations. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues, second edition* (pp. 545-556). New York: Guilford Press.

Prifitera, A., Weiss, L. G., Saklofske, D. H., & Rolfhus, E. (2005). The WISC-IV in the clinical assessment context. In A. Prifitera, D.H. Saklofske, & L.G. Weiss (Eds.), *WISC-IV clinical use and interpretation* (pp. 3-32). Burlington, MA: Elsevier Academic Press.

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

Rourke, B. P. (2000). Neuropsychological and psychosocial subtyping: A review of investigations within the University of Windsor laboratory. *Canadian Psychology, 41,* 34-51.

Rourke, B. P., Ahmad, S. A., Collins, D. W., Hayman-Abello, B. A., Hayman-Abello, S. E., & Warriner, E. M. (2002). Child clinical/pediatric neuropsychology: Some recent advances. *Annual Review of Psychology, 53*, 309-339.

Saklofske, D. H., Rolfhus, E., Prifitera, A., Zhu, J., & Weiss, L. G. (2005). Clinical Interpretation of the WISC-IV FSIQ and GIA. In A. Prifitera, D.H. Saklofske, & L.G. Weiss (Eds.), *WISC-IV clinical use and interpretation* (pp. 33-65). Burlington, MA: Elsevier Academic Press.

Sattler, J. M., & Hardy-Braz, S. T. (2002). Hearing impairments. In J. M. Sattler (Ed.), *Assessment of children: Behavioral and clinical applications* (4^{th} ed., pp. 377-389). La Mesa, CA: Jerome M. Sattler Publisher, Inc.

Schildroth, A., Rawlings, B., & Allen, T. (1991). Deaf students in transition: Education and employment issues for deaf adolescents. In O. Cohen & G. Long (Eds.), Selected issues in adolescence and deafness [Special Issue]. *The Volta Review, 93*, 41 – 53.

Sisco, F. & Anderson, R. (1978). Current findings regarding the performance of deaf children on the WISC-R. *American Annals of the Deaf, 123(2)*, 115-121.

Slate, J. R., & Fawcett, J. (1995). Validity of the WISC-III for deaf and hard of hearing persons. *American Annals of the Deaf, 140*, 250-254.

Stokoe, W. C. (2001). Deafness, cognition, and language. In M. D. Clark, M. Marschark, & M. Karchmer (Eds.), *Context, cognition, and deafness* (pp. 6-13). Washington, DC: Gallaudet University Press.

Sullivan, P. M., & Brookhouser, P.E. (1996). Intellectually gifted deaf adolescents: Etiological and academic correlates. In P. M. Sullivan & P.E. Brookhouser (Eds.), *Proceedings of the Fourth Annual Conference on the Habilitation and Rehabilitation of Hearing Impaired Adolescents* (pp. 69-80). Boys Town, NE: Boys Town Press.

Sullivan, P.M., & Burley, S.K. (1990). Mental testing of the deaf child. In C. Reynolds & R. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (pp. 761 – 788). New York: Guilford Press.

Sullivan, P. M., & Montoya, L. A. (1997). Factor analysis of the WISC-III with deaf and hard-of-hearing children. *Psychological Assessment, 9,* 317-312.

Sullivan, P. M., & Schulte, L. E. (1992). Factor analysis of the WISC-R with deaf and hard-of-hearing children. *Psychological Assessment, 4*, 537-540.

Sullivan, P. M., & Vernon, M. (1979). Psychological assessment of hearing impaired children. *School Psychology Digest, 8,* 271-290.

Tye-Murray, N. (2004). *Foundations of aural rehabilitation: Children, adults, and their family members* (2nd ed.). Clifton Park, NY: Delmar Learning.

United States Department of Education (2003). *25th annual report to congress on the implementation of the Individuals with Disabilities Act*. Retrieved October 15, 2005 from http://www.ed.gov/about/reports/annual/osep/2003/

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children*. New York: The Psychological Corporation.

Wechsler, D. (1955). *Wechsler Adult Intelligence Scale*. New York: Psychological Corporation.

Wechsler D. (1974). *Wechsler Intelligence Scale for Children—Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—Third edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *WAIS-III WMS-III technical manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003a). *Wechsler Intelligence Scale for Children—Fourth edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003b). *WISC-IV technical and interpretive manual*. San Antionio, TX: The Psychological Corporation.

Weiss, L. G., Prifitera, A., & Saklofske, D. H. (2005). Interpreting the WISC-IV index scores. In A. Prifitera, D.H. Saklofske, & L.G. Weiss (Eds.), *WISC-IV clinical use and interpretation* (pp. 71- 89). Burlington, MA: Elsevier Academic Press.

Zieziula, F. R. (Ed.) (1982). *Assessment of hearing impaired people: A guide for selecting psychological, educational, and vocational tests*. Washington: Gallaudet College Press.

Footnote

[1] The American Psychological Association recommends using person first language (e.g., children who are deaf or hard-of-hearing). However, the term "deaf and hard-of-hearing children" or "D/HOH children" is used in this paper in order to be consistent with the way in which deaf people in North America define themselves. This usage is also consistent with that used in the current literature.

Appendices

Appendix A

Review of Reliability Studies

Study	Sample	Methods	Results	Conclusions
WISC				
Evans (1966)	*N* = unkown; ages 6 – 12 years; deaf	Tested this sample of deaf children with the WISC PS three years apart. Calculated the mean scores and correlated them to get the test-retest reliability.	The first mean WISC PS score was 98 and the second was 97. The correlation between these two scores was statistically significant ($p < .01$).	These findings support the test-retest reliability for the WISC PS with deaf children.
Evans (1980)	*N* = 125; ages 5 – 12 years; prelingually deaf	The internal consistency reliability was estimated for the WISC PS subtests by correlating raw score totals for odd and even items and using the Spearman-Brown correction for whole test reliability. Coding was excluded in the calculation due to the fact that it is a speeded test. Evans (1980) divided the sample into four age groups, calculating the internal consistency reliability coefficients for each.	The reliability coefficients found for the four age groups are the following: 5-6 years ($r = .94$), 7-8 years ($r = .95$), 9-10 years ($r = .91$), and 11-12 years ($r = .88$). All correlations were significant ($p < .01$).	The results compared favorably with the estimates in Wechsler's (1949) original standardization with hearing children and support the reliability of the WISC PS for use with deaf children.

Appendix A (continued)

Review of Reliability Studies

Study	Sample	Methods	Results	Conclusions
WISC-R				
Hirshoren, Hurley & Kavale (1979)	*N* = 59; ages 8 – 13 years; prelingually deaf	Internal consistency was assessed for four of the PS subtests of the WISC-R (i.e., Object Assembly, Block Design, Picture Arrangement, and Picture Completion). Coding was not included due to the fact that speed is a major factor on this subtest. Kuder-Richardson Coefficient 20 was used when test responses were dichotomously scored. Otherwise, Cronbach's coefficient alpha was used.	The reliability coefficients for Object Assembly, Block Design and Picture Completion for this sample did not significantly differ from those of the normative sample. However, the internal consistency reliability coefficient for Picture Arrangement was notably higher ($r = .84$ vs. $r = .73$) than that of the standardization sample. The authors note that this is due to the greater variability of the deaf sample on this subtest ($SD = 4.45$) compared to the standardization sample ($SD = 3.00$).	Overall, Hirshoren and colleagues report that the reliability of the WISC-R PS is satisfactory for assessing the intelligence of deaf children.
WAIS-III				
Kostrubala (1998)	*N* = 30; ages 18 – 65 years; severe to profound, bilateral, prelingual hearing loss	Kostrubala assessed the internal consistency of 13 WAIS-III subtests (i.e., Vocabulary, Similarities, Arithmetic, Digit Span, Information, Comprehension, Picture Completion, Block Design, Matrix Reasoning, Picture Arrangement, and Object Assembly), VIQ, PIQ, and FSIQ using Cronbach's coefficient alpha. All items on the test were translated in ASL.	The reliability coefficients for the 11 subtests, the PIQ, and the FSIQ for the deaf sample did not significantly differ from the reliability coefficients calculated for the hearing sample. However, the reliability coefficients for the VIQ for the deaf sample was significantly different (.94) from the VIQ reliability coefficient for the hearing sample (.83).	Overall, the WAIS-III is acceptable for assessing the intelligence of deaf adults.

Appendix B

Review of Validity Studies

Study	Sample	Methods	Results	Conclusions
WISC				
Murphy (1957)	*N* = 300; ages 6 – 10 years; deaf	Assessed the mean and standard deviation of the WISC PS for this sample of deaf children and compared it to the normative sample.	Murphy found the mean PIQ was 98 and the standard deviation was 15.9, which was not statistically different from the normative sample.	This finding supports the validity of the WISC PS for use with deaf children.
Brill (1962)	*N* =105; ages 8-16 years; congenitally or prelingually deaf	Assessed the predictive test-criterion relationship of the WAIS & WISC with two achievement tests, the Gray-Votaw-Rogers Test and the Stanford Achievement Test (SAT).	The correlations between Wechsler Scales and the Gray-Votaw-Rogers Test ($r = .55$). and between the Wechsler Scales and the SAT ($r = .54$) are nonsignificant.	Although these results are not significant, they are comparable to those found for hearing children and suggest that the WISC and WAIS are relatively strong predictors of achievement for deaf children.
Lavos (1962)	*N* = 59; ages 12 years; prelingually deaf	Assessed the relationship between the WISC PS and the nonverbal battery of the Lorge-Thorndike Intelligence Scale.	The results yield a correlation of .77 between the two sets of test IQ quotients	The correlation presents convergent evidence of the validity of the WISC PS for assessing deaf children.

Appendix B (continued)

Review of Validity Studies

Study	Sample	Methods	Results	Conclusions
WISC				
Evans (1980)	N = 52; ages 5 to 12 years; prelingually deaf	Assessed the predictive test-criterion relationship between the WISC PS and the Certificate of Secondary Education	The results yield a significant correlation between the two measures ($r = .81$, $p < .01$).	The finding provides evidence of the validity of the WISC PS for use with deaf children.
WISC-R				
Sisco & Anderson (1978)	N = 1228; ages 6 – 16 years; prelingually deaf	Compared the mean and standard deviation of the WISC-R PIQ and the 6 PS subtests of the deaf sample with the normative sample.	The mean and standard deviation of the PIQ for the deaf sample ($M = 95.7$, $SD = 17.55$) was significantly different than those of the normative sample ($p < .01$). However, a breakdown of the subtests reveals those deaf and hearing children perform similarly on four of the six PS subtests (i.e., Picture Completion, Block Design, Object Assembly, and Mazes). The differences in scores are mostly due to the Picture Arrangement and Coding Subtests.	Overall, this study fails to support the validity of the WISC-R PS for use with D/HOH children.

Appendix B (continued)

Review of Validity Studies

Study	Sample	Methods	Results	Conclusions
WISC-R				
Hirshoren, Hurley & Kavale (1979)	(1&2) *N* = 59; ages 8 – 13 years; prelingually deaf	(1) Assessed the concurrent test-criterion relationship of the WISC-R PS with the Hiskey-Nebraska Test of Learning Aptitude (H-NTLA) and the predictive test-criterion relationship of the WISC-R PS with the SAT . (2) Compared the mean and standard deviation of the WISC-R PIQ of the deaf sample with the normative sample	(1)The correlations between the PIQ and the H-NTLA Learning Quotient scores (r = .89) and between the WISC-R PIQ and seven SAT subtest grade-based scores (Word Meaning r = .34; Paragraph Meaning r = .31; Vocabulary r = .30; Spelling r = .29; Language r = .09; Arithmetic Computation r = .20; Arithmetic Concepts r = .24) are significant ($p < .05$) with the exception of Language and Arithmetic Computation. (2) The mean PIQ of the deaf sample was (88.07) and the standard deviation was (17.84). The range in PIQs reported was 52 to 129.	(1)The correlations are significant; therefore, it may be concluded that the WISC-R PS is a valid measure to use with deaf children. (2) The authors conclude that their sample shows lower mean scores and greater variability in the WISC-R PS compared to the normative sample.

Appendix B (continued)

Review of Validity Studies

Study	Sample	Methods	Results	Conclusions
WISC-R				
Brooks & Riggs (1980)	*N* = 40; ages 6 – 16 years; deaf and hard-of-hearing	Assessed the validity of the WISC-R PS by evaluating its (a) convergent relationship with the WISC PS and (b) test-criterion relationship with the reading subtests of the Stanford Achievement Test-Hearing Impaired Edition (SAT-HI)	Significant correlations were found between the WISC PS and the WISC-R PS ($r = .85$, $p < .05$) and between each of the corresponding subtests between the two tests ($p < .05$). The PIQs for the WISC and the WISC-R were each correlated with reading achievement, as measured by the SAT-HI($r = .18$ and $r = .19$, respectively) and were found to be nonsignificant.	Overall, there is some evidence to suggest that the WISC-R is a valid measure of intelligence among D/HOH children; however, it does not appear that the WISC-R is able to significantly predict reading achievement among this special group.
Braden (1984)	*N* = 1228; ages 6 – 16 years; prelingually deaf	Assessed the factor structure of the WISC-R PS for the deaf sample and the WISC-R standardization sample.	Factors were extracted from the interrcorrelations of the Performance subtests. Only one factor emerged for each group (i.e., deaf sample and standardization sample). The sources of variance were virtually identical in the two groups.	The WISC-R measures the same construct for hearing and deaf children.
Braden & Paquin (1985)	*N* = 32; mean age at WISC-R testing 14 years, 8 months; mean age at WAIS-R testing 18 years, 1 month; deaf	Assessed the validity of the WISC-R PS by evaluating its convergent relationship with the WAIS-R PS.	Correlations were calculated between the WISC-R PS subtests and the WAIS-R PS subtests and the PIQs for each test. WISC-R and WAIS-R subtest scores and PIQs are all significantly correlated ($p < .008$).	The results indicate that the WISC-R and the WAIS-R PS are comparable within this sample of deaf adolescents.

Appendix B (continued)

Review of Validity Studies

Study	Sample	Methods	Results	Conclusions
WISC-R				
Braden (1989)	Study 1: N = 33; grades 1 –8; prelingually deaf Study 2: N = 64; grades 1-8; prelingually deaf	Assessed the test-criterion relationship of the WISC-R PS with the SAT-HI	Study 1: Correlations were calculated between the WISC-R PIQ and seven SAT-HI subtest age-based percentiles (Word Reading $r = .04$; Reading Comprehension $r = .33$; Spelling $r = .36$; Total Reading $r = .02$; Arithmetic Concepts $r = .20$; Arithmetic Computation $r = .05$; Arithmetic Applications $r = .10$). Study 2: Correlations were calculated between the WISC-R PIQ and seven SAT-HI subtest grade-equivalents (Word Reading $r = .27$; Reading Comprehension $r = -.05$; Spelling $r = .22$; Total Reading $r = .37$; Arithmetic Concepts $r = .14$; Arithmetic Computation $r = -.08$; Arithmetic Applications $r = .20$). All correlations were nonsignificant ($p > .05$).	The findings suggest that the WISC-R PS is not a powerful predictor of deaf children's achievement. Overall, these results failed to support the validity of the WISC-R PS for deaf children.
Kelly & Braden (1990)	N = 83; ages 7 – 16 years; prelingually deaf	Assessed the test-criterion relationship of the WISC-R PS with the five subtests from the SAT-HI.	Correlations were calculated between (a) WISC-R PIQ and the five SAT-HI subtest scaled scores (Reading Comprehension $r = .32$; Spelling $r = .24$; Concept of Number $r = .46$; Math Calculation $r = .31$; Math Applications $r = .41$) and (b) WISC-R PIQ and the five SAT-HI percentile ranks (Reading Comprehension $r = .39$; Spelling $r = .33$; Concept of Number $r = .57$; Math Calculation $r = .42$; Math Applications $r = .52$). All correlations were significant ($p < .01$).	These findings support the validity of the WISC-R PS for use with deaf children.

Appendix B (continued)

Review of Validity Studies

Study	Sample	Methods	Results	Conclusions
WISC-R				
Sullivan & Schulte (1992)	*N* = 368; ages 6 to 16 years; D/HOH	Assessed the factor structure of the WISC-R for deaf and hard-of-hearing children separately.	The factor analyses for the deaf and hard-of-hearing children were identical and, therefore, combined. Two factors emerged: Language Comprehension and Visual-Spatial Organization. The Freedom of Distractibility factor did not emerge as it does for hearing children.	The WISC-R does not measure the same constructs for D/HOH and hearing children.
WISC-III				
Maller & Braden (1993)	(1 & 2) *N* = 30; mean age 13.63 years; prelingually deaf	(1) Assessed the predictive test-criterion relationship of the WISC-III with three subtests from the SAT-HI.	(1)Found that the PIQ had low to moderate correlations with the SAT-HI subtests (Total Reading $r = .46$; Total Language $r = .54$; Total Math $r = .63$) and that the VIQ had moderate to high correlations with the SAT-HI subtests (Total Reading $r = .80$; Total Language $r = .85$; Total Math $r = .83$). All correlations were significant ($p < .01$).	(1)These findings support the validity of the WISC-III for use with deaf children. The findings also support the hypothesis that the WISC-III VS are better predictors of academic achievement (with the exception of math achievement) for deaf children.

Appendix B (continued)

Review of Validity Studies

Study	Sample	Methods	Results	Conclusions
WISC-III				
Maller & Braden (1993) (continued)	(1 & 2) *N* = 30; mean age 13.63 years; prelingually deaf	(2)Compared group means of the WISC-III PIQ, VIQ, and FSIQ of the deaf sample with those of the normative sample	(2) The mean PIQ was not statistically different from the normative sample mean. The mean VIQ was significantly below the normative mean ($p < .01$), but was not significantly different from one standard deviation below the normative mean. The FSIQ was within the average range, but was significantly below the normative sample mean ($p < .05$). In addition, there was greater variability in the IQ scores for this sample than compared to the normative sample.	(2) The calculated PIQ and VIQ means are consistent with expectations. The WISC-III PS can be used as an estimate of the cognitive abilities of D/HOH children.
Braden, Kostrubla, & Reed (1994)	*N* = 21; ages 8 – 15 years; prelingually deaf	Compared group means of the WISC-III PIQ of the deaf sample with the mean PIQ of a similar sample of hearing children	The PIQs for the deaf children (*M* = 102.32, *SD* = 11.03) and the hearing children (*M* = 104.55, *SD* = 10.75) were not significantly different.	This finding supports the validity of the WISC-III PS for assessing the cognitive abilities of deaf children

Appendix B (continued)

Review of Validity Studies

Study	Sample	Methods	Results	Conclusions
WISC-III				
Slate & Fawcett (1995)	*N* = 47; mean age at WISC-R assessment 9.9 years; mean age at WISC-III assessment 12.9 years; met Arkansas state criteria for hearing loss	(1) Assessed the relationship between the WISC-III PS and the WISC-R PS as well as the test- criterion relationship between the WISC-III PS and the Wide Range Achievement Test – Revised (WRAT-R). (2) Analyzed the factor structure of the WISC-III PS with core and supplemental subtests.	(1)Correlations revealed that the WISC-R and the WISC-III PIQs were highly related ($r = .93$, $p < .01$). The relationships between the WISC-III PIQ and the three WRAT-R subscales were also all statistically significant, $p < .01$ (Reading $r = .41$; Spelling $r = .48$; Arithmetic $r = .64$). (2) Two factors were identified, Perceptual Organization and Processing Speed. However, Coding (a core subtest) did not load on the Perceptual Organization factor, but Symbol Search (a supplemental subtest) did. Due to this finding, the validity cannot be entirely supported.	(1)These findings yield evidence to support the validity of the WISC-III PS for D/HOH children (2) In general, these findings support the validity of the WISC-III PS for use with D/HOH children.
Maller (1996)	N = 110; ages 8 – 16 years; prelingually deaf	Item level data for the WISC-III VS were analyzed using a one-parameter Item Response Theory, or Rasch, Model. Results of the sample were compared to the WISC-III standardization sample.	Many items that did not fit a Rasch Model for the deaf children, did fit the Model for the hearing children. This suggests that response patterns between the two groups differ. Item difficulty estimates for the deaf children contrast with response patterns for the hearing children.	The WISC-III VS does not measure the same construct with equal accuracy across deaf and hearing children.

Appendix B (continued)

Review of Validity Studies

Study	Sample	Methods	Results	Conclusions
WISC-III				
Maller (1997)	*N* = 110; ages 8 – 16 years; prelingually deaf	Item level data for the WISC-III were analyzed using a one parameter Item Response Theory, or Rasch, Model.	Differential item functioning, or poor fit, was found for the four VS subtests (i.e., Information, Similarities, Vocabulary, and Comprehension) as well as the PS subtest, Picture Completion.	The results support that finding that the WISC-III VS does not measure the same construct equally in deaf and hearing children. However, these findings also challenge the belief that all PS subtests are valid for deaf children.
Sullivan & Montoya (1997)	*N* = 106; 6 – 16 years; hearing loss of 45 dB or greater	Assessed the factor structure of the WISC-III for the sample.	Two factors were extracted from the WISC-III for this sample of children (i.e., children with moderate, severe and profound degrees of hearing loss). These factors were Language Comprehension and Visual-Spatial Organization. The Freedom of Distractibility and Processing Speed factors do not occur among D/HOH children on the WISC-III when both the VS and PS are administered.	The WISC-III does not measure the same constructs in D/HOH and hearing children.

Appendix B (continued)

Review of Validity Studies

Study	Sample	Methods	Results	Conclusions
WAIS-III				
Kostrubala (1998)	*N* = 30; ages 18 – 75 years; deaf with the primary mode of communication ASL	Assessed the mean and standard deviation of the WAIS-III subtest scores, IQ scores, and Index scores.	The mean PIQ score (103.2) fell in the average range. The mean VIQ score (82.7) fell in the low average range and was statistically lower than the PIQ score. The standard deviation of the IQ scores are all similar to those in the general population. At the Index level, the POI and PSI fell in the average range. The VCI fell in the low average range and was statistically lower than the POI and PSI. All the mean PS subtests scores fell within the average range.	These findings support the use of the WAIS-III PS for use in assessing the cognitive abilities of deaf adults.
Meta-Anaylsis				
Braden (1990)	21 studies were included that met the following criteria (a) deaf examinees were given a version of the Wechsler PS, (b) means were reported for at least 5 subtests, and (c) did not include re-administration before a year	Calculate the unweighted and weighted scaled score means for five Subtests of the Wechsler PS (i.e., Picture Completion, Picture Arrangement, Block Design, Object Assembly and Coding/Digit Span)	The unweighted and weighted scaled score means for each of the five Wechsler PS are respectively as follows, Picture Completion (9.40, 9.41); Picture Arrangement (9.36, 9.26); Block Design (9.87,, 9.89); Object Assembly (9.96, 9.92); Coding/Digit Span (8.77, 8.77). The means for deaf individuals in within the average range for all substests. However, the Coding/Digit Span subtest mean scaled score is noticeably lower than the other mean scaled scores.	In conclusion, deaf individuals have mean Wechsler PIQs slightly below the normal hearing means, but within the average range.

Appendix C

Newsletter Posting to Listservs

The WISC-IV with Deaf and Hard-of-Hearing Students

Jeffery P. Braden, PhD
Hailey Krouse, BA
NC State University

The Wechsler Intelligence Scale for Children--Fourth Edition (WISC-IV) is a newly updated version of the most popular intelligence test used with deaf and hard-of-hearing (DHOH) students in the US. Unfortunately, although the technical data presented with the WISC-IV describes the instrument's characteristics with many special populations, no data are provided for DHOH students.

I invite psychologists and other professionals, who are members of national interest groups focusing on DHOH students, to provide me with anonymous archival data describing the performance of DHOH students on the WISC-IV. We will send participating professionals an Excel spreadsheet already formatted to receive data. Professionals will insert data, but NOT identifying information, into the spreadsheet and return it to me via email for analysis. Although the scope of such a study is typically beyond any one professional or entity, a collaborative effort could produce a large and varied sample of DHOH students. Note that, because the study will collect anonymous, archival data, it will NOT be necessary to obtain permission from individual students or their parents/guardians for participation.

This information will be useful to professionals (e.g., psychologists, speech therapists) who use the WISC-IV with DHOH populations, and will indirectly benefit DHOH students by providing these professionals with a better understanding of the strengths and limitations of the WISC-IV when used with DHOH populations. Please contact me (jeff_braden@ncsu.edu) to participate in this project.

Thanks!

Jeff Braden, PhD
Dept. of Psychology
Box 7650
North Carolina State University
Raleigh, NC 27695-7650
Phone: 919-513-7393
Fax: 919-515-1716
jeff_braden@ncsu.edu

Appendix D

Posting for Websites

We are seeking data on the clinical use of the WISC-IV with deaf and hard-of-hearing (DHOH) populations. We are posting this announcement on this Listserv because it is likely to reach qualified professionals who use the WISC-IV with DHOH populations.

The Wechsler Intelligence Scale for Children--Fourth Edition (WISC-IV) is a newly updated version of the most popular intelligence test used with DHOH students in the US. Although the technical data presented with the WISC-IV describes the instrument's characteristics with many special populations, no data are provided for DHOH students.

We invite psychologists and other professionals, who are members of national interest groups focusing on DHOH students, to provide us with anonymous archival data describing the performance of DHOH students on the WISC-IV. Although the scope of such a study is typically beyond any one professional or entity, a collaborative effort could produce a large and varied sample of DHOH students.

This information will be useful to professionals (e.g., psychologists, speech therapists) who use the WISC-IV with DHOH populations, and will indirectly benefit DHOH students by providing these professionals with a better understanding of the strengths and limitations of the WISC-IV when used with DHOH populations.

If you are interested in participating or would like to learn more about this study please view the attached consent form. If you would like to contribute to this study, please print a copy of the consent form, fill it out, and send it to the address designated at the bottom of the form. Once we have received a copy of your consent form, we will email you an Excel Spreadsheet in which we will ask you to fill in subject information and email it back to us. The North Carolina State University Institutional Review Board (IRB) has approved this study. Thank you so much for your time and contribution, this study would not be possible without your help!

Thanks!

Jeff Braden
Dept. of Psychology
Box 7650
North Carolina State University
Raleigh, NC 27695-7650
Email: jeff_braden@ncsu.edu

Hailey Krouse
B.A. Psychology
North Caroline State University
Email: hekrouse@ncsu.edu

Appendix E

Conference Recruitment Posting

Are WISC-IV Scores Valid Predictors of NC Test Scores?
Wanna Find Out??

Jeffery P. Braden, PhD
NC State University

The WISC-IV was published in 2003 to upgrade and improve its predecessor, the WISC-III. There are a number of differences in the WISC-IV, not least of which is the elimination of Verbal and Performance IQs in favor of Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed Indexes. These changes reflect an attempt to align the WISC-IV with contemporary research in cognitive abilities and neuropsychology. The test publisher (The Psychological Corporation) provides substantial data to justify this shift, and to show that WISC-IV composites are related to academic achievement. However, the WISC-IV documentation does not include information linking these new composites to achievement data unique to North Carolina such as the tests used in our ABCs assessment program.

In collaboration with NCSPA and NC State University's Psychoeducational Clinic, I would like to recruit North Carolina psychologists to contribute WISC-IV scores and NC ABCs test results to a common data base. We will use this "data cooperative" to explore the degree to which the WISC-IV relates to achievement data specific to North Carolina. We will send participating psychologists an Excel spreadsheet already formatted to receive data. Psychologists would insert data, but NOT identifying information, into the spreadsheet and return it via the Internet to me for analysis. Although the scope of such a study is typically beyond any one psychologist or entity, a collaborative effort could produce a large and varied sample of North Carolina students. Note that, because the study would collect anonymous, archival data, it would NOT be necessary to obtain permission from individual students or their parents/guardians for participation.

We will hold a meeting at NCSPA's Fall, 2004, annual meeting in Wilmington to discuss the nature of the study, solicit feedback, and recruit interested participants. Following the meeting, I will draft a proposal for review by NC State's Institutional Review Board; their approval will be required prior to any data collection. Remember, the obligation to develop validity data is shared by the test developer and the test user; let's do our part to study how the WISC-IV relates to North Carolina achievement tests, and in so doing, inform and improve professional practice in our state. I hope you will join me for a discussion or contact me (jeff_braden@ncsu.edu) for information about this project.

Appendix F

Consent Form

The WISC-IV with Deaf and Hard-of-Hearing Students

Why do this study?

The WISC-IV was released without any data describing the performance of DHOH children on the scale. There are two reasons why this is a problem: (a) previous versions of the Wechsler Scales are the most popular method for assessing the intelligence of DHOH clients, and (b) the WISC-IV (and in particular, the language-reduced scale) is changed substantially from previous editions. This means that the characteristics of the new version of the Wechsler are unknown, and changes from previous editions are sufficiently substantial to draw caution to assumptions that the new edition will have the same (excellent) characteristics as its predecessors when used with DHOH students.

What do you want from me?

I would like interested professionals who already have WISC-IV data on DHOH clients to provide item, subtest, and scale data to me. By combining data across many professionals, we can overcome the problems associated with the low incidence of deafness and obtain a sufficiently large sample to conduct meaningful psychometric analyses of the WISC-IV (e.g., reliability, validity, utility). I estimate that you would need about 15 minutes per client to enter all the test and demographic data into Excel spreadsheets. I do NOT want identifying information for clients; all data will be sent to me in two encrypted files (one for demographic data, and one for test data). I will send volunteers the encrypted Excel files, random participant ID numbers, and unique passwords to ensure confidentiality of data.

What will you do with the data?

I will aggregate the data across all contributing members of the study group. Once the data are aggregated, I will conduct reliability analyses for subtests and scales. If there are enough data, I will also conduct item analyses to explore statistical bias issues. I will ask The Psychological Corporation to supply me with a matched group of normal-hearing participants to compare and contrast the test's characteristics with DHOH vs. normal-hearing children. I plan to present and publish the results of this study at relevant professional meetings, journals, etc.

Appendix F (continued)

Consent Form

Why should I participate?
There are two reasons why you should contribute data to this study: (a) professional ethics, and (b) professional recognition. With respect to ethics (e.g., AERA, APA, NCME, 1999 Standards for educational and psychological testing 3rd ed.), test users must use data from relevant populations to inform test use. When those data are lacking, the user (and the publisher) have an obligation to collect data to understand the psychometric characteristics of the test when used with special populations. Therefore, ethics suggests that we share an obligation to get data on the WISC-IV with DHOH populations, because we are using the WISC-IV with those populations. Second, with respect to recognition, I will publish all studies under corporate authorship (Professionals Serving Deaf and Hard-of-Hearing Clients). I will list the names, titles, and institutional affiliations of all who contribute to the study in an appendix to all publications, presentations, or other dissemination of the data. If you do not want to be recognized, you may contribute data and I will list you under anonymous contributors (who will be counted, not named). There will be no other reward for participation.

What if I don't want to participate?
That's fine—don't do anything. There are no negative consequences for declining participation, and you can withdraw at any time without penalty. An electronic copy of the results of the study will be available free of charge to anybody who asks, whether or not they decide to participate. I will post the copy to the listserv host site when the analyses are complete.

OK, I'm interested. What do I do next?
Complete this letter of consent and send it to me (see below) via standard mail. I will reply

If you have any questions or concerns that arise in connection with your participation in this study, you may contact the Principal Investigator, Dr. Jeff Braden, at (919) 513-7393, jeff_braden@ncsu.edu or Hailey Krouse at hekrouse@ncsu.edu. This study has been reviewed and approved by the NCSU Institutional Review Board (IRB). If you have questions or concerns about your rights or the rights of the DHOH students, please contact the NCSU IRB administrator, Ms. Debra Paxton, at (919) 515-4514 or debra_paxton@ncsu.edu.

Appendix F (continued)

Consent Form

I have read the above and have been given the opportunity to ask questions. I agree to participate in this research with the understanding that I may withdraw without penalty at any time. I have kept a copy of this letter for my records.

_________________ ____________________________________
Date Signature

____
Password (Please type or print. Minimum 6 characters with at least 1 digit and 1 capital letter.)

____
Please type or print the e-mail address to which we should send your Excel files above.

Please send a copy of your signed/dated letter with your password to:
Jeff Braden, PhD
Dept. of Psychology
Box 7650
North Carolina State University
Raleigh, NC 27695-7650

Appendix G

First Page of WISC-IV Spreadsheet

Examinee Information									
SUBID	**Grade**	**Setting**	**Referral**	**AgeYear**	**AgeMo**	**Gender**	**Ethnic**	**PrimCom**	**HIDegree**
101M	11	1	3	16	10	0	1	0	4
102M	6	1	3	11	6	1	6	0	3
103M	8	1	3	14	11	0	3	0	4
104M	5	1	3	11	5	0	3	0	4
105M	5	1	3	11	10	0	5	0	4
106M	4	1	3	11	1	0	2	1	4
107M	4	4	3	11	6	1	3	2	2
108M	8	4	3	14	9	0	1	0	4
109M	3	4	3	9	11	0	1	0	4
110M	3	4	3	8	9	0	1	0	4
101C	9	1	2	14	11	1	3	0	4
102C	3	1	5	9	0	1	1	0	4
103C	9	1	2	14	11	1	3	0	4
104C	9	1	2	14	7	1	3	0	3
105C	7	1	2	13	0	1	3	0	4
106C	5	1	5	11	0	0	1	0	4
107C	6	1	2	11	5	0	3	0	4
108C	6	1	5	11	3	1	1	0	3
109C	9	1	5	14	8	1	1	0	4

Appendix H

Achievement Test and Other Intelligence Test Spreadsheet

	Reading Achievement Scores				Arithmetic Achievement Score			
SUBID	**Test**	**Test Date**	**Subtest/Cluster**	**SS**	**Test**	**Test Date**	**Subtest/Cluster**	**SS**
103I	KTEA-II	03/08/06	Letter & Word Recognition	67	KTEA-II	03/08/06	Math Concepts & Applications	66
	KTEA-II	03/08/06	Reading Comprehension	58	KTEA-II	03/08/06	Math Computation	72
104l					WJIII	05/02/06	Calculation	103
104I					WJIII	05/02/06	Applied Problems	116
104I					WJIII	05/02/06	Quantitative Concepts	106
102A	WJIII	09/09/04	Letter-Word Identification	66	WJIII	09/09/04	Calculation	78
102A	WJIII	09/09/04	Passage Comprehension	57	WJIII	09/09/04	Applied Problems	75
102A	WJIII	09/09/04	Reading Fluency	74	WJIII	09/09/04	Math Fluency	67
102A	WJIII	09/09/04	Broad Reading	64	WJIII	09/09/04	Broad math	73
103A	WJIII	10/20/04	Letter-Word Identification	48	WJIII	10/20/04	Calculation	75
103A	WJIII	10/20/04	Passage Comprehension	49	WJIII	10/20/04	Applied Problems	65

Appendix I

IRB Letter of Approval

North Carolina State University is a land-grant university and a constituent institution of The University of North Carolina

Office of Research
and Graduate Studies

Sponsored Programs and
Regulatory Compliance
Campus Box 7514
1 Leazar Hall
Raleigh, NC 27695-7514
919.515.7200
919.515.7721 (fax)

From: *Debra A. Paxton, Regulatory Compliance Administrator*
North Carolina State University
Institutional Review Board

Date: *July 14, 2004*

Project Title: *The WISC-IV with Deaf and Hard-of-Hearing Students*

IRB#: **165-04-7**

Dear Dr. Braden:

The research proposal named above has received administrative review and has been approved as exempt from the policy as outlined in the Code of Federal Regulations (Exemption: 46.101.b.4). Provided that the only participation of the subjects is as described in the proposal narrative, this project is exempt from further review.

NOTE:

1. This committee complies with requirements found in Title 45 part 46 of The Code of Federal Regulations.
   For NCSU projects, the Assurance Number is: FWA00003429; the IRB Number is: IRB00000330

2. Review de novo of this proposal is necessary if any significant alterations/additions are made.

Thank you,

Debra Paxton
NCSU IRB

Appendix J

IRB Letter of Addendum Approval

North Carolina State University is a land-grant university and a constituent institution of The University of North Carolina

Office of Research and Graduate Studies

Sponsored Programs and
Regulatory Compliance
Campus Box 7514
1 Leazar Hall
Raleigh, NC 27695-7514
919.515.7200
919.515.7721 (fax)

From: *Debra A. Paxton, IRB Administrator*
North Carolina State University
Institutional Review Board

Date: *November 9, 2004*

Project Title: *The WISC-IV with Deaf and Hard-of-Hearing Students*

Dear Dr. Braden:

Your addendum to the study named above has been reviewed by the IRB office, and has been approved. The addendum does not change the exempt status of your research. If you have any questions please do not hesitate to contact the IRB office at 919.515.4514.

Thank you,

Debra Paxton
NCSU IRB

Appendix K

Description of Achievement and Other Intelligence Test Data

In an attempt to meet sample size requirements, the achievement test data were aggregated in multiple ways. First, as planned, achievement scores were combined from multiple tests that claimed to measure the same academic skill. For example, all subtests the claimed to measure reading comprehension were aggregated across test batteries. Although scores were combined from multiple test batteries, the sample size across all skill domains still did not meet the minimum requirement (i.e., n = 28) for further calculations. Although these correlations were not reported, the sample sizes, means and standard deviations for the grouped achievement test scores are reported in Appendix L.

In a second attempt to meet sample size requirements, the achievement test data were aggregated across all skill domains for reading, mathematics and written language. One score from each examinee for each academic area (i.e., reading, mathematics and written language) was chosen. Scores were strategically chosen so that scores measuring broad academic skills were chosen first. For example, when aggregating scores across reading skills, if available, a broad/composite score of reading was chosen first. However, if a broad/composite reading score was unavailable for a particular examinee, than a reading comprehension score was chosen. When aggregating math scores, the scores were chosen in the following order: math composite scores, math reasoning scores, applied problems scores, math concepts scores and lastly, calculation scores. When aggregating written language scores, the scores were chosen in the following order: written expression scores, writing samples scores, and spelling scores. Although scores were aggregated across skill domains for reading, mathematics and written language, all sample sizes failed to meet the minimum requirement (i.e., n = 28). Therefore, analyses involving achievement scores could not be calculated. Although correlations were not calculated, the means and standard deviations for the aggregated scores for reading, mathematics and written language are reported in Appendix M.

As planned, intelligence test scores that claimed to measure the same construct were aggregated across test batteries. However these aggregated scores failed to meet sample size requirements (i.e., n = 28); therefore; analyses involving other intelligence test scores could not be calculated. Although correlations were not calculated, the means and standard deviations for the aggregated scores are reported in Appendix N.

Appendix L

Sample Sizes, Means and Standard Deviations of Grouped Achievement Test Domains

Academic Skill / Subtest / Composite	Test	*N*	*M*	*SD*
Word Reading		**17**	**79.35**	**19.20**
Word Reading	WIAT-II	7		
Letter-Word Identification	WJ-III	5		
Letter-Word Recognition	KTEA-II	1		
Word Reading	WRAT-4	2		
Reading	WRAT-3	2		
Psuedoword Decoding		**2**	**114**	**4.24**
Psuedoword Decoding	WIAT-II	2		
Reading Fluency		**5**	**78.80**	**8.29**
Reading Fluency	WJ-III	5		
Reading Comprehension		**18**	**89.67**	**18.56**
Reading Comprehension	WIAT-II	12		
Passage Comprehension	WJ-III	5		
Reading Comprehension	KTEA-II	1		
Reading Composite		**10**	**72.70**	**17.10**
Reading Composite	WIAT-II	4		
Broad Reading	WJ-III	6		

Appendix L (continued)

Sample Sizes, Means and Standard Deviations of Grouped Achievement Test Domains

Academic Skill / Subtest / Composite	Test	*N*	*M*	*SD*
Math Calculation		**21**	**89.38**	**17.13**
Numerical Operations	WIAT-II	11		
Calculation	WJ-III	6		
Math Computation	KTEA-II	1		
Math Computation	WRAT-4	2		
Arithmetic	WRAT-3	1		
Math Fluency		**4**	**84.50**	**15.29**
Math Fluency	WJ-III	4		
Math Reasoning		**13**	**78.31**	**18.13**
Math Reasoning	WIAT-II	9		
Math Reasoning	WJ-III	4		
Applied Problems		**1**	**N/A**	**N/A**
Applied Problems	WJ-III	1		
Math Concepts		**2**	**86.00**	**28.28**
Math Concepts and Applications	KTEA-II	1		
Quantitative Concepts	WJ-III	1		

Appendix L (continued)

Sample Sizes, Means and Standard Deviations of Grouped Achievement Test Domains

Academic Skill / Subtest / Composite	Test	*N*	*M*	*SD*
Math Composite		**10**	**78.40**	**19.14**
Math Composite	WIAT-II	5		
Broad Math	WJ-III	5		
Spelling		**11**	**94.55**	**15.31**
Spelling	WIAT-II	6		
Spelling	WJ-III	2		
Spelling	WRAT-3	1		
Spelling	WRAT-4	2		
Written Expression		**4**	**92.00**	**8.76**
Written Expression	WIAT-II	3		
Writing Samples	WJ-III	1		
Written Language Composite		**5**	**81.80**	**25.27**
Written Expression Composite	WIAT-II	1		
Written Expression	WJ-III	4		
Listening Comprehension		**2**	**58.50**	**9.19**
Listening Comprehension	WJ-III	2		
Oral Expression		**1**	**N/A**	**N/A**
Oral Expression	WIAT-II	1		

Appendix L (continued)

Sample Sizes, Means and Standard Deviations of Grouped Achievement Test Domains

Academic Skill / Subtest / Composite	Test	*N*	*M*	*SD*
Oral Language Composite		**1**	**N/A**	**N/A**
Oral Language Composite	WIAT-II	1		
Academic Fluency		**1**	**N/A**	**N/A**
Academic Fluency	WJ-III	1		
Academic Applications		**1**	**N/A**	**N/A**
Academic Applications	WJ-III	1		

Note. WIAT-II = Wechsler Individual Achievement Test-Second Edition; WJ-III = Woodcock Johnson Tests of Achievement, Third Edition; WRAT-3 = Wide Range Achievement Test, Third Edition; WRAT-4 = Wide Range Achievement Test, Fourth Edition; KTEA II = Kaufman Test of Educational Achievement, Second Edition.

Appendix M

Sample Sizes, Means and Standard Deviations of Grouped Achievement Test Skills

Academic Domain / Test: Subtest	*N*	*M*	*SD*
Reading	**24**	**76.42**	**19.33**
WJ-III: Broad Reading	6		
WIAT-II: Reading Composite	4		
WRAT-3: Reading	2		
WRAT-4: Reading	2		
WJ-III: Reading Comprehension	1		
WIAT-II: Reading Comprehension	8		
KTEA-II: Reading Comprehension	1		
Mathematics	**22**	**83.27**	**19.27**
WJ-III: Broad Math	5		
WIAT-II: Math Composite	5		
WIAT-II: Math Reasoning	5		
WJ-III: Applied Problems	1		
KTEA-II: Math Concepts & Applications	1		
WRAT-3: Arithmetic	1		
WRAT-4: Math Computation	2		
WIAT-II: Numerical Operations	2		

Appendix M (continued)

Sample Sizes, Means and Standard Deviations of Grouped Achievement Test Skills

Academic Domain / Test: Subtest	*N*	*M*	*SD*
Written Language	**13**	**85.77**	**15.30**
WJ-III: Written Expression	4		
WIAT-II: Written Language	1		
WJ-III: Writing Samples	1		
WIAT-II: Written Expression	1		
WIAT-II: Spelling	3		
WRAT-3: Spelling	1		
WRAT-4: Spelling	2		

Note. WIAT-II = Wechsler Individual Achievement Test-Second Edition; WJ-III = Woodcock Johnson Tests of Achievement, Third Edition; WRAT-3 = Wide Range Achievement Test, Third Edition; WRAT-4 = Wide Range Achievement Test, Fourth Edition; KTEA II = Kaufman Test of Educational Achievement, Second Edition.

Appendix N

Sample Sizes, Means and Standard Deviations of Grouped Intelligence Test Domains

Skill / Subtest / Composite	Test	*N*	*M*	*SD*
Memory		**2**	**82.00**	**8.49**
Memory Quotient	UNIT	2		
Reasoning		**1**	**N/A**	**N/A**
Reasoning Quotient	UNIT	1		
Symbolic Reasoning		**1**	**N/A**	**N/A**
Symbolic Quotient	UNIT	1		
Nonsymbolic Reasoning		**1**	**N/A**	**N/A**
Nonsymbolic Quotient	UNIT	1		
Performance		**1**	**N/A**	**N/A**
Performance IQ	WASI	1		
Composite		**1**	**N/A**	**N/A**
Full Scale IQ	UNIT	1		

Note. UNIT = Universal Nonverbal Intelligence Test; WASI = Wechsler Abbreviated Scale of Intelligence.

Appendix O

Restriction of Range Formula

$R_{corrected} = r_{xy} (S_x/s_x) / \sqrt{} (1 - r_{xy}^2 + r_{xy}^2 (S_x/s_x)^2)$

Where:

R_{xy} = uncorrected correlation coefficient

S_x = population standard deviation

s_x = sample standard deviation