2010

Examining the appropriateness of nonverbal measures of intelligence with deaf and hard-of-hearing children: a critical review of the literature

Martha S. Cook Klassen

Pepperdine University

Graduate School of Education and Psychology

EXAMINING THE APPROPRIATENESS OF NONVERBAL MEASURES OF INTELLIGENCE WITH DEAF AND HARD-OF-HEARING CHILDREN: A CRITICAL REVIEW OF THE LITERATURE

A clinical dissertation submitted in partial satisfaction

of the requirements for the degree of

Doctor of Psychology

by

Martha S. Cook Klassen

December, 2010

Shelly P. Harrell, Ph.D. – Dissertation Chairperson

This clinical dissertation, written by

Martha S. Cook Klassen

under the guidance of a Faculty Committee and approved by its members, has been submitted to and accepted by the Graduate Faculty in partial fulfillment of the requirements for the degree of

DOCTOR OF PSYCHOLOGY

Doctoral Committee:

Shelly P. Harrell, Ph.D., Chairperson

Susan Himelstein, Ph.D.

M. Natasha Kordus, Ph.D.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Shelly P. Harrell, Ph.D. for her support and guidance throughout this challenging endeavor. I also thank Susan Himelstein Ph.D. and Natasha Kordus, Ph.D. for their contribution as committee members. The expertise of my committee was vital to the completion of this document.

I additionally thank my family and friends for their support and encouragement that has helped me achieve this goal. I am particularly grateful to my husband, James, for his continued support and sacrifice over the years. He has had to forgo many activities while I spent time in front of the computer or at the library. I owe much to my parents for instilling in me the value of education and the belief that one can achieve anything through hard work and determination. Finally, I owe thanks to my brother-in-law, Kenneth Klassen, Ph.D., who contributed his proofreading skills to this project.

VITA

Martha S. Cook Klassen

Education

Pepperdine University, Expected Fall 2010 – Doctor of Psychology
Pepperdine University, 1997 – Master of Arts, Psychology
University of California, Irvine, 1991 – Bachelor of Arts, Psychology; minor in Comparative Cultures

Clinical Experience

Brookdale Hospital, Department of Psychiatry, Brooklyn, New York
Psychology Intern
July 2000 – June 2001

- Clinical Psychology Internship in psychiatric department of hospital, AAPIC approved site
- Provided psychotherapy services for adults, children, and families in outpatient setting
- Administered psychological and educational assessment batteries to adults and children
- Conducted individual and group therapy sessions for adults in inpatient hospital setting
- Participated in non-public school, inpatient, and emergency room rotations

Catholic Charities, Center for Psychological Service, Van Nuys, California
Practicum Student
August 1999 – July 2000

- Conducted family, individual, conjoint therapy, and intake interviewing and assessment in an outpatient psychotherapeutic setting
- Performed psychoeducational assessments for children experiencing learning challenges

Pasadena City College Disabled Student Programs and Services, Pasadena, California
Practicum Student
September 1998 – July 1999

- Administered and scored psychoeducational assessment batteries to adult students with learning or coping problems
- Interpreted test results and helped develop appropriate recommendations for the Student Educational Contract

Pepperdine Psychological and Educational Clinic, Culver City, California
Practicum Student
January 1998 – July 1999

- Conducted individual, conjoint, and family therapy in an outpatient psychotherapeutic setting
- Additional responsibilities included education therapy

Employment

Advancement for Behavior and Educational Development Interventions, Irvine, California
Case Consultant
June 2010 to Present

- Provide consultation to parents for in-home positive support intervention programs based on principals of Applied Behavior Analysis (ABA)
- Conduct Functional Behavior Assessments to identify functions of undesired behaviors and create plans to teach replacement behaviors and improve positive skills
- Lead direct staff on implementation of positive support programs

Beacon Autistic Spectrum Independence Center, Orange, California
Clinical Director
June 2008 – June 2010

- Directed clinical services for company providing in-home, ABA-based behavior intervention programs to Regional Center consumers and families
- Developed Functional Behavior Assessments and implemented positive support programs
- Worked with clients and families to train parents and caregivers on behavior intervention techniques
- Supervised and evaluated performance of case managers and interventionist staff members

Beacon Day School, Orange, California
Clinical Coordinator
January 2006 – January 2010

- Provided individual counseling and group social skills training to students diagnosed with Asperger's Syndrome and High-Functioning Autism
- Participated in California Department of Education on-site review of non-public school

M. J. Lang, Ph.D., Inc., Orange, California
Psychological Assistant
August 2004 – May 2008

- Participated in conducting neuropsychological assessments for children and adults in private practice setting
- Contributed to report development and writing

Levy and Levy, a Psychological Corporation, Los Angeles, California
Psychological Assistant
June 2003 – May 2004

- Provided individual psychotherapy to adults at inpatient rehabilitation residential setting
- Completed intake assessments and termination reports for patients

Eric and Joan Erikson Center for Adolescent Advancement, Van Nuys, California
Child Care Worker, Relief Staff
February 1997 – August 1999

- Worked with severely emotionally disturbed adolescent residents in a milieu treatment setting

- Assisted with daily activities and provided emotional and environmental support for residents

Therapeutic Education Center, Santa Ana, California
Program Aide
September 1994 – August 1995
- Assisted at a non-public school for severely emotionally disturbed, special education students, grades K-12
- Coordinated nutrition and transportation programs
- Supported teachers during classroom activities and field trips

Olive Crest Treatment Centers, Santa Ana, California
Child Care Worker, Primary Staff
May 1990 – September 1994
- Provided emotional and environmental support for severely emotionally disturbed adolescents in 6-bed resident group home
- Coordinated independent living skills and nutrition programs

Additional Professional Training
Behavior Intervention Specialists, Inc.
- Functional Analysis and Treatment of Severe Behavior Disorders, December 2009

Pyramid Educational Consultants
- PECS Basic Training, May 2009

S.U.C.S.E.S.S. Project 2006-2007, Orange County Department of Education
- Thinking about You, Thinking about Me – Presented by Michelle Garcia Winner, March 2007
- Overview of Inside Out: What Makes the Person with Asperger's or High-Functioning Autism Tick? – Presented by Michelle Garcia Winner, March 2007
- Implementing Social Thinking Concepts and Vocabulary into our Schools and Home – Presented by Michelle Garcia Winner, March 2007

The Institute for Applied Behavior Analysis (IABA)
- 2005 Summer Institute on Assessment and Analysis of Severe and Challenging Behavior

ABSTRACT

There are a variety of assessment instruments available today that are widely used to assess the intellectual abilities of children. Specific considerations should be made when using these instruments to assess the deaf and hard-of-hearing population. This critical review of the literature begins with a brief overview of the D-HH population, a general history of intellectual assessment, and assessment considerations that are specific to the D-HH population. Information obtained from available literature regarding the internal consistency of six assessment instruments is presented. The instruments reviewed include the UNIT, the Leiter-R, the WISC-IV, the SB5, the CTONI, and the CAS.

The results indicated that all of the measures examined show sufficient reliability and validity when applied to the general population. However, Braden (2005) has suggested that a measurement instrument is appropriate for a particular group when similar reliability and validity values are found for that group as for the general population. Of the measures examined, CTONI reported internal consistency studies with a subgroup of D-HH children in the manual. Additional independent research on the internal consistency of the UNIT and the WISC-IV is available. The literature review suggests that there are several factors that can influence test results when working with the D-HH population that have not been examined through independent research to date. Areas of interest for future research are presented.

Chapter I. Introduction and Background Literature

A variety of tests that assess intellectual ability for diagnostic and placement purposes are invaluable for use in psychological and educational settings serving the general population. However, there are fewer assessment measures that are appropriate for administration to deaf and hard-of-hearing (D-HH) individuals. Because many intellectual assessment instruments rely heavily on verbal communication abilities, their application with the D-HH population has been questioned. It has been suggested that other measures that rely on a reduced amount of or no verbal language are more appropriate for use when assessing the intellectual abilities of the D-HH. Some current instruments available for use with the D-HH population include the following: Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998), Leiter International Performance Scale – Revised (Roid & Miller, 1997), Wechsler Intelligence Scale for Children, Fourth Edition (Wechlser, 2003), Stanford-Binet, Fifth Edition (Roid, 2003), Comprehensive Test of Nonverbal Intelligence (Hammill, Pearson & Wiederholt, 1997), and the Comprehensive Assessment System (Naglieri & Das, 1997). Relatively more research has been conduced to examine the performance scales of earlier versions of the Wechsler test, which found that similar scores were obtained between D-HH and hearing subjects. Relatively fewer studies have been conducted on the other assessment measures many of the results have been inconclusive (Maller, 2003b). The purpose of this proposed dissertation is to provide a critical review of the available literature that pertains to these assessment instruments and to determine whether they are applicable for use with the D-HH population.

The current chapter presents a summary of the preliminary literature review in

order to provide the background and foundation for the proposed study. To begin, some of the factors and issues related to the D-HH population, particularly those related to the assessment of intellectual abilities, will be defined and explored. This will be followed by a brief general overview of the history of intellectual assessment, followed by specific consideration of the D-HH population. Finally, issues associated with nonverbal assessment and the determination of the reliability and validity of test measures will be discussed. The chapter will conclude with a statement of rationale for the study and delineation of specific research objectives.

The Deaf and Hard-of-Hearing Population

It is difficult to determine exactly how many people experience hearing loss, making up the deaf and hard of hearing (D-HH) population. National surveys estimate that there are approximately 20 million people in the United States with hearing impairment (Henwood & Pope-Davis, 1994; Holt, Hotto & Cole, 1994). The National Center for Health Statistics reported that there are 32.5 million adults (approximately 15% of the population) who have some type of hearing difficulty (National Center for Health and Statistics [NCHS], n.d.). This suggests that hearing loss is the most widespread disability in this country and probably also in the rest of the world. This hearing loss ranges on a continuum from people who have no hearing to those with mild hearing losses that may interfere with conversation but not impair the use of a telephone.

A Gallaudet Research Institute study, using data from 1990-1991, developed rough statistical estimates of the number of hearing impaired people in the United States by grouping individuals by level of hearing ability. The results suggested that there are approximately 20,295,000 individuals with “hearing problems” among the United States

population. Of those considered to be deaf, 421,000 fell into the "deaf in both ears" category; 552,000 of them "cannot hear and understand any speech"; and 1,152,000 were those who "at best, can hear and understand words shouted into the better ear." The Gallaudet Research Institute also reported that 968,000 D-HH children between the ages of three and 17 years-old were living in the United States, based upon the 1990-1991 data (Harrington, 2004). Of those children, over 135,000 have hearing loss that may hinder academic success (NCHS, n.d.).

The varied group of people with some type of hearing loss can differ on four main levels: medical and audiological conditions, communication abilities and preferences, educational settings and achievement, and sociocultural characteristics and behaviors (Brauer, Braden, Pollard & Hardy-Braz, 1998). Medical and audiological differences are often described in terms of cause, onset, severity, and type of deafness. Deafness is usually caused either by genetic or chromosomal conditions, or by disease or trauma. If deafness is present at birth, it is termed congenital, and deafness that occurs after birth is identified as adventitious (Brauer et al., 1998). Deafness that occurs prior to acquiring basic verbal language and speech is considered prelingual, and deafness that occurs after the acquisition of basic verbal language and speech is called postlingual (Brauer et al., 1998; Henwood & Pope-Davis, 1994). Prevocational deafness refers to hearing loss after one acquires verbal language skills but before the age of 19 years (Henwood & Pope-Davis, 1994).

The degree of hearing loss ranges from mild to profound. The decibel (dB) is the unit of measurement for the loudness, or intensity, of a sound. The frequency, or pitch, of a sound is measured in Hertz (Hz). The degree of deafness is measured as the decibel

level required to recognize a sound at any given frequency (Brauer et al., 1998; Marschark, 1997). Normal hearing is defined as having losses up to 25 dB in the better ear (Marschark, 1997). Mild hearing loss (25-40 dB) suggests difficulty only with hearing faint speech. Moderate loss (40-55 dB) describes difficulty with understanding normal speech. Moderately severe hearing loss (55 to 70) indicates frequent difficulty with hearing loud speech. Severe loss (70-90 dB) suggests a person can only understand shouted speech. Finally, profound hearing loss (90-110 dB) indicates an inability to understand speech at all (Brauer et al., 1998). The degree of hearing loss can vary across different frequency levels. Humans can usually hear sounds ranging from 20 to 20,000 Hz. Hearing losses that affect the range of 500-2000 Hz are often the most troublesome because that is the frequency range of the most important aspects of spoken language (Marschark, 1997).

A conductive hearing loss occurs from damage to or a malfunction of the middle ear. This results in an inability to transmit vibrations through the middle ear mechanisms. A sensorineural impairment is caused by permanent damage to the inner ear, typically involving the cochlea or its connection to the auditory nerve. A central hearing loss occurs from permanent damage to the central nervous system involving the auditory centers of the brain, or the "brain end," (p. 28) of the auditory nerve (Henwood & Pope-Davis, 1994; Marschark, 1997). Hearing loss can also involve one or both ears, resulting in unilateral or bilateral impairment (Henwood & Pope-Davis, 1994).

Communication, Education, and Identity in the Deaf and Hard-of-Hearing Population

Modes of communication among the D-HH vary. The most common mode of

manual communication used by the D-HH is American Sign Language (ASL). It is a language with a unique syntax and grammatical structure. It employs the use of conceptual signs that are not associated with English or any spoken language (Henwood & Pope-Davis, 1994). "ASL also makes use of the particular capabilities of physical communication in 3-D space" (Brauer et al., 1998, p. 299).

Other verbal communication systems include oralism, total communication, signed English, and pidgin signed English. Oralism includes emphasis on understanding speech by lip reading, writing, reading, and the use of hearing aids to amplify sound. Total communication uses methods associated with oralism combined with sign language and finger spelling. The goal of oralism is to integrate the D-HH person into the mainstream hearing society. The total communication approach uses any and all methods available to help the D-HH person communicate more effectively. Signed English is another system of manual communication, using finger spelling and signs that are directly related to spoken English in grammar and syntax. Another system is pidgin signed English, which uses aspects of ASL and signed English (Henwood & Pope-Davis, 1994).

Written English ability and speech intelligibility varies among deaf people, and communication preferences vary. Approximately 95% of deaf people are born into and raised by hearing families who use oral and vocal languages. As a result, the mastery of the English language, written and spoken, is a major developmental challenge to most deaf people, and deaf children's English proficiency is often limited. This should not reflect negatively on deafness or a deaf person's intelligence; rather it indicates that

inadequacies in family and school settings to nurture linguistic and cognitive development exist (Brauer et al., 1998).

The educational development of D-HH people is strongly affected by the onset and degree of deafness. Prelingual deafness often results in limited spoken English abilities among preschoolers. Syntactical and grammatical English ability among deaf children usually lags behind vocabulary acquisition. Whether this is viewed as a linguistic difference or linguistic deficiency, the acquisition of reading, writing, spelling, and mathematical skills is limited and often continue into adulthood. While there may be a lag occurring, the mastery of ASL is often as advanced as the mastery of English among hearing individuals (Brauer et al., 1998).

The education of a child with mild hearing loss may require only supplementary support in a regular classroom. More severe deafness requires more intensive educational support. Three distinct approaches are used to educate D-HH children: the oral method, total communication method, and bilingual-bicultural (bi-bi) method. The oral method emphasizes the acquisition of speech and discourages the use of sign language. The total communication method emphasizes the acquisition of language through English signs and speech, often simultaneously. The bi-bi method rejects spoken and signed English and emphasizes the use of ASL. Written and signed forms of English are introduced after a child has mastered ASL. Controversy exists over which method is best, and outcome studies have shown minimal or inconsistent differences among the methods (Brauer et al., 1998).

Socioculturally, deaf people may identify more or less with the Deaf community. Most people object to the term “hearing impaired” because it implies an abnormality.

Many also do not use the term "hearing loss" (p. 301) because there has been no loss from the perspective of a congenitally deaf person. Many use the term "deaf" to describe themselves (Brauer et al., 1998, p. 301). This group includes people with audiological conditions, as well as people who identify themselves as members who share common communication and culture (Steinberg, 1991). ASL is their primary mode of communication; identity is founded in the shared linguistic, historical, and cultural traditions based on ASL. They are active in the social and civic events of the Deaf community, and they would describe themselves as members of the Deaf community (Brauer et al., 1998). On the other hand, many people who are audiologically deaf do not identify with the Deaf community. This group may describe themselves as "hard of hearing." Speech is often their primary mode of communication, and identity is based upon the shared linguistic, historical, and cultural traditions of the "normal" hearing community. In summary, "deaf" refers to those who generally have profound to severe hearing loss, "hard of hearing" refers to those with mild or moderate deafness and identify with the hearing society, and Deaf refers to those who are typically deaf and identify with the Deaf community. An additional group, the "late deafened," (p. 302) includes people who experience severe or profound hearing loss later in life (Brauer et al., 1998).

Assessment and Psychological Services for the Deaf and Hard-of-Hearing

Recent years have shown a progression toward providing unbiased psychological services for culturally diverse groups. Psychological training should involve increased consideration of cultural specificity, individual uniqueness, and a notion of human universality (Henwood & Pope-Davis, 1994). Cultural diversity has traditionally been

defined in terms of ethnic and racial backgrounds. However, more recent definitions have expanded to include gender, sexual orientation, and religion. Henwood and Pope-Davis (1994) further suggest that clients with disabilities should be included as a culturally diverse group that also needs particular attention from psychologists.

According to Pollard (1996), approximately 40,000 deaf people in the United States suffer from serious psychopathology, but only about 2% of those who need mental health services receive them (Leigh & Pollard, 2003; Pollard, 1996). Fortunately, psychology is evolving to better serve deaf consumers. One contributing factor to better services was the acknowledgment of ASL as a legitimate language, thus changing the manner in which deaf people's linguistic, intellectual, and psychological characteristics were viewed (Pollard, 1996).

Other important factors in improving services for the D-HH population resulted from changes in legislation recognizing the importance of services for the d-hh. They started with the Rehabilitation Act of 1973, particularly section 504, and the Education for All Handicapped Children Act (PL 94-142) in 1975. The combination of these laws created legislation to assure free and appropriate public education (FAPE) for children with disabilities. Later amendments of PL 94-142 occurred with the Education of the Handicapped Amendments of 1986 (PL 99-457) and the 1990 Individuals with Disabilities Education Act (IDEA) (PL 101-476). IDEA is now used in reference to the entire PL 94-142 sequence of legislation. These laws along with other requirements mandate the early identification of hearing losses among school-aged children. They also ensure unbiased evaluation of deaf children using alternative and appropriate modes of communication (Marschark, 1997).

A third contributing factor occurred in 1990 when the American Psychological Association's Division 22 (Rehabilitation Psychology) recognized the first special-interest section focusing on the specific population of those who are d-hh. As a result, the number of deaf and hearing professionals with the clinical, linguistic, sociocultural, and ethical knowledge required to work with the D-HH is increasing (Brauer et al., 1998; Pollard, 1996).

Standardized Assessment of the Deaf and Hard-of-Hearing

Providing appropriate and effective psychological services often requires the administration of standardized tests. The variations among the D-HH people must be considered when selecting, administrating, and interpreting tests (Brauer et al., 1998). One factor that can affect psychological services for the D-HH population is the difference in communication that usually exists between the psychologist and the client. The use of interpreters can be intrusive and alter the dynamics of psychological work (Henwood & Pope-Davis, 1994). Familiarity and comfort with the modes of communication used by the D-HH is recommended to promote accurate assessment (Steinberg, 1991). In addition to fluency in sign language, one should be knowledgeable about deafness and its many implications (Pollard, 1996).

Along with improving psychological services for the d-hh, there is an urgent need for developing improved assessment and treatment resources for use with this population (Pollard, 1996). Because formal psychological assessment tools rely heavily on verbal abilities, the use of such tools may be inappropriate or lead to incorrect assessments of the hearing impaired client (Henwood & Pope-Davis, 1994). Problems with language communication can often result in the misdiagnosis of D-HH individuals, and at times

have resulted in individuals incorrectly being diagnosed as mentally retarded or psychotic. In addition, the use of tests which rely heavily on verbal abilities often results in the assessment of the deaf person's language difficulties rather than provide valuable psychological information (Vernon & Andrews, 1990).

Over the years, tests of intellectual ability that were developed for use with the general population have been applied, with varying degrees of success, to the assessment of deaf and h-h individuals. However, the D-HH population has been found to have similar scores on performance measures, leading to the development of nonverbal assessment instruments (Maller, 2003b). As the need to provide quality services for diverse populations has increased, interest in nonverbal assessment has grown (Athansiou, 2000; McCallum, Bracken & Wasserman, 2001). Nonverbal assessment instruments are those that involve test administration requiring no receptive or expressive language demands from either the examinee or the examiner. Unfortunately, many tests described as "nonverbal" are actually language-reduced instruments, which still require verbal directions for the examinee (McCallum et al., 2001).

Nonverbal Measures

It has been a long-held opinion among researchers that verbally loaded intellectual assessment instruments may skew the results of individuals whose backgrounds differ from the norm. Verbally loaded measures assume the examinee has been adequately exposed to and developed the use of a standard form of the dominant language. However, when individuals do not meet these assumptions, such as the D-HH population, the use of nonverbal, or language reduced, intellectual assessment measures is indicated (Braden & Athansiou, 2005).

Braden and Athansiou (2005) identified several issues related to the use of nonverbal intellectual assessment instruments. For example, many measures that have been characterized as "nonverbal" involve little or no language for understanding of directions, have limited linguistic content, and allow for a nonverbal response to test items. However, A true nonverbal test is one that reduces or altogether eliminates the need for examinees to use verbal language when understanding, processing, or responding to test items. Few of these truly nonverbal tests are in existence today (Braden & Athansiou, 2005).

It has been argued whether nonverbal measures assess nonverbal intelligence, or if they measure intelligence nonverbally. Some have proposed that the cognitive processes underlying nonverbal tasks are different than those employed during verbal tasks. However, factor analysis has not supported this difference between verbally mediated cognitive processes and visual reasoning. It has instead supported the idea that the cognitive processes underlying intelligence are consistent and independent of their language loading (Braden & Athansiou, 2005).

The decision to use nonverbal assessment instruments can occur a priori, or be based upon information other than test results, such as the hearing status of an examinee. The decision can also be made a posteriori, or after scores from other tests, generally more language-loaded tests, have been obtained. For instance, inconsistent results may be obtained between a language-loaded and a nonverbal test. This may be interpreted as an indication that a combined score may not truly reflect an individual's intellectual level, and the higher of the two scores may be used to estimate one's ability (Braden & Athansiou, 2005).

Another issue that Braden and Athansiou (2005) identified is related to the lack of evidence related to response processes. "Response process," as identified in the *Standards for Educational and Psychological Testing* (1999), means that there is evidence of the fit between the construct measured by an assessment instrument and the nature of the performance or response in which the examinee is actually engaged. Analysis of individuals' responses and questioning performance strategies and response development can yield evidence for response process. This issue is relevant to nonverbal assessment measures because an individual might respond to a nonverbal task through verbally mediated strategies. However, there is little report of evidence-based response process information related to nonverbal measures (Braden & Athansiou, 2005).

The use of nonverbal assessment instruments may narrow the intended construct of an assessment. For example, the three-tiered conceptualization of cognitive abilities has placed general intellectual ability at the top of the hierarchy (*g*), followed by second-order factors (crystallized, fluid, visualization, and long-term retrieval abilities) and then a variety of specific abilities. When using only nonverbal assessment instruments, the examiner must remove tasks measuring crystallized ability. Therefore, performance on nonverbal intellectual instruments may be less representative of one's general intellectual ability (Braden & Athansiou, 2005).

The use of nonverbal intellectual measures can prevent the incorrect diagnosis of individuals who may appear to be intellectually deficient when performing language-laden tasks. However, it has also been suggested that the results on nonverbal tests may overestimate one's ability to function in language-oriented environments and lead to inappropriate academic or vocational placement. It is necessary for clinicians to consider

such issues when conducting language-loaded or language-reduced assessment procedures and their potential consequences on future placements (Braden & Athansiou, 2005).

Finally, Braden and Athansiou (2005) indicated that examinee characteristics that may influence performance on language-laden or language-reduced tasks must be carefully considered when choosing which instruments to administer. For example, lack of fluency may impact performance on language-laden tasks, and visual impairment may impact performance of a language-reduced task. In addition, the administration of nonverbal measures involves the use of gestures. Examinees from cultures in which gestures are common or the D-HH population may be more comfortable with gestural administration while others may find the process disconcerting. Examiners must also consider the potential for misunderstanding or error when using what may be novel procedures of nonverbal measures (Braden & Athansiou, 2005).

The development of nonverbal measures of intelligence has continued in order to improve with respect to validity and reliability. It has been generally accepted that D-HH and hearing people tend to obtain similar scores on the Wechsler Performance measures. These tests have been the most widely used with the D-HH population in North America. Results from other measures when used for the D-HH have shown more variable results, but whether characteristics of the measurement process or of the sample group itself are influencing the results is unknown. Possible reasons for these differences might be that the sample group does not accurately reflect the D-HH population, that there may be additional unidentified disabilities among the sample group, and that the meanings of the test items may be different for D-HH people due to different learning opportunities and

exposure. The D-HH have also been shown to generally perform better on tasks that require the manipulation of objects, but have been shown to score consistently lower than hearing counterparts on motor-free nonverbal measures. Possible explanations for this include a better understanding of the task when manual dexterity is required and materials can be manipulated. and the use of verbal mediation when solving motor-free tasks (Maller, 2003b).

Verbal measures of intellectual functioning have generally been regarded as inappropriate for use with the D-HH population due to concerns about the validity of such measures, as well as test and item bias. However, many psychologists continue to give verbal measures to D-HH people to obtain clinical information, and the results are often described in reports. Performance on verbal measures has also been shown to be a better predictor of academic achievement than performance test results, and verbal tests can identify verbal strengths and weaknesses within individual D-HH persons (Maller, 2003b).

General History of Intellectual Assessment

A brief history of intellectual assessment will provide a larger context for understanding the development of attempts at intellectual assessment with the D-HH population. The earliest attempt at intellectual assessment dates back to imperial China, where a standardized civil service testing program was used (Anastasi, 1997; Thorndike, 1997; Gregory, 1996). Because there was not a hereditary aristocracy, a measure of human cognitive abilities was needed (Thorndike, 1997). Rudimentary testing by the Chinese emperor of his officials, performed every third year to determine their fitness for office, dates back to 2200 B.C. Over the centuries, the testing was refined, and written

exams were developed during the Han dynasty, from 202 B.C. to 200 A.D. The tests covered five areas, including civil law, military affairs, agriculture, revenue, and geography. In about 1370, the Chinese developed a final testing system in which candidates for public office spent days and nights in isolated booths composing essays on assigned topics. In addition to being a grueling process, this selection process was never validated by the Chinese. This examination system was eventually abolished by royal decree in 1906 in response to widespread discontent (Gregory, 1996).

About the time the civil services testing program ended in China, intellectual assessment was underway in Western civilization. Several trends during the late 19th century were precursors to the development of intellectual measurement as it is known today. One trend influencing the development of intellectual measurement was the increasing interest in the humane treatment of mentally retarded and insane individuals. In prior times, such people were neglected and ridiculed, and even tortured. However, as concern for the proper care of this group increased, the need for uniform criteria to identify and classify such individuals was needed. As described by Anastasi (1997) Esquirol, a French physician, published a two-volume work in 1938 describing what is today identified as "mental retardation" (p. 33). He attempted to develop several procedures to identify the presence and degree of mental retardation, but concluded that one's use of language is the most dependable criterion of an individual's intellectual level. Verbal ability continues to be considered an important part of the concept of intelligence, and many of today's intelligence tests involve much verbal content. Another French physician, Seguin, pioneered methods for educating individuals identified as mentally retarded, which involved sense-training and muscle-training techniques. Some

of these techniques were incorporated into later performance or nonverbal tests of intelligence (Anastasi, 1997).

A second trend influencing the development of intellectual measurement was the movement toward universal compulsory education. This was concurrent with the idea that only an educated population could make strong self-governing decisions. By the end of the 19th century, state laws had been passed in the United States that enacted public education and provided modest funding. In Europe, countries such as France had enacted compulsory education by 1880 (Thorndike, 1997).

The universal compulsory education movement introduced education to children of families who would not have previously sought education. In the United States, compulsory education was also seen as a way to "make Americans" of the many immigrants entering the country. The heterogeneity among the children served by compulsory education was great, and the failure rate was as high as 50% at times. The high failure rate was viewed as a waste of educational resources, so ways to identify those who would be best served by education were sought. Intelligence testing was seen as a means of determining who would be appropriate for successful education (Thorndike, 1997, p. 4).

A third trend of the times which influenced intellectual measurement was that psychology was becoming a quantitative science, based upon the model of physics. Work by psychologists suggested that it was possible to quantify various psychological characteristics (Thorndike, 1997). However, the early experimental psychologists were not particularly concerned with the measurement of individual differences, and developing generalized descriptions of human behavior was typically the goal.

Individual differences were either ignored or viewed as an error that made the resulting generalizations approximate rather than exact. This scientific approach to psychology led to an emphasis on sensory phenomena, which will be seen later in the first psychological tests. It also led to a standardization of procedure, which is particularly important in psychological testing today (Anastasi, 1997).

Several individuals are recognized for their significant contributions to the early development of intellectual measurement. The English biologist, Galton, was primarily responsible for launching the testing movement (Anastasi, 1997). Galton developed a theory of human ability and its measurement, based upon his idea that each person is a blank slate at birth and that knowledge is acquired through sensory experience. He believed that a person with greater sensory acuity and faster sensory information processing would be able to gain more from sensory experience. Therefore, measuring sensory acuity and reaction times should produce an index of intelligence (Thorndike, 1997). Galton was a pioneer in the application of rating-scale and questionnaire methods and in the use of the free association technique used for a variety of psychological purposes. In addition, Galton developed statistical methods for the analysis of data on individual differences (Anastasi, 1997).

Cattell, an American psychologist, coined the term "mental test" (p. 4) in 1890 and brought Galton's ideas to the United States. He proposed a program of mental testing in order to establish a standard metric for intellectual ability assessment (Thorndike, 1997). Cattell shared Galton's view that intellectual functions could be measured through tests of sensory discrimination and reaction time. Cattell's tests were comparable to other test series developed during the 1890's. They were preferred

because simple functions could be precisely and accurately measured, and objective measures for more complex functions seemed almost hopeless at that time. However, Cattell's tests were shown to have little correspondence from one test to another, and there was little or no relation to other estimates of intellectual level based on teachers' ratings or academic grades (Thorndike, 1997; Wasserman & Tulsky, 2005).

At this time, Binet was beginning to study the development of intelligence in France. Using his two daughters as subjects, he created a series of brief games, which were tasks of graded difficulty intended to measure intellectual development. He published his results in a series of papers in which he criticized the Galton/Cattell approach to measuring intelligence. Binet proposed that it is necessary to observe the performance of complex mental acts in order to measure such complex mental processes (Thorndike, 1997).

Binet became involved in a group called the Free Society for the Psychological Study of the Child, a group of concerned parents and professionals interested in improving the effectiveness of the schools by identifying the causes of school failure. Two types of failure situations were identified: children who could learn the material but would not do so, and children who could not learn. These children were identified as "malicious" and "stupid," (p. 5) in respective order, and the goal of the Society was to differentiate the two groups (Thorndike, 1997). In 1904, Binet was assigned by the Minister of Public Instruction to study procedures for the education of retarded children. This work, with the help of his colleague, Simon, led to the development of the first formal measure of intelligence. The first Simon-Binet Scale was called the 1905 scale (Gregory, 1996; Wasserman & Tulsky, 2005).

The 1905 Scale differed from previously developed measures of intelligence in several ways. First, it attempted to determine a child's general level of mental development, using a heterogeneous group of tasks. As a result, the goal was classification rather than measurement. The 1905 scale was also brief and practical, taking less than an hour to administer. In addition, Binet and Simon developed the 1905 scale to directly measure what they viewed as the essential factor of intelligence, which is practical judgment. The test did not examine lower level abilities related to sensory, motor, or perceptual skills. Another difference was that the items on the 1905 scale were arranged by level of difficulty rather than content. The scale could be used to assess levels of intelligence from severe mental retardation to the highest levels of giftedness. Finally, the tests were very verbally laden, which also reflects Binet's departure from Galton's ideas (Gregory, 1996).

A revision was made in 1908, which almost doubled the number of problems from the 1905 scale. The tests were also grouped into age levels, based upon the performance of about 300 normal children of ages three to 13 years. For example, the three-year level was identified as all tests passed by 80 to 90% of normal three-year-old's, and so on. The performance of a child on the 1908 scale could then be described as a "mental level." As this concept underwent various translations and adaptation, the term "mental age" became a common substitute for "mental level" (p. 37). This idea was easy for the public to grasp and probably led to the popularization of intelligence testing (Anastasi, 1997). And, despite Binet's emphasis that a child's mental level should not be seen as an absolute measure of intelligence, the concept would influence the character of intelligence testing for the rest of the century (Gregory, 1996). A third revision of the

Binet-Simon scale occurred in 1911, which was also the year of Binet's untimely death (Anastasi, 1997). The revision in 1911 extended the application of the test into the adult range (Gregory, 1996).

During the years in which the Binet-Simon scale was being developed in France, additional advances in intellectual assessment were being made in other parts of the world. For example, in the United States, Boas had measured 1,500 American school children on various traits, and compared the results to teacher reports of intellectual ability. A developmental influence on the quality of intelligence was also being identified. For instance, Ebbinghaus performed a study that showed older children were better able to complete mutilated sentences. Binet's test integrated the ideas of the time, and created an empirically based measuring device that related increasing intellectual ability to maturation (Thorndike, 1997).

In the United States, the pursuit of a measure of intellectual ability was strong, being supported by the American Psychological Association (APA) with an appointment of a committee on testing in 1885. Thorndike, a student of Cattell, was one early contributor who would continue to study intelligence and measurement for the next 40 years (Thorndike, 1997). Additionally, Goddard, who was responsible for bringing the Binet-Simon test to the United States, was the research director for the Vineland Training School in 1905 (Thorndike, 1997). He studied ways to classify and educate "feebleminded" children (Gregory, 1996, p. 17). He went to France in 1906 to meet with Binet and become familiar with the Binet-Simon scale. In 1908, when Binet published one of his revisions, Goddard promoted an American version for wide use (Thorndike,

1997) by translating the scale with some minor changes to make it more appropriate for use with children in America (Gregory, 1996).

Using the translated Binet-Simon scale, Goddard tested 378 residents of the Vineland Training school, terming those whose mental age was 2 years or lower as "idiots," those with mental ages of three to seven years as "imbeciles," and those with mental ages of eight to 12 years as "feebleminded." He also tested 1,547 normal children with the same scale. He termed children whose mental age was four or more years behind their chronological age as "feebleminded." He also found that this group comprised 3% of the tested sample and recommended that these children should be segregated and prevented from "contaminating society" (Gregory, 1996, pp. 17-18).

In 1910, Goddard was invited to Ellis Island to help examine immigrants entering the United States and became one of the most influential psychologists in America of the early 1900's. Goddard and his assistants administered English translations of the Simon-Binet scale to newly arrived immigrants. This was done in order demonstrate that feeblemindedness among immigrants occurred at a higher rate than among the general population in America and that their average intelligence level was low. The tests were administered through translators; the immigrants were assessed very soon after landing in the United States; and norms based upon the original French test were used to calculate the results. Some obvious problems with this method were that the immigrants were likely feeling tired, frightened, and confused; many of them had little formal education in their homelands, and then, upon entering a new country, they were required to complete an intellectual assessment instrument. Based on the conditions, it is not surprising that so many recent immigrants were identified as "feebleminded" (p. 18). Goddard's work and

its outcome were strongly influenced by the social ideologies of the time. His contribution to psychology can now serve as a reminder of how psychological tests can, even with good intentions, be misused (Gregory, 1996).

Another American, Terman, had been working in his doctoral research on tasks similar to Binet's. The two corresponded between 1904 and 1906, and Terman adapted some of the tests from Binet's 1905 version for his own studies. There also seems to be some mutual influences on the development of both men's pursuits. Terman later accepted a position at Stanford University, where he continued developing tests similar to Binet's for use in the schools. In 1916, he produced the Stanford revision of the Binet-Simon scale, which later became known as the Stanford-Binet in its 1937 revision (Thorndike, 1997). The Stanford-Binet was a substantial revision of the Binet-Simon scale. For example, Terman was the first to use the abbreviation IQ after he suggested that the Intelligence Quotient be multiplied by 100 in order to remove fractions. The number of items on the test was also increased, and the instructions were clear and organized for the administration and scoring of the test. In addition, the test was developed for use with the mentally retarded, normal children, normal adults, and "superior" adults. Finally, the standardization of the test was improved by using a carefully selected representative sample. However, the Stanford-Binet continued to rely heavily on verbal skills (Gregory, 1996).

In the early 1900's, many psychologists believed that the Stanford-Binet scales were not "entirely appropriate for non-English speaking subjects, illiterates, and the speech and hearing impaired" (Gregory, 1996, p. 19). As a result, several performance

scales developed in the 1910's continue to influence many instruments and subtests to this day.

Early evaluators used form board tasks, based upon the Seguin Form Board test, in which the subject arranged a variety of blocks into cut-out shapes as quickly as possible (Dearborn, Anderson & Christiansen, 1916; Pintner & Patterson, 1916; Thorndike, 1997). A similar task with the added component of blindfolding the examinee continues to be used today as a subtest of the Halstead-Reitan neuropsychological test battery (Gregory, 1996). In addition, Knox in 1914 developed several nonverbal tests for use with Ellis Island immigrants that required no verbal responses. The instructions of each task were also demonstrated nonverbally to ensure the subject understood. Knox's tests included a digit-symbol substitution task that continues to be seen on most Wechsler tests today. Additional nonverbal performance tests were developed by Pintner and Paterson in 1916. They developed a series of 15 performance scales that used form boards, puzzles, and object assembly tasks (Pintner & Patterson, 1916). The object assembly tasks have continued to be important components of intelligence measures (Gregory, 1996). Kohs developed the Block Design test in 1920 (Kohs, 1920), which has been a component of the Wechsler scales (Gregory, 1996). The Porteus Maze task was also developed in the early 20th century. This task is an effective assessment tool, although it is not widely used today (Gregory, 1996).

The Stanford-Binet remained the standard of intelligence testing for decades, and new tests were validated through correlational studies with this assessment. The most recent revision of the Stanford-Binet occurred in 1986. However, the Wechsler scales also became a popular alternative. Unlike the Stanford-Binet that only provided a global

IQ score, the Wechsler scales provided a Full Scale IQ as well as a Verbal IQ and a Performance IQ (Gregory, 1996).

At the time of WWI, Yerkes, a Harvard psychology professor, proposed to the U.S. government that its 1.75 million military recruits should undergo intelligence assessment for appropriate classification and assignment. Yerkes chaired a committee to develop such an instrument along with members including Goddard and Terman. The resulting instruments were the Army Alpha and the Army Beta tests, which were to have a significant influence on intelligence measures for the following decades (Gregory, 1996).

The Army Alpha was comprised of eight subtests that relied heavily on verbal skills. These were designed to evaluate average to high-functioning recruits who were fluent in English and could read and write. "The eight tests were: (1) following oral directions, (2) arithmetical reasoning, (3) practical judgment, (4) synonym-antonym, (5) disarranged sentences, (6) number series completion, (7) analogies, and (8) information" (Gregory, 1996, p. 21; Wasserman & Tulsky, 2005).

The Army Beta consisted of nonverbal tasks such as visual-perceptual and motor testing for people who were illiterate or whose first language was not English (Wasserman & Tulsky, 2005; Gregory, 1996). This measure included seven tasks such as tracing a path through a maze and visualizing how many blocks were represented in a three-dimensional image. The instructions of the Army Beta tests were explained through pictorial and gestural methods to reduce the effects of English difficulties. These were given by an examiner and an assistant who were positioned atop a platform.

However, the recruits were often sitting in such a way that they could neither see nor hear the instructions (Gregory, 1996).

While many Army Alpha and Army Beta tests were conducted, it is unclear if the army really used the information for the placement of new recruits. Yerkes proposed in his memoirs that the Army could have functioned with increased efficiency if they had used the test results. However, many army officials questioned the validity of the results due to the unclear and often confusing instructions. Many earned a score of zero, not because they were unable to perform the test, but because they were not familiar with the new instruments. Some were noted to have fallen asleep during the administration because they were unable to make sense of the instructions to complete this new type of testing (Gregory, 1996).

One positive result of the Army Alpha and Army Beta tests was that psychology gained experience in the psychometrics of test construction. This facilitated numerous correlation coefficients and the use of multiple correlations in test data analysis. In contrast, in his book *A Study of American Intelligence*, published in 1923, Brigham examined test results from ethnic and immigrant groups and used his data to promote the idea that African-American, Mediterranean immigrants, and Alpine immigrants were intellectually inferior. His conclusion was that racial intermixture would cause American intelligence to deteriorate. This is another example of the misuse of assessment instruments through inappropriate use with a population and/or flawed statistical analysis (Gregory, 1996).

After the Army's use of group tests, schools and colleges were eager to learn more about these tests that could be both administered to any group of people with almost

anyone administering and scoring. The Army Alpha and Army Beta tests became the template that would influence many intelligence, achievement and college entrance tests for years to come. In 1916, Terman developed the Stanford revisions of the Binet-Simon tests, which resulted in the Stanford Revision and Extension of the Binet-Simon Scales. This scale would become the most popular assessment measure for decades for several reasons. The Stanford-Binet was the most extensive and thorough revision of the Simon-Binet scale and the standardization procedure was the most ambitious and rigorous at the time. In addition, a comprehensive examiner's guide aided with the ease of administration, and the use of the intelligence quotient (IQ) became the new standard for intelligence tests. Terman again revised the Stanford-Binet scales in 1937, and the scale has been revised four additional times since his death in 1956 (Wassserman & Tulsky, 2005).

Another well-known measure that can trace its roots to the Army Alpha and Army Beta tests is the Wechsler scales (Gregory, 1996; Wasserman & Tulsky, 2005). Wechsler's scales surpassed the Stanford-Binet scales as the most widely used intellectual assessment measure in the 1950s and 1960s. However, most of the tasks included in the Wechsler scales were novel or original, and Wechsler's strength seemed to be in synthesizing testing materials that were already in existence. His initial scale, the Wechsler-Bellevue Scale, developed in 1939, quickly gained popularity. This was partly due to the lack of tests for use with adults and the integration of verbal and performance tasks into one test battery. In addition, Wechsler co-normed the scales with other commonly used tests and he used a normative sample procedure that was sophisticated for the time. Wechsler also emphasized psychometric rigor, which introduced the

deviation IQ, which allowed for the ranking of an individual's performance relative to others in the same age group. Wechsler later introduced the Wechsler Intelligence Scale for Children in 1949, the Wechsler Adult Intelligence Scale in 1955, and the Wechsler Preschool and Primary Scale of Intelligence in 1967. All of these tests have been revised over the years and are currently in use (Wasserman & Tulsky, 2005).

History of Intellectual Assessment with the Deaf and Hard-of-Hearing Population

Many interesting historical trends have influenced the specific area of assessing the intelligence of the d-hh. Some of these have been beneficial to the D-HH population, while others have seemed detrimental. Because the estimation of one's intellectual capabilities have traditionally been associated with one's performance on standardized intellectual measures, the appropriateness of such instruments with this specific population has often influenced the perception that psychology as a field has held regarding this group.

Moores (1996) has suggested that the historical evolution of research related to the intellectual functioning of the deaf and hard-of-hearing (D-HH) has occurred in three phases. The first phase described the "deaf as inferior" (p. 160) to hearing counterparts. This was largely based upon research studies by Pintner in the early 20th century, which suggested that D-HH groups tended to score relatively lower on measures of intellectual functioning than hearing samples. Next came the "deaf as concrete" (p. 160) phase. Support of this phase was based upon the work of Myklebust, who concluded that, although differences in performance on intellectual measures did not support inferior intelligence, differences in the development of verbal language would alter the perceptual and conceptual functioning of the d-hh. As a result, their reasoning abilities would be

qualitatively different. The third stage, "deaf as intellectually normal," (p. 161) developed after research by Vernon determined that the D-HH population did rather well on measures of intellectual functioning and performed at or above the mean or median scores of control groups (Moores, 1996). A growing interest in cognitive psychology at this time also led to a greater interest in the cognitive functioning of the D-HH and a reduction on the focus of deafness.

"Deaf as Inferior" stage. One of the earliest articles published on the topic of intellectual assessment of the deaf and hard-of-hearing was titled "Doubtful Cases" by Greenburger (1889). Greenburger was the principal of the Institute for the Improved Instruction of Deaf-Mutes in New York and was involved in the assessment of children for placement in the school. He cited several cases in which children who were initially assessed to be suitable for admission to an "asylum for idiots" (p. 98) began to show considerable improvements after several months of instruction. Later evaluation of the children determined that they were not mentally deficient, but really belonged in a school for the deaf. Although such cases were rare, Greenburger identified and described procedures that would aid in determining which students were good candidates for the school for the deaf and which students did not have sufficient intellectual ability for placement.

His methods for assessing the deaf children were primitive by today's standards, but Greenburger (1889) outlined several methods that he employed. To determine hearing ability, Greenburger recommended that the examiner stand where the examinee could not see the mouth of the examiner while the examiner vocalized a list of speech sounds and sound combinations. The examinee was then asked to repeat the sounds.

Greenburger also recommended that a box of marbles or tiles of different shapes or colors could be used to determine if the deaf child could identify different quantities of the items or sort by color or shape.

Although Greenburger's (1889) recommended procedures for evaluation were not standardized, he did present a process of evaluation that appeared to be based upon extensive personal experience. From a historical perspective, Greenburger's article can be viewed as an interesting precursor to the research and debates on effectively evaluating the intellectual abilities of the deaf population that would come in the following century. Although he believed that the deaf population could benefit from specialized instruction, the message that the deaf probably would not attain the same ability levels as those in the general hearing population was evident.

Other early articles would recommend the use of measurement instruments for the assessment of the deaf population. Dearborn et al. (1916) described several performance instruments that were recommended for use when examining the abilities of immigrants or people speaking a "foreign language," who were "deaf and dumb," (p. 445) or suffering from other speech defects. Tests including the Color-Form Test, various form-board tests, the "Triangle" Performance Test, and the Chair Construction Test were recommended as supplementary measures of the Binet-Simon Scale, which was in wide use at the time. The authors divided the subjects into groups based on age ranges and then measured and compared the average time required to complete the various tasks. These measures were "tried out on a few normal children," (p. 446) and they presented their findings. These tests were administered to hearing subjects using verbal instructions, however, and there was no attempt to administer the tests to the

recommended groups or administer them through a nonverbal mode of communication (Dearborn et al., 1916).

Early researchers, like Dearborn et al. (1916), tended to make the assumption that performance tests, by nature, were similarly applicable to the D-HH population as to the general population. These assumptions tended to be based upon research using hearing subjects. Pintner and Patterson (1916) attempted to address the issue of accurately and effectively assessing the D-HH population by using scores obtained on the Seguin Form Board test administered directly to D-HH subjects. The task was explained through hand gestures and the deaf subjects were reported to easily understand the instructions. A hearing group were also administered the same test and the instructions were verbally presented. The examiners divided the children into groups based on year of age and hearing status. They then measured the average number of errors and the average time required to complete the tasks. The standard deviations were also computed compared along with the average scores between the groups.

Pintner and Patterson (1916) found that the deaf subjects were approximately one year behind the hearing subjects. After one year the deaf children showed relatively fewer errors and smaller average deviations from the group mean. However, the hearing children showed improved completion times. As a result, it was concluded that the deaf children showed greater relative improvement in form board ability and had more homogenous scores, but they were still "backward" (p. 237) compared to the group of hearing children.

In 1919, Pintner presented a non-language intelligence test that he was developing, which would evolve into the Pintner Non-Language Test. His goal was to

develop a "set of tests that involved no language…" so that "…the illiterate, the foreigner and the deaf will all be given an equal chance with the hearing English-speaking literate individual" (p. 199). He used the followings tests: Knox Cube Test, Easy Learning Test (much like the Wechsler Digit-Symbol Coding subtest), Hard Learning Test, Drawing Completion Test, Reversed Drawing Test, and the Picture Reconstruction Test. Each of these tasks had a time limit.

Although Pintner (1919) expressed his intention of developing a non-language test that could accurately measure the intellectual abilities of the D-HH population, among others, he used English speaking children and university students as subjects in his study. Pintner, and most of his contemporaries, continued to find that the D-HH population performed at a lower level than hearing counterparts.

In a later article, published in 1931, Pintner administered the Pintner Non-Language Mental Test to deaf and hearing subjects using gestures and completion of examples to communicate the instructions for all subjects. A review of the Pintner Non-Language Test indicated that it was developed to have no written or spoken language requirements for use with kindergarten, first- and second-grade children. The reliability and validity information was described as "limited." However, a split-half reliability of 0.90 was reported using scores from 111 first grade and 186 second grade children. In addition, a validity of 0.61 with the Stanford-Binet mental ages, using scores from 80 kindergartners, was reported (Whitmer, 1949).

Pintner (1931) found that the scores from the deaf sample were similar to those of the hearing sample. This was "in marked contrast to most comparisons between the deaf and hearing on group intelligence tests, where the deaf are usually very far behind the

hearing" (p. 362). He speculated that the deaf children included in the sample might have been above average in ability, or that the results from the hearing sample were below the norm. Pintner recommended further study to establish the reason for this outcome. He also proposed that the deaf students may have an advantage over the hearing students in understanding the non-language, pantomimed instructions because they naturally develop the skills to attend to non-verbal cues, such as gestures and facial expressions. As a result, the deaf children may have benefited more from the pantomimed instructions than the hearing children (Pintner, 1931).

MacKane (1933), a researcher in the United States, replicated a study conducted by Drever and Collins in Scotland. MacKane reported that the Drever and Collins study found little difference, if any, between the intellectual ability of the deaf and hearing children included in their study. MacKane administered a group of performance tests to deaf students. He also gathered a group of hearing students that were matched by gender, within one month of the same chronological age, of the same racial origin and socio-economic status, and their parents or grandparents were of the same nationality. MacKane then compared the results from the deaf and hearing groups.

MacKane's results supported the previous findings by Drever and Collins that the deaf subjects were no more than one year "retarded" in any age group. However, he did not find superior performance by the deaf subjects at any age level, as was found by Drever and Collins (MacKane, 1933). Other contemporaries of MacKane found that the difference in test scores seen between deaf and hearing subjects was less than previously indicated.

In her paper about the measurement of the mental and educational abilities of the

deaf, Lane (1938) proposed that the ability to measure the mental level of the deaf depends largely upon the definition of intelligence accepted by the examiner. Lane suggested that when intelligence is defined as:

> …the ability to think abstractly and to manage ideas and symbols, we shall probably never find an adequate measure for the deaf child, because any test of this kind involves linguistic ability. All school grades…are weighted on the side of this *abstract* intelligence." She suggested that intelligence instead be defined as, "the ability to use judgment in adjusting to various situations presented in the environment… (p. 169)

Lane (1938) reported having positive results using the Randall's Island Performance Series when assessing young deaf children, and the deaf group had been found to have equal ability as a similar hearing group. Lane also indicated that intelligence levels among the deaf were distributed along the normal curve with a median quotient at approximately 100. She acknowledged that her conclusion went against most other investigators at the time, but she was of the opinion that the retardation found by other researchers was due to the use of tests that were not entirely non-verbal, the effects of group test administration procedures, or the examiner not being familiar with the deaf child. Lane also questioned how the concept of speed and the need to work as quickly as possible on a timed task could be adequately communicated to a deaf child during an assessment.

Despite the work by MacKane (1933) and Lane (1938), most researchers continued to find that the D-HH tended to perform at a lower level on performance tasks than hearing counterparts. In 1939, Zeckel published a study in which he reported below

average performance by deaf children on a performance task up to the age of 12 years, while the hearing children began to show average performance at the age of 10 years. However, older children from both groups did not tend to show below-average scores. He indicated that, even without the verbal element present in a measure, the deaf children showed a "backwardness of intelligence" (p. 122) when compared to the hearing.

Zeckel (1939) purported that language represents symbols for concepts, and congenital deafness eliminates this ability to symbolically express objects or ideas. This lack of language would prevent the verbal intellect from developing, or keep it at a lower level. This lack of practice in the development of the verbal intellect would also "blunt the intellect" (p. 123). For the hearing children, the natural practice of translating concepts into symbols is simultaneously exercising abilities in all other areas of the intellect.

"Deaf as Concrete" stage. Other researchers would continue to develop the use of non-verbal measures for the assessment of intelligence among the D-HH population. The administration procedures, reliability, and validity of tests would continue to be explored and refined. Hiskey (1941) also developed a scale, the Hiskey Test of Learning Aptitude for Young Deaf Children. This test would be unique because it compared the performance of D-HH children to that of the normative sample of other D-HH children. A review of the test indicated that the norms were based on scores from 466 deaf children who attended residential schools in the Midwest region. The correlations of each item group to the entire scale ranged from 0.63 to 0.84, and the split-half reliability correlation was 0.96. A validity of 0.829 was reported when comparing scores from 380 children with scores obtained from the Stanford-Binet Test (Sloan, 1959).

Eventually, it would be generally accepted that the intellectual ability of the D-HH population was not significantly different than that of the hearing population. In the 1950's the focus of research on the intellectual assessment of the D-HH shifted to understanding the conceptual and perceptual abilities of the deaf, rather than only reporting the performance of the D-HH on various tasks. This would lead to the next stage in research that was suggested by Moores (1996), the "Deaf as Concrete" Stage. This stage was largely based upon the work of Helmer Myklebust. Researchers during this stage suggested that differences in performance on intellectual measures did not mean the deaf were intellectually inferior. However, differences in the development of verbal language would alter the perceptual and conceptual functioning of the deaf and result in qualitatively different reasoning abilities.

The relationship between the development of the human mind and the development of language ability was outlined by Theo Irion (1941). He wrote that, in the past, humans were viewed as having minds, which given "…proper stimulation, would develop the human language within the individual" (p. 364). It had been proposed "…that the development of human language is, itself, the development of the human mind" (p. 365). Irion suggested that providing the deaf with some form of language was doing more than merely providing a language, but in addition was actually "building their mentality" (p. 365). This building of mentality would allow the surroundings and the world around an individual to develop significance and meaning for that individual.

Irion (1941) purported that meanings are first understood concretely, as in "object-situations" (p. 366) connections. However, humans soon learn to allow a symbol or sign to represent an object. When this occurs, nouns like *book* or *hammer* can

substitute for the actual article. For example, one can know the meaning of the word *hammer* without seeing a hammer through the understanding of the symbol, sign or word *hammer*. The ability develops to use symbols or signs for other meanings, such as the relationships between objects, activities or experiences. These signs and symbols are called "words" (p. 367). One does not need to have an actual object or experience to be aware of its meaning because the signs or symbols can be used to develop meaningful reactions to them. Language then becomes necessary to develop mentality. One then has the ability to move away from concrete objects and concrete experiences and perform mental reflection. Language is then not just a tool for expressing thoughts or ideas but is also a tool used in the development of thoughts and ideas.

With language, one can take from individual experiences certain qualities or relationships that do not exist by themselves but exist as part of a larger experience. Though language, these qualities and relationships can be treated as if they were independently existing phenomena or abstractions. An example to illustrate this notion was provided by Irion (1941). He explained that one would never be able to locate a "piece of whiteness" (p. 367) anywhere. However, one can experience a white cloud, white paper, white cloth, or some other concrete thing that is white. After whiteness is experienced in many concrete situations, whiteness can then be discussed as if it was an independent entity, and the abstraction has been created. This higher-level mentality largely involves abstractions, which would be impossible to conduct without the development of signs, symbols, or language.

Once one has the ability to become conscious of facts, concepts, abstractions and relationships, these mental constructs can be put together in various ways and reacted to

by creating new experiences or insights. This process is called thinking and reasoning, which is another form of experiencing the signs, symbols, or language. Once the thinking and reasoning process is completed, one usually wants to check one's thoughts against reality for verification. Irion (1941) concluded that thinking and reasoning were dependent upon language, and mentality was dependent upon thinking and reasoning. Therefore, mentality is dependent upon the development of language.

Irion (1941) suggested that the deaf individuals could develop a sign system other than through vocalization and sound. However, this type of "language symbolism" (p. 371) system was crude and did not allow for the fine thought discriminations that were possible with language symbols or signs. Irion concluded his paper by encouraging teachers working with the deaf to do everything possible to provide enriched environments for the deaf, "who in many ways can appear to be dull, and by slow and tedious process finally develop in them the sign and symbol experience" (p. 371). This language experience opens to them a wide range of mental possibilities.

Oleron (1950) conducted a study on the abstract reasoning of the deaf to explore the idea that the verbal basis of abstract reasoning prevented the deaf from obtaining the same level of intellectual ability as the hearing. He published a report of his study in which the performance of deaf students on the Raven's Progressive Matrices, 1938 edition, was evaluated. The directions of the matrices could be easily understood through pantomime, and there was no time limit to the test.

Oleron (1950) reported that the scores from the deaf subjects were inferior to the normative (hearing) sample. Oleron also found that the mental development of the deaf subjects appeared to cease at the age of 18 years, that the mental growth of the deaf

children tested was slower than average, and the differences from the norm increased with age. From the results, he concluded that the general lowering of scores was based upon the inability of the deaf students to complete an "abstract" mental task, and that the mental development of abstract thinking ability did not continue to increase over time as he theorized occurring in hearing individuals.

Oleron (1950) compared the performance of the deaf on the Raven's Progressive Matrices test and the age of onset of deafness. He found that nine of the ten subjects who became deaf after the age of five years scored above the median of the sample. From this he concluded that the benefit of exposure to spoken language became significant at the age of five or six years. He pointed out that this age of onset of deafness is also the age at which a child can continue to retain the use of oral speech. Oleron theorized that there was a certain degree of maturity reached at this age that allowed for certain attainments to become fixed in the individual. Finally, Oleron did not find an effect of residual hearing on test scores, and he was unable to offer a possible reason for this result (Oleron, 1950).

Based upon the results from the Raven's Progressive Matrices test administration to deaf individuals, Oleron (1950) concluded that the deaf do show inferiority in the development of abstract thought. This difference was attributed to the close connection between language and abstract thought. However, Oleron cautioned that this inferiority did not cover the entire field of mental abilities, and results from past studies showing consistent abilities on performance tasks were not contradicted by his results. He suggested that the deaf had difficulty solving tasks such as the Raven's Progressive Matrices because the abstract task of deducing a principle was required. Performance tasks, in contrast, present all relevant clues needed to solve the task.

In 1957, Goetzinger and Rousey conducted a study in which the Wechsler Performance Scale and the Knox Cube test were administered to deaf adolescents. They began their report by stating that, although research has shown variability in results between one type of measurement instrument administered to the deaf and another, there was a general consensus among educators and psychologists that the deaf were intellectually within normal limits. They explained that the new trend in research was to explore the conceptual and perceptual abilities of deaf children. It was purported that because hearing loss was believed to limit the development of reasoning ability, the concept formation and reasoning abilities of the deaf were inferior to those of hearing counterparts. However, there seemed to be variation in results related to the specific area of visual perception ability.

Goetzinger and Rousey (1957) found that the deaf group produced lower scores on the Wechsler Picture Arrangement and Picture Completion subtests than on the other three subtests (Block Design, Digit-Symbols, and Object Assembly). They hypothesized that the Picture Arrangement and Picture Completion subtests require subvocalization to complete. Therefore, this reduction in scores was attributed to the limited language concepts and usage abilities among the deaf subjects.

In his book *The Psychology of Deafness: Sensory Deprivation, Learning, and Adjustment*, Myklebust (1964) further explored the issue of the relationship between deafness and intelligence. Myklebust questioned the impact that the verbal and non-verbal experiences of the individuals who are deaf early in life had on the actualization of intellectual potential. He suggested that deafness did not impact all abstract processes in a uniform manner and that some types of abstract reasoning processes did not appear to

be affected by deafness. Nevertheless, Myklebust asserted that intelligence was related to the development of abstraction, and this relationship appeared to be closely associated with the limitations in verbal language that resulted from deafness. The inferiority in abstraction seen among the deaf was a secondary and reciprocal condition to limited verbal language abilities and not an indication of true mental retardation.

"Deaf as Intellectually Normal" stage. Moores (1996) based the third stage of research with the deaf, on work by McKay Vernon. At that time, there was increasing interest in cognitive psychology, which moved to focus of research toward a better understanding of the cognitive functioning of the D-HH and less focus on the effects of deafness.

Prior to Vernon, Hans Furth (1964) suggested that the deaf do not lack reasoning abilities when compared to the hearing. However, due to variations in experience, the deaf develop different reasoning abilities. Furth summarized that language had not been shown to influence intellectual development in any direct, general, or otherwise decisive manner. He also suggested that the influence of language, direct or indirect, might accelerate the development of intellectual ability. Language might provide opportunities for additional experience by allowing for the communication through ready symbols (words) and linguistic habits for specific situations. Based on his assumptions, Furth suggested that individuals who have limited linguistic experience are not permanently retarded in intellectual ability. They may, however, be temporarily retarded in a developmental phase due to lack of sufficient general experience, and may be retarded on certain specific tasks that would otherwise be facilitated by the availability of word symbols or linguistic habits.

Furth (1964) then explained that the successful performance of deaf persons on intellectual tasks indicated that there was efficient functioning of a symbolic system that did not rely on verbal symbols. In addition, a deaf person might act in a manner that appears to be unintelligent to a hearing person, but the action is reasonable and based on the deaf individual's different type of experience. Furth suggested that experiential interaction with the environment is more responsible for intellectual development than language. As a result, a hearing child may simply have increased opportunities through language to interact with the environment. Thus, language affords more opportunity for more experience, but is not the only component necessary for the development of intellectual abilities. If language was the sole contributor to intellectual development, then people deprived of language during their formative years would remain permanently intellectually delayed. Furth also suggested that future research should focus on children's "nonverbal" cognitive development. He indicated that this would draw the focus away from the presumption that linguistic ability is necessary for the development of intellectual ability.

McCay Vernon (1967) examined the results from studies examining the performance of deaf and hard-of-hearing children on 16 different performance scales. He reported that in two studies that compared the results of congenitally deaf with adventitiously deaf children, both groups performed equally well. Vernon suggested that, when important factors, such as degree of hearing loss, are held constant, and the age of onset of deafness is varied, the levels of cognitive functions were found to be similar. He then concluded that the thinking process is not related to the level of language

development because the language ability of the adventitiously deaf was superior to that of the congenitally deaf.

Four studies included in Vernon's (1967) report compared performance test scores between hearing preschool children and those who were deaf. One experiment showed that both groups performed equally well, while the other three indicated differences in performance with one favoring the hearing and the other two the deaf. Vernon suggested that these result reflected similar performance among the preschool aged children. He further described that the deaf children included in these studies were totally or almost totally without verbal language ability, while the hearing group had normal linguistic abilities. Again, Vernon concluded that language does not affect cognitive development because there was similar ability seen between the two preschool-aged groups.

When looking at the comparison of results between all hearing-impaired children with control groups or test norms, Vernon (1967) found that the results of seven studies showed similar performance while eleven of the studies indicated inferior performance by the deaf. However, these differences tended to be relatively small. Upon closer examination of the studies, Vernon reported that the research conducted by people who are experienced in working with the deaf population produced unanimous results indicating that the deaf groups performed equally as well as the hearing groups. Vernon also stated that when test scores are comparable between individuals who have some sort of language limitation to those who do not, there is an implication that language is not involved in the overt examples of the thinking process during performance tasks. Finally, Vernon suggested that, if language were a factor in cognition, then the lingually deafened

child, who would have no language upon the start of school, would perform very poorly on performance tasks. These children would then continue to show disproportionately rapid improvement as language abilities developed.

Vernon (1967) ended his report by asserting three conclusions on the relationship of language to cognition. First, there is no functional relationship between verbal language and cognitive thought processes. Second, verbal language is not the mediating symbol system of thought. Third, there is no relationship between the formation of concepts and one's level of verbal development.

Vernon's conclusions were supported by later research by Watts (1979). Watts studied the influence of language on the development of quantitative, spatial, and social thinking of deaf children by using three group of children: Deaf; partially hearing, and normal hearing. The measured intelligence levels of the three groups were controlled in order to represent the normal spread of ability.

To measure quantitative thinking, Watts (1979) administered conservation tasks to the three groups of children. Dissimilar results were found with the hearing group showing superior results over the deaf and partially hearing groups across all ages. Watts suggested that if cognitive development were based upon language, then the partially hearing group would have been expected to perform better than the deaf group.

Spatial reasoning was measured by administering tasks requiring an understanding of horizontal and vertical concepts. The general results showed strong similarities in performance between the deaf and hearing children. Watts (1979) reported that the youngest deaf children were significantly inferior to the hearing children, but only on the first portion of the task. This difference seemed to be due to a lack of

experience, however, and the deaf children readily completed the task after a short demonstration. Watts concluded that the lack of linguistic experience did not substantially influence the children's ability to understand spatial transformations.

Social thinking ability was measured through the arrangement of comic strip pictures. Watts (1979) found that the deaf children required more time to understand what the task required of them. However, once they grasped the idea, they easily proceeded with the test. The deaf children were also described as having a sustained interest in this task because its nonverbal nature allowed them to readily display their knowledge. Watts reported that the hearing group showed lower scores, and the deaf and partially-hearing children in the study showed similar results. He suggested that the superior language capacity of the hearing children did not provide them with an advantage, and their relatively lower scores could be attributed to an adverse affect caused by the dominance of language over their thought. The hearing children's search for the words to describe the nonverbal tasks may have masked the meaning. Watts concluded that these results indicated that development was not primarily based on language ability.

While he did not want to understate the importance of language acquisition for the deaf child, Watts (1979) indicated that he would like to see an increased emphasis in education on developing the ability to think operatively. He suggested that the development of knowledge in young children occurs through actions upon the environment and actions with the environment. Providing active experiences for deaf children leads them to concept formation, which can then be supplemented with functional language skills.

"Different does not Mean Deficient" stage. More recently, Marshark (2003) has proposed that we are currently in a fourth phase of research on the intellectual abilities of the d-hh. He has based this on recent research by Tharpe, Ashmead, and Rothpletz (2002).

Tharpe et al., (2002) conducted a study to explore the importance of early environmental stimulation on the development of functional organization of sensory modalities. They attempted to explore past conflicting results in which deaf individuals were shown to either have better visual scanning ability or to have deficits in visual attention ability when compared to hearing counterparts. Some have reasoned that D-HH individuals might develop better visual attention ability as a result of its extensive use for receptive sign language and speechreading purposes. On the other hand, others have suggested that the lack of normal access to sound may lead to an underdevelopment of certain abilities that require the integration of visual and auditory input.

Tharpe et al. (2002) measured visual attention using three groups of students: profoundly and prelingually deaf children who have had a cochlear implant for an average of three years; prelingually deaf children who use a hearing aide; and a group of children whose hearing was within normal limits. None of the children from any group were born to deaf parents, nor did they have any deaf siblings. A continuous performance task (CPT) of visual attention administered on a computer and a paper-and-pencil letter cancellation task were administered to the three groups of children. Parents and teachers also completed the Child Behavior Checklist (CBCL) for each child. In addition, group mean scores on the Test of Nonverbal Intelligence (TONI-3) were

reported to show no significant difference between the three groups or any pair of groups and all individuals performed within the average range.

The results of their study showed that there were very few differences between groups of children on either of the visual attention tasks (Tharpe et al., 2002). All of the children in all groups performed well on the visual attention tasks. The cochlear implant group showed statistically significantly lower scores on the computer administered CPT than the hearing group, but because all groups performed well, the authors questioned the clinical significance of this difference. In addition, parents of the children with hearing loss tended to report a higher level of behavior problems than the hearing group.

Tharpe et al. (2002) found a significant association between CPT performance and nonverbal intellectual level when the effects of intelligence and age were statistically controlled. This type of analysis was reported to not have been included in previous studies. Tharpe et al. (2002) postulated that possible differences between visual attention ability seen in previous studies may have been due more to intellectual differences between subjects than hearing status.

Tharpe et al. (2002) also reported that their sample size was relatively small (n=28), compared to previous studies. However, upon performing power analysis to determine if their sample size was adequate and found "ample statistical power to find group differences of the size previously reported, if those differences existed" (p. 411).

In addition, Tharpe et al. (2002) further indicated that the difference in performance reported in past results was based on the use of different strategies in monitoring the environment used by the three groups. For example, it has been suggested that children with cochlear implants can better use their hearing to monitor the

environment than children who use hearing aids. Therefore, errors on visual attention measures would occur when children were visually scanning the environment and not attending to the task. Tharpe et al. (2002) reported that most of the children kept their eyes "fixated" on the computer screen, and only one error occurred when a subject looked away from the screen. Tharpe et al. suggested additional research on environmental distractions on the performance of visual attention tasks, but they indicated that their findings did not support past theories.

The difference seen in CBCL ratings showed that parents tended to rate the behavior of deaf children as more problematic than parents with hearing children. However, the teachers' results tended to rate the children's behaviors similarly across all groups (Tharpe et al., 2002). This was attributed to different criteria used between parents and teachers. In addition, no differences were found between results on the visual attention measures and CBCL results.

Tharpe et al. (2002) also concluded that caution must be used when applying their results to the general deaf population. First, the deaf children included in their study were aided relatively early in life and may have followed a "different developmental course" than deaf children who receive hearing aid later on. Second, Tharpe et al. suggested that the variation seen in the past and current results might mean that visual attention of deaf children is task dependent. Different visual attention tasks or batteries of tasks may produce different results. Finally, no association between preferred mode of communication and visual attention ability was found when the researchers co-varied the effects of age and intelligence level. However, Tharpe et al. (2002) indicated that they did not initially control for this when arranging their experimental groups. One study

was reported to have found no relationship between mode of communication and visual attention ability. Tharpe et al. (2002) suggested additional research to explore this as well as the general interaction between hearing ability and visual attention. Focus on general intellectual ability, the role of the environmental context in the performance of a task, and the way children function when performing multiple tasks simultaneously will also become important.

Assessment Considerations Specific to the Deaf and Hard-of Hearing Population

There are numerous issues that must be considered when conducting the intellectual assessment of the deaf and hard of hearing (D-HH) population. Leigh and Pollard (2003) identified five factors that can help determine if a psychological measure is appropriate for use with deaf individuals:

> purpose and goodness of fit to the evaluation question, the way instructions are conveyed, the nature and content of the items or tasks, the response modality, and the scoring methods and norms. The test data collection tool will be biased if, in any of these five areas, there is evidence that hearing loss, fund of information, limited competency in English, or sensory or sociocultural aspects of life as a deaf or hard-of-hearing individual would play an undesirable role. They include the goal of the assessment, preferred mode of communication, the etiology of an individual's deafness and the presence of additional disabilities, as well as other developmental and psychosocial issues. (p. 207)

When assessing D-HH individuals, one of the first issues that must be understood is the goal of the assessment. The evaluation of intellectual,

communicative, and personal/social aspects of D-HH children is usually intended for educational planning or other interventions to facilitate development. In order to effectively accomplish this goal, a researcher must recognize that there are three main discrepancies that should be examined which pertain to the performance of D-HH children on assessment measures (Simeonsson, Wax & White, 2001).

The first discrepancy to identify is the gap between a child's intellectual and academic achievement levels. With many D-HH children, scores on cognitive measures fall in the normal range when compared to hearing peers, but achievement results indicate substantially lower performance. Therefore, one goal when assessing D-HH children is to identify any characteristics associated with measured cognitive ability level and one's effective learning and achievement abilities. The second discrepancy to consider when conducting assessment is the difference between a child's linguistic and cognitive competence. As a result, the performance of D-HH children on verbally laden measures is generally interpreted as language problems and not considered to suggest intellectual deficits. Third, discrepancies in cognitive ability and personal or social functioning may be misunderstood when D-HH children are evaluated by linguistically demanding assessment instruments. D-HH children can produce results that suggest deviant or pathological behavior on measures requiring reading of questions or other verbal abilities (Simeonsson et al., 2001).

Another issue related to the evaluation of D-HH children is related to communication. The preferred mode of communication of each child should be considered, and a method to maximize communication between the examiner and

examinee should be established. Even when oral/aural communication is preferred by the examinee, the receptive language ability of the D-HH child continues to be limited. Thus, caution is advised when conducting assessments using the oral/aural communication method. When children prefer to communicate primarily using sign language, the child's familiarity with spoken/written English language must also be considered along with the accuracy of translation. Many verbally based assessment items are difficult to translate from spoken/written English to ASL. This is particularly true when idioms are present in test materials. There are also differences in the temporal sequencing and grammatical structure of English and ASL, which make accurate translations difficult (Simeonsson et al., 2001). Specific considerations related to mode of communication and the administration of measurement instruments will be explored later in this section.

The etiology of an individual's deafness and the presence of multiple disabilities is another issue that must be considered when conducting the assessment of a D-HH child. For example, performance IQ results among children with known etiologies of deafness, such as illness, have been found to be lower than children whose cause of deafness is unknown. In addition, many D-HH children have additional disabilities. Additional disabilities, such as blindness, present other challenges to the assessment process. Many D-HH children with multiple disabilities often have learning disabilities or other issues related to mental health. As a result, the etiology of deafness and the

presence of additional disabilities must be evaluated when conducting an assessment (Simeonsson et al., 2001).

Finally, the developmental history and psychosocial experiences of D-HH children must be considered when conducting an assessment. For example, intellectual, linguistic, emotional, and social assessment of D-HH children born to deaf parents is usually more similar to the results obtained from hearing peers. However, early experiences common to D-HH children born to hearing parents can sometimes lead to delayed development. In addition, D-HH children often miss incidental learning experiences which hearing peers are able to absorb, so differences in scores on linguistic, social, and occasionally nonverbal intelligence tests can be seen (Simeonsson et al., 2001). As a result, an exploration of a D-HH child's developmental and psychosocial history can yield important information related to assessment results.

Differing modes of communication can impact the process of assessment of D-HH individuals if the standardized procedures must be accommodated or modified. In order to facilitate the administration of assessment measures to individuals with disabilities, test accommodations can frequently be effective. Test accommodations create specialized circumstances to facilitate performance, such as enlarged print of test materials, and do not alter the construct, or ability, that is measured by the test. However, test modifications may change the content of an instrument and therefore may alter the construct. Therefore, when a test is adapted for the D-HH population, it must be determined if the change is an accommodation or a modification of the measure (Maller, 2003b).

As cited by Simeonsson et al. (2001), Sullivan (1982) found that D-HH children

averaged an 18-point increase in scores on the Wechsler Intelligence Scale for Children, Revised Edition (WISC-R) when adapted instructions were presented using total communication than compared to those who received the standardized verbal instructions. Maller (2003b) reports that several attempts to adapt measures of intelligence have been made specifically for the D-HH population, including signed instructions of nonverbal tests and signed translations of verbally laden measures. However, research on the properties of adapted instruments has been scant or questionable, probably due to verbally based inadequate sample size. Research using a sufficient sample size has shown the adaptations of this nature have compromised the validity of the instruments.

There are several guidelines suggested to govern the translation of verbal test content for use with the D-HH population. First, the initial translation of material should be made by a fluent bilingual translator. Second, a blind-back translation should be conducted by another person who is fluent and bilingual. This means the translation is translated back into the initial language. Third, the two versions should then be compared and any discrepancies should be identified. Fourth, the first two steps should be repeated until no discrepancies remain. The fifth and final step is to have the translated version examined by a bilingual review committee. The translated version should also be examined to ensure that the intended construct is still measured through empirical study (Maller, 2003b).

Sign language interpreters have also been employed to aid with the administration of assessment instruments. However, the use of interpreters can confound assessment results in several ways. For example, an interpreter can be distracting to both the

examiner and the examinee. An interpreter's personal perceptions or experiences may also influence results. In addition, interpreters might use signs that give away answers or provide examples for the examinee may confound results (Simeonsson et al., 2001).

Determination of the Construct Validity and Reliability of a Measure for Use with the Deaf and Hard-of-Hearing

The construct of a measure can be described as the trait or ability that is measured by an assessment tool, such as, for example, intellectual ability. Empirical evaluation of a measurement tool should be conducted in order to determine if the construct is maintained after a modification, such as translation of verbal instructions into ASL, has been made. Several methods have been accepted by the scientific community to accomplish this task. They include the development of norm-referenced tests, the examination of the reliability and validity of tests, profile analysis, and differential item functioning.

A norm-referenced test is one in which the individual's result is compared to those obtained from a representative sample from a peer group. When using a norm-referenced test with any individual, but particularly with a person with a disability, how well the sample represents the individual should be considered. In order to better represent people with special circumstances, such as deafness, it has been suggested that norms from a representative subgroup should be used. However, it has been argued that special norms for the D-HH population may not improve the psychometric properties of a measure (Maller, 2003b). For instance, deaf norms were developed for the WISC-R Performance Scale. However, Jeffery Braden (1985) found that the use of deaf norms indeed did not improve the psychometric properties of the test for the D-HH population

and he recommended that their use be reconsidered. In addition, test items may have different meanings for D-HH people compared to hearing peers, so a test may measure something different for each group. As a result, using special norms may mean that the performance of a D-HH individual may be compared to that of other D-HH individuals on a trait that was not originally intended to be measured by the test. There is also much variability among the D-HH population on many important factors, such as degree of hearing loss, hearing status of parents, and mode of communication, and it would be difficult to gather a truly representative sample of the D-HH population on which to base any D-HH norms (Maller, 2003b).

The reliability of an assessment instrument refers to its consistency. This includes test-retest reliability and the internal consistency reliability. The reliability of a measure is described as a coefficient score.

Test-retest reliability refers to the consistency and stability of scores over time. When the test-retest reliability is high, an individual's position in the distribution of scores will be maintained when reexamined with the same test at a different time, or with different sets of equivalent items (Maller, 2003b; Anastasi, 1997). Most test manuals do not include test-retest reliability information as it relates to special populations. As a result, the test-retest reliability of most assessment instruments as applied specifically to the D-HH population is unknown (Maller, 2003b).

The internal consistency reliability of a measure refers to the index of test item homogeneity, or the extent to which test items are interrelated. Only one test administration is required to calculate the internal, or interitem, consistency reliability. Internal consistency tends to increase with the homogeneity of the domain assessed

(Anastasi, 1997; Maller, 2003b). There has been little study of the internal consistency reliability of assessment measures as specifically related to use with the D-HH population. As a result, any differences in the internal consistency reliability of assessment measures between the D-HH population and hearing peers generally remain unknown (Maller, 2003b).

The construct validity of an assessment refers to its ability to measure the traits or skills as claimed. Some tests may measure too narrow a construct, or have construct underrepresentation. Other assessment tools may measure construct-irrelevant variance, or systematically measure factors other than those claimed. For example, assessing a D-HH individual with a verbally laden measure may actual evaluate degree of hearing loss or some other factor that is related to deafness instead of the intended construct of intelligence (Maller, 2003b). There are several other measures of construct validity, including content validity, criterion-related validity and factor analysis.

The content validity of an assessment tool refers to the appearance of validity as determined by individuals who are expert in the related field. However, companies that publish assessment instruments have not typically consulted experts in the area of deafness when developing standardized tests for which use with the D-HH population is included (Maller, 2003b). Content validity is different than face validity. Face validity refers to whether an assessment instrument appears to be valid by the examinee, administrative personnel who might decide on its usage, or other technically untrained individuals (Anastasi, 1997).

Measurement of the criterion-related validity involves the examination of the relationship of a test and some other relevant criterion. Determination of the concurrent

validity involves the comparison of a test to another established instrument that measures the same construct, such as intelligence. In addition, the predictive validity can be determined by comparing a test to an outcome that should be predicted by the results, such as intelligence and academic achievement (Maller, 2003b; Anastasi, 1997). There has been research performed on the concurrent and predictive validity of several measures of intelligence as they specifically pertain to the D-HH population. For example, Braden (1994) examined the criterion-related validity of the Wechsler Intelligence Scale for Children, Revised Edition (WISC-R) and the Stanford Achievement Test, Hearing Impaired Edition (SAT-HI). Stronger correlations reflecting concurrent validity have been found among some measures of intelligence, but the predictive validity of intelligence tests for academic achievement has varied. In addition, strong criterion-related validity does not necessarily mean that an instrument has sufficient construct validity because systematic reasons may influence correlations. For instance, degree of hearing loss may impact scores on intelligence and achievement tests and may have some influence on the values of predictive validity. As a result, direct evidence of a test's construct validity is recommended. This can be accomplished through factor analysis (Maller, 2003b).

Two major types of factor analysis have been used to evaluate tests. They are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA is used when there is not an a priori theory regarding the underlying structure of an instrument, and CFA is used when there is a hypothesized theoretical model to explain the underlying structure (Maller, 2003b). CFA demands that the researcher specify if factors are or are not correlated, and the resulting analysis provides "fit statistics" that indicate if the

specific factor structure is appropriate. EFA can be useful when developing a theory, and CFA is more suited to testing an existing theory (Keith, 2005).

According to Maller (2003b), EFA has been employed in several studies to determine the factorial similarity of measures of intellectual functioning across D-HH and hearing samples. However, Maller suggested that several EFA study results are "questionable" because they used varimax rotation, which is not appropriate when factors are correlated. The EFA studies also used small sample sizes because EFA only requires 10 to 20 cases per variable to obtain stable factor loading estimates. In addition, EFA should only be conducted when there is no a priori theory regarding the factor loadings, which rarely occurs.

Maller (2003b) cites Reynolds' statement that test bias is indicated when it has been shown that a test measures different constructs between groups as well as when it measures the same construct but to different degrees of accuracy between groups. As a result, test scores between groups cannot be interpreted in the same manner. Reynolds suggested the use of CFA as being more sophisticated and useful for determining test bias than the previously used EFA. For instance, CFA can be conducted with as few as 100 cases. The goal of CFA is to evaluate simultaneously across groups the theoretical model upon which an assessment tool is based. Fit of the model across groups indicates that the factor structure is invariant, and it is concluded that the test does not measure the purported construct differently across groups. The CFA method is that specific factor loadings, their associated error variances, and the relationship between factors can be individually evaluated. This enables the determination of specific differences between groups as well as the indication of what features of the test structure differ across groups.

CFA has not been widely used in research on measurement instruments, and Maller (2003b) reported that only one study with a sample of deaf children was published at the time.

Profile analysis of an assessment tool involves the interpretation of specific score profiles on tests in order to find patterns of cognitive strength or weakness within an individual. For example, in 1990, Braden found that deaf children consistently produced lower scores on the Coding and Digit Symbol subtests of the Wechsler Performance Scales (WPS), compared to scores on other subtests on the WPS. Later, Maller (1997) found that a sample of D-HH children were more likely to exhibit unique score profile on the WISC-III than the hearing standardization sample. However, few studies employing this method of evaluating the validity of tests when used with the D-HH population have been conducted to date (Maller, 2003b).

A final way of determining the validity of a test for use with the D-HH population is to analyze the differential item functioning (DIF) of a group. This involves the determination of whether a group has more or less difficulty with specific items due to factors, such as language or gender. To do so, the examiner must calculate the probability of a specific subgroup to be less likely to correctly answer a given item because it is more difficult or discriminating. DIF was previously referred to as item bias. When DIF is present, it indicates that membership in a group accounts for some differences in performance on specific items, and the validity of scores for that subgroup should be interpreted with caution. A relatively large sample size is required to evaluate DIF; the procedure is technical and time consuming; and the process can be expensive (Maller, 2003a; Maller 2003b). As a result, DIF results are rarely reported for D-HH

samples in assessment manuals. However, some independent DIF studies have been performed (Maller, 2003b). For example, in 2000, Maller conducted a DIF study on the Universal Nonverbal Intelligence Test (UNIT) and found that all items were invariant between the deaf and hearing sample. As a result, she purported that only the UNIT can be recommended for use with deaf individuals at this time (Maller, 2003b).

Factors that may Influence Test Scores

As with the general hearing population, there are several factors that may influence mean score results obtained from assessment measures among the d-hh. The numerous factors that may affect the mean scores of any group may include test bias, item bias, differences in learning opportunities, varying exposure to information, and gender. Maller (2003b) included parental hearing status, age of onset of hearing difficulty, presence of an additional disability or medical condition, degree of hearing loss, and educational placement as factors that are especially relevant to the D-HH population. However, there have been few studies to examine mean score differences based upon factors such as these. Maller suggested that influencing factors have not been extensively studied because there have been insufficient sample sizes involved in research studies on D-HH assessment results. The possibility that a D-HH individual's intellectual assessment results may be influenced by any of the factors mentioned above should always be considered.

Ways in which Test Results may be Misused

Maller (2003b) urges practitioners to carefully consider any decision to use an intellectual measurement tool with a D-HH individual, and the social consequences should be considered in particular. This is recommended to avoid the potential misuse of

a test. Maller described several ways in which test results have been misused in the past. These include: (a) translating of test instructions or items without first following the recommended process for translating; (b) using oral, written, or gestural administration of instructions or items without evidence that the validity is maintained; (c) using tests with D-HH individuals who have additional disabilities that would limit the skills necessary to complete the test; (d) reporting verbal intelligence scores within a psychological report; (e) using verbal intelligence tests to measure a construct other than intellectual ability when the test is not intended to measure that construct; (f) not considering the factors that can affect assessment results; and (g) analyzing profiles without using available normative comparisons.

In summary, there are several issues that should be considered when conducting the assessment of D-HH children. There are procedural and statistical methods available to aid with the identification of confounding factors and to establish the reliability and validity of specific measures when used with the D-HH population. In addition, practitioners are cautioned to consider factors specific to the D-HH population that can impact test performance.

Summary and Rationale for the Proposed Research

Many studies have been published and much has been written on the general historical development of assessment measures in regards to application with the general public. However, relatively few detailed explorations of the literature related to assessment instruments as they have specifically been used with the D-HH population currently exist. The D-HH population is a heterogeneous group that shares a common difference from the general hearing population. Over the years, the study of this complex

population has evolved from viewing the D-HH as intellectual inferior to being intellectually different due to variation in development and experience. Research methods as well as the views held about the D-HH by researchers have become more sophisticated and sensitive to the confounding factors that may influence the performance of the D-HH on intellectual measurement instruments, compared to hearing counterparts. The pursuit of understanding the intellectual functioning and abilities of the D-HH population is necessary in order to continue to improve assessment services for this population.

The purpose of this dissertation project was to comprehensively present relevant literature that is available for some of the intellectual assessment instruments currently in use for the evaluation of the intellectual ability of D-HH children. A critical review of that literature was also to be conducted.

There are five specific objectives of the proposed research:

1) To provide an integrated and comprehensive understanding of the history and knowledge of intellectual instruments currently in use for the assessment of the deaf and hard of hearing population. Instruments that will specifically be presented include the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998); Leiter International Performance Scale – Revised (Roid & Miller, 1997); Wechsler Intelligence Scale for Children, Fourth Edition (Wechlser, 2003); Stanford-Binet, Fifth Edition (Roid, 2003); Comprehensive Test of Nonverbal Intelligence (Hammill et al., 1997); the Comprehensive Assessment System (Naglieri & Das, 1997).

2) To review the strengths and limitations of specific instruments which are currently used to assess intelligence in the deaf and hard of hearing population.

3) To identify the theoretical implications obtained through research results as they relate to the use of specific assessment instruments when administered to the D-HH population.

4) To organize and present a brief summary of the research information, as it relates to each assessment instrument, in a table format.

5) To identify recommendations for future research directions on intellectual assessment with the D-HH population.

Chapter II. Review and Analysis Process

This dissertation involved a comprehensive and critical review of existing literature on the intellectual assessment of the Deaf and Hard-of-Hearing (D-HH) population. The use of intellectual assessment instruments with the D-HH population has not received a large amount of attention or study compared to the extensive research with other populations. For example, as of June 2010, if one were to conduct a PsychINFO search using the terms "Deaf" and "Intelligence," a list of only 497 citations would be produced, ranging between the years 1889 through 2010. In addition, there have been few attempts to organize the existing body of knowledge related specifically to the intellectual assessment of the D-HH population. The central purpose of the study was to provide an integrated and organized review of the existing literature as it specifically pertains to the intellectual assessment of the D-HH population.

The overarching goal of this critical review and analysis was to aid practitioners who conduct the intellectual assessments of D-HH individuals by increasing their understanding of how various instruments are more or less effective when applied to the D-HH population. In addition, this review of current literature might assist researchers studying this area by aiding in the organization and understanding of past research, as well as in the formation of future questions for study.

The literature reviewed and analyzed was located through the computer search of databases including but not limited to PsychINFO and Dissertation Abstracts. The PsychINFO database contains a comprehensive collection of references to published literature in the field of psychology since 1889. The Dissertation Abstracts includes reference information pertaining to dissertations that have been completed in the field of

psychology. Other sources, included but were not limited to information obtained from newsgroups and online meeting groups that discuss issues related to the intellectual assessment of the D-HH. Finally, the information reported in assessment instrument manuals was reviewed.

There are several parameters identified that served as criteria for the inclusion or exclusion of the literature reviewed. First, the literature included had to be pertinent to the subject of intellectual assessment with the D-HH population. Literature included described the outcome or factor analysis of a specific intellectual assessment instrument(s) or acted as a theoretical or meta-analysis of specific instrument(s). Literature examining the concurrent and predictive validity of intellectual assessment instruments and studies looking at variables related to intellectual assessment instruments were also included. The usefulness of any intellectual assessment instrument in real-world application is often based upon the comparison of scores to other instruments and/or variables such as achievement. In addition, variables such as type of school placement and parental hearing status have been shown to have some relationship with the outcome of the intellectual assessment of the D-HH population. As a result, literature addressing such factors was included in the proposed review and analysis of the literature.

The dates of publication were not be used as parameters for inclusion or exclusion, because, as stated earlier, there have been relatively few documents published related to the use of intellectual assessment instruments with the D-HH population. In addition, a historical review of the literature also required that documents published at any date be included as needed. Also due in part to the relatively sparse literature

available related to this topic, documents from all types of publications were included in the review. This included documents from peer-reviewed journals, empirical studies, theoretical papers, and dissertations. Of the empirical literature reviewed, all types of studies were included for consideration regardless of the sample size, research design, method of statistical analysis, or other research variables.

Another criterion for inclusion for the proposed critical review of literature related to the intellectual assessment of the D-HH population was that the documents be published in English. While conducting the initial search of the literature for review, the researcher identified many documents that seemed to be very relevant to promoting the understanding of intellectual assessment with the D-HH population, but they were originally published in languages other than English. An attempt was be made to obtain any existing English translations of relevant literature. However, since it is beyond the scope of this study to have materials translated, only literature that is currently available in English was included.

Cultural issues have historically been shown to impact intellectual assessment results among the hearing population. Therefore, information related to cultural issues that may impact the assessment of the D-HH population was reviewed.

Some studies were excluded from the critical review. For example, studies related to the intellectual assessment of the D-HH population who also had additional sensory limitations, such as blindness, were excluded. While these factors can have strong impacts on the assessment of any individuals, they were excluded to limit the subject range of the proposed analysis of existing literature.

The critical analysis of the literature related to the intellectual assessment of the D-HH population was presented in the following manner. Literature was included that refers to the following tests: Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998), Leiter International Performance Scale – Revised (Roid & Miller, 1997), Wechsler Intelligence Scale for Children, Fourth Edition (Wechlser, 2003), Stanford-Binet, Fifth Edition (Roid, 2003), Comprehensive Test of Nonverbal Intelligence (Hammill, Pearson & Wiederholt, 1997), and the Comprehensive Assessment System (Naglieri & Das, 1997). A brief description of each assessment instrument was provided. Then, documents related to the latest editions of each intellectual assessment instrument as they pertain to use with the D-HH population were grouped and discussed. Some documents were discussed in more than one group if multiple intellectual instruments were the focus of the document. The literature for each instrument was presented chronologically simply to provide a standard organization for each grouping.

Within each grouping of literature by assessment instrument, each document was individually reviewed. First, any theoretical basis of the document was examined and evaluated as it pertained to the current theories related to the intellectual assessment of the D-HH population. Next, the methodological qualities of each document were examined. This was important because the strength of a study is often based upon methodological aspects such as sample size or statistical analysis used.

After the relevant aspects of each individual document were examined and evaluated, each group of documents was integrated to provide an overview of the appropriateness of each intellectual assessment instrument as it applies to the D-HH

population. Hypotheses that emerged from the integration of the literature were generated. In addition, recommendations for future study or theoretical development were presented. Finally, the clinical implications of any identified hypotheses and/or recommendations for future study were also presented and examined.

There were also two ways in which the terms *verbal* and *nonverbal* were used when describing the intellectual assessment instruments reviewed. When referring to an assessment instrument being verbal, nonverbal, or some combination of the two, the terms indicated the manner in which test items are presented to the examinee and/or the manner in which the examinee responded to test items. In this sense, a verbal test would be expected to include directions that were presented to the examinee verbally through spoken communication, and/or to which the examinee responded verbally through spoken communication. In comparison, a nonverbal test presented the test content to the examinee through a series of visual examples and/or gestures, and/or the examinee responded to test items by pointing, writing, or by some other means other than using spoken language.

Some instruments contain tests that were designed to evaluate abilities as they relate to the construct of verbal intellectual ability. In this case, the term *verbal* was used to describe tasks that required verbal reasoning and mediation strategies in order to think about and develop a response to test items. The term *nonverbal* was used to describe tests that were designed to measure abilities related to the construct of nonverbal intellectual ability, such as the visual reasoning skills used to complete a matrix task.

Chapter III. Critical Review of Literature

Universal Nonverbal Intelligence Test

Brief description and test development. The Universal Nonverbal Intelligence Test (UNIT) is an individually administered, multidimensional assessment measure for use with children and adolescents ranging in age from five through 17 years. It was developed to measure the general intellectual functioning levels of children and adolescents who might be disadvantaged by testing with more traditional, verbally-laden assessment tests. This included individuals who have speech, language, or hearing impairments, color or vision deficits, varying cultural or language backgrounds, or those who are unable to communicate through verbal language. The UNIT can also be a useful tool when making diagnostic decisions, such as when identifying learning disabilities, mental retardation, or psychiatric disorders. The use of receptive or expressive language is not required by the examiner or examinee when following the standardized administration protocol. In other words, no spoken language is required to administer or take the UNIT, which makes it a truly nonverbal test (Bracken & McCallum, 1998).

The UNIT was developed by Bruce A. Bracken and R. Steve McCallum and was published by The Riverside Publishing Company in 1998. The UNIT consists of six subtests, which include: Symbolic Memory; Cube Design; Spatial Memory; Analogic Reasoning; Object Memory, and Mazes. The subtests can be combined to form the Abbreviated Battery (using the first two subtests), Standard Battery (using the first four subtests) or the Extended Battery (using all six subtests). The UNIT also produces the following five scales: Memory Quotient, Reasoning Quotient; Symbolic Quotient; Nonsymbolic Quotient, and Full Scale Intelligence Quotient (IQ). The subtest scores are

reported as scaled scores with a mean of 10 and standard deviation of 3. The Quotient scores are expressed as standard scores with a mean of 100 and standard deviation of 15 (Bracken & McCallum, 1998).

The Symbolic Memory subtest requires the examinee to view, recall, and recreate sequences of universal symbols (e.g., green boy, black woman). On the Spatial Memory subtest, the examinee must view randomly placed dots on a page and then recreate the spatial pattern by placing chips onto a grid. On the Object Memory subtest, the examinee is shown pictures of common objects. The examinee is then shown a larger array of pictures and chips are placed on the pictures that were previously seen. The Cube Design subtest requires the examinee to reconstruct a design using green and white blocks. The Analogic Reasoning subtest presents the examinee with a matrix analogy using pictures (e.g., hand/glove, foot/____) or geometric figures, and the examinee then indicates the answer from among four options. On the Mazes subtest, the examinee uses a paper and pencil to trace a path from the center starting point to the exit of a maze, without making incorrect decisions en route (Bracken & McCallum, 1998; Bracken & McCallum, 2005).

The UNIT subtests were designed to fit within a two-tiered model of intelligence, which includes memory and reasoning abilities. Within the two-tiered model, the subtests are conceptualized as involving two types of internal mediation processes, or organizational strategies, which are the symbolic and nonsymbolic strategies. The responses to the Symbolic Memory, Analogic Reasoning, and Object Memory subtests can be verbally mediated or symbolically processed through the act of labeling, organizing and categorizing the information. As a result, these subtests are considered to be symbolic tasks. In contrast, the Cube Design, Spatial Memory and Mazes subtests

involve abstract and figural stimuli, and their responses are not readily associated with language. Therefore, these subtests are considered to be nonsymbolic tasks. As a result, the UNIT was described as evaluating a total of four cognitive processes which are operationalized by the six subtests (Bracken & McCallum, 1998; Bracken & McCallum, 2005).

The UNIT developers indicated that support for the four strategies operationalized by the UNIT has been present in the literature pertaining to intellectual assessment for many years. For example, Wechsler has historically emphasized the importance of distinguishing between symbolic (verbal) and nonsymbolic (performance) methods of assessing intellectual functioning. Jensen has also proposed a two-tiered hierarchical model of intelligence that consisted of memory (level I) and reasoning (level II). The authors note that the UNIT memory tasks were developed to assess more complex memory functioning than the level I memory tasks designed to recall relatively simple content associated with Jensen's contribution. The UNIT's theoretical organization is also consistent with the Gf-Gc model of fluid and crystallized intelligence proposed by the Cattell, Horn and Carroll (CHC model). The authors suggested that intelligence is composed of a fundamental ability, *g*, which is the basis from which all unique and specialized skills evolve. They further indicated that it makes little sense to conceptualize intelligence as being either verbal or nonverbal. There are, however, verbal or nonverbal means to evaluate one's intellectual functioning. Therefore, the UNIT should be considered a nonverbal measure of intelligence which was designed to be a strong measure of *g*, and not simply a measure of nonverbal intelligence (McCallum & Bracken, 2005).

The Memory Quotient of the UNIT involves short-term recall and recognition memory for abstract and meaningful material. It reflects memory ability for content, location and sequence of stimuli. The Reasoning Quotient is an index of thinking and problem-solving abilities when working under familiar and novel conditions. It represents the ability to process patterns, understand relationships, and plan. The Symbolic Quotient is related to the ability to solve problems involving stimuli and solutions that can be mediated verbally through labeling, organizing and categorizing. The Nonsymbolic Quotient is an index that reflects the ability to solve problems using abstract stimuli and solutions that are not meaningful or typically verbally mediated. Finally, the Full Scale Intelligence Quotient (FSIQ) is an index that represents the overall cognitive functioning level and is related to an individual's ability to learn and think about familiar and novel information (Bracken & McCallum, 1998).

The subtests of the UNIT were developed to meet several goals, measuring either complex short-term memory or reasoning ability. They were also designed to involve one of the two mediation processes related to symbolic or nonsymbolic thinking when applied to the memory or reasoning task. The individual items contained in the subtests were created to require no spoken language in their administration, require no spoken responses from the examinee, and contain task demands that could be communicated effectively through gesture, demonstration, or the modeling of sample items. In addition, any changes in task demands within a subtest needed to be communicated effectively through non-spoken, gestural means. The stimuli, gestures and models also had to be familiar to individuals from a variety of cultures. In addition, the assessment of intellectual ability, rather than speed, was important because the importance of speeded

responses varies among cultures. As questioned by Lane (1938) and others, the challenges of communicating effectively the need to work quickly on a timed test and the impact this understanding can have on test results has been a concern when assessing the D-HH population.

The test items were developed to be visually stimulating and interesting, and they needed to be appropriate for all examinees across gender, age and culture. The presentation of the items needed to be brief, clear and concise, and all of the items developed needed to consistently reflect the theoretical orientation of the subtest (Bracken & McCallum, 1998).

Normative sample. The UNIT's standardization sample was developed to represent a stratified random sampling that was representative of the general United States' population based on the 1995 U.S. Census data. The sample was matched based on the following demographic categories: gender; race (White, African American, Asian/Pacific Islander, Native American, and Other); Hispanic origin; region (Midwest, Northeast, South, West); urban or rural community setting; classroom placement; special education services, and parental education attainment level. Data was collected from 108 sites located in 38 states. The sample was comprised of 2,100 children and adolescents ranging in age from 5 years, 0 months to 17 years, 11 months, 30 days. The subjects were divided across 12 age groups. Data collected from an additional 1,765 children was used for the reliability, validity and fairness studies (Bracken & McCallum, 1998).

Reliability. Evaluation of the reliability of the UNIT included examination of its internal consistency. Using data from the normative sample, split-half correlations were analyzed to determine the internal consistency. The reliability estimates for each subtest,

index scale, and the Abbreviated, Standard and Extended batteries were computed for each of the 12 age groups. The average subtest reliability coefficient across all ages was reported to be 0.83 for the Standard Battery and 0.80 for the Extended Battery. The average subtest reliability coefficients across the age groups ranged from 0.64, on the Mazes subtest, to 0.91, on the Cube Design subtest. The average reliability coefficients were also examined using data from a clinical/exceptional sample of individuals belonging to the populations for which the UNIT was developed. The average subtest reliability was 0.92 for both the Standard and Extended Batteries. The authors concluded that the UNIT approaches or meets the minimum reliability standards for the normative sample and clinical populations (Bracken & McCallum, 1998).

The average composite scale reliability coefficients were reported to be 0.89 for the Standard Battery and 0.88 for the Extended Battery. In addition, similar quotients for the clinical sample were higher with a reported average coefficient of 0.96. The FSIQ reliability coefficient averaged at 0.91 on the Abbreviated Battery, at 0.93 on the Standard Battery, and at 0.93 on the Extended Battery. Among the clinical sample, the average coefficients were reported to be slightly higher. Again, the authors indicated that these scores were sufficient evidence for internal consistency across the total test (Bracken & McCallum, 1998).

The sums of the relevant scaled scores were distributed and then converted into standard scores to develop the UNIT's index scores. The confidence intervals were based on the estimated true scores and the standard errors of estimation (SEE). This means that the confidence intervals were centered on the estimated true score but were corrected to account for regression toward the mean (Bracken & McCallum, 1998).

The internal consistency of the UNIT at decision-making points was also examined. The decision-making points refer to the scores at which clinical and educational decisions occur, such as when a standard score of 70 is used to determine a mentally retarded decision or a standard score equal to or greater than 130 is used to determine giftedness. The reliability of the UNIT near these points was calculated separately for FSIQ. Scores obtained on the Standard Battery were between -1.33 SD and -2.66 SD from the mean, and between +1.33 SD and +2.66 SD from the mean. The split-half method using data collected from 471 individuals from the standardization sample and the clinical/exceptional sample was employed to develop the reliability coefficients. With the exception of the Mazes subtest's scores from the high ability sample, all of the average obtained and corrected subtest reliability coefficients exceeded 0.80. In addition, the corrected reliability coefficients of the scales, including the FSIQ, exceeded 0.90, and the obtained coefficients were near or above 0.80. This was interpreted to indicate that the UNIT can be considered a reliable measure when making clinical and educational decisions based on test results (Bracken & McCallum, 1998).

The Test-Retest reliability was evaluated by examining data collected from 197 participants. There were approximately 15 individuals in each age group who were administered the UNIT twice, after approximately a three-week interval of time. The ages of the sample were combined to form four age groups: ages 5 to 7; ages 8 to 10; ages 11 to 13, and ages 14 to 17 years. The results indicated that the test-retest coefficients approached or exceeded 0.90 for all ages over 8 years. Practice effects appeared to peak in the 8 to 10 year age group and then drop in older groups. The Object Memory and Mazes subtest scores appeared to be the least stable over time and the Cube

Design subtest score was reported to be the most stable over time. The results also indicated that the Reasoning Quotient was the most stable score examined. Again, the test authors indicated that this was sufficient evidence to support the reliability of the UNIT (Bracken & McCallum, 1998).

Several comparison studies were also conducted by the UNIT developers to evaluate its reliability among a variety of demographic groups. One comparison study compared the performance of deaf and hard of hearing individuals with scores obtained from non-hearing-impaired individuals. The study looked at UNIT scores obtained from 106 individuals who were deaf or hearing-impaired and receiving special services. For this study, the term *deaf* was used to refer to individuals with severe hearing losses who use sign language as their primary mode of communication. The term *hard of hearing* referred to individuals who retained sufficient hearing for communication through verbal language. The majority of the participants were described as deaf and only a few participants were described as having moderate hearing loss. All of the participants were enrolled in a school for deaf or hearing-impaired students where enrollment was dependent on one or more of the following conditions: an inability to communicate effectively due to hearing impairments; an inability to perform academically at a level that was commensurate with the expected level due to hearing problems; a delay in language development due to hearing impairments (Bracken and McCallum, 1998).

The sample group comprised of 60 females and 46 males with an average age of 10.7 years with a standard deviation of 3.3 years. Other demographic characteristics of the sample group included: an ethnic makeup of 85 White, 15 African American, and 6 Other subjects; 7 Hispanic and 99 non-Hispanic assessees, and parental education levels

mainly at the high school level. The participants were matched according to age, gender, race, ethnicity, and parental education level to non-hearing impaired individuals from the standardization sample (Bracken and McCallum, 1998).

The mean UNIT scores obtained from the deaf and hearing-impaired group and the matched hearing group ranged from 3.59 on the Abbreviated Battery to 8.01 on the Extended Battery Full Scale IQ. The Full Scale IQ scores from the Abbreviated, Standard and Extended Batteries were 3.59, 6.20 and 8.01, respectively. All of the differences were in the favor of the non-hearing-impaired group. However, it was reported that the differences seen were less than would be expected on a task that involved more language demands and supported the reliability of the UNIT for use with deaf and hard of hearing individuals (Bracken and McCallum, 1998).

Validity. From the beginning of its development, the UNIT was designed to represent relevant cognitive processes that are related to intellectual ability. Tasks were designed to measure memory and reasoning, two central aspects of intellectual functioning, and do so through symbolic and nonsymbolic internal mediation processes. In addition, the UNIT measures these processes without the demands of verbal receptive or expressive language (Bracken and McCallum, 1998).

The UNIT tasks were initially developed to reduce the influence of culture and other examinee characteristics on performance. To evaluate this statistically, the internal test characteristics were evaluated to determine any bias present in the measure. For example, the content of the UNIT was evaluated by experts, including psychologists, representing a variety of cultural, ethnic, and racial backgrounds. The expert consultants were chosen to represent the perspectives of male and female individuals, African

Americans, Asian Americans, Hispanic Americans, Native Americans and deaf and hearing-impaired individuals. An optometrist who was an expert in color-vision deficiencies was also asked to review the test materials to ensure individuals with common vision-color deficiencies could discriminate the colors used on the stimuli (Bracken & McCallum, 1998).

The fairness of the individual test items used in the UNIT was also evaluated across gender, race, ethnicity and language use. The items were studied individually within the separate groups as indicated by item response theory (IRT). This purports that an individual with a higher ability level will be more likely to successfully answer an item than an individual with lower ability, and any given individual should be more likely to respond correctly to an easier item than a more difficult item. The item characteristic curves were developed and item-fit statistics were calculated using data from the standardization sample. All of the items included in the UNIT were determined to have reasonable fit within the model. When examining the item fit among the subgroups, analysts concluded one item on the Analogic Reasoning subtest was found to have poor fit statistics among the Hispanic American group. However, additional analysis by the bias review experts and other statistical procedures did not indicate bias, so the item was retained (Bracken & McCallum, 1998).

The differential item functioning (DIF) analysis examines the similarities of item functioning across gender, race, ethnicity, and language characteristics. The DIF was examined between several dichotomous groups: male/female; African American/White; Asian American/White; Native American/White; Hispanic/non-Hispanic; hearing-impaired/non-hearing-impaired. The results indicated item differences between two

groups on two items. The Asian American/White group and Native American/White group each produced variation on one item each. Because the differences were small, near the higher level of difficulty, and in favor of the minority groups, the items were retained in the UNIT (Bracken & McCallum, 1998).

The exploratory factor analysis of the UNIT was performed to see what factor models would show the most consistency across methods and samples. A two-factor structure was suggested for the UNIT to represent a general intelligence *g* along with a higher-order factor *g*. Factor analysis using data from the standardization sample indicated that a two-factor structure was the most appropriate. Using the four-subtest Standard Battery, the subtests clustered into a memory factor (I) and a reasoning factor (II). The two factors accounted for 77.5% of the variance. A similar analysis was conducted using scores from the six-subtest Extended Battery and the two-factor solution emerged. However, a third factor was also added to the solution. The Mazes subtest did not correlate to the reasoning factor as had been anticipated, but rather correlated with a third salient pattern coefficient (greater than 0.40). This was associated with a unique form of reasoning, which was identified as planning. The Mazes subtest shared the least amount of common variance with the other subtests on the UNIT and its reliable specific variance is relatively high. As a result, it appeared to provide information about an examinee's intellectual functioning that the other memory or reasoning subsets could not (Bracken & McCallum, 1998).

A higher-order factor-analysis was conducted in which the factors are rotated obliquely so they do not overlap and represent broader areas of generality than just a primary factor. The results of this analysis indicated that a general *g* factor exists over all

of the subsets, but the first-order memory and reasoning factors also emerged. These results were reported to be stable for data obtained from the standardization sample and the clinical/exceptional sample (Bracken & McCallum, 1998).

A third set of factor analyses were conducted using data from the Standard Battery with the ages combined into four groups. These results supported the presence of the two factors, memory and reasoning. In addition, at ages 12, 13, and 14 years, Symbolic Mediation and Nonsymbolic Mediation factors emerged. The examiners concluded that symbolic and nonsymbolic mediation becomes more prominent during puberty and may be related to neurocognitive maturation that takes place during that stage. They also concluded that these results provided additional support for the primary and secondary cognitive constructs which the UNIT was designed to measure (Bracken & McCallum, 1998).

Confirmatory Factor analysis was conducted to provide further information about the factor structure of the UNIT. Data obtained on all six of the subtests from the standardization sample, divided into four age groups, was analyzed. The results indicated that a single general intelligence (*g*) factor was present along with the primary and secondary scales. It was concluded that the UNIT subtests measure abilities as expected and fit into the hierarchical theoretical model under which the UNIT was constructed (Bracken & McCallum, 1998).

To explore the construct validity of the UNIT, scores obtained from the UNIT were compared to those obtained on the Wechsler Intelligence Scale for Children, Third Edition (WISC-III). Data were collected from the following sample groups: examinees with learning disabilities (n = 61); examinees identified as mentally retarded (n = 59);

examinees identified as gifted (n = 43); and Native American examinees (n = 34). The subjects ranged in age from 6 to 16 years of age. The correlation coefficients between the UNIT Abbreviated, Standard and Extended FSIQ scores and the WISC-III FSIQ score were 0.78, 0.84 and 0.83, respectively. Among the sample identified as having metal retardation, lower FSIQ scores were seen on the WISC-III compared to the three UNIT battery FSIQ's. However, when the scores were corrected, the correlations were 0.86, 0.84 and 0.88 for each of the UNIT FSIQ's. Again, this was described as a strong and positive correlation between the two measures. The difference in scores was also reduced when the WISC-III Performance IQ score was used in place of the FSIQ. A similar pattern was seen when the scores from the sample identified as gifted was analyzed, particularly when the Performance IQ scores were compared with those obtained from the three UNIT batteries. Analysis of scores obtained from the Native American sample indicated a higher correlation between the FSIQ scores from the three UNIT batteries with the language-reduced Performance IQ score on the WISC-III (Bracken & McCallum, 1998).

The predictive validity of the UNIT was evaluated by looking at how the UNIT scores correlated with those obtained on achievement tests. The Woodcock-Johnson Tests of Achievement, Revised (WJ-R) was administered to three sample groups: individuals identified as gifted; individuals with learning disabilities; and individuals with mental retardation. The subjects ranged in age from 6 to 16 years. For the gifted group, all three of the UNIT FSIQ scores correlated highly with the WJ-R Broad Mathematics, Broad Knowledge, and Skills Cluster. Lower correlations were seen on the Broad Reading and Broad Written Language clusters. The subjects from the group with

learning disabilities showed similar correlation patterns. In contrast, the scores obtained from subjects in the mental retardation group showed lower correlations between the UNIT and WJ-R scores, which ranged from 0.40 to 0.63 (Bracken & McCallum, 1998).

Independent research. In 2000, Susan J. Maller published the results of a study that examined the differential item functioning (DIF) on items from four subtests of the UNIT, which included Symbolic Memory, Spatial Memory, Analogic Reasoning and Object Memory. The subjects in this study included 104 severely or profoundly deaf participants ranging in age from 5 through 17 years. Each subject required sound within the speech frequencies to be greater than 70 dB, used sign language as their primary mode of communication, were enrolled in self-contained special education classrooms, and presented with no other identifiable disability. In addition, 45 of the participants were males and the remaining 59 were females. The subjects resided in four sites, located in the Southeastern, Western and Midwestern United States. In addition, the subjects also represented several racial/ethnic groups, including 16 African Americans; three Asian/Pacific Islanders, 75 Whites; 7 Hispanics, and three others. Data from 104 hearing counterparts was obtained from the UNIT standardization sample through a matching procedure. The hearing subjects had no other disabilities and did not participate in special education programs. The deaf and hearing subjects were matched by total subtest scores, age, ethnicity and gender.

To start, items were screened for DIF in order to determine if there was a need to identify a set of non-DIF items with which to match the examinees of equal ability. The Mantel-Haenszel (MH) DIF detection method was used, which is approximately chi-square with one degree of freedom. This purification procedure was described as

being unnecessary because no items were determined to exhibit DIF during this initial screening (Maller, 2000).

The next step was to analyze each of the four subtests was analyzed separately for potential DIF. The remaining two subtests (i.e., Cube Design and Mazes) were not included because, due to the time-related bonus points, a much larger sample size and a more sophisticated statistical IRT model would be required. The DIF of within each of the four subtests was analyzed using the likelihood ratio for DIF detection method, which Maller described as "state of the art." The fit of the model to the data was evaluated through the likelihood ration goodness-of-fit statistic (Maller, 2000).

The results of the MH DIF statistics were reported along with the probability values. No items reported exhibited significant MH DIF (all $p > 0.05$). The fit of the items was then examined using four to six items of varying difficulty. All of the likelihood chi-square fit statistics were reported to indicate good fit. Specifically, Symbolic Reasoning, all $p \geq 0.33867$; the Spatial memory, all $p \geq 0.34398$; Analogic Reasoning all $p \geq 0.85134$, and Object Memory, all $p \geq 0.97620$ (Maller, 2000).

The IRT item difficulty estimates for the hearing and deaf subject samples were additionally reported. These ranged from -4 (easy) to +4 (difficult). The item difficulties were found to be similar for both groups and the items tended to be ordered in difficulty in a similar manner between groups. The likelihood ratio DIF tests also indicated that the items were equally difficult between the two subject groups (Maller, 2000).

In summary, Maller (2000) found no items on the UNIT to exhibit DIF. The sample size was relatively large, compared to most studies including deaf subjects, because deafness is a relatively low-incident condition. Maller did indicate that the DIF

tests of significance may have lacked power due to the size of the sample. However, samples of 100 subjects have been considered adequate for MH DIF detection procedures. Other studies including similar sample sizes have been able to detect DIF between samples when examining other measures of intelligence. Maller concludes by suggesting that her study provides additional support that the UNIT produces invariant results when administered to deaf children and is therefore an adequate tool for the assessing the intelligence of members of this group. Further replication studies and examination into the usefulness of the UNIT in making educational decisions for deaf children is encouraged.

In 2004, Krivitski, McIntosh, Rothlisberg and Finch published results from a study in which a profile analysis of children using the UNIT. The purpose of the study was to determine if deaf children performed similarly on the UNIT as hearing children. The study included 39 deaf and 39 hearing children who ranged in age from 5 years to 17 years. The deaf children were identified as being prelingually deaf and having no comorbid conditions. For the study, deafness was defined as a hearing loss of 60 dB or greater, which was classified as severe to profoundly deaf. Some participants met the criteria in one ear and had less than 60 dB hearing loss in one ear, but were still considered to be deaf and were included in the study.

The deaf participants were then matched to hearing counterparts who were included in the standardization sample of the UNIT. Age, gender, Hispanic origin, race and the highest combined parents' education levels were the criteria used for matching subjects. Each group included 18 females and 21 males. In addition, each group was comprised of 27 Caucasians, four Asians, four Hispanics, three Black and one other racial

group member. In regards to parental education level 20 subjects' parents had completed 4 years of college and/or graduate schooling, 10 had graduated from high school, 8 had attended some college, and one had not graduated from high school (Krivitski et al., 2004).

Of the deaf subjects, 21 preferred to communicate solely through American Sign Language (ASL), three used a combination of ASL and voice/speaking, two used Pidgin Signed English (PSE), two used Signing Exact English (SEE), and two used PSE and voice/speaking. Other individual students preferred to communicate through various combinations of ASL, PSE, SEE, Manually Coded English (MCE) and voice/speaking (Krivitski et al., 2004).

The UNIT was administered to the deaf subjects by the senior author of the article, who is proficient in ASL. The introduction and rapport development with the deaf participants was conducted in sign language with care taken to accommodate each subject's preferred method of communication. The Extended Battery of the UNIT was administered using standardized procedures, and required 45 to 60 minutes administration time for each subject (Krivitski et al., 2004).

Three of the 39 deaf subjects chose to discontinue the process before the entire extended battery could be administered so their data was not included in the analysis of the results. Krivitski et al. (2004) reported the mean scores, standard deviations and ranges for the deaf, hearing and combined samples for all subtests and quotient scores. Correlations between the Full Scale, quotient and subtest scores were also reported.

Krivitski et al. (2004) indicated that the correlations among the six subtests of the UNIT did not exceed 0.90. This was interpreted to mean that multicollinearity among the

subtests was not evident. It was further reported that the intercorrelations among the subtests of the UNIT were higher for the sample of deaf children than for the hearing counterparts with the exception of the Mazes subtest. The Mazes subtest correlation with the five other subtests tended to be low and negative for the deaf and combined samples, compared to that of the hearing sample. The correlations among the subtests and the quotient scores for the hearing and combined sample groups tended to be more consistent with those reported for the standardization sample. However, the correlations for the subtest and quotient scores from all three sample groups included in this study were higher than those reported for the standardization sample.

To conduct the profile analysis of the performance of the deaf and hearing samples, the means of the deaf and hearing groups on the six subtests of the UNIT were compared as well as the pattern of means across the six subtests. To test for parallelism, differences in scores, or segments, were identified. A one-way MANOVA was computed using these segments as the dependent variables and the sample group as the independent variable. It was concluded that a statistically significant difference between the two groups on one or more of the segments was present ($F(5, 59) = 2.820$, $p = 0.022$). The Symbolic Memory, Spatial Memory and Mazes subtest means were higher for the hearing group, and the Cube Design mean was higher for the deaf group. The two groups' means were described as being "virtually identical" for the Analogic Reasoning and Object Memory subtests (Krivitski et al., 2004).

To further explore the significant results of the parallelism results, *t* tests for independent samples were computed to compare the mean scores from the deaf and hearing groups on the six subtests. The *t* test results were reported to not be significant;

however, there was a large effect size for the Cube Design subtests. The authors described this difference as being important in practical terms because, although the difference was not significant, there was a difference present between the two sample groups (Krivitski et al., 2004).

When the six subtest means were combined, the test of levels did not show a significant difference. The authors concluded that the two groups performed similarly when the subtest scores were combined (Krivitski et al., 2004).

Finally, the test for flatness was used to assess whether the means of the subtests differed from one another regardless of group membership. The results of this calculation were not found to be significant ($F(5, 69) = 1.807$, $p = 0.123$). This was interpreted to indicate that there were no significant differences among the means of the six subtests when the scores from the combined deaf and hearing subjects were used (Krivitski et al., 2004).

In conclusion, Krivitski et al. (2004) found that the deaf and hearing children performed similarly on the UNIT and did not show a specifically higher or lower performance on specific subtests. However, statistically, their performance did not indicate a parallel pattern when mean differences on subtests were analyzed and a profile difference may exist, as indicated on the Cube Design subtest scores, that can be important at a practical level. Overall, the results of the study did not support earlier research on measures of intellectual functioning that showed a generally lower score among deaf subjects compared to hearing subjects. This was attributed to the characteristics of the UNIT. Namely, the UNIT is an updated instrument with strong psychometric characteristics and a standardized, nonverbal administration procedure.

Krivitski et al. (2004) reported that subject selection to control for confounding variables (e.g., no comorbid conditions present, etc.) may have impacted the results of their study. However, they concluded that the UNIT does not appear to penalize children with hearing or language challenges. Because it does not require an interpreter for administration or verbal responses from the examinee, it was recommended as a useful tool when evaluating children with hearing or language challenges.

Another study by Maller and French, published in 2004, was reportedly conducted to examine the factor structure of the UNIT across samples of deaf and hearing individuals. Data were analyzed from 102 deaf participants ranging in age from 5 to 17 years (see above for details of sample group) who participated in the standardization of the UNIT. The deaf subjects were severely to profoundly deaf and required speech frequencies to be over 70 dB in order to be heard, and had no comorbid conditions identified. The standardization data was obtained from 2,096 children, ranging in age from 5 to 17 years, who were chosen to represent the general population in the United States based on several demographic cosiderations, such as gender, geographic region of residence, special education placement, parents' education levels, race and ethnicity.

The primary (Memory and Reasoning) and secondary (Symbolic and Nonsymbolic) factors, as identified by the developers of the UNIT, were examined separately for each sample. Several models were used to investigate the fit of the models, which included the goodness-of-fit index (GFI), the comparative fit index (CFI) and the root mean squared error of approximation (RMSEA). Maller and French (2004) examined the factor structure by constraining parameters to be equal across groups. Progressively more restrictive models could be tested by adding additional constraining

parameters. If a more restrictive model showed a significant decline in fit, then a difference in the factor model between groups would be indicated. If a difference in the factor structure between groups is evident, then additional follow-up analysis would be performed to determine if the difference was due to variation in ability between groups or if a systematic bias in the test was present.

The results indicated that the primary factor model (Memory and Reasoning) was invariant across groups, with the exception of the Mazes subtest. This indicated that the Mazes subtest may have a different meaning for the deaf group. In the follow-up analysis, the deaf subjects showed lower scores on the Analogic Reasoning subtest compared to the standardization sample. Additional examination of the differences in the Mazes and Analogic Reasoning subtest scores was not conducted because the statistical requirements were not sufficiently met. As a result, Maller and French (2004) could not offer solid conclusions regarding the causes of the differences. However, they did indicate that earlier research found that deaf children can show lower scores on motor-reduced nonverbal intelligence tests, such as non-verbal matrix tasks. The Analogic Reasoning subset of the UNIT does not require the manipulation of any objects, and it can be argued that it requires some verbal mediation when solving the tasks. It was also reported that the standardization sample showed a higher Memory latent factor mean than the deaf group. Because the UNIT memory tasks are complex activities, there may be some verbal mediation required which could account for the lower mean score.

The second factor structure (Symbolic and Nonsymbolic) was generally supported by the results from the deaf sample, but pattern coefficients on three subtests (i.e., Cube Design, Spatial Memory and Mazes) were not invariant across the groups. As a result,

the Nonsymbolic subtest scores may have different meanings for the deaf and standardization sample groups (Maller & French, 2004).

Maller and French (2004) offered other potential explanations for the differences seen between groups. For example, the scores obtained from the deaf group may not be generalizable or there may have been unreported comorbid conditions that could have impacted performance. In addition, it has been suggested that the gestures used during the administration of the UNIT may be confusing to deaf children who are accustomed to using sign language. In general, the authors concluded that each of the factor structures were generally supported for the deaf sample, but the primary factor structure was preferred because only the Mazes subtest showed variance across groups. The UNIT was also described being the only published measurement of intelligence that has been shown to have no DIF and partial support for the factor model for use with deaf children. They caution administrators who work with deaf children to be aware that Analogic Reasoning scores may be lower than expected and deaf children may have some challenges completing tasks of short-term memory (Maller & French, 2004).

Leiter International Performance Scale - Revised

Brief description and test development. The Leiter International Performance Scale – Revised (Leiter-R) is an individually administered, multidimensional assessment measure that can be administered to individuals ranging in age from 2 years, 0 months to 20 years, 11 months. It is a nonverbal measure with standardized administration procedures that require no spoken language ability to administer or take. It it evaluates intellectual, memory, and attention abilities. The Leiter-R was developed for use with individuals for whom traditional measures of intelligence are not appropriate. This

includes individuals with significant communicative disorders, cognitive delay, English used as a second language, hearing impairments, motor impairments, traumatic brain injury, attention-deficit disorder, and other specific types of learning disorders (Roid & Miller, 1997).

Included in the test kit are three stimulus easels, response picture-cards, several manipulatives, and other printed materials. When responding to the Leiter-R, the examinee places picture cards or manipulatives into "slots" on the "frame" that is attached to the base of each easel stimulus book (Roid & Miller, 1997).

The Leiter-R was developed as the need grew for nonverbal cognitive assessment and improved treatment and academic planning for children and adolescents with English as a Second Language backgrounds or communication disorders. Nonverbal intellectual abilities typically include reasoning, spatial and two-dimensional visualization, memory, attention, concentration complex tasks, and processing speed when working on complex tasks. Proficiency in perceiving, manipulating, or reasoning with words or numbers, or any other materials traditionally associated with "verbal" processes should not be required when solving nonverbal tasks. The nonverbal tasks are completed using pictures, figural illustrations, and coded symbols. In addition, the administration of instructions and responses from examinees is adapted to a nonverbal, or gestural/pantomime, format (Roid & Miller, 1997).

Research related to several models of intelligence was considered during the development of the Leiter-R, including that related to Carroll's three-stratum model, Gustafsson's model, and Woodcock's work. The authors indicated that Carroll's three-stratum model of intellectual abilities, with general intelligence, or "*g*", at the first level

provided more detail and verification than the other models. The authors further identified the second level of eight ability domains, similar to those described by Horn and Cattell in 1966. These include fluid reasoning (Fg), crystallized ability (Gc), broad visualization (GV), auditory ability (Ga), two factors of processing (Gs and decision speed), and long-term retrieval (Glr). The third level of Carroll's theory includes more specific abilities, such as inductive reasoning under Gf (Roid & Miller, 1997).

The hierarchical model of the Leiter-R includes increasingly complex subdomains of ability as the age of the examinee progresses. For example, general intelligence, or "*g*", remains as the first factor. For children ages 2 through 5, the second-level factors include reasoning, visualization, attention, memory, memory span, and a recognition memory factor that is included only for children ages 4 to 5 years. For children ages 6 through 10 years-old, the second-level factors include reasoning, visual/spatial, attention, memory, recognition memory, and memory span. For examinees ages 11 through 20 years, the second-level factors include reasoning, visualization, attention, memory and memory span. There are one or more third-level factors that are assessed within each second-level factor for all age groups (Roid & Miller, 1997).

The Leiter-R subtests are arranged into two groups. First, the Visualization and Reasoning (VR) battery contains 10 subtests that measure nonverbal intellectual ability related to visualization, reasoning, and spatial memory. The second, the Attention and Memory (AM) battery contains 10 subtests that measure nonverbal attention and memory functions. Four rating scales that can be completed by the examiner, parents, teachers, and a self-review completed by the examinee are also included to provide behavioral information about the examinee. All of the subtests have been given "game names" in

order to make the process more fun when testing young examinees (Roid & Miller, 1997).

The VR and AM batteries can be administered together or separately, depending on the assessment questions and clinical needs of the individual being assessed. Four subtests of each battery can be administered in approximately 25 minutes to obtain a brief estimate of global intellectual functioning or to rapidly distinguish ADHD and LD in children. When a more comprehensive assessment is desired, such as for treatment or educational planning decisions, the six-subtest administrations of the VR and AM batteries, which each take approximately 40 minutes to administer, are recommended (Roid & Miller, 1997).

The Leiter-R Visualization and Reasoning Battery (VR) subtests include the following: Figure Ground (FG), or The Find it Game, during which the examinee identifies embedded figures or design within a complex stimulus; Design Analogies (DA), or The Funny Squares Game, which is a classic matrix analogies task that requires the mental rotation of figures on more complex tasks; Form Completion (FC), or The Put Together Game, in which one must recognize a "whole object" from a randomly-displayed array of fragmented parts; Matching (M), or The Matching Game, in which one must discriminate and match visual stimuli by selecting the card or manipulative shape that matches what is seen in the easel stimuli; Sequential Order (SO), or The Which Comes Next Game, where logical progressions of pictorial or figural objects is presented, and one must select the stimuli that progresses in the corresponding order; Repeated Patterns (RP), or The Over and Over Game, in which patterns of pictorial or figural objects are repeated, and the examinee must supply the "missing" piece of the pattern;

Picture Context (PC), or The Belongs Together Game, which measures one's ability to recognize the pictured object that has been removed from a larger picture; Classification (C), or The Goes Together Game, which involves the classification of objects or geometric shapes; the Paper Folding Game (PF), which requires the examinee to mentally "fold" an object that is displayed in a two-dimensional, unfolded, state and match it to a target; and the Figure Rotation (FR), or The Turn it Around Game (Roid & Miller, 1997).

Normative process. The Lieter was originally developed by Russel Leiter in 1929 and included research conducted on children from Hawaii from a variety of ethnic backgrounds. Several revisions were published over the subsequent years, and one adaptation that included revisions to the test administration directions by Grace Arthur in 1949 is known as the Arthur Adaptation of the Leiter International Performance Scales. These revisions contained items that were conceptually grouped within the common constructs of nonverbal intelligence, and, while the subgroupings of factors were identified, they were not systematically analyzed for each age group. As a result, the nonverbal nature of the Leiter and its theoretical background has been praised over the years, but its psychometric characteristics have been criticized as being inadequate. The developers of the Leiter-R were aware of this criticism and made particular effort to conduct a complete psychometric analysis using a nationally representative norm sample (Roid & Miller, 1997).

The development of the current version of the Leiter-R began with the creation of the Tryout Edition and the psychometric analysis of the results. The use of the Tryout Edition led to the deletion, modification or addition of what would be the Standardization Edition of the Leiter-R (Roid & Miller, 1997).

The demographic characteristics of the normative sample used in the standardization of the Leiter-R was representative of the 1993 population survey obtained from the United States Bureau of the Census. This group included proportionate samples of Caucasian, African-American, Asian American, and Native American individuals. Children of Hispanic origin, including those who were mixed with other racial categories, were included in a special category. The importance of Spanish being used as a first or primary language made being of Hispanic descent an exclusive category of ethnicity rather than one of ancestral racial origin. In addition, sample members were grouped proportionately by the level of parents' completed education. The normative sample was selected from all four geographic regions in the United States (Northeast, Midwest, South, and West) as well as from representative community sizes (Roid & Miller, 1997).

Due to the length of time required to administer the full batteries, an efficient method of obtaining standardization information was developed. Because the VR Battery is used to determine more important IQ scores, and the AM Battery is generally used as a clinical and diagnostic tool in "ruling out" various cognitive characteristics, the need for a greater sample size for the standardization of the VR Battery was determined. As a result, the VR Battery was administered to a sample of 1,719 typical children, and the AM battery was administered to a subset of 763 of the same children. For the 763 children included in both groups, the VR and AM batteries were administered on two contiguous occasions (Roid & Miller, 1997).

Individuals ranging in age from 2 years to 20 years were included in the sample. Children ages 2 through 5 years were divided into six-month age groupings; children ages 6 through 11 years were divided into one-year groupings; and those ranging in age

from 12 through 20 years were divided into two to three-year groupings. It was indicated that the age at the time of testing was identified by the full month-of-age. For example, children who were 5 years, 6 months and 4 days old were placed in the same age group as children who were 5 years, 6 months, and 25 days old. As a result, scores from children whose birth dates place them near the cut-off point between two normative age groupings should be interpreted with caution (Roid & Miller, 1997).

Additional samples of children were included in the normative process who met atypical, clinical or exceptional category requirements. The inclusion criteria was based on results from other standardized tests, from the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV), and from corroborative confirmation of category inclusion from information obtained from local school districts. These included the following groups of children: 98 children with Severe Speech or Language Impairment; 69 children with Severe Hearing Impairment; 61 children with Severe Motoric Delay or Deviation; four children with Traumatic Brain Injury; 123 children with Significant Cognitive Delay (mental retardation); 112 children with Attention Deficit Disorder w/ or w/o Hyperactivity; 67 children identified as Gifted; 29 children identified with Learning Disability (Verbal > Nonverbal IQ); 39 children with Learning Disability (Nonverbal > Verbal IQ); 73 children identified as English as a Second Language (ESL-Spanish), and 26 children identified as English as a Second Language (ESL-Asian/Other) (Roid & Miller, 1997).

When 80% of the results from the standardization procedure was collected, several analyses were performed to determine the final sets of items for all subtests, to verify reliabilities of proposed versions of the subtests, and to determine the stopping

rules for each subtest. Item difficulties, Rasch fit statistics and calibrations, correlations with other assessment measures, differential item functioning statistics, and chi-square differences between the typical and atypical sample scores were analyzed. Any items that showed poor fit with the Rasch model, differential item functioning between gender or ethnic groups, or the presence of other indicators of poor psychometric properties were removed from the final subtests. The stopping points were determined by a probability that less than 5% of any additional items would answered correctly (Roid & Miller, 1997).

Reliability. The reliability of the Leiter-R was checked using classical test-theory and item-response (IRT) approaches. The average internal consistency estimates for the subtests in the VR Battery across age groups were reported to range from 0.75 to 0.90. The reliability estimates for the IQ and Composite scores from three age groupings (ages 2-5, 6-10 and 11-20) were calculated. The VR Battery IQ and Composite reliability scores ranged from 0.88 to 0.93. The test-retest reliability of the VR Battery was examined using scores obtained from 163 children and adolescents who were administered the VR battery on two occasions. The IQ and Composite test-retest reliability scores ranged from 0.83 to 0.96. The individual subtest test-retest reliability scores ranged from 0.70 to 0.90 (Roid & Miller, 1997).

To examine the consistency of the Leiter-R when used to make educational and clinical decisions, the Brief IQ Screener and Full Scale IQ scores of the 163 children were examined. All of the children with scores below the cutoff score of 70 were identified during both administrations of the Leiter-R. The Standard Errors of

Measurement on the VR Battery subtests by age group were also calculated, which ranged from .96 to 1.50 (Roid & Miller, 1997).

Validity. To determine the content validity of the final version of the Leiter-R, each subtest was evaluated by the 114 examiners who participated in the standardization of the test. This was to assure that the tests and materials could be administered in a nonverbal and nonlanguage mode. The input from the examiners was obtained through a rating process, and only the subtests with uniformly high ratings were included in the final version of the Leiter-R, or they were revised. Many items from the original Leiter were eliminated because they were old-fashioned or unclear, and the other main classifications of acceptable items were expanded into full subtests. In addition, teaching items were developed to ensure that a clear understanding of each task would be conveyed prior to the examinee attempting the tasks. The resulting collection of subtests, particularly those contained in the VR Battery, were designed with the intent of testing as many nonverbal cognitive abilities as possible within the testing time (Roid & Miller, 1997).

IRT methodology was employed to analyze the fit of each item and ensure item and test fairness across groups. This included looking at the fit of each item to the "*g*" domain of the VR Battery and at the absence of differential item functioning (DIF). The final items included were reported to show exceptional fit to the FACETS and conventional Rasch 1-parameter logistics model when one analysis of all 305 items from the 10 subtests of the VR Battery was conducted. When gender and ethnicity was controlled, the final items on the Leiter-R were also reported to show exceptional fit, which the authors purport to indicate fair measurement across groups. All of this

information was further purported to indicate that the Leiter-R does measure cognitive abilities without the interference of other variables (Roid & Miller, 1997).

The criterion-related validity of the Leiter-R was evaluated by determining its ability to accurately classify individuals based on test score interpretation. Cutoff points for the identification of giftedness or cognitive delay were developed by comparing scores from subjects determined to be "typical," or those with no presence of any exceptional or disability condition) with subjects identified as "atypical," or those with a documented exceptionality that has been verified by some criteria that is independent of the Leiter-R. The results indicated that the Leiter-R has a high level of accuracy, over 80% correct classification, in the identification of cognitive delay and a more moderate level of accuracy when making the classification of giftedness. Less sensitivity of the Lieter-R in making decisions related to ADHD and learning disabilities was reported. As a result, the authors suggest that additional measures should be employed when using the Leiter-R to make the classification of giftedness, Attention Deficit Hyperactivity Disorder (ADHD) and/or learning disabilities (Roid & Miller, 1997).

Validity studies looking at the criterion-related validity of the Leiter-R when administered to special groups were also conducted. These groups included the following: severely speech/language impaired; severely hearing impaired; severely motor delayed or deviated; traumatically brain injured; significantly cognitively delayed; exhibiting ADHD with or without hyperactivity; gifted; learning disabled – nonverbal type; learning disabled – verbal type; using English as a second-language (ESL) – Spanish, and ESL – Asian or other. The severe hearing impairment group included children ranging in age from 2 through 18 years with a median age of 8 years, and

approximately half was females and half was males. The results indicated relatively lower mean subtest scores and relatively lower mean Brief IQ and Full IQ scores from the special group with severe hearing impairment. It is suggested that these lower scores may be related to the individuals having histories of schooling difficulties or the presence of additional handicapping conditions among some of the group members. Consistently lower mean scores were found in the group with significant cognitive delay, and higher scores were found for the gifted group, which reportedly provided support for the criterion-related validity of the Leiter-R (Roid & Miller, 1997).

Concurrent validity studies were conducted using several established tests. The correlation of the Leiter-R with the original Leiter was examined by administering both batteries to a group of 124 children and adolescents. The correlation between the IQ scores obtained on the original and revised Leiter tests was 0.85. This was reportedly higher than the correlation results of most test battery revisions with the original battery and indicates a difference in IQ scores of approximately 12 points. However, the authors report that most test batteries show a numerical decrease in scores of approximately 0.3 IQ points per year because individuals are compared to an increasingly difficult standard over time. Since the original Leiter was developed in 1948 and 48 years had passed since the revision, a difference of -13.9 between scores was predicted. The authors suggested that the difference in scores from the original version of the Leiter and the current Leiter-R was due more to historical trends of the normative status of IQ mean scores than to error or low correlation between the two tests (Roid & Miller, 1997).

The correlation of the Leiter-R with other established measures of intellectual ability was examined by comparing the scores from a sample of children who were

administered the Leiter-R and the WISC-III. This sample consisted of 126 children ranging in age from 6 to 16 years, and the majority of the children were from the Midwest and southwest regions. Within the sample, 47% fell into the normal range, 18% were identified as having cognitive delay, 9% were identified as gifted, and 23% were identified as ESL-Spanish. The results indicated that the Brief and Full Scale IQ score obtained on the Leiter-R correlated consistently high with the WISC-III Full Scale IQ and the Performance IQ scores ($r <= 0.85$). The WISC-III Verbal IQ showed a lower correlation with the Leiter-R Brief and Full Scale IQ scores ($r = 0.77$ and 0.80 respectively). It is suggested that this level of correlation with the verbal tasks was higher than expected and suggests there is a strong global "*g*" factor that is common between the two instruments. An additional correlational study was conducted using archival data from 82 children who were administered the Leiter-R and the WISC-III within six months of time. Similar levels of correlation between WISC-III Full Scale IQ and the Leiter-R Brief ($r = 0.82$) and Full Scale IQ ($r = 0.83$) scores were found (Roid & Miller, 1997).

The predictive validity of the Leiter-R was examined by comparing scores with those obtained from a variety of individually administered tests of achievement, including the WIAT Reading Composite, WIAT Math Composite, WJ-R Broad Reading, WJ-R Broad Mathematics, WRAT-3 Work Reading, and the WRAT-3 Arithmetic. All of the correlations with the Leiter-R Brief and Full Scale IQ scores ranged from $r = 0.62$ to $r = 0.82$. The correlations were all above the reported average correlation of 0.60 that most cognitive tests show with tests of achievement (Roid & Miller, 1997).

The construct validity of the Leiter-R was measured through a variety of ways.

First, the developmental trends of an instrument should show increased median raw scores on tests as the age of the subjects increases. The Leiter-R "Growth Scale," which is based on the four subtest scores included in the Brief IQ score, shows an increasing upward trend until the 10-year age group, where the continuing trend increases at a slower rate until the 15-year age group is reached. There is a plateau in scores after that age group (Roid & Miller, 1997).

A test-retest study was conducted using scores from a sample of 22 children with identified cognitive delays and 22 children with no identified cognitive delays. The subjects were randomly assigned to either a control condition, which included only pretesting and posttesting, or to a learning-potential (LP) experimental group who received additional "mediated instruction" to teach the children the skills needed to complete the tasks. The children were also asked to explain their reason behind responses, which allowed for self-reflection about and self-correction of responses. T-tests were then conducted due to the small number of subjects in each of the four groups to examine the test and retest mean scores. Of the experimental subjects, two significant changes in scores were seen among the cognitive delay individuals and three were seen among the subjects without delays. Although the sample size was small, the results were viewed as supportive of the sensitivity of the Growth Scale (Brief IQ) and as an additional indication of the construct validity of the Leiter-R (Roid & Miller, 1997).

Exploratory factor analysis of the Leiter-R using a common-factor model that allowed for correlated factors was conducted. The results indicated the presence of several common factors, which included visualization, reasoning, attention and memory. There was some variability across age groups, which was attributed to the different set of

subtests administered to some age groups. To accommodate this, four age groups were created for the analysis: ages 2-3; 4-5; 6-10, and 11 to 20. The analysis using age groups indicated the stable presence of reasoning and visualization factors across all age groups. Among the younger age groups, the Sequential Order subtest showed a stronger visual factor. A greater loading of the visual factor was also shown on the Form Completion, Matching and Figure Ground subtests among the older age groups (Roid & Miller, 1997).

Confirmatory factor analysis was completed using LISREL and AMOS. Using LISREL analysis, the one factor and two-factor models were identified as being poor fits. Among the 4 to 5 year old group, the four factor model showed the best fit. These four factors included Fluid Reasoning, Visualization, Attention, and Recognition Memory. For the 6 to10-year old age group, a five factor model showed the best fit. These five factors included Reasoning, Visual-Spatial, Attention, Recognition Memory, and a combined Associative/Memory-Span factor. These five factors were also confirmed among the 11-21-year age groups (Roid & Miller, 1997).

Because the Leiter-R is theoretically based on the hierarchical "*g*" theory, the presence of this factor across all subtests was analyzed. Several subtests, including Figure Ground, Form Completion, Sequential Order and Associated Pairs showed high "*g*" loadings for all age groups. Among the age groups above 6 years, high "*g*" loading was also seen on Design Analogies, Repeated Patterns and Visual Coding. Above age 11, the "*g*" loading on Sequential Order and Paper Folding was strong (Roid & Miller, 1997).

The common, specific and error variance was also reported for the Leiter-R. When specific variance exceeds error variance, there is evidence of specificity between

subtests. The general pattern of the Leiter-R scores shows 46% common, 34% specific, and 20% error variance overall. This classic C > S > E pattern is reported to be ideal and comparable to other established measures of cognitive ability. In addition, 8 of the 10 VR Battery subtests show this ideal pattern of variance. Some caution should be used when interpreting the Sequential Order and Paper Folding subtests when working with some age groups (Roid & Miller, 1997).

Cross-Battery correlation and factor analysis studies were conducted using the Woodcock-Johnson Psychoeducational Battery-Revised (WJ-R) and the Wechsler Intelligence Scale for Children, Third Edition (WISC-III). For the WJ-R analysis, scores from the Leiter-R Brief IQ results obtained from 105 children were examined to identify the pattern of correlation and similarity of mean scores between the two measures. The two fluid-reasoning subtests on the WJ-R (Analysis/Synthesis and Spatial Relations) showed the highest correlation with the Brief-IQ four-subtest set and indicated good construct validity of the Leiter-R. In addition, the means of the two batteries were reported to be consistent, which was reported to be supportive of the Growth Scale of the Leiter-R (Roid & Miller, 1997).

The WISC-II and Leiter-R VR Battery results obtained from 126 children were also examined for the purpose of cross-battery factor analysis. The results showed loading on one verbal factor and on one nonverbal/performance factor. A differentiation between the nonverbal factors did not emerge because the WISC-III tends to have a visualization factor, rather than fluid reasoning, in its performance subtests. However, a convergence of nonverbal elements from both measures reported (Roid & Miller, 1997).

The fairness of the Leiter-R when administered to ethnic groups was reported.

The Leiter-R was been found to have few significant differences between scores when results from Caucasian and Hispanic groups or Navajo and Normative samples were examined. In addition, the differential item functioning on the Leiter-R was examined by using Rasch item analysis. It was reported that the 10 VR Battery subtests were found to be "exceptionally free" from differential item functioning between Caucasian and Hispanic samples as well as between Caucasian and African-American samples. When using archival data to predict achievement of subjects, the ability of the Leiter-R results to predict mathematic scores among Caucasian and African-American children was found to be similar. Fairness of the Leiter-R between genders was also found to be consistent with one exception. The Attention Sustained (Attention Battery) scores obtained from females in the 11-20 year age group tended to be slightly higher than those obtained from males (Roid & Miller, 1997).

The Wechsler Intelligence Test for Children, Fourth Edition

Brief description and test development. The Wechsler Intelligence Test for Children, 4th Ed. (WISC-IV) is the current version of the Wechsler instrument for the assessment of intellectual functioning abilities of children and adolescents. It is intended for use with children ranging in age from 6 years, 0 months to 16 years, 11 months of age. It was developed to update the previous edition, the Wechsler Intelligence Scale for Children, 3rd Ed. (WISC-III), based on research in the areas of cognitive and neuropsychological assessment (Zhu & Weiss, 2005).

The WISC-IV test kit includes the administration manual, the technical and interpretive manual, the WISC-IV stimulus book 1, the block design set, the response

books 1 and 2, and scoring templates. It was published by The Psychological Corporation in 2003 (Wechsler, 2003).

The WISC-IV consists of 10 core subtests and five supplemental subtests, all of which produce scaled scores with an average of 10 and standard deviation of 3. The WISC-IV also provided four composite scores that are associated with different areas of functioning: the Verbal Comprehension Index, the Perceptual Reasoning Index, the Working Memory Index and the Processing Speed Index. The composite scores are represented at standard scores with mean of 100 and a standard deviation of 15. When following the standardized administration procedures, the directions for completing the WISC-IV subtests are read to the examinee using spoken language; some of the performance tests include additional demonstration of the tasks for the examinee (Wechsler, 2003).

The Verbal Comprehension Index score represents the child's ability to think and work with verbal concepts, verbal reasoning and comprehension ability, acquired knowledge, and ability to attend to verbal stimuli (Zhu & Weiss, 2005). Within the Verbal Comprehension Index, the Similarities subtest presents the examinee with two words or concepts and asks the examinee to indicate how they are similar. The Vocabulary subtest shows a picture or a written word and then asks the examinee to provide a definition. The Comprehension subtest asks the child to answer verbally presented questions based on his or her understanding of social situations. The supplemental Information subtest requires the examinee to answer questions that tap into one's general knowledge base. Finally, the supplemental Word Reasoning subtest

requires the child to identify a concept based on a series of increasingly specific clues (Wechsler, 2003; Zhu & Weiss, 2005).

The Perceptual Reasoning Index describes fluid reasoning, spatial processing, attention to visual detail and visual-motor integration abilities (Zhu & Weiss, 2005). Among the Perceptual Reasoning Index subtests, the Block Design subtest is a timed task that requires the examinee to view a constructed model or picture and then use red and white stimulus blocks to recreate the model or picture. The Picture Concepts subtest presents the child with rows of pictures and he or she is asked to choose one picture from each row that is related to one from each of the other rows. The Matrix Reasoning subtest requires the examinee to look at an incomplete matrix and then select the missing piece from a group of options. Picture Completion is a supplemental subtest that presents a picture with a missing feature that the child is asked to identify within a specific time limit (Wechsler, 2003; Zhu & Weiss, 2005).

The Working Memory Index is associated with the location at which incoming information is temporarily stored, calculations or transformations take place, and then an output is produced (Zhu & Weiss, 2005). In the Working Memory Index area, the Letter-Number Sequencing subtest requires the examinee to listen to a sequence of random letters and numbers and then recall them in numerical and alphabetical order. The supplemental Arithmetic subtest requires the examinee to mentally solve arithmetic problems (Wechsler, 2003; Zhu & Weiss, 2005).

The Processing Speed Index is associated with the speed at which a child can accurately process simple or routine information (Zhu & Weiss, 2005). Within the Processing Speed Index, all of the subtests are timed. The Coding subtest presents the

child with a key containing geometric symbols and then the child is asked to copy them in the corresponding areas below the symbols. The Symbol Search subtest requires the child to scan a row of symbols and indicate if one or more target symbols is or is not present. The supplemental Cancellation subtest requires the child to scan an array of designs and mark specific target pictures (Wechsler, 2003; Zhu & Weiss, 2005).

Historically, the WISC-IV was derived from the original Wechsler-Bellevue Intelligence scale developed in 1939. The Wechsler tests have been reported to be the most frequently used assessment instruments, and their clinical utility and psychometric properties have been extensively researched. One traditional criticism of the Wechsler scales has been that they have lacked a strong theoretical foundation (Zhu and Weiss, 2005).

While the contents of Wechsler's original scales were based on their clinical utility, Zhu and Weiss (2005) suggested that the theoretical underpinnings of Wechsler's original scales were derived from work by two theorists, Charles E. Spearman and Edward L. Thorndike. Spearman's general intelligence theory, or Spearman's *g*, was evident in Wechsler's belief that intelligence is not equal to one's intellectual abilities. Instead, it is a *global* entity that is a collective of specific, qualitative different abilities. Intelligence is a multifaceted and multidetermined, and it required a global capacity that allows a person to understand and interact with the world. Thorndike influenced Wechsler's decision to include a wide array of subtests in his instruments. Wechsler believed that intelligence can manifest itself in multiple ways and an intelligence scale must employ as many different tests as possible to be both effective and fair.

Wechsler approached the original Wechsler-Bellevue test with the view that

intelligence is a global concept because it is associated with the whole of an individual's behavior and is at the same time specific because it includes elements that are specific and separate from one another (Wechsler, 2003). Wechsler pioneered the practice of grouping subtests into Verbal and Performance scales in addition to providing an overall composite score. These groupings were originally based on clinical and practical reasons. Wechsler did not intend to imply that only verbal and performance abilities were measured by these subtests, but it was one of many ways the subtests could be grouped (Zhu & Weiss, 2005).

In recent years, factor analytic studies have supported the concept that there is a general intelligence as well as provided evidence that intelligence is composed of more specific abilities that combine into several higher-order ability domains. These studies have also supported the theoretical foundation of Wechsler's measures (Zhu &Weiss, 2005). Wechsler (2003) defined intelligence as the capacity an individual possesses to purposefully act, rationally think, and effectively manage his or her environment. Other aspects, like planning and goal direction, enthusiasm, impulsiveness and persistence were also influential on a person's intelligent behavior. While these other aspects of intelligence were not directly measured by his scales, they did influence the individual's performance on the tasks as well as in all other challenges faced during life.

According to Wechsler, the clinician should view each examinee as unique and consider individual attributes other than intelligence when interpreting test results. He believed that what was measured through his tests was simply what was overtly being tested, and the tests themselves were only a means to an end. The basis of what he was measuring was the capacity of an individual to make sense of one's world and develop

the resources to cope with the challenges that world presented. Therefore, the process of assessing a child's intelligence involved more than simply obtaining scores related to performance on a group of tasks (Wechsler, 2003).

As the Wechsler scales have developed, the theoretical basis of the Wechsler scores have been evident in other measures of intelligence. There is a high correlation between these measures and the Wechsler tests, which suggests that they are measuring similar aspects of intelligence. The Wechsler scales have shown themselves useful in identifying several neurodevelopmental challenges, such as learning disabilities and mental retardation. Many of the original subtests continue to be included in the current versions of the Wechsler tests as well as on other intellectual measures. In addition, additional subtests and composite scales have been included as research results have suggested the importance of working memory and processing speed as being important domains of intellectual functioning. Despite its clinical utility, the debate over whether Wechsler was lucky as he developed his tests, or if he made his choices through his keen insight into the nature of intelligence and it's measurement, continues today (Wechsler, 2003).

Normative process. The standardization of the WISC-IV was conducted using data obtained from 2,200 children, ages 6 years and 0 months to 16 years, 11 months, or from children belonging to various special groups. For example, the Arithmetic subtest was normed on a stratified sample of 1,100 children, or 100 per age group, from the general standardization sample. All test administrators and children involved in the norming process were paid for their participation. Some exclusionary conditions were established for participation in the standardization sample, which

included testing on any intelligence measure within the previous six months; uncorrected vision or hearing loss; a lack of fluency in English; inability to communicate verbally; current admission to a hospital or psychiatric facility; current medications that might depress performance, and previous diagnosis with a condition or illness that might depress performance, such as stroke or brain surgery. A representative proportion of children from these special groups were included in the normative group, however, to represent the approximately 5.7% within the population attending school (Wechsler, 2003).

The participants in the standardization sample were matched on several demographic characteristics as indicated for the general population by the 2000 U.S. Bureau of the Census. These characteristics included age, race, gender, parent education level and geographic region of residency. The standardization sample consisted of an equal number of males and females who were divided into 11 age groups with 200 participants in each group. The racial proportions of children in each age group were matched to that of the corresponding age group of the U.S. population according to the March 2000 Census information. The sample was also stratified according to five levels of parent educational attainment. Finally, participants were gathered based on area of residence to correspond to the four major geographic regions of the United States, as specified in the 2000 Census report (i.e., Northeast, South, Midwest and West) (Wechsler, 2003).

During the standardization process, start and discontinue points were developed to reduce the number of items administered. Start points had pass rates of 95% to ensure the majority of children would experience success on the first items and reduce the need

for reversal. The awarding of bonus points on the Block Design and Coding A subtests was also established to allow increased points for decreased completion time. The subtest scores were developed by determining the cumulative frequency distribution of raw scores for each age group, normalizing the distributions, and then calculating the appropriate scaled score for each raw score. The progression of scaled scores within each age group and between each age group was analyzed and minor irregularities were smoothed. To develop the composite scores, the five sums of age-corrected scaled scores within each composite were calculated for each child. The means of the five sums were determined for each age group. These averages reflected a high degree of similarity from age to age within each of the scales. No significant variation was found based on age in the mean sum of scaled scores for each composite. Examination of the sums of scaled scores showed that the sums were normally distributed, and the age groups were combined to construct equivalent composite scores. Then, for each composite, the sums of scaled scores were normalized, and the appropriate composite score was assigned to each of the sums of scaled scores and the score distributions were smoothed. Finally, to develop the age equivalent scores for the subtest raw scores, a total raw score that corresponded to a scaled score of 10 was identified for each age group. It is recommended that age equivalent scores be used with caution because they are commonly misinterpreted and have psychometric limitations (Zhu & Weiss, 2005).

Reliability. The internal consistency of the WISC-IV was examined using data collected from the standardization sample through the split-half method. The reliability coefficients between the scores of the two half-tests were then computed. The Coding, Symbol Search and Cancellation subtest scores were not included in this study because

the natures of the tasks indicated the split-half coefficients would not be adequate estimates of reliability. Average coefficients for the subtest scores across age groups ranged from 0.79 to 0.89. In addition, the subtests that had been included in the WISC-III showed improved reliability. The reliability coefficients for the WISC-IV composite scales ranged from 0.88 (Processing Speed) to 0.97 (Full Scale). The relatively higher coefficients seen on the composite scores were due to the broader sample of abilities represented in the composite scores, compared to the subtest scores (Wechsler, 2003; Zhu & Weiss, 2005).

A split-half method of examining the reliability of the WISC-IV for use among special populations was also conducted. The sample for this analysis included 661 children belonging to 16 groups, including: Intellectually Gifted; Mental Retardation-Mild Severity; Mental Retardation-Moderate Severity; Reading Disorder; Reading and Written Expression Disorders; Mathematics Disorder; Reading, Written Expression, and Mathematics Disorders; Learning Disorder and Attention-Deficit/Hyperactivity Disorder; Attention-Deficit/Hyperactivity Disorder; Expressive Language Disorder; Mixed Receptive-Expressive Language Disorder; Open Head Injury; Closed Head Injury; Autistic Disorder; Asperger's Disorder, and Motor Impairment (Wechsler, 2003). The internal consistency reliability coefficients for these groups were reported to be similar or higher than those obtained in the analysis of data from the normative sample. These results were interpreted to mean that the WISC-IV is useful as a reliable measure of intellectual functioning among the normative sample as well as with individuals belonging to the 16 special populations (Wechsler, 2003; Zhu & Weiss, 2005).

The test-retest reliability of the WISC-IV was additionally examined for the

subtest and composite scores. A sample of 243 children, approximately 18 to 27 from each age group, were administered the WISC-IV twice. The time between administrations ranged from 13 to 63 days with an average interval of 32 days. The test-retest reliability was estimated for five age groups: six to seven years, eight to nine years, 10 to 11 years, 12 to 13 years, and 14 to 16 years. The average corrected stability coefficient for the majority of the subtests fell in the 0.80's. The coefficient for the Vocabulary subtest was 0.92; others fell in the 0.70's. The average corrected coefficients for the composite scores were determined to be in the good to excellent range (e.g., high 0.80's or 0.90's). The results also indicated that the test-retest gains in scores were reduced on the Verbal Comprehension and Working Memory composites when compared to the gains on the Perceptual Reasoning and Processing Speed composites (Wechsler, 2003; Zhu & Weiss, 2005).

To evaluate the interscorer agreement on the WISC-IV, all of the protocols obtained from the normative sample were double-scored by two independent raters. Because most of the subtests have clearly defined and objective scoring criteria, the interrater reliability was found to be very high and coefficients ranged from 0.98 to 0.99. To additionally examine the interscorer reliability on the subtests with more subjective scoring procedures, scores from the Similarities, Vocabulary, Comprehension, Information, and Word Reasoning subtests were obtained from 60 randomly selected protocols within the normative sample. Four raters who had no previous experience with the WISC-IV then independently rescored the protocols. The interrater reliability coefficients were calculated and found to range from 0.95 to 0.98. These results were interpreted to mean that the WISC-IV has high interscorer reliability, even when an

evaluator has little to no prior experience scoring (Wechsler, 2003; Zhu & Weiss, 2005).

Validity. During the revision of the WISC-IV, the test developers made effort to ensure that the subtest and individual items sampled the domains of intellectual functioning that the WISC-IV was intended to measure. Literature and expert reviews related to the content of WISC-III were examined; and during the evaluation of the WISC-IV content, new subtests were also extensively studied through literature reviews, expert opinion, as well as empirical study (Wechsler, 2003).

The response process of the WISC-IV supported the conclusion that the examinee engages in the expected cognitive processes when responding to the test items. The response frequencies for the multiple-choice items were examined to determine if any commonly led to errors by many examinees. If an incorrect response was repeatedly offered as a correct response, it was evaluated to determine if it could plausibly be considered an unintentional acceptable answer by the examinees. The examinees were also questioned to identify the problem-solving approaches used and modifications were made as needed to distractor items (Wechlser, 2003).

Much research has been conducted to examine the internal structure of previous version of the Wechsler scales. For example, research to clarify the third factor, formerly referred to as Freedom from Distractibility, led to the refinement of the third factor, now called Processing Speed, and the identification of the fourth factor, Working Memory. Research on the intercorrelations of the current WISC-IV subtests was conducted with several a priori hypotheses in mind. First, it was expected that all subtests would show low to moderate correlation to one another because they were all measuring some aspect of a general intelligence factor, or *g*. Second, subtests associated with a particular

composite index would correlate at a higher level than with those associated with other indices. Third, as indicated by studies on prior versions of the Wechsler scales, some subtests would correlate more highly with *g* than others. In addition, the subtests with higher correlations to *g* would also correlate more highly with each other, particularly if they were associated with the same composite index (Wechsler, 2003).

The intersubtest correlations and the sums of scaled scores for each composite were calculated for all 11 of the age groups included in the WISC-IV normative sample. All intersubtest correlations were found to be statistically significant, and the pattern of intercorrelations was reported to be similar to that found on the WISC-III and some other Wechsler scales. The subtests associated with the Verbal Comprehension composite correlated most highly with one another, as well as with the Arithmetic and the Picture Completion subtests. The Arithmetic subtest requires a high level of auditory comprehension ability and the Picture Completion subtest has been shown to be related to verbal abilities in the past. A moderate correlation was indicated between the Verbal Comprehension subtests and the Perceptual Reasoning subtests. This was attributed to the relatively high association with *g* that is common among all of those subtests. Another finding that was consistent with studies conducted on the WISC-III was a moderate correlation between the Verbal Comprehension composite subtests and the Letter-Number Sequencing subtest and a slightly lower correlation with the Digit Span subtest (Wechsler, 2003).

The correlations between the Perceptual Reasoning composite subtests were almost as high as those found among the Verbal Comprehension composite subtests, which was attributed to their high correlations with *g* among both subtest groupings, and

the above-mentioned correlation with the Perceptual Reasoning subtest. The Perceptual Reasoning subtest also showed a moderate correlation with the Working Memory subtests, which suggested a likely role of working memory when completing tasks of fluid reasoning (Wechsler, 2003). The subtests related to the Working Memory composite correlated most highly with each other and with the Verbal Comprehension subtests. This was attributed to the auditory comprehension demands associated with the Working Memory tasks (Wechsler, 2003).

The Processing Speed composite subtests were also found to correlate most highly with each other. The Symbol Search and Coding subtests also showed moderate correlations with other subtests, which was attributed to the visual and motor abilities required to perform many of the WISC-IV tasks. This finding was also reported to be consistent with studies conducted on the WISC-III. Cancellation was reported to show the least correlation to *g* due to its minimal correlation to other subtests on the WISC-IV (Wechsler, 2003).

Exploratory factor analysis of the WISC-IV was conducted using two sets of data, which included only the core subtests for one set and the core with supplemental subtests for the other. The core subtest analysis used scores from the 2,200 children included in the normative sample, while the core/supplemental analysis used date from the 1,525 children from the normative sample who were also administered the Arithmetic subtest. The children's scores were divided into four age groups, which included ages 6 to 7, 8 to 10, 11-13, and 14-16. Separate factor analyses were conducted for each of the four age groups (Wechsler, 2003).

The results of the exploratory factor analysis using only the core subtests

confirmed that each subtest loaded most highly on its corresponding factor with one exception. The Picture Concepts subtest loaded almost equally on the Verbal Comprehension and Perceptual Reasoning factors for the 6 to 7 year old age group. However, this split-loading was not evident among older age groups (Wechsler, 2003b). The results from the core/supplemental subtests confirmed the predicted factor structure was evident. However a more complex pattern of secondary loading was seen on some subtests. A split-loading of the Picture Concepts subtest on the Verbal Comprehension and Perceptual Reasoning factors was again found in the 6 to 7 year old group. A small secondary loading of the Picture Completion subtest on the Verbal Comprehension index was seen across all age groups. The Information subtest showed a slight loading on the Working Memory factor among the 6 to 7 and 8 to 10-year-old age groups. In addition, Arithmetic, although clearly associated with the Working Memory factor, showed small factor loadings on the Verbal Comprehension among the 11-13 year old age group and on the Perceptual Reasoning factors among the 14-16 year old age group (Wechsler, 2003).

Confirmatory factor analysis, using models from one to four factors, were conducted to further test the factor structure of the WISC-IV using the core subtests. Analysis was conducted using one, two, three and four factor models. The reported results indicated that the four-factor model was the best fit, compared to the Null Model (i.e., no common factor) and the one-factor model. A similar confirmatory factor analysis was conducted using the core and supplemental subtests. This time, analysis was conducted using up to a five-factor model. These results were reported to support a four- and five-factor model, when compared to the null model and the one-factor model. On the five-factor model, the Arithmetic subtest loaded on the fifth factor. However, it

was reported that the five-factor model did not show a substantial improvement over the four-factor model, where Arithmetic loaded on the Working Memory factor (Wechsler, 2003).

Examination of the relationships between the WISC-IV and other measures was conducted to further establish the validity of the WISC-IV. The WISC-III was included among the other measures examined. Both versions of the Wechsler test were administered to 244 children, ranging in age from 6 to 16 years, in counterbalanced order with an average test interval period of 28 days. The composite scores obtained on the WISC-III were higher than the corresponding WISC-IV composite scores, and the main effect for test was found to be statistically significant. With respect to individual subtest scores, small effect sizes were identified for the Block Design, Similarities, Coding, Symbol Search and Picture Completion subtests (Wechsler, 2003).

The corrected correlation coefficients of the composite scales were reported as follows: WISC-IV VCI correlation with WISC-III VIQ = 0.87; WISC-IV PRI correlation with WISC-III PIQ = 0.74; WISC-IV WMI correlation with WISC-III FDI = 0.72, and the WISC-IV PSI correlation with WISC-III PSI = 0.81. To account for these results, it was reported that there was the greatest amount of change between the WMI and the FDI and the least amount of change between the PSI composites during the revision and development of the WISC-IV. It was also noted that the WISC-IV VCI and WISC-III VCI (r = .88) correlated similarly to the WISC-IV VCI and the WISC-III VIQ. These results were interpreted to indicate that the WISC-IV and WISC-III measured similar constructs (Wechsler, 2003).

Scores obtained from 550 children, ranging in age from 6 to 16, from the

WISC-IV and WIAT-II, administered across an average of 12 days, were analyzed to additionally explore the construct validity of the WISC-IV. The results indicated that the WISC-IV FSIQ score correlated highest with the WIAT-II Total Achievement score ($r = 0.87$), and the WISC-IV PSI composite score correlated the lowest ($r = 0.58$). In addition, the WISC-IV VCI correlated highly with the WIAT-II Reading and Oral Language composite scores and least with the WIAT-II Written Language composite score. The PRI composite correlated highly with the Mathematics composite score; the WMI correlated highly with the WIAT-II Reading composite; and the PSI correlated highly with the WIAT-II Written Language composite. These results were interpreted to indicate that the WISC-IV composite scores showed convergent and discriminate relationships to domains of achievement measured by the WIAT-II composite scores. The pattern of relationship was also reported to be similar to that found between the WISC-III and WIAT-II tests (Wechsler, 2003).

Several special group studies were conducted during the standardization process to examine the clinical utility of the WISC-IV for use with individuals associated with the special groups. These special groups included the following: children identified as gifted; children with mild or moderate mental retardation; children with learning disorders; children with learning disorders attention-deficit/hyperactivity disorder; children with attention-deficit/hyperactivity disorder; children with expressive language disorder; children with mixed receptive-expressive language disorder; children with traumatic brain injury; children with autistic disorder; children with Asperger's disorder, and children with motor impairment. A special group study that focused on the clinical use of the WISC-IV with D-HH children was not reported (Wechsler, 2003).

Independent research. In 2008, Krouse published a thesis on the reliability and validity of the WISC-IV for use with the D-HH Population. WISC-IV test results from 128 D-HH children were obtained for analysis from nine participants (i.e., school psychologists who work with D-HH children) from various areas of the United States. Archival scores obtained were included based on the following criteria: age ranging from six years, 0 months to 16 years, 11 months; significant hearing loss identified as having a hearing disability; prelingual onset of deafness found as occurring prior to age of 5 years; hearing loss as the primary disability if more than one disability was identified; and previous testing as part of a psychological evaluation, such as for educational placement or for clinical diagnosis.

The scores obtained from the D-HH sample were compared with those reported for the normative sample in the WISC-IV Technical and Interpretive Manual. The results suggested that, with the exception of the Block Design and Picture Completion subtests, the internal consistency reliability, obtained using split-half correlations, was significantly more reliable for the D-HH sample than those reported in the test manual (Krouse, 2008).

The validity of the WISC-IV with the D-HH sample on the Perceptual Reasoning Index (PRI) was significantly lower than the normative sample scores. Krouse (2008) indicated that this finding was contrary to past research on earlier versions of the Wechsler Scales from which the performance scales were recommended as the best measure of intellectual functioning of D-HH children. This finding was attributed to sample characteristics or differences in the language demands of the PRI subtests. In addition, Krause suggested that changes in what the PRI is measuring, compared to

previous versions of the Wechsler Scales. For example, the current subtests are more similar to motor-free nonverbal tests and contain fewer demands of manual dexterity, compared to the earlier scales, and the D-HH population has historically obtained relatively lower scores motor-free nonverbal tests. Finally, Krause suggested that the current PRI subtests are more representative of fluid intelligence than crystallized intelligence, compared to previous Performance Scales. Scores obtained from the Verbal Comprehension Index (VCI) were approximately one standard deviation below the mean obtained from the normative population, which was described as being consistent with previous research results. This indicated that the VCI is not appropriate for use with the D-HH population as a measure of intellectual functioning.

The subtest interrelationships were examined and 15 of the 44 correlations were found to be zero. Krouse (2008) interpreted this finding to mean that the validity of the WISC-IV for D-HH children was only partially supported. Additional examination indicated that the majority of these non-significant correlations involved the PSI subtests of Coding, Symbol Search or Cancellation, and the other two were between PSI and VCI subtests. All of the subtests within an index were found to be significantly correlated.

Krouse (2008) suggested additional research to further examine the reliability and validity of this measure with the D-HH population. However, several implications for practitioners who work with the D-HH population were proposed. One implication was that the WISC-IV is a reliable measure of intellectual functioning with the D-HH population. However, Krause warned that the current study did not determine the internal consistency reliability for all subtests and indices. Another implication was that sufficient support for the validity of the WISC-IV with this population was not

determined. Krause further suggested that the WISC-IV PRI may not function in the same manner as the WISC-III PIQ, although the publishers suggest that the PRI is similar to the Performance Scales of prior versions of the Wechsler Scales. Finally, the results implied that the WISC-IV VCI should not be used as a measure of intelligence for the D-HH population.

Other independent research studies have been conducted on the Wechsler Intelligence Scale for Children, 3rd Ed. (WISC-III). It has historically been suggested that examiners refrain from using, or use with caution, the Wechsler Verbal scales when testing deaf and hard of hearing examinees. The reason is typically that D-HH children consistently show lower scores on the verbal subtests and composites when compared to scores on the performance subtests and composites. More recently, studies examining the differential item functioning between deaf and hearing subjects have suggested that the Verbal scales are not appropriate for use with the deaf population (Braden, 2005).

Maller published a study in 1996 that compared the WISC-III verbal item results from 110 deaf children with scores from children who were identified as having similar measured ability from the WISC-III standardization sample. Because deaf children typically show lower scores on the WISC-III Verbal tests, the hearing sample was comprised of younger children than the deaf group. Maller questioned if the WISC-III Verbal scale item fit a Rasch Model, if the Verbal item difficulty was consistent between the two groups, if the items retained their order of difficulty across the groups, and if the item calibrations for the hearing sample fit the Rasch Model when using data from the deaf sample. All of the items from the Verbal portion of the WISC-III were translated

into ASL or PSE that is mostly ASL, except for the Vocabulary items, and care was taken to ensure that the deaf participants understood all items presented (Maller, 1996).

Maller (1996) found that not all of the WISC-III items fit the Rasch Model for all groups. Many of the items that did not fit the Rasch Model for the deaf group did appear to fit the model for the younger hearing group. This indicated that the response patterns of the two groups differed. Maller suggested that items not fitting the Rasch Model for the deaf sample should not be used when estimating ability levels of deaf examinees. Maller additionally found that many of the items showed poor fit across samples which suggested that the item difficulty estimates for hearing children were inconsistent with the response patterns of the deaf children (Maller, 1996).

The item difficulty across samples was not supported, and many of the items administered earlier in the test showed DIF against the deaf children, and many items administered later showed DIF in the favor of the deaf children. Although the mean scores for the tests were matched, these results further suggest that the pattern of differential item difficulty is not the same for the deaf and hearing children. The item difficulty was also found to be more spread out for the hearing subjects, indicating that the items were more discriminating for the hearing children (Maller, 1996).

Maller (1996) concluded that using the WISC-III Verbal scale with deaf children would yield questionable results. The author indicated that two questions should first be asked: (a) "What do deaf children know, and what should they know?" and; (b) "What are deaf children exposed to in school and at home?" (p. 163). Only after identifying the answers can the appropriateness of item content for deaf examinees be discussed.

Included in the WISC-IV administration manual is extensive information to assist

examiners who assess D-HH children. Braden (2005) provided detailed descriptions when hearing loss was the primary disability of the steps that examiners should take when administering the WISC-IV to D-HH children. They include evaluation of the needs of the examinee; evaluation professional expertise; determination of whether the examiner has the knowledge and ability to guide the assessment decisions; determination of whether the examiner has expertise in the examinee's preferred mode of communication; determination of whether the assessment of the examinee needs to be conducted with appropriate accommodations; determination of whether the examinee's behavior and test results suggest valid outcomes: and interpretation of the scores in light of the examinee's unique characteristics and the availability of relevant research and theory. Braden also indicated that the information related to each step is not based on research using the WISC-IV, but is based on research associated with the WISC-III and logical, although subjective, analyses.

Stanford-Binet Intelligence Scales, 5th Edition

Brief description and test development. Stanford-Binet Intelligence Scales, 5th Edition (SB5) is the latest revision of the measure originally developed by Binet and Simon in 1908, by Terman in 1916, and by Thorndike, Hagen and Sattler in 1986. The current edition was developed by Gale H. Roid and published by The Riverside Publishing Company in 2003 (Roid, 2003). The SB5 is an individually administered, multidimensional test of cognitive abilities for individuals ranging in age from 2 to 85+ years. The SB5 is unique in that each of its five factors are measured in both the Verbal and Nonverbal domains. Therefore, half of the test is language-reduced, or nonverbal. The Nonverbal section employs hand-on tasks that readily engage the

examinee, particularly young children and individuals who have lower cognitive functioning levels. To follow the standardized protocol for administration, some receptive language ability is required for the nonverbal section because a limited amount of spoken instructions are presented. However, the nonverbal subtests allow for nonverbal, or non-spoken, responses from the examinee. The Verbal tasks require receptive language ability to understand the spoken instructions, expressive language ability to provide responses, as well as some reading ability (Roid & Pomplun, 2005).

The SB5 produces Full Scale IQ, Nonverbal IQ and Verbal IQ scores, and the five subtest factor index scores for Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual-Spatial and Working Memory. These scores are represented as standard scores with a mean of 100 and standard deviation of 15. There are 10 subtests and two routing subtests administered at the start of the SB5 to determine developmental starting points for the remaining subtests. The subtest scores are expressed as scaled scores with a mean of 10 and a standard deviation of 3. There is also an Abbreviated Battery IQ that is derived from scores obtained on the two routing subtests and can be used as supplemental information to a battery of tests that have been administered or when a brief screening measure of intelligence is desired (Johnson, D'Amato, & Harrison, 2005).

The SB5 is theoretically based on the five-factor hierarchical model developed by Carroll in 1993, which was an extension of the work performed by Cattell and Horn. This model is now referred to as the Cattel-Horn-Carroll, or CHC, theory. Carroll suggested a three-stratum theory that built upon Cattel and Horn's work. Carroll suggested that the Cattell-Horn theory covered all of the major areas of intellectual functioning, but it needed to provide for a third-order *g* factor that accounted for the

correlations among the broad second-order factors (Roid, 2003).

After extensive review of the literature related to assessment, as well as discussion with experts in the fields of giftedness, special education, preschool assessment and adult clinical disorders, five factors were chosen for the basis of the SB5. The five factors included in the SB5 and their corresponding factors from the CHC theory are Fluid Reasoning (Fluid Intelligence or *Gf*); Knowledge (Crystalized Knowledge or *Gc*); Quantitative Reasoning (Quantitative Knowledge or *Gq*); Visual-Spatial Processing (Visual Processing or *Gv*), and Working Memory (Short-Term Memory or *Gsm*). These five factors have been identified as having the highest *g* loadings in the CHC model. They have also been shown to be predictive of academic achievement (Roid, 2003).

Fluid Reasoning ability is associated with the solving novel problems. Within the Fluid Reasoning factor is the nonverbal Object-Series/Matrices subtest. This subtest is also the first routing subtest administered. The verbal Fluid Reasoning subtests include Early Reasoning, Verbal Absurdities and Verbal Analogies (Roid & Pomplun, 2005).

The Knowledge factor is associated with one's fund of general information accumulated over time through experiences. Within the Knowledge factor are the nonverbal Procedural Knowledge and Picture Absurdities subtests and the verbal Vocabulary subtest. The Vocabulary subtest is also the second of the routing subtests administered (Roid & Pomplun, 2005).

Quantitative Reasoning is related to solving numerical problems, managing number concepts and solving word problems. Within the Quantitative Reasoning factor are the Nonverbal Quantitative Reasoning subtest and the Verbal Quantitative Reasoning subtest (Roid & Pomplun, 2005).

Visual-Spatial Processing is associated with the ability to see relationships among figural objects, determine spatial orientation, identify a whole from among diverse parts and see general patterns among visual stimuli. The Visual-Spatial Processing factor contains two nonverbal tasks, the Form Board and Form Patterns subtests, and one verbal task, the Position and Direction subtest (Roid and Pomplun, 2005).

Working Memory refers to the ability to hold information in short-term memory and mentally transform it in some manner. The Working Memory factor contains two nonverbal tasks, the Delayed Response and the Block Span subtests, and two verbal tasks, the Memory for Sentences and the Last Word subtests (Roid & Pomplun, 2005).

The nonverbal aspect of the SB5 was developed to address the needs of the multicultural nature of today's society. Efforts were made to develop equally balanced tasks that contained high verbal demands with those that minimized the need for verbally expressive language. While the nonverbal tasks do involve some brief statements used by the examiner when presenting the tasks and may involve internal verbal mediation, the responses require nonverbal pointing, movements, or assembly of tangible objects. Alfred Binet was aware that intelligent actions could exist without the use of language and that intelligence could be conceptualized as being without images or words. He believed that thought could occur with or without conscious elements (Roid, 2003).

Normative sample. The SB5 was standardized on a normative sample that consisted of 4,800 individuals ranging in age from 2 years to 85+ years. Subjects resided in four geographical United States Census areas, which included the northwest, midwest, south and the west. Subjects from urban and rural areas were selected within each geographical area. Care was made to avoid including more than two closely related

subjects, and approximately 5% of the subjects who were school-aged were enrolled in special education programs and were mainstreamed in regular education classrooms for more than 50% of their day. No modifications or adaptations of the standardized test administration were made, although they were made for some of the special studies (e.g., deaf or hard of hearing). Criteria to exclude participants included their having severe medical conditions, limited English language proficiency; possessing severe sensory or communication deficits, exhibiting severe behavioral or emotional disturbances, enrollment for those of school age in special education programs for more than 50% of the school day (Roid, 2003).

The standardization sample was stratified across several demographic variables to resemble the general population as reported by the U.S. Census Bureau in 2001. These stratification variables included age, gender; race/ethnicity, geographic residential region, and socioeconomic level. Among the sample, 30 age groups were developed. The genders within each age group were divided evenly, except among the elderly where higher percentage of females among the general population exists (Roid, 2003).

Several special groups were also tested when developing the SB5, including individuals with Attention-Deficit Hyperactivity Disorder, Autism, developmental disabilities, along with those identified as gifted or learning disabled, deaf or hard of hearing (Roid, 2003).

Reliability. The reliability of the SB5 relates to its ability to measure the true attributes of an individual and its consistency across items or time. The split-half method, corrected by the Spearman-Brown formula, was computed for the 10 subtests, each of the four IQ scores (Full Scale, Nonverbal, Verbal and Abbreviated Battery), and

the five factor index scores for each age group as well as the average coefficients across all age groups. The coefficients for the Full Scale IQ scores were high (0.97 to 0.98) and were shown to be consistent across all age groups. Because this is a sum of all 10 subtest scores, it was expected to be higher. Reliabilities for the Abbreviated Battery were found to average 0.91, which Roid (2003) indicated was excellent, particularly as the Abbreviated Better is comprised of only two subtest scores. The Verbal and Nonverbal IQ scores also showed excellent reliability (average of 0.95 and 0.96, respectively). In addition, the factor indexes produced average coefficients above 0.90 and had higher coefficients (averages ranging from 0.84 to 0.89) than the individual subtests (Roid, 2003).

In addition to the reliability coefficients described above, the standard error of measurement (*SEM*) was computed for each scaled score, IQ score and the Factor Indexes. However, instead of using the conventional approach of adding and subtracting the *SEM* from the score, the SB5 developers constructed confidence intervals around estimated true scores based on the standard error of estimation (*SEE*). This creates an asymmetrical interval because the *SEE* accounts for the regression of scores toward the mean of 100. In addition, the item response theory suggests that the precision of a test for estimating each ability level can be plotted as a curve. The curves for the Verbal, Nonverbal and Full Scale IQ scores show high levels of precision, particularly throughout the average age-range of the test. In addition, the precision at the advanced levels of performance also indicated that the SB5 is particularly useful for the screening of gifted individuals (Roid, 2003).

The test-retest reliability of the SB5 was evaluated by administering the same test

to a sample on an additional occasion. Four samples of subjects were administered the SB5 over two different administrations. The groups included children ages 2 through 5 years, individuals ranging from 6 through 20 years of age, adults ranging from 21 to 59 years of age, and individuals over 60 years of age. Among the groups, there were somewhat more females than males, and there was a relatively lower percentage of ethnic minorities in the group over age of 60 years, compared to the reported population in the United States (Roid, 2003).

It was expected that the test-retest coefficients of the subtest scaled scores, IQ standard scores and factor index standard scores would be reasonably high because the SB5 measures skills that are believed to be relatively stable over time. However, because there are situational factors that can impact the testing environment and conditions, the correlations were not expected to be as high as the internal-consistency correlations described above. In addition, the evaluators expected the mean scores to increase with the second administration as a result of test practice and familiarity with the measure. The test-retest correlations were corrected to account for the variability (Roid, 2003).

The resulting coefficients for the scaled subtest scores ranged from a low of 0.66 on the Nonverbal Working Memory, for the 21 to 59 year age group, to a high of 0.93 on the Verbal Knowledge, for the 21 to 59 year age group. The median correlations for the four age groups were 0.82, 0.87, 0.79 and 0.86, respectively. The correlations for the Abbreviated Battery IQ ranged from 0.84 to 0.88, and the Factor Index correlations ranged from 0.79 to 0.95. The Nonverbal IQ and Verbal IQ correlations were strongly correlated over time with coefficients of 0.89 and 0.95 respectively. Finally, the Full Scale IQ coefficients ranged from 0.93 to 0.95, and mean differences ranged from two

points in the 60+ age group to four points among the six to 20 year age group (Roid, 2003).

To evaluate interscorer reliability of the SB5, approximately 40 protocols from the normative sample having a score of 0, 40 protocols having scores of 1, and 40 protocols having scores of 2 were selected randomly until each subtest had at least 120 responses to study. Two trained examiners then rescored the record forms to create a second and third set of scores for each item to compare to the original examiner's scores. Each pair of item scores was then compared and a Pearson correlation was calculated for each subtest. The interscorer correlations ranged from 0.74 to 0.97 with a median interscorer coefficient of 0.90. The test developers interpreted these results to mean that the SB5 a high level of reliability that is comparable to other published measures of intellectual functioning (Roid, 2003).

Validity. To determine the validity of the SB5, the content validity was evaluated by obtaining feedback on the SB5 items from numerous researchers throughout the development of the test , both experts on assessment as well as clinical examiners. To evaluate the fairness of each item, the differential item functioning (DIF) was calculated statistically and experts from different points of view reviewed the items. The reviewers were from five racial/ethnic/linguistic groups (Black or African American, American Indian and Alaskan Native, Asian, Hispanic, white or Anglo American), five religious groups (Buddhist, Christian, Jewish, Hindu and Muslim), the two genders, and two groups of individuals who have disabilities (deaf and hard of hearing and another general disability category). More than 400 standardization items were studied and only five were deleted due to significant DIF results. Four were verbal items that had significant

DIF between the Black or African American and White or Anglo-American groups and one nonverbal item had significant gender DIF (Roid, 2003).

The internal consistency of the SB5 was also evaluated to determine if there was any bias in the construct validity. Using data obtained from the normative sample, the internal consistency reliability was calculated for Black or African American, Asian, Hispanic and White or Anglo-American groups among two age groups: 6 to 10 and 11 to 16 years. This resulted in 363 and 421 randomly selected cases in each age group respectively. The split-half subtest scores for the verbal and nonverbal subtests were correlated. There were no significant results among the comparisons between the minority and majority subgroups. The alpha coefficient was higher for the Hispanic group, ages 6 to 10, when compared to the White or Anglo-American group, which favored the Hispanic group (Roid, 2003).

The age trends of the SB5 were examined by looking at the difference in performance across age groups. Intellectual ability is generally considered to increase from birth to adulthood as the brain matures, followed by a gradual decline during the elderly years. Analysis of the SB5 raw scores showed increasing mean raw scores until the 50- to 60-year-old age range on the Verbal Knowledge factor index (i.e., accumulated knowledge). In contrast, the Nonverbal Visual-Spatial factor index average raw scores peaked at the 20- to 25-year-old age range, which was consistent with research on visual-spatial cognition abilities (Roid, 2003).

The factor structure of the SB5 was analyzed by examining scores from four age groups: 2 to 5 years, 6 to 10 years, 11 to 16 years, 17 to 50 years, and 51 to 85+ years of age. For the subtest correlations, each subtest was removed from the IQ score to compute

a corrected for inflation value of the coefficients that would result by leaving the subtest score in the IQ scores. The resulting coefficients were reported to be positive and uniform, as expected from a multifactor cognitive battery and supportive of the SB5's general construct validity (Roid, 2003).

To confirm the five-factor model of the SB5, split-half subtest scores from the normative sample, arranged into five age groups, were analyzed. The number of individuals in each age group ranged from 514 to 1,400. Confirmatory factor analysis was conducted to compare the results of one-factor through five-factor models. The results indicated that the best fit was with the five-factor model across all age groups, and the fit statistics improved as the number of factors increased (Roid, 2003).

Another study of the construct validity and fairness of the SB5 involved an examination of the factor structure across groups. Scores on the subtests from the four ethnic subgroups obtained from the normative sample were analyzed. None of the resulting *chi*-square values were significant at the recommended level. The SB5 developers concluded that the correlation matrices were similar across the four major ethnic groups studied (Roid, 2003).

To evaluate the criterion-related validity of the SB5, performance on the SB5 was compared to that of the SB-IV. A counter-balanced design was used, so approximately one half of the subjects were administered the SB5 first and the others were administered the SB-IV first. The correlations between the each corresponding factor score ranged from 0.64 (Working Memory) to 0.79 (Abstract/Visual Reasoning). The Full Scale IQ from the SB5 and the Composite SAS from the SB-IV correlation coefficient was 0.90 (Roid, 2003).

Performance by individuals on the SB5 and the Wechsler Intelligence Scale for Children, 3rd Ed. (WISC-III) was examined to further explore the criterion-related validity. Sixty-six children and adolescents, ages 6 to 16 years, evenly distributed by gender and from diverse racial and ethnic backgrounds, were administered both tests. The difference between the average SB5 and WISC-III Full Scale IQ scores was five points, and the overall Full Scale IQ correlation coefficient was 0.84. The correlation between the Verbal scales was found to be at 0.85. There were relatively lower correlations found between the Visual-Spatial and Working Memory factor index scores, which fell at 0.42 and 0.46, respectively. The SB5 developers indicated that this likely reflects the variation in the tasks and the manner in which they are scored, such as time bonuses present on the WISC-III but not on the SB5 (Roid, 2003).

The SB5 was additionally compared to the Wechsler Adult Intelligence Scale, 3rd Ed. (WAIS-III). A sample of 87 adults, ages 16 to 84 were tested using both measures. The majority of the sample were females and there was a relatively lower representation of minority ethnic groups, compared to the U.S. Census information. The correlation coefficient for the Full Scale IQ average scores was found to be 0.82, and the Verbal factor index correlation was 0.81. In addition, the Visual-Spatial and Performance IQ average score correlation coefficient was 0.72. The test developers indicated that this may have been due to the wide age spread of the sample or difference in the correction for variability. However, they concluded that these results indicated evidence for criterion-related validity between two independently developed measures that have different factorial structures (Roid, 2003).

An additional study, which included 29 deaf and hard of hearing individuals,

compared their results on the SB5 with scores obtained on the Universal Nonverbal Intelligence Test (UNIT). The sample group consisted of children and adolescents ranging from 6 to 19 years of age, 60% of which were female. The sample was also comprised of 58% White or Anglo-American, 21% Black or African American and 21% Hispanic individuals. The sample also represented a variety of parental education levels. The correlation coefficient of 0.57 was found between the UNIT Full Scale IQ and the SB5 Nonverbal IQ scores, and a coefficient of 0.60 was reported between the UNIT Full Scale IQ and the SB5 Visual-Spatial Processing factor index. These scores were considered to be sufficient evidence of concurrent validity considering the SB5 requires some receptive verbal ability from the examinee (Roid, 2003).

Predictive validity refers to how well a measure can predict performance in other areas. Performance on the SB5 and the Woodcock-Johnson Tests of Achievement, 3rd Ed. (WJIII-Ach) was compared between a group of 472 students ranging in age from 6 to 19 years. The sample included slightly more females and individuals from a variety of racial and ethnic backgrounds and parental education levels. The corrected correlation coefficients were found to range from 0.50 to 0.84. The lowest correlations were on the Basic Reading Skills and the Written Expression subtests, which the author indicated is often found when comparing intelligence and achievement test results. The highest correlations were found between the SB5 Verbal and Full Scale IQ scores and the more complex areas of achievement. These areas included Reading Comprehension, Math Reasoning and Academic Applications (Roid, 2003).

Another study involving the Wechsler Individual Achievement Test, 2nd Ed. (WIAT-II) examined the predictive validity of the SB5. Eighty children ranging in age

from 6 to 15 years, including slightly more females than males and representing a variety of ethnic backgrounds, were administered both measures. The correlations for the SB5 Verbal IQ and the WIAT-II Total Composite score was 0.83. Lower correlations tended to be found when the SB5 scores were compared to the WIAT-II reading and writing scores, and higher correlations were evident when the SB5 scores were compared to the math and oral language WIAT-II results. The author also indicated that the median correlation of 0.60 is consistent with the expected average correlation between IQ and achievement measures generally reported in the literature (Roid, 2003).

Comprehensive Test of Nonverbal Intelligence (CTONI)

Brief description and test development. The Comprehensive Test of Nonverbal Intelligence (CTONI) is a multi-dimensional, individually administered assessment instrument used to measure reasoning skills of individuals ranging in age from 6 years to 89 year, 11 months. The CTONI is a nonverbal test with standardized administration procedures that do not require spoken language from the examiner or examinee. Instructions are given through gestures and examples, and examinees point to answers when responding to items. It requires no manipulation of objects, reading, or writing. It contains six subtests that measure an individual's ability to find relationships among pictures of familiar items as well as unusual geometric designs. The subtests include Pictorial Analogies, Geometric Analogies, Pictorial Categories, Geometric Categories, Pictorial Sequences, and Geometric Sequences (Pearson, 2003).

The CTONI test kit includes three picture books containing the stimulus items that are presented to the examinee. The examiner documents responses on record forms. It was designed by Hammill, Pearson and Wiederholt and was published by PRO-ED in

1997. When developing the CTONI, reviewers examined a collection of 36 nonverbal tests. These included unidimensional nonverbal tests and specific nonverbal subtests from larger, multidimensional test batteries. The test developers determined three principles which emerged from this review. First, they believed that a nonverbal test should be administered either orally or in pantomime, based on the examiner's judgment and the given assessment situation. Second, the nonverbal test should measure three types of thinking ability: analogic reasoning; categorical formulation, and sequential reasoning. Third, the abilities should be measured through pictured objects as well as through geometric designs. Using these three principles as guides, the six subtests of the CTONI were developed. Developing a test that includes both pictorial and geometric stimuli that was consistent with that of other nonverbal tests was indicated by the authors to contribute to the content validity of the CTONI (Pearson, 2003).

The scores from the subtests are described as standard scores with a mean of 10 and a standard deviation of 3. The subtest scores are combined to develop three composite quotients: Pictorial Nonverbal Intelligence Quotient; Geometric Nonverbal Intelligence Quotient, and an overall Nonverbal Intelligence Quotient. The composite scores are described as standard scores with a mean of 100 and a standard deviation of 15 (Hammill et al., 1997).

Hammill, Pearson and Weiderholt theoretically based the CTONI on the work of several theorists, including Horn and Cattell, Das, Jensen and Wechsler. Cattell and Horn proposed their model of fluid and crystallized intelligence. Fluid intelligence is related to nonverbal mental tasks that tend to be less related to culture. Crystallized intelligence is related to acquired skills taught directly or indirectly through one's

environment. The CTONI tasks are more closely aligned with the construct of fluid intelligence (Hammill et al., 1997).

Das' model of intelligence was based on neuropsychological contributions of Luria. Using this model, intelligence is categorized as involving simultaneous processing, in which stimuli are arranged in a parallel manner to facilitate decision making, or as involving successive processing, in which the stimuli must be ordered sequentially to facilitate decision making. The CTONI subtests can also be understood in terms of Das' simultaneous and sequential processing theory (Hammill et al., 1997).

Jensen's model proposed a two-level theory or intelligence which included the associative and cognitive levels. Abilities at the associative level have a high degree of correspondence between the form of the stimulus and the form of the response. Abilities associated with the cognitive levels require some transformation of the stimulus before a response is made. All of the CTONI subtests can also be associated with Jensen's cognitive level of intelligence (Hammill et al., 1997).

Finally, Wechsler proposed no particular theory of intelligence, but did promote adherence to Spearman's concept of general intelligence, or *g*, and his subtests were grouped in such a manner to imply this theoretical orientation. In his test design, he attributed each subtest to a Verbal Scale or a Performance Scale, which developed a verbal-nonverbal dichotomy. The CTONI subtests are all easily recognized as being similar to tasks affiliated with Wechsler's Performance Scale (Hammill et al., 1997).

Of the CTONI subtests, the Pictorial Analogies subtest measures the ability to recognize the relationship between two objects and then identify a similar relationship between to different objects. The Geometric Analogies subtest involves a similar task

using geometric designs. The Pictorial Categories subtest measures one's ability to examine two related pictures and then select from among a different group of pictures the one that is related to the first two. The Geometric Categories subtest is a similar task that involves identifying the related geometric design. The Pictorial Sequences subtest measures the ability to select from a group the one picture that completes a sequence of actions depicted in three other pictures. The Geometric Sequences subtest involves a similar task using geometric designs (Hammill et al., 1997).

Normative sample. The CTONI was normed on a sample of 2,901 individuals who resided in 30 states and the District of Columbia. Efforts were made to gather a sample group that was representative of the United States' population as a whole. Subjects were included based on geographic region, and controlled for gender, race, residence, ethnicity, family income, educational attainment of parents, and disabling conditions. The sample was additionally stratified by age. The percentage of individuals for the controlled characteristics were matched to that published by the U.S. Bureau of the Census in 1997 for the school-aged population (ages 6 through 18 years, 11 months), and individuals 19 years and older (Pearson, 2003).

Reliability. The internal consistency reliability of the CTONI was examined using Chronbach's coefficient alpha method. Using scores from the normative sample, the coefficient alpha's for each subtest and composite score were calculated for the 19 age intervals. Reported results indicated that all but two of the coefficients from the subtests rounded to or exceeded 0.80. In addition, the composite coefficients were all greater than 0.90. To further investigate the internal consistency reliability, similar calculations were made for 10 selected subgroups within the school-aged sample. These groups included

Caucasoids, African Americans, American Indians, Panamanians, Asians, speakers of English as a Second Language, students diagnosed with learning disabilities, students who were deaf or hard of hearing, males, and females. Similar coefficients were found, with all subtest coefficients rounding to or exceeding 0.80 and the composite score coefficients all being greater than 0.90 (Hammill et al., 1997).

Test-retest reliability method was used to evaluate the stability of the CTONI. Scores from 30 students in the third grade and 30 students in the 11th grade were studied. Pantomime instructions were used during the first administration and oral instructions were presented during the second administration. With one exception, the reported test-retest coefficients for the subtests were greater than 0.80. In addition, the coefficients for the composite scores rounded to or exceeded 0.90 (Hammill et al., 1997).

Interscorer differences for the CTONI were also evaluated. Two individuals from the PRO-ED research department independently scored 50 protocols that were randomly selected among students ranging from 14 to 17 years of age. The resulting correlation coefficients for the subtest scores rounded to or exceeded 0.95 and the coefficients for the composite scores rounded to or exceeded 0.98. This, along with the reliability studies reported above, is considered by the test developers as evidence that the CTONI has a high level of reliability, and the CTONI was shown to be reliable for use with the ten subgroups' studies. In addition, when examining English-speaking, general education students, either the pantomime or the oral instruction method can be effectively used (Hammill et al., 1997).

Validity. The content validity of the CTONI was described previously in the manner in which the subtests align with earlier work and theories of intelligence from

Horn and Cattell, Das, Jensen and Wechsler. Additional statistical analysis has been performed to evaluate other types of validity for the CTONI (Hammill et al., 1997).

Two methods were used to explore the bias of the CTONI. The Item Response Theory (IRT) approach was used to compare the performance between five dichotomous groups: male/female, African American/non-African American, American Indian/non-American Indian, speakers of English as a Second Language (ESL)/non-ESL, and learning disables/non-learning disabled. Performance on less than 5% of the items across all of the group comparisons was significantly different at the 0.001 level of statistical significance. Because no test can ever be completely free of all bias, the authors suggested that the relatively low level of item bias found on the CTONI is within the acceptable range (Hammill et al., 1997).

The second approach used to explore bias in the CTONI was the Delta Scores approach developed by Jensen. These are linear transformations of the *z* scale. The procedure was applied to the five dichotomous groups identified above along with a deaf/hearing dichotomous group. The correlation coefficients for the six groups on the six CTONI subtests ranged from 0.97 to 0.99. The magnitude of these coefficients is considered to be very high and provide additional evidence that the CTONI contains little to no test bias (Hammill et al., 1997).

Criterion-prediction validity studies were reported by the CTONI authors. This terminology was used in place of criterion-related validity because the procedures evaluate the ability of a test to "predict" performance on other activities. One study compared CTONI scores with those from three criterion tests that were already in available from 43 elementary students identified with learning disabilities. The tests

included the WISC-III, the TONI-2 and the Peabody Picture Vocabulary Test, Revised (PPVT-R). The resulting correlations with the WISC-III ranged from 0.51 to 0.81, with the highest correlation between the CTONI Nonverbal IQ and the WISC-III Full Scale IQ scores. The correlation between the CTONI Nonverbal IQ and the TONI-2 Quotient was reported to be 0.82. Finally, the correlation between the CTONI and the PPVT-R Quotient was reported to be 0.74. These correlations were interpreted as suggesting that the CTONI is a valid measure of intellectual functioning, particularly when used for children with learning disabilities (Hammill et al., 1997).

Another study correlated CTONI scores with scores from the WISC-II Performance Scale and subtest scores obtained from 32 deaf students ranging in age from 8 to 18 years. The CTONI Pictorial Nonverbal IQ correlated with the WISC-III Performance scale coefficient was 0.87, the Geometric Nonverbal IQ coefficient was 0.85, and the CTONI Nonverbal IQ correlation coefficient was 0.90. Again, the authors reported that this supplied evidence of the validity of the CTONI as a measure of intellectual functioning, particularly among a group more likely to be assessed by a nonverbal measure (Hammill et al., 1997).

A third study correlated the CTONI with scores from the TONI-3 from 550 normal, non-disabled adults ranging from 20 through 89 years of age. The demographics of the adults in the third study were representative of the U.S. population in 1997 for race, ethnicity, social class and gender. The resulting correlation coefficients between the CTONI Nonverbal IQ and the TONI-3 Form A and Form B Quotients were 0.77 and 0.75 respectively. Again, the authors suggested that these results lend further support to the validity of the CTONI (Hammill et al., 1997).

The construct validity of the CTONI was evaluated through calculating the coefficients describing the relationship between performance of tasks and age of the examinee. The coefficients at each of the 19 age groupings of the CTONI reported indicated that the subtests are strongly related to age during the school years as the mean scores become larger as the subjects' ages increase. This pattern is consistent with current theories of intellectual development. The pattern of scores among adults are also consistent with the patterns reported on other tests of intellectual functioning, like the TONI-3 and the WAIS-III. The CTONI scores for adults tend to level between ages 19 and 59 years, and then show a decrease after age 60 (Hammill et al., 1997).

The performance on the CTONI between several groups was also evaluated. Using the standard scores from the school-aged members of the normative sample, a comparison was made with scores from 11 subgroups. The subgroups included males, females, Caucasoids, African Americans, Hispanic Americans, American Indians, Asians, ESL, learning disabled, deaf, and mental retardation. The mean score from each subgroup were supportive of the construct validity of the CTONI. The group with mental retardation scored more than two standard deviations below the average. The Caucasoid group scored slightly higher than the average. However, in all cases, the subtest standard scores rounded to 10, and the composite quotients were 102 or 103. The minority subgroups were reported to perform relatively well on the CTONI with average scores for subtests, with the exception of two, within the standard error of measurement for the test. In addition, the composite scores were within the average range. The authors purport that these results provide evidence that the CTONI contains little bias (Hammill et al., 1997).

Factor analysis was also conducted to explore the CTONI's construct validity.

The principal-components method was used for this process. Because the CTONI subtests measure some feature of nonverbal ability, it is expected that all of the subtests would load on a single factor. In addition, that factor would measure general nonverbal intellectual ability. The factor analysis did indicate that all six subtests loaded on a single factor, which is described as the Nonverbal Intelligence Quotient. The subtest factor correlations were reported to range from 0.50 to 0.71 (Hammill et al., 1997).

Cognitive Assessment System

Brief description and test development. The Cognitive Assessment System (CAS) was created by Jack A. Naglieri and J. P. Das in 1997. It is a multi-dimensional assessment instrument that is individually administered to children ranging from 5 years to 17 years of age. It consists of a total of 12 subtests that are organized into four scales. Each of the four scales is associated with the four areas of the PASS theory. The CAS produces a total Full Scale score and four PASS scale scores. These scores are represented as standard scores with a mean of 100 and standard deviation of 15. Two forms of the CAS can be administered: The Basic Battery, which consists of eight subtests (two per PASS scale); and the Standard Battery, which consists of all 12 subtests (Naglieri, 2005). The CAS test kit includes stimulus books and examinee response forms. The CAS was published by Riverside Publishing in 1997.

The CAS developers believed in the importance of linking psychological practice with theory. The specific cognitive processes that are associated with the PASS theory of psychological processes includes Planning, Attention, Simultaneous, and Successive cognitive processes. The CAS subtests were created to relate the PASS theory of

cognitive processes, and each task relates to some aspect of the PASS theory. (Naglieri & Das, 1997; Naglieri, 2005).

The PASS theory is based on four psychological processes derived from the neuropsychological work of A. R. Luria. The PASS theory emphasizes the examination of cognitive processes as they relate to performance, rather than the general verbal-nonverbal model of intellectual functioning seen in many measures of intellectual functioning. The PASS cognitive processes represent the foundations of intellectual functioning that together form an interrelated cognitive system that interacts with an individual's knowledge base and existing skills. By looking at a child's performance on the four PASS scales, information can be obtained that describes how a child thinks, identifies a child's individual strengths and areas of need (which can contribute to making effective differential diagnoses), and assess fairly for a child that which assists in the selection or design of effective interventions (Naglieri, 2005).

Planning is a skill that is critical to all activities in which an individual must solve a problem. It is the mental process by which one determines, applies and evaluates solutions to problems. To successfully complete CAS tasks associated with Planning, a child must develop a plan of action, evaluate the method of approach, monitor the effectiveness of the method, and revise or reject the plan as task demands alter. The child must also control the impulse to act without forethought (Naglieri & Das, 1997; Naglieri, 2005).

Attention is the mental process by which one selectively focuses attention on relevant stimuli while inhibiting responses to non-relevant, competing stimuli over time. Several types of attention have been identified. Focused attention requires attention

directed at a particular activity. Selective attention is required to inhibit responses to potentially distracting stimuli. Sustained attention is related to the variation of performance over a period of time. This can be altered by variations in the amount of effort required for problem solving over time (Naglieri & Das, 1997; Naglieri, 2005).

Simultaneous processing refers to the mental activity through which one integrates separate stimuli into a conceptual whole. Many simultaneous processing tasks contain spatial aspects. However, simultaneous processing tasks can be verbal or nonverbal in nature as long as the cognitive demand consists of the integration of information. Simultaneous processing tasks include the integration of parts into a single whole with an understanding of the logical relationships among the parts and the meaning of the whole (Naglieri & Das, 1997; Naglieri, 2005).

Successive processing involves working with stimuli in a specific order that forms a chain-like progression. Each element must be specifically related to those that precede it and the stimuli are not otherwise interrelated. Tasks of successive processing require the use, repetition, or comprehension of information based on the order of the information (Naglieri & Das, 1997; Naglieri, 2005).

The four PASS theory processes are interrelated constructs that function as a whole. Most cognitive tasks require most, if not all, of the PASS processes. However, not every PASS process will be equally involved in every task. Because the processes are interrelated, the CAS was developed to provide a comprehensive understanding of a child's ability in all areas in order to assist with educational planning (Naglieri, 2005).

The CAS contains three subtests that are associated with each of the PASS areas of cognitive processes. Among the Planning subtests, the Matching Numbers subtest is a

timed task that requires the examinee to identify two matching numbers within a row of numbers. The Planned Codes subtest is a timed task that presents a legend of codes and the examinee is required to enter the appropriate codes into the empty boxes on the page below. The Planned Connections subtest is a timed task that requires a child to connect numbers or number and letters in a sequential order (Naglieri & Das, 1997; Naglieri, 2005).

In the area of Attention, the Expressive Attention subtest requires the child to state a feature of a pictured object (e.g., size or color) based on the real aspects of the item and not on the manner in which it is portrayed. For example, the child might be asked to state whether an animal is big in real life, while the provided image of the animal is smaller than others. The Number Detection subtest requires the child to only underline numbers that are printed in a particular manner. The Receptive Attention subtest requires the examinee first to underline pairs of pictures that identical, and then to underline pairs of pictures that are lexically similar but are depicted differently (Naglieri & Das, 1997; Naglieri, 2005).

Of the Simultaneous subtests, the Nonverbal Matrices subtest requires the examinee to determine how the parts of shapes or designs are related and then chose from six options the one that also relates best. The Verbal-Spatial Relations subtest presents the child with six drawings that depict objects and shapes in specific configurations followed by a printed question. The examiner reads the question and the child indicates which drawing best matches the verbal description included in the question. On the Figure Memory subtest, a child is shown a figure for five seconds. The child is then shown the same design embedded in a larger and more complex design, and the child is

asked to trace the original design (Naglieri & Das, 1997; Naglieri, 2005).

The Successive subtests include the Word Series subtest. On this task, the child is asked to repeat words in the same order presented by the examiner. On the Sentence Repetition subtest, the examinee listens to sentences that contain color words (instead of nouns, verbs, etc.) that the child is then asked to repeat verbatim. The Speech Rate subtest is only administered to children ages five through seven years. On this task, the child is asked to repeat a three-word series of high-imagery, single- or double-syllable words in the same order, 10 times. The Sentence Questions subtest is administered to children ages eight through 17 years. This task requires the child to listen to a sentence containing color words, similar to those used in the Sentence Repetition subtest. The child is then asked questions about the sentences (Naglieri & Das, 1997; Naglieri, 2005).

Normative sample. When developing the normative sample group for the CAS, Naglieri and Das made effort to match the demographics of the sample to that of the 1990 U.S. census information. A total of 3,072 children, ages 5 through 17, were administered the CAS during the standardization process. This included 2,200 children whose scores made up the normative sample, and 872 children whose scores were also used in the reliability and validity studies. Data was collected from individuals residing in 68 cities across the United States. Participants were selected based on the following demographics: age, gender, race, Hispanic origin, region, community setting, classroom placement, educational classification, and parental educational attainment level. Data from an equal number of males and females was obtained across the entire age range. The percentage of each subject group belonging to the identified demographics was similar to those in the general population (Naglieri & Das, 1997).

Reliability. The reliability of the CAS was examined in several ways. For the Simultaneous and Successive subtests (except speech rate), subtest reliability coefficients were calculated using the split-half method. For the Planning and Attention subtests, as well as the Speech Rate subtest, the test-retest reliability method was used. This method was chosen because these tasks involve timing the examinee. The Full Scale and PASS Scale reliability coefficients were calculated by using the formula of reliability of linear combinations. Reliability coefficients calculated for each one-year age group and the average for each subtest, PASS Scale and Full Scale from the Basic and Standard Batteries were reported. The average reliability coefficients for the Planning subtests ranged from 0.75 to 0.82. The Attention subtest average reliability coefficients ranged from 0.77 to 0.80. Among the Simultaneous scale subtests, the average reliability coefficients ranged from 0.83 to 0.89. Finally, the Successive scale subtest average reliability coefficients ranged from 0.81 to 0.85 (Naglieri & Das, 1997).

On the Basic Battery, the Full Scale average reliability coefficient was 0.87. The average reliability coefficient for each of the PASS scales was 0.85, 0.84, 0.90, and 0.90, respectively. On the Standard Battery, the Full Scale average reliability coefficient was 0.96. The average reliability coefficients for each of the PASS scales were 0.88, 0.88, 0.93, and 0.93, respectively. The authors note that the greater reliability was found on the Full Scale and Pass Scales because there was a greater amount of information than on the individual subtest calculations. However, they note that all of the reliability coefficients obtained indicated that the CAS has high internal reliability (Naglieri & Das, 1997).

The test-retest reliability of the CAS was based on scores obtained from 215 children from the standardization sample. The interval of administration ranged from

nine to 73 days with an average time of 21 days. The children were grouped into three age groupings of 5 to 7, 8 to 11, and 12 to 17 years of age. The average corrected stability coefficients for the CAS subtests across the three age groups was 0.73. The average corrected stability coefficients for the Full Scale and the Basic and Standard Battery PASS Scales was 0.82 (Naglieri & Das, 1997).

Validity. As stated in the CAS manual, to maintain content validity, the test items on the CAS were developed through task analysis and experimental examination. This was an effort to make the tasks efficiently reflect the processes associated with the PASS theory. The Planning Scale subtests include relatively easy tasks but require the child to make a decision about how to approach the novel tasks. The strategies chosen are observable and can be used to enhance interpretation. The Attention Scale subtests all require the examination of the features of the stimuli as well as a decision to respond to one feature and not respond to another competing feature. The Simultaneous Scale subtests require the child to synthesize separate elements and relate them as a group using verbal and nonverbal information. The Successive Scale subtests require the examinee to manage stimuli that are presented in a specific order and the meaning is dependent on that order (Naglieri & Das, 1997).

The construct validity of the CAS was measured by examining the progression of scores obtained by children across age groups. The scores were expected to increase when the raw score of a task was based on the number correct or a ratio of the number correct as well as utilized time. Conversely, they were expected to decrease when the raw score of a task was based on time alone. According to the manual, raw scores were

computed for the entire standardization sample and demonstrated appropriate changes as the age of the examinee increased (Naglieri, 1997).

The correlations of the PASS subtests were computed, along with the correlations between each subtest and the Full Scale and PASS Scales on the Basic and Standard Batteries. These results indicated that each subtest from each PASS Scale correlated highest with its associated scale and lowest with the unassociated scales. The authors concluded that these results showed sufficient evidence for convergent and discriminate validity respectively of each subtest (Naglieri & Das, 1997).

Factor analysis was also conducted to provide additional support for the validity of the CAS. Confirmatory factor analysis was conducted across four age groups. For each age group, the PASS model was specified so each subtest was only allowed to load on its respective factor. The authors then evaluated the model through several fit indices. The reported results indicated that there was as good fit between the PASS model and the scores obtained from each of the four age groups. The Goodness-of-Fit Index results were above 0.90 and the Adjusted Goodness-of-Fit Index results were above 0.80. In addition, the Root Mean Squared Residual values were small. The authors reported that these results were similar to those obtained on other measures of cognitive ability, including the WISC-III, Differential Ability Scales, and the Woodcock-Johnson Psycho-Educational Batter-Revised. The maximum-likelihood factor loadings from the analysis also showed appropriate loadings of the subtests to their assigned factors (Naglieri, 1997).

Exploratory factor analysis was conducted using the same four age groupings. The results indicated that the one- and two-factor solutions were insufficient for all age

groups. The three-factor solution was sufficient for the 8 to 10 year group and the 11 to 13 year group, and the four-factor solution was sufficient for the 5 to 7 year group and the 14 to 17 year group. The authors concluded that the exploratory factor analysis provided some support for a four-factor solution and some support for a three-factor solution. The difference was related to the issue of combining or separating the planning and attention factors. The authors reported that the rationale to maintain four factors was based theoretically on the PASS theory, as well as on an empirical and clinical foundation. They believed that there is adequate theoretical support to maintain that planning and attention are interrelated but distinct constructs. In addition, there is clinical utility in separating the two constructs (Naglieri & Das, 1997).

As described in the manual, to explore the criterion-related validity of the CAS, the relationship of CAS results were compared to those of the Woodcock-Johnson-Revised Tests of Achievement (WJ-R). To do so, the CAS and selected WJ-R achievement subtests were administered to 1,600 children ages 5 through 17 years of age. The group of children was developed to match the demographic characteristics described in the 1990 U.S. census. The sample was divided into four age groups and median correlations within each age group were calculated and compared. The correlation of the WJ-R Cluster scores with the CAS Full Scale Standard Battery score was 0.73, and the correlation with the CAS Basic Battery Full Scale score was 0.74. The authors concluded that the PASS cognitive processes abilities are related to achievement (Naglieri & Das, 1997).

In another correlation study, the CAS scores and WJ-R scores were compared to those obtained from the WISC-III. Three samples of children were studied, which

included regular education students (n = 46), children with identified learning disabilities (n unreported), and children identified as being mentally retarded (n = 80). The reported results indicated that the CAS scores and PASS processes were similarly correlated to achievement across the three groups as was found by the WISC-III (Naglieri & Das, 1997).

The CAS results from the three groups of students listed above were compared with the scores obtained from the WISC-III to determine how the two measures of intellectual functioning correlated. The authors reported that among the regular education students the CAS Full Scale and the WISC-III Full Scale scores were comparable and the Scale scores showed a similar range. In addition, there was a significant correlation between the CAS Planning Scale and the WISC-III Performance IQ Perceptual Organization and Processing Speed Index scores. The CAS Simultaneous Scale score correlated significantly with all of the WISC-III Index scores, and the Successive correlated significantly with all but the Processing Speed Index score. Finally, the CAS Attention Scale score correlated significantly with the Performance and Processing Speed scores from the WISC-III. The authors concluded that for regular education students the CAS Simultaneous and Successive scores correlated most highly with the WISC-III, and the Planning and Attention scores were least correlated. Similar analysis with the learning disabled group of students showed that the CAS Simultaneous scale correlated most highly with the WISC-III. In addition, it was found that for the mentally retarded group of children the PASS processes were less related to the WISC-III scores (Naglieri & Das, 1997).

Studies in which black and white children were matched on as many demographic

variables as possible have also been conducted to see if there is any test bias based on the race of the examinee. The groups' mean scores were then examined and the effect sizes (differences between the means divided by the group's average standard deviation) were compared. Results from a variety of intellectual ability measures were reported. As indicated, tests with greater verbally-laden tasks showed greater effect sizes, compared to those that measure cognitive processing (e.g., CAS). Among the tests reported, the CAS was shown to have the least effect size of 0.26 (Naglieri, 2005).

One study to examine the predictive validity of the CAS compared scores on the CAS with those on the Woodcock-Johnson Tests of Achievement, Third Edition (WJ-III ACH) obtained from 119 children, ages 6 through 16. The scores obtained from the WJ-III ACH and the CAS were found to correlate at 0.69. When the scores were corrected for restrictions in age ranges, the correlation between scores rose to 0.83. The developers of the CAS suggested that performance on the CAS is useful in predicting academic achievement. High scores on the CAS suggest greater cognitive functioning which should be linked to higher academic performance. In addition, because tasks on the CAS do not rely on acquired knowledge as does most other measures of intellectual achievement, the CAS can be useful in identifying cognitive strengths for children who may come from disadvantaged environments or who have experienced academic challenges in the past (Naglieri, 2005).

Independent research. In a dissertation published in 1992, Ojile reported the results on a study in which the performance of approximately 50 D-HH children on the CAS was compared to the performance of a group of hearing children. The groups were divided into younger children, average of approximately 9 years old, and older children,

approximately 13 years old. Ojile found that on the simultaneous and successive tasks the younger D-HH children obtained relatively lower scores than the hearing group on the verbal tasks, while the older D-HH children obtained relatively lower scores than the hearing group on both verbal and nonverbal tasks. He also found that both age groups performed relatively lower on the planning tasks, both verbal and nonverbal. Ojile suggested that the lack of strategies and failure to identify important cues may have hindered the performance of the D-HH subjects. Ojile indicated that additional study with a larger sample size was necessary to further explore the performance of D-HH children on the CAS.

Chapter IV. Discussion

Summary of Results

The review of the literature pertaining to the use of the UNIT, LEITER-R, WISC-IV, SB5, CTONI and the CAS indicates that they are all well-developed measures of intellectual functioning. Each test is based on theoretical foundations that are accepted by the professional community. In addition, spoken language is not required for the administration of the UNIT, LEITER-R, and CTONI, and examinees are not required to respond using spoken language. The CAS includes nonverbal demonstration of instructions in addition to a verbal presentation, and the examinee can respond verbally or nonverbally. The WISC-IV and SB5 employ verbal presentation of test items and a variety of verbal and nonverbal responses from the examinee.

Each test in the standardization of measurements obtained a large sample of individuals who were matched for age, gender, race/ethnicity and parent education levels as indicated on U.S. Census data that was current at the time of standardization. The UNIT, LEITER-R, SB5 and CTONI all reported the inclusion of individuals who were D-HH in the normative sample group. The UNIT and the CAS reported a group of D-HH individuals who were included in the reliability and validity studies. The WISC-IV included individuals with corrected hearing loss in the standardization sample, but the proportion of the sample group who met this criteria was not indicated in the technical manual.

Bracken (1987) published an article that described standards for establishing adequate reliability of measurement instruments. He suggested that the internal consistency estimates for all subtests on a measure should meet or exceed 0.80 and that

the total test reliability should meet or exceed 0.90. Athanasiou (2000) further supported this criteria for establishing adequate reliability. All of the assessment manuals report adequate test reliability for subtest scores, although not all of the scores met or exceeded 0.80. The majority of the measures showed reliability scores on the various IQ, Battery or Composite scores that rounded up to or exceeded 0.80. Again, however, there was reported variation within the measures on different scales and with different age groups. The WISC-IV, SB5 and CTONI appeared to have the strongest composite reliability scores.

Bracken (1987) additionally suggested that the test-retest reliability of a measure should meet or exceed 0.90, which was also supported by Athanasiou (2000). With the exception of the LEITER-R, the reported test-retest reliability scores of the tests reviewed exceeded 0.80, and most approached or exceeded 0.90. The UNIT, WISC-IV, and CTONI showed the strongest test-retest reliability on reviewed composite scores of the measured samples.

Three of the measures reported additional studies that compared the performance of D-HH subjects with the similarly matched hearing subjects. The UNIT study, which included 106 D-HH children, found variation in mean scores that favored the hearing subjects, but the differences were less than expected on tasks with increased language demands. The LEITER-R found relatively lower Brief and Full Scale IQ scores among the deaf subjects. This was attributed to schooling difficulties or additional handicapping conditions among the D-HH sample. The CTONI manual reported a study that examined the internal consistency of the test, using scores from a D-HH subgroup. The study found

composite coefficients to exceed 0.90, which was similar to the coefficients found with the standardization sample.

Six aspects of construct validity that should be adequately evaluated for all assessment measures were described by Messick (1995) and were additionally supported by Athanasiou (2000). First, the content aspect of validity refers to how relevant and representative the content of an instrument is to the construct it purports to measure. Typically, construct validity is obtained through examination by a panel of individuals identified as expert in the area of concern. Second, the substantive aspect of validity refers to the foundation of a measurement on valid theory and that theoretical processes are engaged in by the examinee during the testing process. Third, the structural aspect of validity describes how the item content of each measure aligns with its theoretical foundation. It is generally identified through factor analysis, item analysis, and subtest patterns. Fourth, the generalizablity aspect of validity is associated with how well test results can be generalized across groups, time, settings and tasks. This is measured by analyzing criterion results across groups, comparing factor structures across groups, and by correlating scores from one instrument with another instrument that purports to measure the same construct. Generalization across raters is also part of the generalizability aspect. Fifth, the external aspect is measured through study of the convergent and discriminate validity. It refers to how scores relate to scores on other measures and reflect the direction and strength that would be expected given the construct being measured. Finally, the consequential aspect of validity is associated with the value and social implications and consequences of how test scores are used and interpreted.

As described by Braden and Niebling (2005), the American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME) published *The Standards for Educational and Psychological Testing (3rd Ed.)* in 1999. Within this revised publication, five sources of evidence that a measurement instrument is valid were described. These include test content, response processes, internal structure, relations to other variables, and consequences of testing. The specific standards used to determine validity should be chosen by the test developer to best meet needs. In addition, the term "construct" has been broadened to describe the concept or the characteristic that a test was designed to measure. The idea of validity has also been broadened to describe the degree to which the scores obtained on tests reflect the construct that the test claims to measure. The distinct content, construct and criterion-related validity that were previously sought have been replaced by the notion that a test requires multiple sources of validity (Braden and Neibling, 2005).

Each of the intellectual assessment measures reports that the test content was reviewed by a panel of experts. In addition, all of the intellectual assessment measures were described as being developed under the guidance of an accepted theory of intellectual functioning. Further, the subtest and composite groupings were additionally developed to facilitate reasoning and other thought processes that would be consistent with what would be expected by each measure's theoretical foundation. However, the test manuals do not describe investigation recommended by Messick (1995) and the AERA standards into the actual thought processes occurring within examinees while performing assessment tasks.

All of the measures reviewed report information related to validity of the tests as indicated by some combination of several methods of analysis, such as IRT, DIF, criterion-related validity, and confirmatory and exploratory factor analysis. Results from studies examining the construct and predictive validity of the test instruments were also reported in the manuals. In general, all six of the intellectual assessment measured provide adequate information regarding the reliability and validity of each test to merit publication and are accepted for use among members of the assessment community.

Are Measures Reliable and Valid for the Deaf and Hard-of-Hearing Population?

As described by Braden (2005), the AERA document developed in 1999 included guidelines for modifying accommodations of testing situations when required for examinees that have some sort of disability that would prevent them from receiving the standard administration of a test. These included making modifications to presentation format, response format, timing, test setting, only using portions of a test, and using substitute or alternate assessments. According to Braden (2005), the first four accommodations have a relatively small impact on the standardization of the instrument and are generally accepted by the testing community. The last two accommodations, however, can alter the content of the assessment measure, and some see this as sufficiently significant to make the testing process invalid.

Braden (1999) also reviewed two issues that can compromise test validity when making assessment accommodations for examinees with disabilities. One is construct under-representation. This occurs when an individual is assessed, but the construct or constructs of interest are not sufficiently represented, i. e. "under-represented", by the assessment. The second is construct-irrelevant variance. This occurs when an

assessment is actually examining performance that is not related to the domain of interest. Construct-irrelevant variance often makes a test too difficult or too easy for the test taker. For example, giving a test in English to an individual who is not an English speaker would make the test more difficult for that person. It would also introduce the irrelevant construct of knowledge and understanding of the English language to the test. An example of a test being made too easy would be when multiple-choice questions are written so that the response "All of the above" is only an option when it is actually the correct answer. An individual may become aware of that feature and be able to correctly respond to questions when the actual answer is not known (Braden, 1999).

Some have suggested that norms specific to the performance of D-HH examinees should be developed for each test and that D-HH individuals should be included in the general standardization sample. However, Braden (2005, 1985) suggests that specific norms for a group tend to produce trivial differences and are generally irrelevant. The inclusion of D-HH individuals in the general standardization sample group, or not, would have little impact on the overall validity of the test.

Braden (2005) indicated that the usefulness of a test with a specific group is evidenced by the identification of similar reliability and validity characteristics of the test when used with that particular group. However, at this time there are limited studies to examine the reliability and validity characteristics of particular tests when they are used with the D-HH population. Of the six test instruments examined in this document, only the CTONI reported on internal consistency studies with a subgroup of D-HH examinees. Additional independent research on the reliability and validity of the UNIT when administering it to D-HH examinees was also available: Maller (2000) found no

significant DIF between deaf and hearing groups; Krivitski et al. (2004) found no differences between scores from deaf and hearing groups, although some differences may exist on the Cube Design subtest; and Maller and French (2004) found support for the two primary factors (Memory and Reasoning) and the second factors (Symbolic and Nonsymbolic), although the Nonsymbolic factor may have a different meaning for the deaf group. Independent research on the WISC-IV by Krause (2008) found split-half reliability scores from a sample of D-HH children to be similar to those obtained from the normative sample. However, the validity of the PRI scores obtained from the D-HH sample were found to be different than those obtained from the normative group. It was suggested that the revised PRI subtests may measure different abilities compared to previous versions of the Wechsler Performance Scales, or that the subtests associated with the PRI may more resemble motor-reduced nonverbal tasks than past versions. Caution against using the VCI as a measure of intellectual ability for the D-HH population was supported by the results.

The studies by Maller (2000) and Maller and French (2004) analyzed test results that were obtained by the UNIT publishers during the standardization process, and the study on the WISC-IV by Karuse (2008) used data submitted by a variety of individuals who work with the D-HH population. Although this allowed for rapid access to data on a larger number of subjects, it also limited the control the authors had over the administration and scoring procedures. This lack of experimental control over the administration procedures may impacted the results obtained, particularly those obtained by Kraus who had to trust the professionals who submitted the archival data that the instruments they used were administered in a valid and reliable manner.

In addition, the two independent studies with results that suggested that the content of the measures may have different meanings for the D-HH population compared to the hearing population were obtained by Krivitski et al. (2004) and Kraus (2008), who did not use test data on D-HH subjects obtained from the publishers' standardization sample. Krivitski et al. used test scores that were obtained from tests administered by the primary author, who is fluent in ASL, and then compared these results to hearing counterparts who were obtained from the standardization sample. As reviewed above, Kraus obtained scores from a variety of professionals who work directly with the D-HH population. This also suggests that the administration procedures used when administering assessment measures to the D-HH population may have impacted the individual test results, which then influenced the data analysis outcomes.

Additional Factors that can Influence Deaf and Hard-of-Hearing Population's Performance on Measures of Intellectual Ability

In addition to concerns about the reliability and validity of a test when it is used for the D-HH population, there are several demographic characteristics within the D-HH group that have been found to impact test scores. The information related to these demographic characteristics is gathered from past research studies that did not include the six measures of intellectual ability reviewed above.

It has been found that verbal subtest and composite scores are correlated with the level of hearing loss for the individual. In contrast, performance-based subtest and composite scores do not show the same correlation (Braden, 2005). In addition, D-HH examinees often obtain Verbal IQ scores that are one standard deviation below the mean of hearing examinees, and lower scores are still evident for D-HH examinees on some

nonverbal tests of intelligence (Maller, 2003b). Braden states that many researchers are looking at the reduced relationship between degree of hearing loss and performance on language-reduced tests as an indication that the language-reduced tests are more equitable when assessing the intellectual abilities of D-HH individuals (Braden, 2005).

School placement of D-HH children has been linked to differences in performance on intellectual measures. It had been purported that deaf children who attend residential schools have lower IQ scores when compared to deaf children who attend non-residential school programs. Braden, Maller and Paquin (1993) conducted a longitudinal study on IQ scores obtained over a period of approximately three and one-half years. The subjects included a group deaf students with severe to profound hearing impairment who attended a school for the deaf as either residential students or as commuter students living nearby. No students at the residential school were hearing. Additional scores IQ scores were obtained from a group of D-HH students who attended a day program for D-HH students, located in several schools that include hearing students.

Among the groups, the D-HH children attending the day program showed little change in IQ scores over time. The deaf students attending school as residents showed an increase in scores to a level that was close to that of the group of students attending the day program. In addition, the group of commuter students attending the residential school showed an increase in scores that exceeded the level of the group of students attending the day program. Although this study involved a relatively small sample of D-HH children and there were many variables and selection criteria that could not be controlled, the authors suggest that residential school placement for D-HH children is not

necessarily as detrimental to the intellectual development of the student as was previously believed. The authors referred to the increased availability of D-HH role models, D-HH peer groups, and social interactions that promote fluent communication typically available at the residential school setting; they suggested these may have an impact on cognitive development when children are allowed to spend a length of time in a residential program (Braden et al. 1993).

Another variable that has been linked to variation in intellectual test scores is familial hearing status. Higher scores on IQ tests have been obtained from D-HH children with parents and/or siblings who are also D-HH, compared to D-HH children with hearing parents. Maller (2003b) outlines potential reasons for this difference in scores. First, it was believed that deaf children with deaf parents (DCDP) were exposed to meaningful communication earlier than deaf children with hearing parents (DCHP). However, deaf children with deaf siblings (DCDS) also showed higher IQ scores than DCHP. Then, the lower scores obtained from DCHP was thought to be related to additional coexisting disabilities that were impacting intellectual development. In addition, the scores from DCDP and DCDS were reportedly higher than hearing children's scores, so a generally higher IQ ability obtained through heredity was believed to be influencing the outcomes.

In Braden's meta-analysis, published in 1994, the results from numerous studies associated with IQ, age of onset of deafness, and degree of hearing loss were analyzed. He concluded that children who were prelingually deaf showed lower scores on verbal IQ measures, when compared with children who became deaf after approximately the age of 5 years. In addition, the degree of hearing loss was not correlated with performance on

nonverbal measures of intelligence. However, it was moderately to highly correlated with measured verbal intelligence ability (Braden, 2005; Braden 1994). As a result of this information, many researchers now view language-reduced intellectual measures as more equitable and fair when assessing the abilities of D-HH children (Braden, 2005).

There have been very few studies examining score differences between male and female D-HH children. Braden (1994) analyzed six existing studies that reported separate results of male and female D-HH individuals on a variety of tests. He found that younger females showed some improved performance on untimed tests (e.g., Draw-a-Person), and younger males showed better scores on mechanical aptitude tests. Among adolescent and adult subjects, females performed better on speeded pencil-paper tasks, like the Coding and Symbol Search subtests on the Wechsler scales. However, these differences were not statistically significant. Braden suggested that differences on individual subtests with larger sample sizes could show significant differences, but these differences were not sufficient to impact overall IQ scores or to influence the factor structure of the assessment measures. He concluded that the differences between the performance of male and female D-HH population are small and inconsistent, and similar to the differences seen between the male and female hearing population.

Slate and Fawcett (1996) administered the WISC-III and an achievement test to 47 D-HH students with sensorineural hearing loss ranging from mild to severe-profound and to two students with mixed hearing loss. The children were evaluated for three-year academic assessment purposes. The administration modes used included oralism and Total Communication. The results indicated that the boys' scores were almost one standard deviation above those obtained from the girls on the Performance Scale tests,

but the same difference was not seen on the achievement tests. Slate and Fawcett offered that the factor structure of the Performance Scale is different for boys than for girls. They also offered the possibility that girls who do well on tasks associated with the WISC-III Performance Scales are less likely to have additional challenges in the classroom and are therefore less likely to be referred for assessment.

In addition, some comparison studies on the performance on IQ measures between African-American and white D-HH children have been reported. As is typically reported among the hearing population, scores from African-American D-HH children showed lower performance IQ scores, which were approximately one standard deviation below scores obtained from white peers (Braden, 1994).

Differences in scores within the D-HH population have been found on nonverbal tests that require the manipulation of materials, when compared to nonverbal tests that are motor-free. (Braden, 2005). D-HH individuals tend to score in the normal range on nonverbal performance tasks that require the examinee to manipulate objects, such as puzzle or block design tasks. In contrast, D-HH scores tend to be about one-third of a standard deviation below average on motor-reduced tasks, such as matrices (Braden, 2005; Maller, 2003b). It has been suggested by Braden that manual dexterity helps D-HH individuals obtain bonus points for speed of completion when performing many of the object-manipulation tasks (Braden, 2005).

Recommendations for Future Research

The above review of six intellectual assessment measures and their use with the D-HH population has been an effort undertaken to organize the available literature on the historical considerations of assessment with this population and evaluate the

appropriateness for each instrument's use with this population. However, it has also raised many questions that remain unanswered.

One of the outcomes of this review indicates the limited amount of independent research that has been conducted on these assessment instruments with the D-HH population. There is a strong need for additional studies to help answer the questions related to the reliability and validity of each test when applied to the D-HH population. At this time, there does not appear to be sufficient research to fully support any of the six assessment instruments reviewed as reliable and valid for use with this group. Where limited research is available, there have been no studies to replicate or refute the results. As indicated by Braden (2005), it is necessary to determine that an instrument is functioning for the D-HH population in the same manner as with the general population on which it was standardized. Without information on the reliability and validity of a test when administered to any member of a subgroup, the interpretation of results and application of their meaning will be questionable for that individual. For example, it has been suggested that D-HH individuals perform differently on motor-reduced performance tasks than hearing counterparts, which may in turn impact the validity of an assessment instrument.

In addition, the determination of the reliability and validity of a measurement instrument appears to be more important than the development of specific D-HH norms when interpreting test results for this or for any subgroup. The inclusion of D-HH individuals in the standardization process of a measure is important as efforts are made to make the normative sample an accurate representation of the general population.

However, inclusion of group members in the standardization sample is not sufficient to ensure a test is appropriate for use with that group (Braden, 2005).

There are several factors associated with the D-HH population that may have a significant impact on test performance by members of this group and are worthy of future research. For instance, there may be effects from the various modifications and accommodations (i.e., presentation format, response format, timing, and test setting) that may be made during test administration. Although these modifications and accommodations are not perceived as having a significant impact on test results, there is minimal research to independently examine their potential effects. The alignment of the mode of communication used by the examiner and the preferred mode of communication of the examinee is also an area of question. In addition, the mode of communication used during test administration and its potential impact on test outcomes might vary from one assessment instrument to another. For instance, past research has indicated that it is difficult to communicate to D-HH individuals the need to work quickly on timed tasks.

More research is needed to further the understanding of how the type of school program that a D-HH individual attends impacts test results. In addition, the length of time that an individual has attended a residential or day program that is specifically for D-HH students has been shown to influence test results, and more study is needed to fully understand this variable (Braden et al., 1993).

Although the D-HH population share some common characteristics related to hearing loss, it is a heterogeneous group in most other areas. These areas of within-group variability have been shown to impact test outcomes and are in need of additional research. These areas of variability include, but are not limited to degree of hearing loss,

age at which hearing loss occurred (e.g., prelingually or postlingually deaf), etiology of hearing loss (e.g., congenital or adventitious), additional disabilities present, and having parents or other family members who are also D-HH. The variables that are present among the general population, such as racial/ethnic background and gender, are additional variables that may impact test results for the D-HH population. There is a current need for research on any or all of these variables in order to further understand how they may or may not affect the performance of D-HH individuals on measures of intellectual assessment.

Finally, there are questions that arise from the methods by which research on the D-HH population is conducted. For example, the few independent studies available for review described varying methods for obtaining subjects and data for analysis. There may be some impact on research results when data is obtained from archival scores or other sources compared to when tests are administered directly by the researchers. Examining the impact that research methodology has on results can aid with the development of future studies, as well as with the comparison and contrast of research results.

Meaning to Professionals and Deaf and Hard-of-Hearing Examinees

What does the information presented above mean to the professional who is required to provide a fair and accurate assessment of intellectual functioning for a D-HH individual? What does it mean to the D-HH individuals who wish to obtain a fair and accurate assessment of intellectual functioning? The six measures reviewed are the latest versions of each test and all are accepted by the professional testing community as being reliable and valid for use with the general population. However, there is a lack of

research and study on each of the six instruments for use specifically with the D-HH population. The UNIT, with three available published studies by independent researchers, shows the most support for being a valid measure among the six tests examined. The results obtained from independent research using D-HH samples shows support for the reliability and validity of the UNIT with that group. The reported independent results were also aligned with the information reported by the test publishers. One located study on the WISC-IV concluded that the reliability of the WISC-IV when administered to D-HH children was similar to that published in the test manual, and that the VCI was not appropriate for use with D-HH children. However, the study did not fully support the validity of the WISC-IV for use with this population and suggested that the test content might have different meaning for the D-HH population or that the latest revision of the measure may have emphasized motor-reduced tasks on which the D-HH population has been shown to obtain relatively lower scores. This may have been due to variables inherent in the study, such as using using data collected from a variety of professionals and a limited amount of control over test administration.

In 2007, a message was posted on the DeafEval listserve (http://health.groups.yahoo.com/group/Deaf-Eval) requesting information from researchers who might be in the process of conducting studies on nonverbal assessment measures and their use with the D-HH population. At that time, Ms. Hailey Krouse, who was a School Psychology Doctoral Candidate at North Carolina State University, responded (personal communication, April 3, 2007). She indicated that she was conducting a master's thesis on the topic of the reliability and validity of the WISC-IV with the D-HH population and was in the process of collecting sufficient data for analysis during the upcoming years.

This thesis was completed in 2008 and the results were described previously in this document. A similar request for information was again posted in 2010, but no responses associated with current research were obtained.

There is a dearth of research on the use of intellectual assessment measures and their specific reliability and validity when used with the D-HH population. The significant influence the test results have on the examinees' current and future lives requires clinicians to approach cognitive testing with this population with thoughtfulness and care. At this time, as indicated by following the suggestions made by AERA in 1999 appear to be the best-practice procedures for clinicians who must make accommodations when administering assessment measures to D-HH individuals. As measures are updated and new instruments are developed, it is this author's hope that research specifically related to test usage and the D-HH population will become a natural part of test development process.

REFERENCES

Anastasi, A. (1997). Historical antecedents of modern testing. In A. Anastasi & S. Urbina (Eds.), *Psychological Testing* (pp. 32-45). Upper Saddle River, NJ: Prentice-Hall, Inc.

Athanasiou, M. S. (2000). Current nonverbal assessment instruments: A comparison of psychometric integrity and test fairness. *Journal of Psychoeducational Assessment*, *18*(3), 211-229.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 5* (4), 313-326. doi: 10.1177/073428298700500402

Bracken, R. S. & McCallum, B. A. (1998). *Universal Nonverbal Intelligence Test.* Itasca, IL: Riverside.

Bracken, R. S. & McCallum, B. A. (2005). The universal nonverbal intelligence test. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 141-161). New York, NY: Kluwer Academic/Plenum.

Braden, J. P. (1985). WISC-R deaf norms reconsidered. *Journal of School Psychology, 23*, 375-382.

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

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

Braden, J. P. (1999). Accommodations in testing: Methods to ensure validity. *Assessment Focus Newsletter, 8*(1), 1-3. Retrieved from

http://www.pearsonassessments.com/NR/rdonlyres/2D0DA1D0-B37E-4365-B7A2-F1A785C9F9F1/0/Assess_Focus_Spring_99.pdf

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). San Diego, CA: Elsevier Academic Press.

Braden, J. P. & Athanasiou, 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, 2nd Ed.* (pp. 557-577). New York, NY: The Guilford Press.

Braden, J. P., Maller, S. J., & Paquin, M. M. (1993). The effects of residential versus day placement on the performance IQs of children with hearing impairment. *The Journal of Special Education, 26*(4), 423-433.

Braden, J. P. & Neibling, B. C. (2005). Using the joint test standards to evaluate the validity evidence of intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues, 2nd Ed.* (pp. 615-630). NY: The Guilford Press.

Brauer, B. A., Braden, J. P., Pollard, R. Q., & Hardy-Braz, S. T. (1998). Deaf and hard of hearing people. In J. H. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 297-315). Washington, DC: American Psychological Association.

Dearborn, W. F., Anderson, J. E., & Christiansen, A. O. (1916). Form board and

construction tests of mental ability. *The Journal of Educational Psychology, 7*(8), 445-458.

Furth, H. G. (1964). Research with the deaf: Implications for language and cognition. *Psychological Bulletin, 62*(3), 145-164.

Gallaudet University Library, Deaf-related Resources. (2004, July). Frequently asked questions. Retrieved from http://library.gallaudet.edu/dr/faq-statistics-deaf-us.html

Goetzinger, C. P. & Rousey, C. L. (1957). A study of the Wechsler Performance Scale (Form II) and the Knox Cube Test with deaf adolescents. *American Annals of the Deaf, 102*, 388-398.

Greenburger, D. (1889). Doubtful cases. *American Annals of the Deaf, 34*(2), 98-99.

Gregory, R. J. (1996). *Psychological testing: History, principles, and applications.* Needham Heights, MA: Allyn & Bacon.

Hammill, D. D., Pearson, N. A., & Wiederholt, J. L. (1997). Comprehensive test of nonverbal intelligence. Austin, TX: PRO-ED.

Harrington, T. (2004). Statistics: Deaf population of the United States. Retrieved from http://library.gallaudet.edu/dr/faq-statistics-deaf-us.html.

Hiskey, M. S. (1941). A new performance test for young deaf children. *Educational and Psychological Measurement, 1*, 217-232.

Henwood, P. G. & Pope-Davis, D. B. (1994). Disability as cultural diversity: Counseling the hearing impaired. *The Counseling Psychologist, 2*(3), 489-503.

Holt, J., Hotto, S., & Cole, K. (1994). Demographic aspects of hearing impairment: Questions and answers. Retrieved from:

http://gri.gallaudet.edu/Demographics/factsheet.html

Irion, T. W. H. (1941). The place of language in mental development. *American Annals of the Deaf, 86*, 364-373.

Johnson, J. A., D'Amato, R. C., & Harrison, M. L. (2005). Stanford-Binet intelligence scales, fifth edition. In R. A. Spies & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook* (pp. 975-984). Lincoln, NB: The University of Nebraska Press.

Keith, T. Z. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues, 2nd Ed.* (pp. 581-614). New York, NY: The Guilford Press.

Kohs, S. C. (1920). The block design tests. *Journal of Experimental Psychology, 3*, 357-376.

Krause, H. E. (2008). *The reliability and validity of the WISC-IV with deaf and hard-of-hearing children.* Retrieved from http://www.lib.ncsu.edu/resolver/1840.16/2149

Krivitski, E. C., McIntosh, D. E., Rothlisberg, B., & Finch, H. (2004). Profile Analysis of deaf children using the Universal Nonverbal Intelligence Test. *Journal of Psychoeducational Assessment, 22*(4), 338-350.

Lane, E. S. (1938). Measurement of the mental and educational ability of the deaf child. *Journal of Exceptional Children, 4*, 169-173.

Leigh, I. W. & Pollard, R. Q. (2003). Mental health and deaf adults. In M. Marschark & P. E. Spencer (Eds.), *Deaf studies, language, and education* (pp. 203-215). New York, NY: Oxford University Press.

MacKane, K. (1933). A comparison of the intelligence of deaf and hearing children. *Teachers College Contributions to Education, 585*, viii-47.

Maller, S. J. (1996). WISC-III verbal item invariance across samples of deaf and hearing children of similar measured ability. *Journal of Psychological Assessment, 14*, 152-156.

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

Maller, S. J. (2000). Item Invariance in four subtests of the Universal Nonverbal Intelligence Test (UNIT) across groups of deaf and hearing children. *Journal of Psychoeducational Assessment, 18*(3), 240-254.

Maller, S. J. (2003a). Best practices in detecting bias in nonverbal tests. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 23-47). New York, NY: Kluwer Academic/Plenum.

Maller, S. J. (2003b). Intellectual assessment of deaf people: A critical review of core issues and concepts. In M. Marschark & P. E. Spencer (Eds.), *Deaf studies, language, and education* (pp. 451-462). New York, NY: Oxford University Press.

Maller, S. J. & French, B. F. (2004). Universal Nonverbal Intelligence Test factor invariance across deaf and standardization samples. *Educational and Psychological Measurement, 64*(4), 647-660.

Marschark, M. (1997). *Raising and educating a deaf child.* New York, NY: Oxford University Press.

Marschark, M. (2003). Cognitive functioning in deaf adults and children. In M.

Marschark & P. E. Spencer (Eds.), *Deaf studies, language, and education* (pp. 451-462). New York, NY: Oxford University Press.

McCallum, S., Bracken, B., & Wasserman, J. (2001). *Essentials of nonverbal assessment.* New York, NY: John Wiley and Sons, Inc.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from a person's responses and performances as scientific inquiry into scores meaning. *American Psychologist, 50*(9), 741-749. doi: 10.1037/0003-066X.50.9.741

Moores, D. F. (1996). *Educating the deaf: Psychology, principles, and practices.* Boston, MA: Houghton Mifflin.

Myklebust, H. R. (1964). *The psychology of deafness: Sensory deprivation, learning, and adjustment.* New York, NY: Grune & Stratton, Inc.

Naglieri, J. A. (2005). The Cognitive Assessment System. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues, 2nd Ed.* (pp. 441-460). New York, NY: The Guilford Press.

Naglieri, J. A. & Das, J. *P. (1997). Cognitive Assessment System.* Itasca, IL: Riverside.

National Center for Health Statistics (NCHS). (n.d.). *Fast facts A to Z.* from http://www.cdc.gov/nchs/fastats/disable.htm

Oleron, P. (1950). A study of the intelligence of the deaf. *American Annals of the Deaf, 95*(2), 179-195.

Ojile, E. O. (1992). A preliminary investigation of the use of the Cognitive Assessment System (CAS) with the deaf and hearing. (Doctoral dissertation). University of Alberta, Canada. *Dissertation Abstracts International: Section A,* 53/06, 1849.

Pearson, N. (2003). Comprehensive Test of Nonverbal Intelligence. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 141-161). New York, NY: Kluwer Academic/Plenum.

Pintner, R. (1919). A non-language group intelligence test. *The Journal of Applied Psychology, 3*(3), 199-214.

Pintner, R. (1931). A group intelligence test suitable for younger deaf children. *Journal of Educational Psychology, 22*, 360-363.

Pintner, R. & Paterson, D. G. (1916). The form board ability of young deaf and hearing children. *The Psychological Clinic: A journal for the study and treatment of mental retardation and deviation, 9*, 234-237.

Pollard, R. Q. (1996). Professional psychology and deaf people: The emergence of a discipline. *American Psychologist, 51*(4), 389-396.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales (5th ed.) Technical Manual.* Itasca, IL: Riverside.

Roid, G. H. & Miller, L. J. (1997). *Leiter International Performance Scale – Revised.* Wood Dale, IL: Stoelting Co.

Roid, G. H. & Pomplun, M. (2005). Interpreting the Stanford-Binet Intelligence Scales, (5th ed.). In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues, 2nd Ed.* (pp. 325-343). New York, NY: The Guilford Press.

Simeonsson, R. J., Wax, T. M., & White, K. (2001). Assessment of children who are deaf or hard of hearing. In R. J. Simeonsson & S. L. Rosenthal (Eds.), *Psychological and developmental assessment: Children with disabilities and*

chronic conditions (pp. 248-266). New York, NY: The Guilford Press.

Slate, J. R. & Fawcett, J. (1996). Gender differences in Wechsler Performance scores of school-aged children who are deaf or hard of hearing. *American Annals of the Deaf, 141*(1), 19-23.

Sloan, W. (1959). Hiskey test of learning aptitude. In O. K. Buros (Ed.), *The fifth mental measurements yearbook* (pp. 409-410). Highland Park, NJ: The Gryphon Press.

Standards for educational and psychological testing (3rd ed.). (1999). Washington, DC: American Educational Research Association.

Steinberg, A. (1991). Issues in providing mental health services to hearing-impaired persons. *Hospital and Community Psychiatry, 42*(4), 380-389.

Sullivan, P. M. (1982). Administration modifications on the WISC-R Performance Scale with different categories of deaf children. *American Annals of the Deaf, 127*(6), 780-788.

Tharpe, A. M., Ashmead, D. H., & Rothpletz, A. M. (2002). Visual attention in children with normal hearing, children with hearing aids, and children with Cochlear Implants. *Journal of Speech, Language, and Hearing Research, 45*, 403-413.

Thorndike, R. M. (1997). The early history of intelligence testing. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 3-16). New York, NY: Guilford Press.

Vernon, M. (1967). Relationship or language to the thinking process. *Archives of General Psychiatry, 16*, 325-333.

Vernon, M. & Andrews, J. F. (1990). *The Psychology of Deafness: Understanding deaf*

and hard-of-hearing people. White Plains, NY: Longman.

Wasserman, J. D. & Tulsky, D. S. (2005). A history of intellectual assessment. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues, 2nd Ed.* (pp. 3-22). New York, NY: The Guilford Press.

Watts, W. J. (1979). The influence of language on the development of quantitative, spatial and social thinking in deaf children. *American Annals of the Deaf, 124*(1), 46-56.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children (4th ed.)*. San Antonio, TX: Harcourt Assessment.

Whitmer, D. A. (1949). The Pintner non-language primary mental test. In O. K. Buros (Ed.), *The third mental measurements tearbook* (pp. 237-238). Highland Park, NJ: The Gryphon Press.

Zeckel, A. (1939). A comparative intelligence test of groups of children born deaf and of good hearing, by means of the Porteus Test. *American Annals of the Deaf*, *84,* 114-123.

Zhu, J. & Weiss, L. (2005). The Wechsler scales. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues, 2nd Ed.* (pp. 297-324). New York, NY: The Guilford Press.

Table 1
Summary of General Information

	UNIT	Leiter-R	WISC-IV	SB5	CTONI	CAS
Age Range	5-17 years	2 years, 0 months to 20 years, 11 months	Six years to 16 years, 11 months	Two through 85+ years of age	Six through 89 years, 11 months	Five through 17 years of age
Administration Mode	Nonverbal	Nonverbal	Verbal	Verbal	Verbal or Pantomime, Administrator's choice	Combination of Verbal and Nonverbal Actions
Response Mode	Nonverbal	Nonverbal	Verbal with nonverbal responses on some subtests	One half of tasks require nonverbal response, one half require verbal	Nonverbal	Verbal or Nonverbal
Scores Produced	Six subtests; Abbreviated Battery; Standard Battery; five Quotient Scales – Memory, Reasoning, Symbolic, Nonsymbolic and Full Scale Intelligence (IQ)	Visualization and Reasoning Battery – 10 subtests; Brief IQ from group of four subtests; Full Scale IQ from group of six subtests	Ten core subtests and five supplemental; Verbal Comprehension, Perceptual Reasoning, Working Memory and Processing Speed Index scores, Full Scale IQ score	Ten subtests; two routing subtests; Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual-Spatial and Working Memory Factor Index scores; Nonverbal IQ, Verbal IQ and Full Scale IQ	Six subtests; three composite quotients including Pictorial Nonverbal IQ, Geometric Nonverbal IQ and overall Nonverbal IQ	12 subtests; four scales, Full Scale score and four PASS scale scores; Basic Batter of 8 subtests, Standard Battery of 12 subtests

(table continues)

	UNIT	Leiter-R	WISC-IV	SB5	CTONI	CAS
Theoretical Foundation	Two-tiered model of intelligence of memory and reasoning, with symbolic and nonsymbolic mediation within each tier; consistent with Jensen's two-tiered hierarchical model and the Cattell, Horn and Carroll CHC model	Carroll's three-stratum model of intelligence with general intelligence "*g*" at first level, eight ability domains at second level, more specific abilities at third level	Spearman's general "*g*" intelligence theory; Thorndike's influence on using array of subtests to measure "*g*"	Cattell-Horn-Carroll (CHC) five factor theory of Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual-Spatial Processing and Working Memory; also a third-order *g* factor to account for correlations among broad second-order factors	Horn and Cattell's theory of crystallized and fluid intelligence; Das' model of intelligence which is based on neuropsychological contributions of Luria of simultaneous and successive processing of information; Jensen's two-level theory of associative and cognitive levels of intelligence; Wechsler's theory of general intelligence or "*g*"	PASS theory includes: Planning; Attention; Simultaneous, and Successive cognitive processes; based on neuropsychological work by Luria; information on how the child performs the tasks provides information on how the child thinks
Normative Sample Group	2,100 children matched for 1995 U.S. Census data	VR Battery administered to 1,719 children matched to 1993 U.S. Census data	2,200 children matched for 2000 U.S. Census data	4.800 individuals ranging in age from two years to 85+ years matched to 2001 U.S. Census data	2,901 individuals grouped by age from six years to 18 years, 11 months and 19 years and older, matched to 1997 U.S. Census data	2,200 children ages five through 17 years of age matched for 1990 U.S. Census data;

(table continues)

	UNIT	Leiter-R	WISC-IV	SB5	CTONI	CAS
D-HH Subjects Included in Normative Sample Group	0.2% of normative sample identified with hearing impairment; additional 1,765 children for reliability, validity and fairness studies,	Included samples of children with clinical/exceptional characteristics, including 69 children with severe hearing impairment	Hearing impaired with correction included in normative group	D-HH subjects included in normative group	Deaf individuals included in normative group	None in normative group

Table 2
Summary of Reliability Results

	UNIT	Leiter-R	WISC-IV	SB5	CTONI
Internal Consistency/ Split-Half Method	Average coefficients approach or exceed minimum reliability standards (0.80) for normative sample and clinical populations	Average coefficients for IQ and Composite scores ranged from 0.88 to 0.93	Average coefficients from 0.79 to 0.89 on subtests, 0.88 to 0.97 on composite scores	Full Scale IQ coefficient ranged from 0.97 to 0.98 across all ages; average Abbreviated Battery coefficient of 0.91; Verbal and Nonverbal IQ coefficients were 0.95 and 0.96	All but two of coefficient alphas across age groups rounded to or exceeded 0.80; Composite coefficients all greater than 0.90
Test-Retest Reliability	Coefficients approached or exceeded 0.90 for all ages	IQ and Composite coefficients range from 0.70 to 0.90 across all age groups	Average coefficients for Composite scores in good to excellent ranges of high 0.80's to 0.90's	Abbreviated battery coefficients ranged from 0.84 to 0.88 across age groups; Factor Index correlations ranged from 0.79 to 0.95; Nonverbal IQ and Verbal IQ were 0.89 and 0.95; Full Scale IQ coefficients ranged from 0.93 to 0.95	60 students administered CTONI with pantomime first time and verbal instructions second time; all but one reported test-retest coefficients on subtests greater than 0.80 and composite coefficients rounded to or exceeded 0.90

(table continues)

	UNIT	Leiter-R	WISC-IV	SB5	CTONI
Comparison Studies including D-HH sample	106 D-HH with normative sample – All differences in mean scores favored the non-hearing-impaired group but were less than expected on tasks with increased language demands	Relatively lower Brief IQ and Full IQ scores for sever hearing impairment group, attributed to schooling difficulties or additional handicapping conditions	None reported	None reported	Additional internal consistency study with subgroups (included deaf subgroup) found coefficient alphas across age groups rounded to or exceeded 0.80; Composite coefficients all greater than 0.90

Table 3
Summary of Validity Results

	UNIT	Leiter-R	WISC-IV	SB5	CTONI	CAS
Expert Review	Yes	Yes	Yes	Yes	Yes	Yes
Item Response Theory (IRT) – assumes that items invariant across groups, and items are unidemensional	All items determined to have adequate item-fit statistics	All items reported to have exceptional fit	Not reported	Not reported	Found low item bias on less than 5% of items across all group comparisons	Not reported
Differential Item Functioning (DIF) – differences in the statistical properties of an item between groups of examinees of equal ability	Differences seen between two groups on two items in favor of minority groups	All items reported to be fair across gender and ethnicity	Not reported	Four verbal items and one nonverbal item removed due to DIF	Delta Scores approach with groups, which included a deaf/hearing group comparison, found correlation coefficients ranging from 0.97 to 0.99 across all groups to suggests little to no test bias	CAS was among a group of measures administered to black and white children, CAS showed least effect size of 0.26

(table continues)

	UNIT	Leiter-R	WISC-IV	SB5	CTONI	CAS
Criterion Related Validity		Sensitive to identifying giftedness, less sensitive to ADHD and learning disabilities;	Compared performance of 244 children on WISC-IV and WISC-III; found higher composite scores on WISC-IV; correlations of corrected composite scales ranged from 0.72 to 0.87	Compared to SB4 results, correlations between four Factor scores ranged from 0.64 to 0.79; Full Scale IQ of SB5 correlated with SB-IV Composite SAS r = 0.90: No significant results found on minority or majority group	WISC-III correlations ranged from 0.51 to 0.81 with highest between CTONI Nonverbal IQ and WISCIII FSIQ; TONI-2 correlation was 0.82; PPVT-R correlation was 0.74	WJ-R Ach Cluster scores correlated with CAS Full Scale Battery score at 0.73, Basic Battery Full Scale score at 0.73; indicates PASS cognitive processes are related to achievement; additional study to compare CAS, WJ-R Ach and WISC-III with regular education, learning disabled and mentally retarded children showed similar correlations across age groups with WISC-III; Additional study showed CAS Full Scale and WISC-III FS scores comparable and had similar ranges

(table continues)

	UNIT	Leiter-R	WISC-IV	SB5	CTONI	CAS
Exploratory Factor Analysis	Two-factor structure on Standard Battery; third factor emerged on Extended Battery – Mazes subtest correlated to third factor of planning	Found four-factors of visualization, reasoning, attention and memory; some variability across age groups	Core and Supplemental subtests load on predicted four factors; some split-loading on subtests for certain age groups			One- and two-factor models insufficient at all age groups; three-factor solution sufficient for 8-10 and 11-13 year age groups; four-factor solution sufficient at 5-7 and 14-17 year age groups; conclusion that planning and attention are interrelated but distinct constructs
Additional Factor Information	General "*g*" factor exists over all subtests, but first-order memory and reasoning factors also emerged	Four-factor model for ages four to five years; five-factor model for ages six through 21 years of fluid reasoning, visualization, attention and recognition memory	All subtests show low to moderate correlation with each other because all measure some aspect of "*g*"; moderate correlation between Verbal Comprehension and Perceptual Reasoning subtests			Correlations of each subtest with Full Scale and PASS Scales on Basic and Standard Batteries showed sufficient convergent and discriminative validity patterns for each subtest

(table continues)

	UNIT	Leiter-R	WISC-IV	SB5	CTONI	CAS
Confirmatory Factor Analysis	A single general intelligence "*g*" factor present, thus fits into hierarchical theoretical model under which test was developed	Theoretical "*g*" loading higher on Figure Ground, Form Completion, Sequential Order and Associated Pairs for all ages; on Design Analogies, Repeated Patterns and Visual coding above age six; on Sequential Order and Paper Folding above age 11	Four-Factor model was best fit; five-factor model, with Arithmetic loading on fifth factor, was not substantially better than four-factor model	Best fit using five factor model for all age groups using split-half subtest scores; also compared four ethnic subgroups with normative sample and found similar correlations matrices across groups	All subtests load on a single factor, the Nonverbal Intelligence Quotient	Four factors of PASS model showed good fit statistics when each subtest allowed to load on its respective factor
Predictive Validity	WJ-R Achievement Broad Mathematics, Broad Knowledge and Skills Cluster correlated highly with UNIT FSIQ; lower correlations seen on WJ-R Broad Reading and Broad Written Language clusters	WIAT Reading Composite, WIAT Math Composite, WJ-R Broad Reading, WJ-R Broad Mathematics, WRAT-3 Word Reading and WRAT-3 Arithmetic correlated r = 0.62 to r = 0.82	WIAT-IV Total Achievement score correlated with WISC-IV FSIQ r = 0.87; lowest correlation with WISC-IV PSI r = 0.58	WJ-III Ach correlation coefficients ranged from 0.50 to 0.84; WIAT-II Total Composite score correlation of 0.83	CAS and WJ-III Ach correlation of 0.60; increased to 0.83 when corrected for restricted age ranges	WJ-R scores from regular education, special education and mentally retarded groups were compared with CAS and WISC-III scores; found CAS and PASS processes were similarly correlated to achievement across groups as the WISC-III scores

(table continues

	UNIT	Leiter-R	WISC-IV	SB5	CTONI	CAS
Construct Validity	WISC-III FSIQ corrected correlations with UNIT Abbreviated, Extended and FSIQ, r >= 0.84	WISC-II FSIQ and Performance IQ scores correlated highly with Leiter-R Brief and Full IQ, r >= 0.85; lower correlation with WISC-III Verbal IQ	WISC-III composite score correlations ranged from 0.72 through 0.87;	WISC-III with 66 children found 5-point difference in Full Scale IQ and overall correlation r = 0.84; Verbal Scale correlation 0.85; lower Visual-Spatial and Working Memory correlations r = 0.42 and 4 = 0.46; due to differences in scoring methods: With WAIS=III with 87 adults found Full Scale IQ r = 0.82; Visual Spatial with PIQ r = 0.72; similar criterion but different factor structure: 29 D-HH children and adolescents; UNIT FSIQ and SB5 nonverbal IQ r =0.57; UNIT FSIQ and SB5 Visual-Spatial Factor index r = of 0.60	Found subtest scores for children were strongly related to age of examinee which is a pattern consistent with intellectual theory, patterns for adults showed decrease after 60 years of age; Subgroup comparisons with normative scores showed lower scores among group with mental retardation, slightly higher scores among Caucasoid group, minority groups showed average scores on subtests with the exception of two	Progression of scores across age groups followed appropriate changes as the age of examinee increased

Table 4
Summary of Independent Research Results

	UNIT	Leiter-R	WISC-IV	SB5	CTONI	CAS
Independent Research	Maller (2000) examined DIF with 104 profoundly deaf subjects, no items with significant DIF	None found	Krause (2008) compared scores from 128 children to scores reported from normative sample; found support for reliability with group of D-HH children; Limited support for validity of PCI may be due to differences in what is measured or due to effect of motor-reduced nonverbal tasks; consistent limited support for using VCI with D-HH population	None found	None found	Olije (1991) found lower successive and simultaneous scores for younger D-HH children on verbal tasks; lower simultaneous and successive scores for older children on verbal and nonverbal tasks; lower scores for both for younger and older D-HH children on planning tasks
	Krivitski, McIntosh and Finch (2004) conducted profile analysis with 39 deaf and 39 hearing children, found no higher or lower performance between groups but some differences may exist on Cube Design subtest		Maller (1997) found different item responses between 110 deaf subjects compared to matched-ability but younger hearing subjects from standardization sample on WISC-III			

(table continues)

	Maller and French (2004) studied factor structure of UNIT with 102 deaf subjects, found support for primary factor model (Memory and Reasoning) with exception of Mazes subtest, second factor structure (Symbolic and Nonsymbolic) was supported but suggest nonsymbolic factor may have different meaning for deaf group		Braden (2005) described practices that may be useful to examiners who use the WISC-IV to assess D-HH children			