THE BAYLEY-III: A CONCURRENT VALIDITY STUDY FOR CHILDREN 24- THROUGH 42-MONTHS-OLD

Seraphim Mork

A thesis submitted in partial fulfillment
of the requirements for the
Master of Arts

Department of Psychology

Central Michigan University
Mount Pleasant, Michigan
October, 2013

ACKNOWLEDGEMENTS

I would like to express my gratitude to my thesis chair, Sharon Bradley-Johnson, Ed.D for her faith, understanding, support, compassion and assistance during this project. I would like to express my sincere thanks to Sandra Morgan, Ph.D. and Frimpomaa Ampaw, Ed.D for their patience, support and useful comments. I would also like to thank Barb Houghton for her wealth of information about resources and deadlines.

I would like to thank my family; particularly my mother and my sister for their support and undying faith in me.

ABSTRACT

THE BAYLEY-III: A CONCURRENT VALIDITY STUDY FOR CHILDREN 24- THROUGH 42-MONTHS-OLD

by Seraphim Mork

The Bayley Scales of Infant and Toddler Development –III (Bayley-III; Bayley, 2006) is the most commonly used measure of cognitive development for children from 1-42 months. However, there is limited data to examine its concurrent validity with other non-vocal cognitive measures. Because of the minimal data on concurrent validity, another comparison study with a test that has considerable support for its validity, such as the Cognitive Abilities Scale-2 (CAS-2; Bradley-Johnson & Johnson, 2001) would be beneficial. A sample of 22 children ages 2-0 to 3-6 residing in central Michigan were used for this study. The Bayley-III and the CAS-2 were administered by the author to the children at their day care centers and preschools according to the directions in the examiner's manuals. Both the CAS-2 Nonvocal Cognitive Quotient (NVCQ) and the overall General Cognitive Quotient (GCQ) which required speech on some items, were compared with Bayley-III results. Corrected correlations were determined. Both the CAS NVCQ and GCQ were significantly higher than Bayley-III results. Thus, the Bayley-III results are measuring skills similar to those of the CAS-2 although the two tests differ significantly in their results.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

The primary reason for conducting a cognitive assessment is to determine a child's current psychological and educational functioning and to use this information to make decisions regarding instruction based on the results. These decisions include diagnosing a child with a condition and intervening early to help meet a child's unique needs (Nagle, 2007).

Early intervention is important because research has shown that it improves the cognitive abilities of low-functioning children. The seminal longitudinal research investigation known as the "Carolina Abecedarian Project" examined early educational intervention effects for children from low socio-economic status backgrounds. Results indicated that the intellectual and academic achievement of the children increased as a result of the early intervention (Campbell, Pungello, Miller-Johnson, Burchinal, &Ramey, 2001 as cited in Parke & Gauvain, 2009). The importance of early academic intervention was also demonstrated by comparing the long-term intellectual and academic benefits gained by students from low-income families who were given preschool intervention followed by early elementary treatment, preschool treatment only, early elementary school treatment only, and students who did not receive any treatment. The findings showed that preschool treatment was relatively more effective than interventions that did not occur until early elementary school (Campbell & Ramey, 1995).

However, to determine which young children would benefit from such interventions requires accurate assessment procedures. Unfortunately, the current measures for assessing cognitive development for young preschoolers typically have not been good predictors of later childhood performance. According to Sattler (2008), generally, intelligence results obtained before the age of 5 are not reliable and should be interpreted cautiously. He also noted that some intelligence tests do not address the multidimensional nature of intelligence and thus, they provide a limited description of intellectual performance.

One important set of factors to consider when choosing an instrument to assess young children's cognitive development is the test's technical adequacy. Technical adequacy is usually evaluated by examining information on a test's standardization, reliability, adequacy of the test's floors and item gradients, provision of information on validity as well as sensitivity and specificity.

Standardization

Standardization refers to the methods used to select items for a test, administration of the items to a representative sample, analysis of the results, development of age norms and rules for the administering and scoring of responses to the items (Hegde & Pomaville, 2006).

Reliability

Reliability is the extent to which a child's scores on a test remain stable each time performance is measured (Anastasi & Urbina, 1997). Test-retest reliability describes the stability of performance on the test over time. Internal consistency reliability has to do with the degree to which items on a test measure the same dimension of behavior throughout (Nagle, 2007). When more than one form of a test is available, alternate-form reliability information describes the stability of results across the different forms. Finally, inter-examiner reliability describes the extent to which different examiners agree in scoring a test (Sattler, 2008).

Test Floors

According to Nagle (2007), a test has an adequate floor if standard scores extend at least two standard deviations below the mean. When tests lack a sufficient number of easy items, it may not be possible to differentiate among children's different performance levels on the test. If a test lacks an adequate floor, results may be inflated and thus, misleading (Bracken & Walker, 1997). Thus, with young children, those with developmental delays may be missed.

Item Gradients

Adequacy of item gradients is indicated by how rapidly standard scores increase as a result of a child's success or failure on a single test item (Bracken, 1987). If the change in standard scores is substantial, results may not discriminate well among various levels of children's performance.

Validity

Validity information describes how well a test measures the skills it is said to measure. Content validity, describes the degree to which test items are related to the skill being measured (Hegde & Pomaville, 2006). Construct validity is the extent to which an instrument measures a particular psychological construct; whereas concurrent validity describes the extent to which examinees' scores on a test correlate with their scores on another test considered to be valid that measures the same construct (Sattler, 2008). Predictive validity describes how well a test result predicts an examinee's future performance.

Sensitivity and Specificity

Sensitivity indicates how well a test correctly identifies individuals with a disorder. According to Glaros and Kline (1988), sensitivity indicates the ability of an assessment instrument to produce positive results for an individual who has the condition of interest. In contrast, specificity refers to the ability of an assessment instrument to give a negative result for an individual who does not have the condition of interest (Glaros & Kline, 1988). That is, specificity indicates how well a test identifies individuals as not having the disorder of interest.

Current Cognitive Options for Assessing Young Children

The following information describes the technical adequacy of norm-referenced options for assessing the cognitive development of children as young as 2 and 3 years of age. Because many of these children may not have developed intelligible speech yet, may not be able to speak, or their speech may be difficult to understand, the options chosen for this review include tests that provide non-vocal results or require vocalization on relatively few items. Although some of the tests reviewed cover areas besides cognitive development, only the cognitive scales will be reviewed because this study addresses the assessment of cognition. In addition, only the technical adequacy information for children ages 2 and 3 were reviewed. This is because that is the age group of interest for this project.

Salvia and Ysseldyke (2007) suggested that a minimum for acceptable reliability is .90 when results are used to make important decisions about children. This is the criterion that was used to evaluate the reliability of the following measures.

Bayley Scales of Infant and Toddler Development-Third Edition (Bayley-III)

According to Anderson et al, (2010), the most commonly used developmental measure for young children in research and clinical work is the Bayley-III (Bayley, 2006). The test has scales for Cognitive, Language, Motor, Social-Emotional and Adaptive behavior for children from 1-42 months.

Memory, exploratory behavior and sensorimotor development are some of the skills measured on the Cognitive Scale. Twenty-three percent of the items are timed; the use of timed items is a concern because children at this age are easily distracted and have a short attention span. Also, using timed items interferes with an examiner's ability to observe a child during testing. Further, timed items may not be appropriate for some children with motor impairments. Directions for the Bayley-III Cognitive Scale are given orally, but the test has no items that require a child to vocalize. Cognitive Scale results can be expressed as quotients (*M* = 100, *SD* = 15), scaled scores (*M* = 10, *SD* = 3), percentiles, age equivalents or growth scores (*M* = 500, *SD* = 100).

The test was normed on 1,700 children from January to October, 2004 with 100 children in each age group. The sample was similar to the 1997 U.S. Census data in terms of geographic distribution, parent education level, race/ethnicity and sex. No information was given on urban/rural distribution. Although children with abnormalities were not initially part of the standardization sample, a "representative proportion" of such children who took part in validity studies were later included in the sample. However, no descriptive information was provided for these children. Thus, a large, nationally representative sample was used to develop the norms for this test.

In terms of reliability, the internal consistency for 2- and 3-year-olds was given in 3 month intervals and the range was .92-.97. For test-retest reliability, the correlations pertaining to 2- and 3-year-olds was .86 for children from 19-26 and 33-42 months. The average test-retest interval was only 6 days (range 2-15 days). According to Bradley-Johnson and Johnson (2007) a short retest interval is acceptable because of the fast rate of growth in children at this age. No information on inter-examiner reliability is provided for the Cognitive Scale.

To address content validity, items were said to be selected based on a review of items on the second edition of the scale as well as information from the literature. Because the item sample is very small, the Bayley-III should be used with caution when planning instruction. Also some items such as completing pegboards within a particular duration are not of educational importance. Construct validity information shows that the Cognitive Scale correlates with the Bayley Motor and Language Scales at .39 to .51 respectively. When special populations including children with Down Syndrome, Pervasive Developmental Disorder, Cerebral Palsy, Specific Language Impairment, those at risk for Developmental Delay, those with asphyxiation at birth, Prenatal Alcohol Exposure, those Small for Gestational Age, and Premature or Low Birth Weight children were assessed using the Bayley-III, as expected, these children scored lower than their normal counterparts. Exceptions were those who were premature or those who were small for their gestational age.

In terms of concurrent validity, the Cognitive Scale of the Bayley-III was compared with the Mental Scale of the Bayley-II (Bayley, 1993); the correlation was .60. With the -Full Scale IQ of the Wechsler Preschool & Primary Scale of Intelligence-Third Edition (WPPSI-3; Wechsler, 2002), the correlation was .79. The Bayley-II Cognitive Scale results were compared with results from areas related to cognition as well. The Total Language Composite of the Preschool Language Scale-Fourth Edition (Zimmerman, Steiner, & Pond, 2002) correlated at .57 with the Cognitive Scales; the Peabody Developmental Motor Scales-Second Edition (Folio & Fewell, 2000), had correlations with the Bayley-III ranging from .25-.51, and the Adaptive Behavior Assessment System-Second Edition (Harrison & Oakland, 2003) had a correlation range of .04-.42. No information was available on predictive validity.

Adequate floors begin at 16 days and there are no problems with item gradients. Overall the Bayley-III is a relatively new version of a very frequently used instrument. The test has a large representative norm sample and the Cognitive section does not require vocalization.

In terms of reliability, the internal consistency for 2- and 3-year- olds is high. Test-retest reliability is somewhat low and there are no data on inter-scorer reliability. The test has some timed items and some of the items are not of educational importance. Also, the Bayley-III should be used with caution when planning instruction because it has a small item sample for various skills. The Bayley-III has demonstrated modest correlations with the Bayley-II and a high correlation with the WPPSI-III and low to modest correlations with language, motor and adaptive instruments. However, no information was found on predictive validity for this edition of the test.

Cognitive Abilities Scale-Second Edition (CAS-2)

The CAS-2 (Bradley-Johnson & Johnson, 2001) measures cognitive ability in children from 3 months through 3 years. The test consists of two forms: the Infant Form (3-23 months) and the Preschool Form (24-47 months). Only the Preschool Form will be reviewed.
The Preschool Form covers the areas of oral language, reading, mathematics, handwriting and enabling behaviors. The overall scores are expressed as the General Cognitive Quotient (GCQ; *M* = 100, *SD* = 15). For children who are unable or unwilling to speak or whose speech cannot be understood, there is a Non-vocal Cognitive Quotient (NCQ; *M* = 100, *SD* = 15) based on test items that do not require vocalization. Directions for the test are given orally by the examiner. Both results can also be expressed as percentiles or age equivalents. Because of the irregularity of young children’s responses, and to gain enough information for planning instruction, a number of items are assessed three times each.

In terms of standardization, the test was normed on 1,106 children from October 1997 through August 1999 with 248 through 305 children for each one-year age level. The sample was similar to the 1997 U.S. Census data in terms of geographic distribution, gender, race, ethnicity, urban/rural residence, and educational background of the parents. The sample included children with disabilities; also the children's demographic characteristics are representative for the different age levels.

Information on internal consistency was presented at 3-month intervals with correlations for the GCQ being .93 for children from 36-41 and 42-47 months and .94 for children at 24-29 and 30-35 months. Correlations for the NCQ were .88 for children who were 36-41 months old, .89 for those who were 42-47 months old, .90 for children who are at 30-35 months and .93 for those at 24-29 months. Information on test-retest reliability was presented at 1- year intervals. The correlations were .96 and .98 on the GCQ and NCQ respectively for 2-year-olds and .94 and .92 on the GCQ and NCQ respectively for 3-year-olds. For inter-scorer reliability, correlations were also presented by 1-year intervals. Seventy nine children were assessed and the resulting correlations were .99 for 24 to 35-month-olds and .99 for 36 to 47-month-olds.

Content validity was addressed by selecting items which are important for the intellectual development of young children. Timed items were excluded because the speed with which a child can perform a skill does not seem to provide important information for a young child and timed items limit the degree to which the examiner can observe the child during assessment. To help with planning instruction and to make test administration easier, items were organized into three sections. Items selected were shown to discriminate well using item analysis. When the preschool form of the CAS-2 GCQ and NCQ were compared with the Bayley Scales of Infant Development-Second Edition (Bayley, 1993), the concurrent validity correlations were .82 and .86 respectively. Correlations with Pictorial Test of Intelligence-Second Edition (French, 2001) were .67 for the GCQ and .80 for the NCQ. Comparisons with the Performance subtests of the Wechsler Preschool and Primary Scale of Intelligence-Revised (Wechsler, 1989) produced results of .77 and .87 respectively. The GCQ correlations with the Detroit Test of Learning Ability-Primary: Third Edition (Hammill & Bryant, 2005) and the Test of Early Language Development-Third Edition (Hresko & Hammill, 1999) are .86 and .77 for children from 3-0 to 3-10.

In terms of construct validity, the scores of participants were shown to increase with age. The mean scores of European American, African American, and scores of children of both genders were all average. The mean scores for children with physical disabilities were in the average range, and those for children with cognitive disabilities were lower than average. For 3-year-olds, achievement scores on the CAS-2 correlate highly with the Test of Early Reading Ability-Second Edition (Reid, Hresko, & Hammill, 1989) and the Test of Early Mathematics Ability-Second Edition (Ginsburg & Baroody, 1990). In 2008, Swanson, Bradley-Johnson, Johnson and O'Dell also evaluated the concurrent validity of the CAS-2 for 2-year-olds with the Bayley Scales of Infant Development-Second Edition (Bayley, 1993) and the construct validity of the CAS-2 for 2-year- olds with the Adaptive Behavior Assessment System-Second Edition (Harrison & Oakland, 2003). The correlations were .63 and .65 respectively. For 3-year-olds, they evaluated the concurrent validity of the CAS-2 with the Detroit Test of Learning Ability-Primary: Third Edition (Hammill & Bryant, 2005) and the construct validity of the CAS-2 with the Test of Early Language Development-Third Edition (Hresko & Hammill, 1999). The correlations were .86 and .77 respectively. They also found that the CAS-2 could predict the performance of 2- and 3-year-olds who were tested 6 years later on the Wechsler Intelligence Scale for Children-Third Edition (Wechsler, 1991); the correlation was .72.

The floors of the CAS-2 are adequate from 3 months and above and the test does not have problems with item gradients.

Overall the CAS-2 is considered a good measure of cognitive development for 2- and 3-year-olds, especially those who will not or cannot vocalize or produce intelligible speech because it has both the GCQ and NCQ. It has a nationally representative sample and good internal consistency and inter-scorer reliability for each age level. Reliability coefficients for stability are .92 or higher for ages 2 and 3. In terms of validity, the CAS-2 correlates well with other measures such as the Bayley Scales of Infant Development-Second Edition and Pictorial Test of Intelligence-Second Edition. It also has demonstrated good predictive validity. The test does not have problems with floors or item gradients. The test also does not have timed items, enabling examiners to be more attentive to the child.

Leiter International Performance Scale-Revised (Leiter-R)

The Leiter-R (Roid & Miller, 1997) Visualization and Reasoning (VR) Battery is a nonverbal cognitive measure for ages 2-0 through 20-11. The VR Battery can be used for individuals who are not fluent in English, those who have problems hearing and moving and also for individuals with problems communicating.

The VR Battery produces a Full Scale IQ and composite scores for Fundamental Visualization, Fluid Reasoning, and Spatial Visualization. The scores can also be expressed as standard scores, age or grade equivalents, percentiles, and normal curve equivalents. Standard scores in the form of composites have a mean of 100 and a standard deviation of 15. Standard scores t in the form of subtest standard scores have a mean of 10 and standard deviation of 3.

Standardization included a stratified sample of 1,719 students. However, information on the dates on which the data were collected was not given in the manual. The sample was similar to the 1993 U.S. Census data in terms of gender, race, ethnicity, parental education, and geographic distribution. For ages 2 through 11, there were at least 100 participants for each 1 year age level. Children with disabilities were not added to the sample although the quotients go down to 30. This implies that the low scores are not based on empirical information.

In terms of reliability, the correlation for internal consistency for 2 to 5-year-olds was .92 for the Full Scale IQ. For the subtests, the range was .71-.91 for 2-year-olds and .73-.92 for 3-year- olds. For test-retest reliability, data were presented for three age groups rather than age levels. For 2- to 5-year-olds the correlation was .90 for the Full Scale IQ.

Content validity was addressed through the use of item-response-theory and factor analysis. Items on this test were selected based on opinions from experts, relevant literature and information on internal consistency. The criterion-related validity was evaluated by demonstrating the degree to which the Leiter-R could correctly group children who had previously been classified as mentally retarded or gifted. A correlation of at least .75 is considered acceptable for positive prediction (Carran & Scott, 1992; Gredler, 2000). The Leiter-R appears to be useful in differentiating between children who had previously been classified as mentally retarded and those who were not because the Full IQ had a correlation of .84 for sensitivity and .97 for specificity. To evaluate construct validity, the VR battery results were compared with the Wechsler Individual Achievement Test (Psychological Corporation, 1992), Woodcock-Johnson Psychoeducational Battery-Revised (Woodcock & Johnson, 1989) and Wide Range Achievement Test-Third Edition (Wilkinson, 1993). The correlations ranged from .63-.83.

To address concurrent validity, the Leiter-R was compared with the Leiter (1948). The resulting correlation was .85 for the Full IQ. When compared with the Wechsler Intelligence Scale for Children-Third Edition (Wechsler, 1991), the correlation was .86. When compared with selected subtests from the Stanford-Binet: Fourth Edition (Thorndike, Hagan, & Sattler, 1986), the Wide Range Assessment of Memory and Learning Test (Adams & Sheslow, 1990), and the Test of Memory and Learning (Reynolds & Bigler, 1994) the correlations ranged from .38-.86.

The Leiter-R was administered to 11 special groups of students, including those with severe speech/language impairment, severe hearing impairment, severe motor delay, traumatic brain injury, significant cognitive delay, ADHD and ADD, learning disability, English as a second language and gifted children and these children performed on the test as expected, e.g., children with cognitive impairments had low scores and gifted children had high scores. In terms of predictive validity, no information was provided.

For the VR battery, adequate floors start from at 2-6 and there are no item gradient problems.

In conclusion, the Leiter-R is a non-vocal option for children who cannot or will not speak. Its internal consistency and the test-retest for the full scale IQ are high and its validity has been demonstrated in a number of studies. However, it has several problems including having too few participants at some age levels, providing low scores that are not based on empirical data, and having no children with cognitive impairments included in the sample. The dates for data collection for the norm sample and data on predictive validity also were not provided.

Merrill-Palmer-Revised (M-P-R)

The M-P-R (Roid & Sampers, 2004) covers cognitive, gross motor, adaptive, social-emotional and language development. The age range of the test is 1 month through 6 years, 6 months.

The cognitive battery is made up of the cognitive, fine motor and the receptive language scales and the overall score on these scales is described as a developmental index (DI). Results can be expressed as quotients (*M* = 100, *SD* = 15), percentiles, age equivalents or Rasch Growth scores. The cognitive and the fine motor scales require speech on only a few items; however, a DI cannot be obtained if only these two scales are used.

The test was standardized using a sample of 1,068 children with at least 145 participants for each 1-year age interval. The sample was similar to the 2000 U.S. Census data in terms of gender, parent education, and racial/ethnic background. No information was provided to describe the children's urban/rural residence; however, it was noted that participants from rural areas performed significantly better than those from urban areas on the test. Children who were excluded from the sample were those with physical, mental, or emotional abnormalities, those who were born prematurely or experienced severe trauma at birth, those who had other biological risk factors, those who had difficulty understanding English, and children who did not reside with either or both parents. However, it is unclear whether these special groups were later added to the normative sample or not. A group of children whose primary language was Spanish were also included in the sample. If children with cognitive abnormalities were excluded from the sample, the mean may be an over estimate and the standard deviation may be restricted.

The internal consistency coefficient for the DI for ages 13-24 months is .97 and for ages 25- 48 months, .98. For ages 13-24 months the correlations for the Cognitive, Receptive Language, and Fine Motor scales were .93, .93 and .90 respectively. For children 24-48 months olds, they were .95, .96 and .92 respectively. The correlations for test-retest reliability were based on a sample of 41 children ranging from 3- 70 months of age and a retest interval averaging 3 weeks was used. The correlations were .89 for the DI, .87 for Cognitive, .90 for Receptive Language, and .90 for Fine Motor. Thus, the DI and Cognitive Scales results were somewhat low. Because the data were not presented by age level, it is unclear whether results were stable for ages 2 and 3. No information was provided on inter-scorer agreement.

In terms of content validity, Uzgiris and Hunt's (1975) exploratory play model and Carroll's (1993) model of cognitive abilities served as guides in developing items. Information from experts as well as the research literature on infant development, item-response theory analysis and scaling verification were also used. Criterion-related validity was analyzed by comparing the M-P-R with the Bayley Scales of Infant Development-Second Edition (Bayley, 1993) for children from 1- to 39- months-old. Correlations were .92 for the DI, .76 for the Cognitive Scales, .86 for the Fine Motor Scales and .92 for the Receptive Language Scales. The criterion-related validity was also examined by comparing results with the Stanford-Binet Intelligence Scale-Fifth Edition (Roid, 2003). For children aged 2-3 years, the correlations were .75-.83 for the Cognitive scales. In terms of construct validity, as age increased the raw scores also increased. Factor analysis also supported the combination of the three scales of the cognitive battery.   No information was provided on the long-term predictive validity of the test.

The M-P-R had adequate floors at all ages; however, problems exist with item gradients for low raw scores.

Strengths of the M-P-R include the fact that the cognitive battery is almost completely non-vocal; the test has acceptable internal consistency as well as adequate floors. However, the test has problems with item gradients for low scores and test-retest reliability data were not presented by age level making it difficult to evaluate the test's reliability for particular age levels. The lack of data on inter-scorer reliability is problematic because the tests' directions are complex. In terms of standardization, the test may not have included children with cognitive abnormalities, which according to Salvia and Ysseldyke (2007), introduces systematic bias to test norms. Although the standard scores extend as low as 10 for certain ages these low scores appear to be based on extrapolation. In addition, data on predictive validity have yet to be provided.

Purpose of the Study

The Bayley-III is the most commonly used measure of cognitive development (Anderson et al, 2010). However, there have not been any studies done on concurrent validity with other non-vocal cognitive measures. Because of the minimal data on concurrent validity, another comparison study with a test that has considerable support for its validity, such as the CAS-2 would be beneficial. The CAS-2 was chosen for this study because it correlated highly with the Bayley-II and is the only non-vocal measure that has been shown to have good long-term predictive validity for young children. Both the CAS-2 General Cognitive Quotient (GCQ) that requires speech and the Nonverbal Cognitive Quotient (NCQ) were compared to the Bayley III overall Cognitive Composite.

CHAPTER II

METHOD

Participants

A sample of 22 children ages 2-0 to 3-6 residing in the central Michigan area was used for this study. The age of children in the study was at most 3-6 because the age range for the Bayley-III only goes up to 3-6. The children were recruited from local daycare and preschool programs.

Table 1. *Demographic Characteristics of the Sample* (*N* = 22)

Characteristics	Percentage of the Sample
Gender	
Boys	77.30
Girls	22.70
Ethnicity	
Asian/ Caucasian	4.50
Caucasian	91.00
Hispanic	4.50
Educational Attainment of Parent	
High school	8.33
4-year college	33.33
Post graduate	58.33

Thirty-one out of about 50 forms were returned. Seven of the returned forms were excluded because the children were older than 3-6. The remaining two forms were excluded because they did not want to be assessed.

Instrumentation

The Bayley Scales of Infant and Toddler Development-Third Edition (Bayley, 2006) and the Cognitive Abilities Scale-Second Edition (Bradley-Johnson & Johnson, 2001) were administered by the author to the children according to the directions in the examiner's manuals.

Procedure

Directors of daycare and preschool programs in the central Michigan area were contacted regarding the study. They were given a written consent form describing the study. (See Appendix A.) If they agreed to participate, information explaining the study as well as consent forms were sent to them to be distributed to the parents of 2- and 3-year-olds who attended their programs. (See Appendix B for this form.) Only children whose parents signed and returned the consent forms were included in the study.

The Bayley-III and the CAS-2 were individually administered at the day care centers and preschools by the author. The testing was done at a table within the child's classroom. Testing took place over two weeks on average; the first test was administered on week one and the second, administered on week two. Each assessment session lasted for about 25-30 minutes. Which test was administered first was determined randomly by rolling a die. Even numbers indicated the Bayley-III was given first and odd numbers indicated that the CAS-2 was given first. Randomizing the test order, controlled for any order effects on the resulting test scores.

Although all of the children were vocal, the CAS-2 NVCQ as well as the GCQ was obtained for all participants. The NVCQ was used because all Bayley-III items are also non-vocal. The GCQ was used was used because it encompasses the entire CAS-2 test and has more items.

Parents were given a summary report describing skills their child demonstrated on the CAS-2 as well as skills that would be appropriate to practice next with their children. No information regarding a child's score was given to the parents or the daycare/preschool staff.

CHAPTER III

RESULTS

To examine the inter-examiner reliability, the examiner and a colleague conversant with both tests, independently scored 25% (5) of the protocols for each measure. The inter-examiner reliability was determined using item-by-item comparisons. The inter-examiner reliability was found by determining the total number of agreements divided by the number of agreements and disagreements for each test. The average inter-examiner agreement was found to be 98.5% for the Bayley-III and 98.8 % for the CAS-2.

The means (*M*) and standard deviations (*SD*) for the Bayley-III and CAS-2 appear in Table 2. The data show that the sample fell within the average range on both measures.

Table 2. *Means, Standard Deviations and Ranges for Bayley-III and CAS-2 (N=22)*

Test	*M*	*SD*	*Range*
Bayley-III			
Cognitive Composite	93.64	6.21	80-110
CAS-2			
GCQ	102.45	16.71	65-123
NVCQ	116	10.96	95-132

Table 3 presents the standard scores obtained on the tests for each participant.

Table 3. *Individual scores for the Bayley-III Cognitive Composite and CAS-2 GCQ and NVCQ (N=22)*

Participant	Bayley-III Cognitive Composite	CAS-2 GCQ	CAS-2 NVCQ
1	95	97	124
2	95	112	117
3	110	105	125
4	95	123	117
5	95	122	125
6	100	77	128
7	80	102	100
8	95	85	118
9	95	65	95
10	90	83	109
11	90	107	100
12	90	118	119
13	95	101	132
14	85	108	117
15	100	83	95
16	95	116	121
17	85	104	124
18	100	121	125
19	85	82	103
20	90	119	125
21	95	105	113
22	90	107	119

This table shows that the Bayley-III (mean = 93.64) for each participant was generally lower than the CAS-2 GCQ (mean = 102.45) and the NVCQ (mean = 116.00).

Table 4 presents Pearson Product Moment correlations for the Bayley-III Cognitive Composite and the CAS-2 GCQ and NVCQ. Because of the restricted range observed (i.e., standard deviations less than 15 for both the CAS-2 NVCQ and the Bayley-III), Guilford and Fruchter's (1978) formula was used to correct the correlation coefficients.

Table 4. *Corrected and Uncorrected Pearson Product Moment Correlations for the Bayley-III and the CAS-2 Results* (*N=22)*

Test	Bayley-III Cognitive Composite	
	r	r_c
CAS-2 GCQ	.42	.75**
CAS-2 NVCQ	.37	.69**

*p<.05, **p<.01

The resulting correlations suggest that the two tests are closely related and thus, are measuring the same construct. However, the correlations are not so high as to suggest that the tests are interchangeable. The correlation between the Bayley-III Cognitive Composite and the CAS-2 GCQ and were somewhat higher than the correlations between Bayley-III Cognitive Composite and the CAS-2 NVCQ.

Results from a two-tailed paired t-test indicated that the Bayley-III Cognitive composite was significantly lower than the CAS-2 GCQ , t = -2.72 (*df* =21 *), p < .05).* Results from a two-tailed paired t-test indicated that the Bayley-III Cognitive composite also was significantly lower than the CAS-2 NVCQ , t = -10.06 (*df* =21 *), p < .05)* . The Bayley-III (mean = 93.64) was significantly lower than both the CAS-2 GCQ (mean = 102.45) and the NVCQ (mean =

116.00). A mean difference of 8 points was found between the Bayley-III and the CAS-2 GCQ and a mean difference of 22 points was found with the CAS-2 NVCQ and the Bayley-III results.

CHAPTER IV

DISCUSSION

The 22 children in the sample all obtained results within the normal range on at least of the measures. Whether similar results would be obtained for children functioning well below the average range remains to be examined.

The moderate correlations obtained between the Bayley-III results and the CAS-2 results suggest the tests measure the same construct (i.e., cognition), but do so in somewhat different ways. Surprisingly, the degree of correlation between the two tests was less for the CAS-2 NVCQ than for the CAS-2 GCQ despite the fact that, like the Bayley-III, the NVCQ does not require speech. This result may be due to the fact that there were fewer items for the NVCQ (52) than the GCQ (88).

The Bayley-III results were significantly lower than both the CAS-2 GCQ and NVCQ. Why this was the case is unclear, but several factors may have contributed to this result. One factor may be the Flynn effect, the phenomenon whereby higher scores typically are obtained on intelligence tests that are considerably older than other tests (Flynn, 1987). The CAS-2 (1997) norms are 7 years older than those of the Bayley-III (2004) norms. Generally newer tests produce lower scores than older tests (Flynn, 1987). Another possibility is that, in the examiner's opinion, overall the CAS-2 appeared to be more engaging than the Bayley-III which may have resulted in children demonstrating better performance on the CAS-2. For example, the children enjoyed items on the CAS-2 such as naming objects shown in pictures as well as playing with the car and the tea set. Their enjoyment was evident because they seemed to smile more than they did to items on the Bayley-III and were quicker to answer questions. Also, the scoring of some Bayley-III items was ambiguous. For example, it was difficult for the examiner to differentiate between the different types of play required on 3 items which may have caused more conservative scoring of these items, leading to lower scores on the Bayley-III. Also, because of the timed items on the Bayley-III, some children may have lost points because they did not complete items quickly enough to receive credit, despite the fact that they may have had the ability to demonstrate the skills. For items such as competing a peg board or puzzles, some children were able to perform those tasks on the Bayley-III but only after the allocated time had lapsed. Thus, they did not receive full credit for these items. This occurred with about three children. The CAS-2 does not have timed items. Results also could have been affected by a combination of these factors.

Correlations for the Bayley-III with the CAS-2 GCQ (r = .75) as well as the CAS-2 NVCQ (*r* =.69) are lower than the correlations for the Bayley-II and the CAS-2 GCQ (*r* = .82) and the CAS-2 NVCQ (*r* = .86) reported in the examiner's manual, suggesting that the Bayley-II is more similar to the CAS-2 than Bayley-III is to the CAS-2.

Conclusion

The Bayley-III measures the same construct of intelligence as the CAS-2 although the two tests differ significantly in their results. The Bayley-III Cognitive composite produces a significantly lower score than either the CAS-2 GCQ or NVQ.

Limitation

The sample size of 22 is small so results may not generalize to the broader population. The sample was made up Asians/ Caucasian, Caucasians only and Hispanics only; therefore, this was not a nationally representative sample in terms of race/ethnicity or geographic region. This fact limits generalization of the results.

Future Studies

Future studies can focus on determining if similar results will be obtained for individuals functioning well below the normal range as well as how these two tests compare for children of different racial/ethnic backgrounds.

APPENDICES

APPENDIX A

LETTER TO THE DIRECTORS OF PRESCHOOLS/DAYCARES

Dear Daycare Provider,

I am a doctoral student working under the supervision of Dr. Sharon Bradley –Johnson, a faculty member in the Psychology Department at the Central Michigan University. For my thesis project, I would like to carry out a study to examine how the results of a recently revised cognitive test compare to another cognitive test that is well established. The aim of the study is to determine whether or not the two tests provide similar results for children from 24 through 42 months of age.

Because children in your program fall in this age range, I would appreciate your help with this research.

This project has been approved by the Institutional Review Board (IRB) at the Central Michigan University. I have prepared permission forms to be sent to parents describing participation in the study. Results obtained from testing will not be available to your program, but parents will receive a written summary of skills their child excels in, as well as skills that would be helpful to work on next. Testing will require 2 separate sessions which should last no longer than 30 minutes each. The tests will be administered individually to each child in the child's classroom. The tests include tasks that children typically enjoy, and involve toys and materials that are colorful and interesting. Times for testing will be arranged with your staff so testing does not interfere with the children's schedules.

If you are not satisfied with the manner in which this study is being conducted, you may report (anonymously if you so choose) any complaints to the Institutional Review Board by calling 989-774-677, or addressing a letter to the Institutional Review Board, 251 Foust Hall Central Michigan University, Mt, Pleasant, MI 48859.

If you have any question regarding this study, please feel free to contract me via phone: (313)728-1002, or email: mork1sp@cmich.edu.

Or you may contact my supervisor: Dr. Sharon Bradley Johnson Psychology Department, Sloan 232, Mt. Pleasant, Michigan 48858 (989)854-2740

____

Initial

If you are willing to participate, please sign below. Thank you for your time and consideration in helping with this project.

Sincerely,

Seraphim Mork

Doctoral Student, School Psychology

________________________________ __________________

Daycare Center Director Signature Date

____

Initial

APPENDIX B

PARENT CONSENT FORM

Dear Parent/Caregiver,

My name is Seraphim Mork and I am currently a second-year school psychology doctoral student at Central Michigan University. I am conducting this study to fulfill the requirements for my doctoral degree.

Your child is invited to participate in this study because he/she is between 24 to 42 months of age. The following information is provided to help you make an informed decision about your child's participation. Your child's participation or non-participation will not affect his/her daycare/preschool program. If you have any questions, please call Seraphim (313) 728-1002 or Dr. Bradley-Johnson (989) 854-2740.

Title of project: The Bayley-III: A concurrent validity study for children 24- through 42-month-olds.

Name of investigator: Seraphim Mork **Phone :**(313) 728-1002

Supervisor: Sharon Bradley-Johnson, Ed.D . **Phone**: (989) 854-2740

Professor of Psychology

My study is aimed at finding out if a recently revised test, the *Bayley Scales of Infant and Toddler Development-Third Edition (Bayley-III)* provides similar results to the *Cognitive Abilities Scale-Second Edition (CAS-2)*; a well-established test of cognitive development for young children.

If you decide to have your child participate in this study, I will give the two tests to your child in 2 testing sessions. These sessions will take place at your child's daycare/preschool. Each session should take about 30 minutes. During testing your child will engage in several tasks such as exploring new toys, imitation of actions and simple problem solving tasks for preschoolers.

____

Initial

If you choose to allow your child to participate, a summary report of the test results will be provided to you describing your child's strengths and skills to work on next. No test scores will be available to you or to the preschool/daycare staff. If you want to share the report with the preschool/daycare staff, that will be up to you.

You will be asked to fill out the attached information sheet including basic background information regarding you and your child such as gender, race, ethnicity, and your education level. This information will not be used to select children for the study. It will only be used to describe the group of children who will participate.

Participation is voluntary. No negative effects are anticipated for the children in the study, and most children enjoy the chance to explore new toys and other testing materials. You have the right to withdraw your child from the study at any time without penalty. If you agree to allow your child to participate, but your child expresses any discomfort or is unwilling to participate during the session, the session will be discontinued immediately.

To ensure confidentiality no names will appear on test records. Instead a number will be used. The list of children's names and their assigned number will be kept separate from record forms. Though this study may be published in a journal article, no names of children will appear in the paper.

If you agree to let your child participate please complete and return this form to me, and if you have any questions please contact me via phone: (313) 728-1002 or email: mork1sp@cmich.edu. You will be given a signed and dated copy of this form to keep. Thank you very much for your time. Your help is appreciated.

If you are not satisfied with the manner in which this study is being conducted, you may report (anonymously if you choose) to the Institutional Review Board by calling 989-774-6777, or a letter to the Institutional Review Board, 251 Foust Hall Central Michigan University, Mt, Pleasant, MI 48859.

____

Initial

Please indicate your highest educational level: _______

Please indicate your child's gender:

______Male ______Female

Please indicate your child's primary racial/ethnic background:

______African American ______Caucasian _____Hispanic

______Asian ______Native American ______Other

Consent to participate:

I understand the purpose of the study and issues around consent. I have read the above information and agree to participate in this study and allow my child to participate. I have received a copy of this consent for my own records

Name of Parent(s) or guardian(s): Date:

__________________________ ______________________

Parent or guardian's signature:

____

Child's Name:

____

Examiner's name: Examiner's Qualification:

__________________________ _____________________

In my judgment, the participant's parent(s) or guardian(s) has/have voluntarily and knowingly given informed consent for their child to participate in this research.

Name of researcher: Date:

________________ ____________________

Signature of researcher: Date:

________________ ____________________

____

Initial

APPENDIX C

TEST REPORT

Dear Parents:

Thank you very much for allowing B to assist me with my research project for my thesis. The following is a summary of your child's performance on the Cognitive Abilities Scale-Second Edition. Results include examples of skills B demonstrated and examples of skills that would be appropriate to work on next. If you have questions about the results, please call ___________; I will be happy to discuss them with you. I hope you find the results helpful.

Language. Examples of skills B demonstrated in this area include understanding the position words *in, , up, down, top, next to, away from, in front of* and *around*, labeling pictures of objects such as a *glasses, wagon* and *hammer*, understanding and using several pronouns, using noun-verb combinations, regular plurals (e.g., cups), and *-ing* verbs such as *running* in her speech

Language skills she is likely to learn next include more frequent use of pronouns in her speech (e.g., she and him), and understanding the position words *together*, and describing events in order (e.g., answering the question "What do you do when you wash your hands?") and use of possessive with 's(e.g., daddy's).

Early Reading. She turned books right side up to look at them, turned book pages one-at-a-time, and remembered an idea from a story read to her.

Skills to develop next are pointed to pictures named by the examiner, learning letter names and letter sounds. Learning letter names is useful, but it is the letter sounds that are helpful for reading.

Early Math. B understood the math concepts *all, big, little, empty, full, tall, short, same,* and *different,* correctly held up her fingers to show her age, counted to 5 in imitation, and matched pictures of quantities of 1 and 2 objects.

A skill to learn next include understanding more math concepts such as *more* and *few,* counting different groups of objects and answering the question, “How many?” To do this the child must recognize that the last number stated when counting answers the question.

Early Handwriting. She held the pencil with her fingers rather than her fist and correctly copied a circle, a vertical line, a horizontal line, and a plus sign from a picture. Thus, she has good control of the pencil.

Enabling Behaviors. This section assesses physical and verbal imitation and memory, skills important for school success. B readily imitated both words and gestures and exhibited a very good memory for information she hears. Imitation helps children learn new skills and practice others.

B’s skills important for later school success are developing well and it was a joy to work with her.

Cordially,

REFERENCES

Adams, W., & Sheslow, D. (1990). *Wide Range Assessment of Memory and Learning* Wilmington, DE: Wide Range Inc.

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

Anderson, P. J., De Luca, C. R., Hutchinson, E., Roberts, G., Doyle, L. W., & the Victorian Infant Collaborative Group. (2010). Underestimation of developmental delay by the new Bayley-III scale. *Archives of Adolescent Medicine, 164*, 352-356.

Bayley, N. (1993). *Bayley Scales of Infant Development-Second Edition* San Antonio, TX: The Psychological Corp.

Bayley, N. (2006). *Bayley Scales of Infant Development-Third Edition.* San Antonio, TX: The Psychological Corp.

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

Bracken, B. A., & Walker, K. C. (1997). The utility of intelligence tests for preschool children.In D. P. Flanagan, J.L. Genshaft, & P. C. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 484-502). New York: Guilford.

Bradley-Johnson, S., & Johnson, C. M. (2001). *Cognitive Ability Scale- Second Edition.* Austin, TX: PRO-ED.

Bradley-Johnson, S., & Johnson, C. M. (2007). Infant and toddler cognitive assessment. In B. A. Bracken, & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (pp.29-48). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Carron, D.T., & Scott, K.G. (1992). Risk assessment in preschool children: Research mplications for the early detection of educational handicaps. *Topics in Early Childhood Special Education, 12,* 196-211.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. Cambridge: Cambridge.

Campbell, F. A., & Ramey, C. T. (1995). Cognitive and school outcomes for high-risk African-American students at middle adolescence: Positive effects of early intervention. *American Educational Research Journal, 32,* 743-772.

Campbell, F. A., Pungello, E. P., Miller-Johnson, S., Burchinal, M., & Ramey, C. T. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood educational experiment. *Developmental Psychology, 37*(2), 231–242.

Folio, M.R., & Fewell, R.R. (2000). *Peabody Developmental Motor Scales-Second Edition.* Austin, TX: PRO-ED.

French, J. (2001). *Pictorial Test of Intelligence* (2nded.). Austin, TX: PRO-ED.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical psychology*, *44*(6), 1013-1023.

Ginsburg, H. P., & Baroody, A.J. (1990). *Test of Early Mathematical Ability-Second Edition.* Austin, TX: PRO-ED.

Gredler, G. R. (2000). Early childhood screening for developmental and educational problems. In B. A. Bracken (Ed.), *Psychoeducational assessment of preschool children* (pp. 399-411). Boston: Allyn & Bacon.

Guilford, J. P., & Fruchter, B. (1978). Fundamental Statistics in Psychology and Education. New York: McGraw-Hill.

Hammill, D. & Bryant, B. (2005). *Detroit Test of Learning Aptitude-Primary: Third Edition*. Austin, TX: PRO-ED.

Harrison, P. L., & Oakland, T. (2003). *Adaptive Behavior Assessment System-Second Edition*. San Antonio, TX: Harcourt Assessment, Inc.

Hegde, M., & Pomaville, F. (2006). *Assessment of communication disorders in children.* San Diego, CA: Plural Publishing.

Hresko, W. P., & Hammill, D. (1999). *Test of Early Language Development-Third Edition*. Austin, TX: PRO-ED.

Leiter, R. G. (1948). *Leiter International Performance Scale.* Wood Dale, IL: Stoelting.

Nagle, R. J. (2007). Issues in preschool assessment. In B. A. Bracken, & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (pp.39-48). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Parke, R. D., & Gauvain, M. (2009). *Child psychology: A contemporary viewpoint* (7th ed.). New York, NY: McGraw-Hill.

Reid, D. K., Hresko, W. P., & Hammill, D. (1989). *Test of Early Reading Ability-Second Edition.* Austin, TX: PRO-ED.

Reynolds, C.R., & Bigler, E.D. (1994). *Test of Memory and Learning.* Austin, TX: PRO-ED.

Roid, G. H. (2003). *Stanford –Binet Intelligence Scales* (5thed.). Itasca. IL: Riverside Publishing Co.

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

Roid, G. H., & Sampers, J. L. (2004). *Merrill-Palmer-Revised: Scales of Development*. Wood Dale, IL: Stoeling.

Salvia, J., & Ysseldyke, J. E. (2007). *Assessment: In special and inclusive education* (10th ed.). Boston, MA: Houghton Mifflin.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations*. (5th ed.). San Diego, CA: Author.

Swanson, J. R., Bradley-Johnson, S., Johnson, C. M., & O'Dell, A. R.(2008). Studies of concurrent criterion-related, construct, and predictive criterion-related validity. *Journal of Psychoeducational Assessment*, *27*(1), 46-56.

The Psychological Corporation. (1992). *Wechsler Individual Achievement Test*. San Antonio, TX: Author. Relocate this reference under P for Psychological.

Thorndike, E.L., Hagen, E.P., & Sattler, J.M. (1986). *Stanford-Binet Intelligence Scale-Fourth Edition.* Chicago, IL: Riverside.

Uzgiris, I., & Hunt, J. McV. (1975). *Assessment in infancy: Ordinal scales of psychological development*. Urbana, IL: University of Illinois Press.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence-Revised.* San Antonio, TX: The Psychological Corp.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children-Third Edition.*San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2002). *Wechsler Preschool & Primary Scale of Intelligence-Third Edition.* San Antonio, TX: The Psychological Corp.

Wilkinson, G. S. (1993). *The Wide Range Achievement Test: Manual* (3rd ed.). Wilmington, DE: Wide Range.

Woodcock, R.W., & Johnson, M.B. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised*. Chicago, IL: Riverside.

Zimmerman, I.L., Steiner, V.G., & Pond, R.E. (2002). *Preschool Language Scale-Fourth Edition.* San Antonio, TX: The Psychological Corp.