CHAPTER ONE

Introduction

Background

With increasing pressure from the accountability movement (Hatch & Grieshaber, 2002), the academic standards for American students have been set higher than ever before. The No Child Left Behind legislation signed by President Bush requires all students to score at or above the proficiency level established by their states. Schools or programs that fail to meet the criteria will take negative consequences in various ways. The call for more testing and higher standards has pushed the accountability pressure down to the early childhood field (Harbin, Rous & McLean, 2004; Hatch & Grieshaber, 2002; Hatch, 2002). In the Bush Administration's "Good Start Grow Smart" initiative, the accountability issue was also brought up. The initiative pointed out that some children are not receiving high quality care because early childhood programs are seldom evaluated based on how they prepare children to succeed in school. Federal and state governments also have increased their investments in early childhood programs for preschool children in order to improve the program quality and better serve all children. Child outcome is one indicator for program quality so that it becomes one part of the accountability requirement. In order to align children's early experience with what will happen in K-12 classrooms, researchers and administrators have looked to the K-12 accountability system as a resource when setting standards for early childhood programs (Harbin et al.). Once the standards are set, how to measure child outcomes to hold programs accountable becomes an important question.

The original forms of measuring young children's outcomes started hundreds of years ago. Early discussion of the necessity of intelligence tests was generated by two influences in society. Both practical need and theoretical interest stimulated people's interest in testing individual's intelligence.

In the social venue, attention had been paid to the defective and delinquent population. Due to the disadvantage of mental make-up, the defective and delinquent population had been treated unfairly. During the 16th century when a sense of social justice to all classes was developed, the physicians began the real study of insanity in order to fight for a better treatment of the insane (Pintner, 1923).

In the theoretical venue, the development of experimental psychology, the study of individual differences, the growth of eugenics, and the development of anthropological measurement all longed for the investigation of human intelligence (Pintner, 1923). The discipline of experimental psychology intended to investigate human intelligence to seek to establish general laws of the normal human minds; the school of eugenics intended to use human intelligence to prove that individual differences were derived by inheritance; and the development of anthropological measurement needed the study of human intelligence to link human ability with physical characteristics (Pintner). However, not until the late 19th century, did interest in studying young children's intelligence emerge in response to the initial recognition of childhood as a separate period in the life cycle (Wortham, 1995). Early publications by famous scholars such as John Locke, Rousseau, and Frederick Froebel reflected concerns for appropriate practice and education of young children (Wortham). Along with these concerns, came the movement of measuring young children's ability.

The first round of intelligence tests for young children was designed around the beginning of the 20th century (Standardized Tests and Our Children: A Guide to Testing Reform, 1990). The development of measurement for young children started with Cattell's mental test.

Cattell’s focus on individual differences led to his development of a mental test. In the article published in *Mind* (1890), he first used the term mental test and described 50 different measures which, for the most part, assessed sensory and motor abilities. He also predicted the practical use of mental tests for training and for diagnostic evaluation. As Cattell was trying to prove psychology to be scientific, he pleaded for standardization of methods and procedures, and urged the necessity for the establishment of norms. Even though most of the tests that Cattell described had little predictive validity for educational achievement or for other aspects of intellectual functioning, he did make a significant contribution to bring the assessment of mental ability out of the field of abstract philosophy and proved that mental ability could be studied experimentally and practically.

Around the same time Cattell described these mental measurements, Wissler (1901) and Sharp (1898) concluded that most sensorimotor tests had poor predictive validity for school achievement and could not effectively distinguish bright from dull children. Therefore, at the turn of the century, Emil (1855/1926) introduced complex tests for measuring mental functioning, and many of his tests were based on abilities needed for daily life. In addition, Kraepelin recognized the need to examine an individual enough times to reduce chance variation. At the same time Alfred Binet (1857/1911), Victor Henri (1872/1940) and Theodore Simon (1873/1961) argued that the key to the measurement of intelligence was to focus on higher mental processes instead of simple sensory functions. Their work culminated in the development of the 1905 *Binet-Simon Scale* (Binet & Simon, 1905), which might be considered the first practical intelligence test. It is in this scale that norm and standardization were first established. Children’s “mental age” was calculated in order for the examiner to use results from this scale to determine educational placement. Also, standardization was used so that examiners could compare their results. This scale served the purpose of objectively diagnosing degrees of mental retardation and became the prototype of subsequent scales for the assessment of children's mental ability.

While Binet and Simon successfully measured intelligence in general terms and abandoned the attempt to analyze intelligence in its component parts, there were many other specialized tests that had been developed to evaluate specific facets of cognitive ability. The definition of cognitive intelligence became heatedly debated.

During a symposium conducted in 1921, thirteen psychologists gave thirteen different views about the nature of intelligence. To name a few, Terman (1916) defined intelligence as the ability to carry on abstract thinking, Binet defined intelligence as a collection of judgment, practical sense, initiative, and ability to adapt oneself to circumstances, and Wechsler recognized that intelligence was affected by the way the abilities are combined and by the individual's drive and motivation. To sum up, intelligence was recognized by a majority of experts as the ability to solve problems, think abstractly, acquire knowledge, and adapt to the environment.

Statement of the Problem

Standardized Tests

Once the concept of intelligence was defined, tests were designed to measure the intelligence of young children. Tests such as the Binet-Simon Intelligence Scale (Binet & Simon, 1905), Woodcock-Johnson Test of Achievement (Woodcock & Johnson, 1977), Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983), and McCarthy Scale of Children's Abilities (McCarthy, 1972) were developed. They were all standardized with a definite method of administering the tests and the norms for the interpretation of results. Standardized, or norm-referenced tests, are commercially published tests that contain a set items and have a uniform procedure for administration and scoring (Anderson, Hiebert, Scott &

Wilkinson, 1985; Popham, 1999). They provide a comparison of individual performances to that of state or national samples (Elliott, Ysseldyke, Thurlow & Erickson, 1998).

The public began to use standardized tests for selection and retention purposes at the beginning of the 1900s (Meadows & Karr-Kidwell, 2001). The use of standardized tests has been increased since 1950s when the failure of science education in the United States was given a large amount of attention. In order to improve the educational quality across the nation, measuring children's progress became necessary. Due to the fact that there was no other assessment measure was available at that time, the common approach of documenting and reporting children's growth relied on the use of standardized, norm-referenced assessments. During the 1980s, the most frequently used measures for kindergarten screening included the Brigance Screening (Brigance, 1985), Batelle Development Inventory (Newborg, Stock, Wnek, Guidubaldi, Svinicki, 1984), Developmental Indicators for Assessment of Learning (Mardell & Goldenberg, 1975)**,** and Gesell Developmental Observation (The Gesell Institute, 1985).

However, with the increased use of standardized tests, researchers indicated no evidence of positive change in child learning even though children's scores on standardized tests might have increased. Except for preparing children score better on the post-test, standardized tests do not deal directly with teacher effectiveness or student motivation (Stiggins, 1999). On the contrary, when using standardized tests as the measure for accountability, it added stress on both teachers and students (Sack, 2000). The pressures to do well on these high-stakes tests can have opposite effect from the one we seek (Stiggins). The pressures can possibly add anxiety levels of teachers and it can also have negative impact on child's self-confidence. Therefore, concerns such as the unintended consequence (Stiggins) of the way child's information were collected and reported were raised (Ratcliff, 1995; Roe & Vukelich, 1994).

Even though concerns for standardized tests grew stronger during the past three decades, the use of standardized tests on young children persists in school districts and states (Meadows & Karr-Kidwell, 2001; Olson, 2001; Smith, 1999). Since the late 1990s, the public's push for using standardized tests to hold schools accountable has become a noticeable phenomenon (Dorn, 1998), and test-based accountability models have received widespread support from the business sector (Kornhaber, 2004).

With the increased concern of threatening young children by conducting too many standardized tests (Hatch, 2002; Hatch & Grieshaber, 2002), scholars and organizations such as National Association for the Education of Young Children [NAEYC], National Association of Early Childhood Specialist in State Departments of Education [NAECS/SDE], and National Association of School Psychologists [NASP], have called for a stop to norm-referenced and standardized tests (Shepard, Tylor & Kagan, 1996; NAEYC, 2003; NAECS/SDE, 2003; NASP, 2002) and suggested alternatives to evaluate and observe young children as they are learning.

Authentic Assessment

Authentic assessment, sometimes referred as naturalistic assessment (Barnett & Macmann, 1992), play-based assessment (Bufkin & Bryde, 1996), contextualized assessment (Bell & Barnett, 1999), or performance assessment (Moorcraft, Desmarais, Hogan, & Berkowitz, 2000; Klein & Estes, 2004), is a type of assessment that assesses children in their natural environment. It is defined as the process of gathering data by systematic observation for making decisions about an individual (Berk, 1986), and "active student production of evidence of learning" (Mitchell, 1995, p.2). It is so named because these assessments are meant to reflect practices and performances that actually occur within a broader environment rather than those found in the context of a specific test (Black & William, 1998; Wiggins, 1998). Authentic assessment includes methods such as work samplings, anecdotal notes, portfolios, checklists, rating scales, and teacher-designed classroom observations (Janesick, 2001). Authentic assessment provides information for program planning, is linked to curriculum, and is flexible in terms of the way data are collected. It can serve as an alternative method for reporting young children's outcome data.

In order to provide young children and their families the best services, single forms of assessment are insufficient (Greenspan, Meisels, et al.., 1996). Both standardized and authentic assessments should be used (Wortham, 2008). For example, standardized, norm-referenced assessment should be used to identify children for special needs because it allows the individual child's performance to be compared to a representative sample of children who are progressing normally, while a curriculum based, authentic assessment is appropriate for providing information about children's learning because it measures individual children's strengths and weaknesses in some determined objectives and it indicates ability with respect to certain skills (Bailey & Nabors, 1996). With the emerging of authentic assessment the assessment paradigm has gone through a change that moves from the traditional psychometric paradigm to the postmodern contextual paradigm (Berlak, 1992). Authentic assessment, representing the new paradigm, values plural meanings, value-laden assessment techniques, inseparability of cognitive from affective learning, and local control. The use of authentic assessment has been recommended by many researchers (Appl, 2000; Atkins-Burnett, Xue, Nicholson, Bickel & Hee son, 2003; Hatch, 2002; Kornhaber, 2004; Meadows & Karr-Kidwell, 2001; Miesels, 1996; Ratcliff, 2001).

However, even though authentic assessment is advocated for its relevance to young children's learning processes and their natural environment, the technical soundness of this type of assessment is questioned. Authentic assessments are challenged with the concern that it cannot provide data for comparison and eligibility purposes (Herman, 1992; Kornhaber, 2004; Sanders & Horn, 1995; Pierson & Beck, 1993; Wiggins, 1992). It is also questionable for accountability and high-stake purposes. Authentic assessment poses challenges for teachers to collect data reliably (Bergen, 1993).

Research Objectives

In order for the public to feel confident using data from authentic assessment for program accountability provide instructions and make decisions about young children, the challenges mentioned above should be appropriately addressed. If authentic assessment demonstrates its reliability and validity evidences, it will become a dependable alternative for assessing young children's developmental outcomes. This study addressed one of the concerns about authentic assessment regarding the technical adequacy issues of the authentic assessment. This study is a validation study. For a test to be acceptable, validation of the test is a necessary step. Validation is a process of developing scientifically sound test scores and their relevance to the proposed use (American Educational Research Association [AERA], 1999). During the validation process, evidence is collected to provide a sound scientific basis for the proposed score interpretations (AERA).

As an example of authentic assessment, the Assessment, Evaluation and Programming System, 2nd Edition (AEPS®) is a curriculum-based assessment that assesses children in their natural environment. The data of the AEPS® is collected through ongoing observation of children in the classrooms, and it provides information for developing goals and objectives for individual child and program planning. In this study, the AEPS® was validated by comparing its scores with the criterion measure, and teachers' perceptions of the authentic assessment were explored to provide evidence of the social validity of authentic assessment.

Justifications of the domains chosen

With the amount of attention given to the reading and math results of the American students (Anderson, Hieber, Scott & Wilkson, 1985; Langer, 1984; National Research Counsil, 1998; Purves, 1984), it is not hard to imagine how much attention has been paid to children's reading, writing and math abilities. The National Education Goals announced in 1989 indicated that American students would demonstrate competency in English, mathematics and science. By the year of 2000 U.S. students would be first in the world in science and mathematics. Also, under the No Child Left Behind Act (Elementary and Secondary Education Act, 2001) states were required to participate in the reading and mathematics assessment in order to hold schools accountable for making sufficient progress. In responding to these aggressive goals and accountability movements, the Good Start Grow Smart legislation specifically asked early childhood programs to voluntarily develop guidelines on pre-reading, language and writing skills for young children that align with State K-12 standards (Harbin et al.., 2004).

Besides the political needs for emphasizing young children's literacy and pre-math development, scholars also indicated that children begin to learn to read and write during their early years. Piaget (1954) and the generation of researchers that followed also shifted the focus of research to the development of mathematical knowledge prior to formal schooling in mathematics. The importance of early literacy and pre-math was also indicated in numerous studies (Greenwood, Hart & Carta, 1994; Manset-Williamson, John, Hu & Gordon, 2002; Tudge & Doucet, 2004; Walker, Whitehurst & Lonigan, 2001). Therefore, young children's skills in early literacy and pre-math are of upmost interest to teachers, researchers and policy makers. Early childhood programs are required to provide sufficient instruction on these skills.

Because of the importance of early literacy and pre-math stated above, and because of the public's special interests in these two areas, children's performance on early literacy and pre-math skills were chosen in this study. If evidence can be collected to support the use of literacy and pre-math results generated from authentic assessment, the credibility of using authentic assessment to collect data for accountability and reporting purposes is strengthened.

Definition of Terms

Validity

A variety of sources of evidence may be used in evaluating validity of a measure. These sources represent different aspects of validity. These sources of evidence includes: 1) evidence based on test content, referred to content validity, 2) evidence based on response processes, 3) evidence based on internal structure, referred to construct validity, 4) evidence based on relationship to other variables, referred to criterion-related validity, and 5) evidence based on consequences of testing, referred as social validity (AERA, 1999).

Criterion-related validity. In this study, criterion-related validity was explored. The source of validity for criterion-related validity comes from the relationship between the measure to be validated and other variables. These variables include measures of the criteria the test claims to measure, as well as other tests that measure the same constructs as the test claims to measure. Evidence of this type of validity usually comes from comparison with other measures and usually involves correlational evidence. Two designs have been historically distinguished for evaluating such kind of validity. They are called predictive validity and concurrent validity. A predictive validity provides evidence for the accuracy of how the test data can predict criterion score that are obtained at a later time. Concurrent validity, which avoids time delay, is useful to investigate alternative measures of some specific constructs (AERA,1999). In this study, the concurrent validity was examined.

Social validity. Social validity examines feedbacks from consumers to guide program planning and evaluation (Wolf, 1978). This type of validity focuses on the intended or unintended consequences of test use as well as the acceptability of the test. Investigation into the sources of consequences provides evidence for social validity. For example, if differences in scores from an early childhood measure are due to the unequal distribution of skills the test claims to measure then the finding of the differences per se do not imply any lack of validity. However, if the differences in test scores are due to the test's sensitivity to some children's characteristics not intended to be part of the test construct, the social validity of the measure needs to be questioned. Recent development of philosophy of assessment has highlighted the importance of investigating the consequences of assessment use (Moss, 1992), therefore, the examination of the social validity of the alternative measure for accountability purpose was included in this study.

Early Literacy and Pre-math

According to Whitehurst and Lonigan (1998), emergent and conventional literacy includes children's understanding of vocabulary, narrative, conceptual knowledge, story schema, phonemic awareness, and letter recognition. According to researchers working in the area of mathematic development (Baroody, 1992; Clements, Swaminathan, Hannibal, & Sarama, 1999; Newcombe & Huttenlocher, 2000; Starkey & Cooper, 1995), pre-math development includes children's skills such as counting, arithmetic problem solving, spatial reasoning and geometric knowledge. In both Assessment, Evaluation, and Programming System, 2nd Edition [AEPS®] and

Battelle Developmental Inventory, 2nd Edition [BDI-2] test, items related to children's early literacy and pre-math competency are covered under cognitive and communication domain, therefore, these two domains are chosen for this validation study.

Research Questions and Hypotheses

Research Questions

Three research questions were studied in this study: 1) Is the AEPS® a concurrently valid measure for assessing young children's competence in cognitive domain for accountability purpose? 2) Is the AEPS® a concurrently valid measure for assessing young children's competence in communication domain for accountability purpose, 3) what are teachers' perceptions on using authentic assessment and standardized tests? Both quantitative and qualitative approaches were used to answer the research questions.

Hypotheses

In response to the three research questions, three hypotheses were made: 1) the AEPS® is concurrently valid when used for assessing young children's competence in cognitive domain for accountability purpose, 2) the AEPS® is concurrently valid when used for assessing young children's competence in communication domain for accountability purpose, and 3) teachers perceive authentic assessment as a useful tool to measure young children's progress and plan for curriculum.

CHAPTER TWO

Literature Review

Assessment

The definition of assessment formally appeared in literature around the early 1980s when Goodwin and Goodwin (1982) described it "process of determining, through observation or testing, an individual's traits" (p.523). In the following years, numerous other scholars have given assessment definitions. Collectively, it is defined as a comprehensive process of obtaining information about children in order to make evaluations and decisions (McAfee, Leong, & Bodrova, 2004; McLean, 1996; Meisels & Fenichel, 1996; Meisels, 1994; Meisels, 2001; Nicholas & Sandra, 1995; Salvia & Ysseldyke, 1995). However, the methods or approaches of obtaining information have been debated for decades (Notari-Syverson & Losardo, 2004; Wortham, 2008). The debate revolves around the use of standardized test and authentic assessment (Elliott, Ysseldyke, Thurlow & Erickson, 1998; Kornhaber, 2004; Notari-Syverson & Losardo, 2004). There were advantages and disadvantages for each method. Based on different use of the information, each approach has its own application.

Assessment Purposes

In the field of early childhood education, assessment is needed for four reasons: 1) assessments is used to identify children with special problems and determine the need for additional service, 2) assessment is used to determine the level and rate of young children's learning, 3) assessment is used to monitor children's progress, and 4) assessment is also used to evaluate the program quality (Appl, 2000; Grisham-Brown, Hemmeter, & Pretti-Frontczak, 2005; NAEYC, 2003; Nagle, 2000; Shepard, Lynn, Kagan & Wurtz, 1998; Wortham, 2008).

Purpose 1: Identify children who may need special services and instructions. In 1986, Individual with Disabilities Education Act [IDEA] required states to systematically locate, identify, and evaluate children in need of special education services (PL 94-142). Researchers (Wortham, 1995; Culberston & Willis, 1993) also suggested that since the developmental change in young children is rapid, there is a need to assess children regularly to determine whether they are progressing normally. If the results from assessment suggest that the child is not developing at a normal pace, additional testing may be needed and appropriate intervention services should be provided during the critical years of childhood. Therefore, there is a need for careful screening and extensive testing before the selection of a combination of intervention programs and other services.

Purpose 2: Support learning and instruction. In a democratic society where one of the goals of public education is to assist every individual to develop to his or her full potential, assessing a child's strengths and potential has assumed a great deal of importance (Hynd & Semrud-Clikeman, 1993). Assessment gives teachers and parents information on what a child can do and what a child is ready to learn next. Only after children's strengths and needs are identified, can teachers and related personnel plan for more effective programs. For young children with special needs, the process of identifying their individual goals and intervention strategies also require decisions about children's strengths and their needs (Appl, 2000). Therefore, assessment is needed for planning purposes.

Purpose 3: Monitoring progress. Once the intervention and instructional strategies are in place, the child's progress toward his or her goals and objectives need to be monitored. Monitoring children's progress assists teachers and other professionals in determining whether the intervention and instructional strategies are working for the child (McLean, 1996). Once the data was collected for this purpose, teachers and related personnel can determine if the child is being appropriately challenged, and make predictions about future plans (Appl, 2000).

Purpose 4: Program evaluation and high-stakes accountability. Whether a program is providing best services to its children partially depends on its children's outcomes, either compared to children in other programs or same children at previous time periods. Assessment from an outside agency provides un-biased data for the program to demonstrate its accountability (Shepard, Kagan & Wurtz., 2001). By reporting children's outcomes, the outside authority links teachers and schools to high stakes decisions, in some cases, losing funding or facing school take-over (Elementary and Secondary Education Act, 2002). In this case, assessment serves a critical criterion for program evaluation and program accountability.

Assessment Methods

Based on different purposes of the assessment, different methods should be used to meet the needs for the purpose it serves (Shepard, et al.., 2001; Appl, 2000). Assessment methods include observation, interview, documentation of children's work, checklists, rating scales, and portfolios, as well as direct testing (NAEYC, 2003). For identification purposes, norm-referenced, standardized tests serve this purpose better because norm-referenced tests provide norms for comparison. However, norm-referenced tools are questioned for their fairness (Biggar, 2005; Wesson, 2001; Wortham, 2008). Bagnato and Neisworth (1995) also documented the extent to which standardized tests failed to serve the purpose of determining eligibility. Therefore, researchers are now proposing that norm-referenced measure should not be the only source used to identify children, instead, multiple assessments and assessors are needed for the identification and referral purposes (Blair, 2003; DEC, 2001; Goldsmith, Davidson, & Rickman, 2002; NAEYC, 2003; Shepard et al..; Wakschlag, Leventhal, Briggs-Gowan, Danis, & Keenan, 2005; Wortham, 2008).

When assessments are used as part of the teaching-learning process, the content of assessments should be closely related to what children are learning and what teachers are teaching. Standardized test has been criticized for its lack of educational relevance (Grisham-Brown et al.., 2004), therefore, assessment that is conducted by observing children during an instructional activity would serve this purpose better (Shepard, et al.. 2001; Shepard, et al.., 1998).

For the purpose of monitoring progress, a measure that can be implemented across time and reflect progress is preferable. Standardized test which is composed of non-functional and narrowly defined skills makes it difficult to demonstrate progress over time (Grisham-Brown et al.., 2005). Assessment that is ongoing, linked to functional skills, and embedded in children's daily schedule and activities are more appropriate for this purpose.

When assessment is used for the accountability purpose, it must be sufficiently reliable and valid so that it will be fair to schools. Researchers indicated that the younger the child, the more difficult it is to use standardized tests to get valid and reliable results (Kamii & Kamii, 1990; Perrone, 1990; Meisels, 1987). Standardized tests used for preschool children have always been questioned for its reliability and validity (Appl, 2000; Brown, 1993; Gnezda and Bolig, 1988; James & Tanner, 1993; Kohn, 2001; McLean, 1996; Nagle, 2000; Ratcliff, 1995). As President Bush made the announcement in 2003 that all Head Start students would be given a national standardized test to account for Head Start's accountability, the reliability and validity issues were raised again. The National Reporting System (NRS), a standardized test, used for Head Start programs was reported lack of reliability and validity (Government Accountability

Office [GAO], 2005). The technical adequacy aspect of the NRS was examined by interviewing all of the members of the Technical Work Group- a team of experts that assisted the Head Start Bureau, and the results indicated that the agency had not shown the NRS to be valid and reliable overtime. Crawford (2005) pointed out in a paper that the because of the lack of reliability and validity, the NRS was not to be used for accountability purposes. In this case, an alternative method of assessing preschool children that is reliable and valid is needed.

Besides the mismatch between the standardized test and the assessment purposes, the social validity of standardized test has also been an issue. Negative impact of standardized test on young children and program quality were identified by a number of researchers (Andersen, 1998; Shepard, Taylor, & Kagan, 1996; Kellaghan and Madaus, 1991; Shepard, 1991). Test results from standardized tests for preschool or kindergarten children often were misused for retention and denying children to enter kindergarten or first grade (Shepard et al..; Smith, 1999; Wortham, 2008). When using standardized tests to report accountability data, pressures forced teachers and school administrators to spend a lot of planning time preparing children to do well on test items instead of devoting time to teach children competencies that are beneficial to their lives in real world. Using standardized tests for accountability narrowed the curriculum by overemphasizing basic skills and neglecting higher-order thinking (Shepard et al..; Kellaghan and Madaus; Shepard; Smith and Rottenberg, 1991). Furthermore, such narrowing is likely to be greatest in program that serving at-risk and disadvantaged children where there is the most pressure to improve outcomes (Shepard et al..; Herman and Golan, 1991). Standardized assessment also causes unnecessary stress and unrealistic expectations for children, especially for young children (Andersen, 1998).

The above discussion of the purposes and methods of assessing young children suggested that if an assessment is comprehensive enough to serve all the above purposes, it must involve multiple documentation methods, multiples assessors, is reflective of what children are learning, is embedded into children's daily schedule, and is technically adequate. As the traditionally used standardized tests have been criticized for their static and rigid nature, as well as its bias against children from different groups and lack of educational relevance (Daniels, 1999; James & Tanner, 1993; Meadows & Karr-Kidwell, 2001; Nagle, 2000; Notari-Syverson & Losardo, 2004; Ratcliff, 1995; Wortham, 2008), a call for an assessment paradigm shift was raised by scholars.

Characteristics and Uses of Authentic Assessment

As authentic assessment has emerged in response to the need of appropriate assessments for young children, it holds the promise to accomplish the goal of education which is to improve child's learning. Authentic assessment supports children to construct, integrate and apply their knowledge and to think critically and creatively (Xue, Meisels, Bickel, Nicholson, & Atkins-Burnett, 2000).

Representing the new assessment paradigm in early childhood education, authentic assessment has four unique traits that are different from traditional standardized tests: 1) conducted under naturalistic environments, 2) process oriented, 3) directly linked to curriculum, and 4) multi-disciplinary and comprehensive. In authentic assessment, children must complete tasks within the context of either a real or a stimulated exercise (Stiggins & Bridgeford, 1986). Authentic assessment is not a one time snapshot and involves the use of the prior knowledge base, in-depth understanding and production of knowledge in an integrated form (Newmann & Archbald, 1992). Wiggins (1989) also described authentic assessment as "contextualized complex challenges, not fragmented static lists of tasks" (p.71). In authentic assessment, students apply knowledge they have mastered to complete the tasks and most of their responses involve multiple skills and knowledge (Moss, 1992; Stiggins & Bridgeford, 1986). The difference between traditional standardized assessment and authentic assessment lies in the fact that the response from standardized test items has to be interpreted to be inferred as the competence in certain domain areas, while the response of authentic assessment items is the competence of domain achievement. The interpretation of the item or instrument is not necessary because the assessment is the product itself (Shepherd, 1991).

Even though authentic assessment is complicated in terms of procedure, it provides more useful information about children (Moss, 1992). It also has more meaningful implication on children's real life for professionals to set goals and to implement personalized instruction to better prepare young children (Daniels, 1999). Authentic assessment has its aesthetic, utilitarian or personal value apart from documenting the competence of the learner and beyond school (Newman and Archbald, 1992). The competence of the learner (know) has to be reflected through learner's performance (do), therefore, authentic assessment goes beyond measuring what children know to measuring what children can do (Pierson & Beck, 1993).

Using authentic assessment practice in classroom has increased children's self-esteem through allowing students to be active participants in the evaluation of their work (Craig & McCormick, 2002; Biondi, 2001). Authentic assessment also leads to better understanding of student progress for both teacher and student than standardized methods (Biondi, 2001). When authentic assessment was used to provide information for curriculum, teachers were able to design more effective instructional programs for students (Grisham-Brown, Hallam, and Brookshire, 2006; Fuchs and Fuchs, 1996), provided better classroom environments (Hallam, Grisham-Brown, Gao, & Brookshire, 2007) and improved child outcomes (Vanderheyden & Burns, 2005).

Researchers also have examined people's attitudes towards authentic assessment and their perceived outcomes of implementing authentic assessment. Different types of authentic assessment such as teacher observation, checklists and event tasks were frequently used by teachers in physical education, and these teachers perceived that the use of authentic assessment enhanced students' self-concept, motivation and skill achievement (Mintah, 2003). In Meisels, Xue, Bickel, Nicholson & Atkins-Burnett (2001)'s study, 246 parents were surveyed on their attitudes toward the work sampling system and the results suggested that parents held positive attitudes toward the work sampling system and two-third of the parents preferred the work sampling system to a conventional reporting card.

Reliability and Validity of Authentic Assessment

Authentic assessment has been linked to positive classroom and child outcomes and favorable teacher and parent feedbacks. However, there is a gap in the literature on the technical soundness of authentic assessment. The technical soundness of a measure relies on its reliability and validity. Maurer (1996) proposed that authentic assessment needs to be reliable, valid and generalizable in order to gain credibility with parents and the public. But it is not easy to collect the evidence of reliability and validity of authentic assessment. All types of authentic assessment have a number of reliability and validity problems that cannot be easily handled with traditional approaches and criteria for validity studies (Brandt, 1992; Moss, 1992). The lack of standardization in authentic assessment is considered one of the main concerns ("Authentic Assessment", 1992). After all, if the assessment processes and the interpretation of assessment results are not rigorous, the assessment itself provides no data for analysis or comparison

(Pierson & Beck, 1993). Also, teachers are sometimes confused about how to reconcile the traditional testing procedures with the dimension of "authentic" performance assessment (Bergen, 1993). Although teachers are now being asked to plan and carry out "authentic" assessments, the major "outcome" variables receiving attention are still children's scores on standardized tests. For most teachers, how to plan and carry out authentic assessments and how to record and report the results remain challenging to them (Bergen, 1993).

Reliability

Researchers have been making efforts in establishing standards for reliability and validity for authentic assessment (Gilbert, 1990). Reliability is one of the factors that can affect the validity of the measure (Salvia & Ysseldyke, 1995). Validity is the extent a test measures what it is supposed to measure. Reliability is the measure of a test's dependability, accuracy, stability, and predictability (Wiersma & Jurs, 1985). If the measurement itself cannot achieve its internal consistency or cannot be scored consistently across time and by different individual, it cannot provide valid information about the extent to which it measures the correct behavior. Therefore, in order for an authentic assessment to be valid, the reliability must be achieved first (Salvia & Ysseldyke, 1995). According to Wiersma & Jurs (1985), reliability is usually determined in one of four ways: test-retest, alternative form, split half, and inter-rater reliability. Test-retest reliability means a measurement is administered twice in a relatively short period of time. Alternative form reliability means that two equivalent forms of the same instrument are administered. Split-half reliability means that test items are divided into two parts and each part is administered separately. Inter-rater reliability means that different observers collect data through direct, ongoing observation of a child's behavior. In authentic assessment, indictors are presented to each child in different ways and the observation is usually ongoing, therefore, it is not possible to construct parallel forms of authentic tasks to establish the traditional test-retest or split-half reliability (“Authentic Assessment”, 2004). However, inter-rater reliability can be and needs to be achieved. A systemic way of evaluating every child’s performance is therefore critical (Bergen, 1993). Scoring criteria, procedure, and behavior to be assessed have to be delicately defined (Pierson and Beck, 1993; Wiggins, 1989).

Studies have been conducted on teachers’ scoring accuracy on authentic, curriculum-based assessment. Studies found out that high levels of reliability can be achieved with sufficient trainings (Khattri, Reeve & Kane, 1998; Jaeger, Mullis, Bourque & Shakrani, 1996; Shavelson, Baxter & Gao, 1993). Meisels et al.. (2001) have also documented that teachers’ judgments on curriculum-based assessment are trustworthy. With these findings, the concerns regarding teachers’ judgment or inter-rater reliability of the authentic measure could be relieved. While the reliability issue of the authentic measure is being addressed, validity of authentic assessment remains another concern for some of its critics.

Validity

Validity refers “degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests” (AERA & APA, 1999, p.9). If the measure is not valid, it cannot provide theoretically sound interpretation of the results. Therefore, the concern regarding the validity of assessment has a critical impact for accountability purposes (Zatta & Pullin, 2004) as well as other uses of the results. A study examining the criterion validity of the Work Sampling System was conducted (Meisels et al.., 2001). The results demonstrated that the work sampling system was correlated well with a standardized battery, and was a reliable predictor of achievement ratings in k-3 grade level. The score generated from Work Sampling System also had significant utility for discriminating accurately between children who were at risk and those who were not at risk. The study shed lights on using authentic assessment as an alternative method to record child outcomes, but more studies need to be done on the validity of authentic assessment in order to convince the public that the results generated from authentic assessment are trustworthy. This leads to the emergent needs of research on the validity of authentic assessment.

CHARPTER THREE

Conceptual Framework

Assessment of children from birth through the preschool year is different from assessment for children of older age. The contemporary models of child development have moved towards a more comprehensive approach (Losardo & Notari-Syverson, 2001). Young children develop in a complex and dynamic process determined by multiple biological and environmental factors which interact with each other in a continuous and reciprocal manner (Meltzer, 1994). As young children are highly sensitive to their environment, any change in any of the factors can lead to a different developmental outcome. Scholars in the 1970s and 1980s also realized that young children are: a) unable to maintain attention for a long period of time, b) their behavior is highly variable from day to day, situation to situation, and even minute to minute; c) they have not sufficiently developed social awareness to want to do best to please the examiner, d) they often turn their attention to things that they are interested in instead of maintaining focus on the assessment tasks, and e) they can be uncooperative; and they are sometimes fear of stranger or strange situations (Peterson, 1987; Johnson-Martin, 1985; Dunst & Rheingrover, 1981; Dubose, 1979, 1981; Harbin, 1977; Bayley, 1969). All of these characteristics put limit on the choice of an assessment measure for young children (Shepard, Kagan & Wurtz, 2001; Bracken, 2000; Bracken & Walker, 1997; Wortham, 1995). Also, due to young children's rapid development paces, one snapshot of their status cannot capture the nature of their development.

Shift in Assessment Paradigm

Since the publication of *A Nation at Risk*, a report of the National Commission of Excellence in Education (1983), a change in assessment paradigms has occurred. Representing the traditional assessment paradigm, standardized tests reflect the idea of tests being used for elimination and retention purposes. Also, the traditional assessment paradigm assumes that child's competence can be demostrated through their one time preformance on test items. With the conceptualization of early childhood assessment changing dramatically over the years. "The early childhood assessment has gradually moved out of the tester's office into more familiar settings and has begun to use team approaches as well as more innovative and contextually relevant techniques than ever before" (Meisels & Fenichel, 1996, p.32). The changing paradigm shifts the key idea of assessment from examining individual's deficits and limitations to emphasizing individual's positive characteristics, from viewing an individual as a static mix of traits and abilities to viewing an individual as dynamic "wholes" or "system, and from closed, secretive forms to more open and communicative forms (Letendre, 2001). Assessment is thus moving from a norm-referenced, product oriented, and direct approach to a criteria-referenced, process oriented, and indirect approach (Vacc & Ritter,1995). Meisels and Fenichel also pointed out that assessment has changing from a fragmented undermining approach to a more systemic, contextual, and integrated approach which is related to the core process of human growth and development.

Factors Contributing to the Changes

According to McAfee and Leong (2002), changes in perception of assessment, theory of children's learning, composition of preschool population, curriculum goals and instructional strategies have led to the change in assessment paradigms.

Perception of assessment. Historically, assessment for young children was often used to eliminate the disqualified ones (Pintner, 1923). However, since the 1970s, people started to rethink the purpose of assessment. The changing perception of assessment suggests that assessment should be used to help the individual learn and improve rather than to determine whether the child “passes” or “fails” (Tyler & Wolf, 1974, p.170; NAEYC & NAECS/SDE, 2003; McAfee & Leong, 2002, p.3). Instead of considering assessment as a tool to judge a child as “smart” or “dull”, people now look at assessment as a way to provide information about the child’s growth and to support teachers’ planning and children’s learning (Paget & Nagle, 1986; Greenspan & Meisels, 1996; Appl, 2000; Shepard, Kagan, & Wurtz, 2001). The fundamental purpose for assessment is now intervention and instruction (Meisels & Fenichel, 1996).

With the fundamental purpose of assessment changing to intervention and instruction, more emphasis is placed on the child’s strengths, interaction, and the environment within the assessment process. By paying more attention to the context in which the child can learn and perform better, rather than focusing only on the discrete skills that the child fails, teachers and other related personnel can design more effective and individualized curriculum and intervention strategies for this particular child (Shepard, Kagan, & Wurtz, 2001; NAEYC, 2003). Therefore, the changing perception of assessment leads to a contextual paradigm for the emerging preschool assessment.

Theories of learning and development. Since assessment is directly linked to children’s learning outcomes, the way the assessment is designed is inevitably influenced by learning theories. The traditional belief of children’s learning and development includes the views from behaviorism and associationism. Since the simple stimulus-response association is the basic premise of behavioral learning theory (Thorndike, 1922), the traditional view of learning assumes that children simply learn what is taught and reinforced (Thorndike, 1922; Hull, 1943; Skinner, 1954; Gagne, 1965) and that learning only occurs when the child is stimulated and reinforced. Based on this point of view, the test itself becomes the reinforcement of the learned skills, and it is believed that tests on discrete knowledge ensure learning (Hull, 1943; Skinner, 1954; Gagne, 1965). Also, behaviorism asserts that learning is highly sequenced and is not transferable (Hull, 1943; Skinner, 1954; Gagne, 1965), which means mastery of one skill cannot influence the mastery of another skill, and therefore, skills have to be taught separately. Since the traditional assessment paradigm was established during the period when the objective world view dominated and people's deficits and limitations were emphasized, the traditional assessment paradigm took a behaviorist point of view in interpreting the child's learning. Assessment at that time thus served the purpose of eliminating the unqualified (Shepard, 2000).

In contrast to the traditional assessment paradigm, the emerging assessment paradigm arises during the period when the postmodernist world view is dominating. Since postmodernism emphasizes subjectivity, interaction, and flexibility, it is more comfortable for assessment experts to take the constructive and contextual view of learning. Therefore the constructivist, cognitive, and contextual learning theories, which suggest that children actively construct knowledge within a context and make meanings out of activities, provide theoretical foundations for the emerging paradigm (Shepard, 2000).

In detail, cognitive learning theory views learning as an active mental process of acquiring, remembering and using knowledge (Shepard, 2000). Jean Piaget is an important figure in founding the cognitive developmental theory for children. Piaget (1954) proposed that the learning process is a period of knowledge construction that results in a behavioral change after a cognitive scheme has been developed. Piaget's (1978) cognitive theory suggested that learning takes place when knowledge is internally accommodated and personalized. In order for knowledge to be internally accommodated and personalized, the child must develop a cognitive scheme that can make it possible. According to Piaget (1978), children go through different developmental stages in their childhood. These stages are invariant and irreversible for all children. In each stage, one specific cognitive scheme is developed. Even though each child moves at an individual pace, the sequence for these stages is the same for all children. Different characteristics appear at different stages, and different cognitive schemes can lead to different interpretations and responses to the same stimulus. The same behavior could mean different things for children at different cognitive stages. Therefore, Piaget's cognitive learning theory reminds test developers, teachers, and parents to pay attention to the interpretation of child performance. Simple yes/no or multiple choice answers may not be helpful for teachers to interpret the child's competency and learning process. The same correct answer does not necessarily mean that all children are learning the same thing. From the cognitive point of view, the emerging assessment paradigm focuses on individualized tests with specific tasks and the environment for different children.

Constructivists understand the individual as an "active agent seeking order and meaning in the social contexts where his or her unique personal experiences are challenged to continue developing" (Mahoney, 1995, p.5). Constructive learning theory views learning as a complex and active mental process. It assumes that people learn skills and knowledge through being actively involved in the activities. According to constructivists, children do not learn passively but actively. They are meaning-making people who continuously integrate prior experience with new information. Children work on, process, interpret, and negotiate the meaning of information encountered (Newmann, Mark & Gamoran, 1996). Once they do that, they incorporate new skills and knowledge presented into their repertoire. Therefore, offering children opportunities to interact with the immediate environment helps them to learn new knowledge. Besides interacting with the environment, children also actively involve themselves by responding to the feedback they receive during the learning process. A child does things, gets feedback, and then makes modification according to the feedback. After several trials and modifications, the child is able to perform necessary skills and gain knowledge. Teachers' feedback to children is a central part of the social process that mediates children's development of intellectual abilities, construction of knowledge, and formation of identities (Shepard, 2000). Therefore, the interaction between teachers and children is important in facilitating children's learning. Constructivists also believe that the perspective of the observer and the object of observation cannot be separated (Gergen, 1985). Furthermore, since the idea of construction, rather than the retrieving and remembering knowledge, is the critical component of learning, the emerging assessment paradigm looks more into the process than the product. In other words, children's ability to construct knowledge and perform tasks is paid more attention than their ability to remember the information. Following constructivist learning theories, the emerging assessment paradigm values the active involvement and meaningful interaction between teachers and children during assessment.

Contextualists such as Lev Vygotsky view a child-in-context participating in some event as the smallest unit of study (Miller, 1992). According to the contextualists, the interaction among child, object, and another person leads to the learning (Vygotsky, 1978). The child's learning is inherently social (Vygotsky, 1978). The child, the object and the other person are fused in some activities, which is the context. Contexts define and shape any particular child and his or her experience. When the context changes, the child's developmental outcomes change accordingly.

Contextualists also believe that development can only be understood by looking at the process of change, not the static frozen developmental moment. Process is more important than product because it provides more information about children's learning by looking directly at a child’s series of actions and thoughts as he or she tries to solve a problem and advances his or her own thinking. Based on that belief, Vygotsky (1978) examined what a child actually did over time when he or she was involved in an activity with other people and objects, rather than focusing on what concepts he or she possessed. Also, his idea of a zone of potential development brings more insights on how one should assess children. According to the *zone of proximal development* (ZPD), each child has its “potential development level as determined through problem solving under adult guidance or collaboration with more capable peers” (Vygotsky, 1979, p. 86). Therefore, the focus of investigation should be more of a process rather than a static product. In addition, the idea of ZPD reminds us that adult-child or child-child interaction need to be counted in assessing the child’s potential developmental level. Following the contextualists’ learning theory, the emerging assessment paradigm emphasizes interaction, context, and process.

Based on these learning theories, intellectual ability is socially and culturally developed, learners construct knowledge within a social context, new learning is shaped by prior knowledge and culture perspective, and the process tells more about learning and the learner than product (Shepard, 2000). Therefore, an emerging assessment paradigm which adopts these learning theories tends to pay more attention to interaction, context, and process than traditional assessment does.

Change of composition and characteristics of classroom population. The changing nature of classroom population also contributes to the changing paradigm as well. With the increased diversity in racial, ethnic, socioeconomic, linguistic, and educational aspects in early childhood programs, there are more and more demands that assessments fairly reflects diversity. Traditional testing and measurement, which values the objectivity of the world is not always suitable for children from different ethnic and cultural backgrounds (Berlak, 1992). This calls for a new type of assessment that values human subjectivity results in the emerging holistic assessment paradigm in which the child is the focus of the assessment instead of the test.

In addition, since the preschool population has its unique characteristics such as vulnerability, rapid developmental changes, and behavioral fluctuations (Paget & Nagle, 1986), the emerging preschool assessment is more flexible, on-going, and sensitive to individual needs.

Change of curriculum goals and instructional strategies. The advocacy for the connection between assessment, curriculum and instructional strategies moves the assessment paradigm to be more integrated, multi-faceted, and ongoing (Meisels & Fenichel, 1996). According to Bredekamp and Rosegrant (1992), curriculum goals, content, instructional strategies, and assessment should be interrelated. The traditional approach to curriculum and instructional strategies suggested specified educational objectives that needed to be taught, curriculum content should be designed using a utilitarian approach, and different objectives should be taught based on the “innate intellectual ability” of the students (Bobbitt, 1912). Based on this curriculum and instructional approach, assessment was an objective and discrete measurement that eliminated the “unqualified”.

However, with the changing view of children’s learning and development, curriculum goals and instructional strategies have been modified as well. Curriculum and instructional guidelines for what children should learn and how they learn have been developed throughout the nation (NASP, 1999; NAECS/SDE, 2003; NAEYC, 2003). These guidelines suggest “hands-on” and “real-world” focuses on the curriculum and instruction, as well as equal opportunities for all children. Therefore, traditional standardized testing which focuses on discrete skills and the elimination of unqualified students does not adequately reflect the current trends in curriculum and instructional strategies. Instead, recommended assessments emphasize integrated and higher order thinking skills.

All the above factors move the assessment paradigm from a psychometric approach to a contextualistic approach. The move improves the possibility of obtaining more useful information about young children (See Figure 1 and Figure 2).

Recommended Assessment Practices

In response to the paradigm change, the National Association for the Education of Young Children (2003) and the Division of Early Childhood (2005) recommended assessment practices for young children. Both NAEYC and DEC recommend that assessments should be: 1) developmentally appropriate, 2) cultural and linguistically responsive, 3) embedded in daily activities, 4) multidisciplinary, 5) linked to curriculum and instruction, and 6) involving families.

Developmentally Appropriate

Assessments for young children should include measures that address all areas of development including physical well-being and motor development, social and emotional development, approaches to learning, language development, and cognition and general knowledge (NAEYC, 2003; DEC, 2005). Assessment approaches and materials should match children’s interests and developmental status. Contrived tasks and materials, as well as administration by unknown people under unfamiliar circumstances are not preferred. Young children are easily discouraged by strangers and they can be less likely to talk to strangers, which may affect the results of speech and language assessments. Assessments should not only measure children’s immediate mastery of a skill, but also whether a child can demonstrate the skills across settings, activities, and with other people.

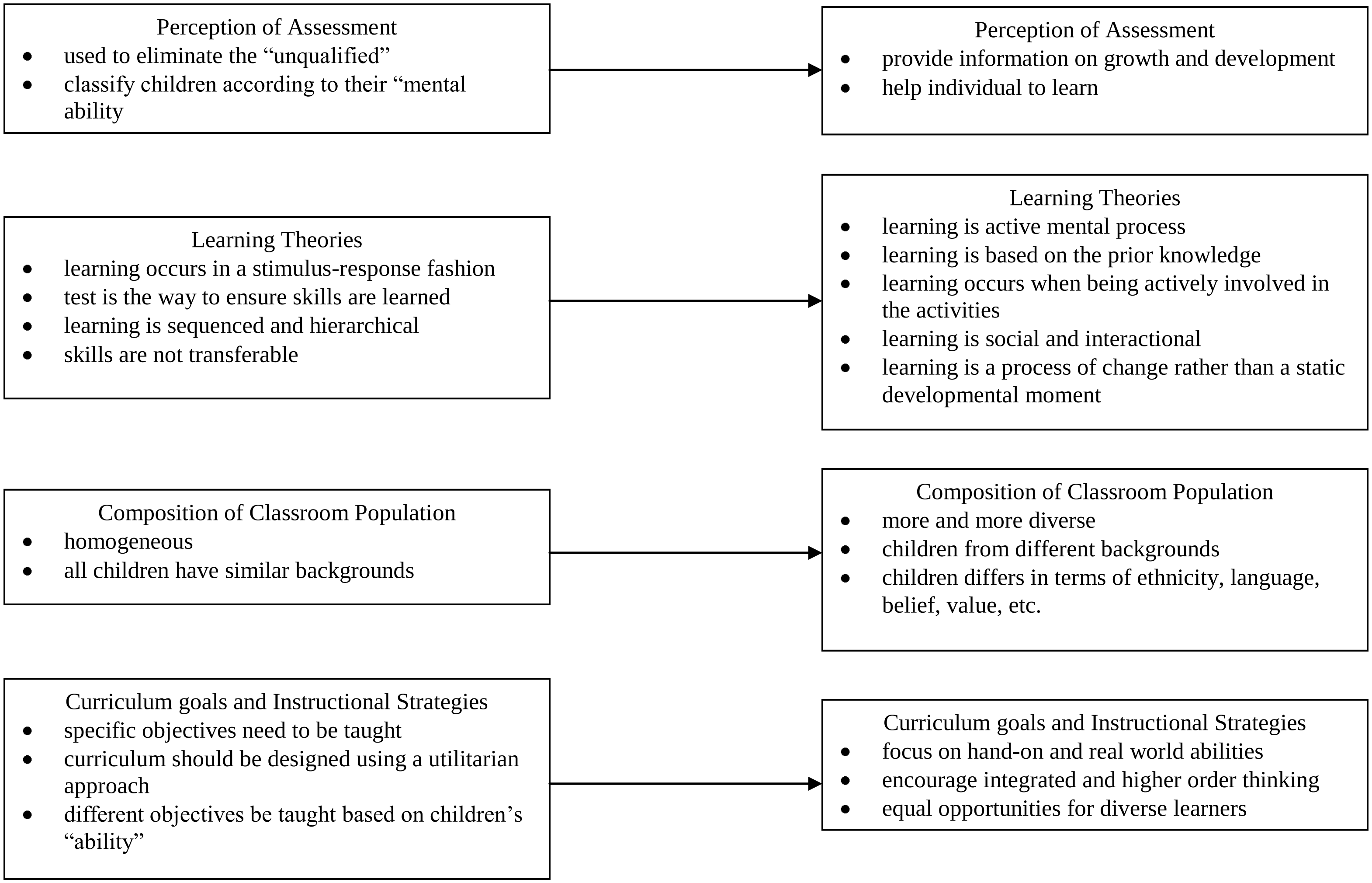
The assessment system should also emphasize repeated observations and documentation in order to reflect the smallest increment of change. Children with more severe delays and impairments especially need assessment that is sensitive to small increment of progress (DEC, 2001).

Culturally and Linguistically Responsive

Except being developmentally appropriate for children of different interest and developmental status, assessment should also be valid for use with children from different backgrounds including age, culture, home language, socioeconomic status, abilities and disabilities (NAEYC, 2003). Assessment should ensure teachers' recognition of similar knowledge and skills across differences in culture representation and incorporate culturally based experiences, including family values and language. Assessment materials and procedures should reflect children's sensory, physical, responsive, and temperamental differences and accommodations should also be made for children with disabilities (DEC, 2001).

Transdisciplinary Assessment

Multiple sources of information should be used to document children's progress and multiple measures should be used to assess children's strengths, progress and needs (NAEYC, 2003; DEC, 2001). This requires a transdisciplinary approach which means to integrate the expertise of different specialists in different areas so that more efficient and comprehensive assessment may be conducted (Bruder, 1994). Information should be gathered from families, professional team members, different agencies, service provides, and other regular caregivers. For example, the arena assessment emphasizes the collaboration between different agents (Parette, Bryde, Hoge & Hogan, 1995). With the involvement of people from different disciplines, across different settings, an assessment can capture a whole picture of a child in order for people to make sound decisions about the child.



Perception of Assessment

- used to eliminate the “unqualified”
- classify children according to their “mental ability

Perception of Assessment

- provide information on growth and development
- help individual to learn

Learning Theories

- learning occurs in a stimulus-response fashion
- test is the way to ensure skills are learned
- learning is sequenced and hierarchical
- skills are not transferable

Learning Theories

- learning is active mental process
- learning is based on the prior knowledge
- learning occurs when being actively involved in the activities
- learning is social and interactional
- learning is a process of change rather than a static developmental moment

Composition of Classroom Population

- homogeneous
- all children have similar backgrounds

Composition of Classroom Population

- more and more diverse
- children from different backgrounds
- children differs in terms of ethnicity, language, belief, value, etc.

Curriculum goals and Instructional Strategies

- specific objectives need to be taught
- curriculum should be designed using a utilitarian approach
- different objectives be taught based on children’s “ability”

Curriculum goals and Instructional Strategies

- focus on hand-on and real world abilities
- encourage integrated and higher order thinking
- equal opportunities for diverse learners

Figure 1: Theoretical foundation for assessment paradigm shift (Shepard, 2000; Wortham, 1996)

Linked to Instruction

Assessment information about child's growth and development should be used to provide information for teachers and related personnel to make decisions regarding changes to the environment, interactions, and experiences that will enhance the child's development (NAEYC, 2003). Based on the assessment results, teachers develop short term and long term plans for each child and the group considering children's knowledge, skills, interests, and other factors. Professionals also should measure the level of support a child needs to perform a task so that when planning for curriculum, teachers can make instructional changes. Assessment that can provide results that are immediately useful for planning is preferred.

Family Involvement

Children, birth to eight, benefit from close partnerships and ongoing communication between their families and their education programs. Therefore, teachers and parents should share information periodically regarding children's progress in all domains. When evaluating a child, professionals and families provide different aspects of information about a child, so they are independent rather than interchangeable raters (Suen, Lu, Neisworth & Bagnato, 1993). Teachers and families should work together to make decisions on children's learning goals and approaches to learning. As families are central partners of their children, professionals should provide easy access to families regarding children's assessment activities. Professionals should collaborate with families to discuss selection of materials, processes, methods, and assessment situations that are best for the child (Grisham-Brown, 2000). Information regarding children's routines, interests, abilities, and special needs should be collected from families (NAEYC, 2003), as well and assessment information should be written in an understandable and family-friendly way (DEC, 2005).

Naturalistic. Young children should be assessed in contexts that are familiar to the child (NAEYC, 2003; DEC, 2001). Assessments should include teachers' observational records of children's experiences during regular classroom time, in a wide variety of circumstances that are representative of the child's behavior in the program over time. The recordings should capture the behaviors children use in routine circumstances instead of artificial situations that impede the usual learning and developmental experiences in the classroom, or divert children from their natural learning processes. By avoiding artificial settings, authentic assessment reduces construct-irrelevant influences on young children's test performance because the test-takers' environment, their physical situations, temperament, as well as examiner's characteristics makes differences in children's responses (Bracken, 2000).

Based on the emerging assessment paradigm stated above, authentic assessment has emerged as an approach for assessing young children. It follows the changing trend. Authentic assessment values different responses from children. It allows different paces and forms of responses. Authentic assessment usually assesses abilities that are meaningful to the society. Items in an authentic assessment reflect the value of the current society and what children are supposed to learn. Authentic assessment also integrates different facets of children's abilities. Local institutions are empowered in the authentic assessment because classroom teachers can make scoring decisions (Berk, 1986; Wiggins, 1989; Stiggins & Bridgeford, 1986; Moss, 1992; Newmann & Archbald, 1992). Therefore, the changing assessment paradigm offers a conceptual foundation for the development and use of authentic assessment.

Curriculum-based assessment, as one type of authentic assessment possess the characteristics of recommended assessment practices. The Assessment, Evaluation, and

Programming System, 2nd Edition (AEPS®) as an curriculum-based assessment responded to the assessment paradigm shift and is designed for eligibility and intervention purposes.

The research goal of this study is to investigate the validity of the AEPS® for accountability purpose.

CHAPTER FOUR

Methods

Research Questions

Research Design

Quantitative. A quantitative approach was used to answer research questions one and two. Concurrent validity, as one type of the criterion-related validity, examines at the correspondence between the test to be validated and the criterion measure that has already been validated (Wortham, 1995). The connection between the criterion measure and the test to be validated is the evidence for criterion-related validity (Nunnally, 1978). The connection between criterion and the assessment is traditionally estimated by correlation coefficient (Carmines & Zeller, 1983; Nunnally, 1978), therefore, correlational analysis is often used in concurrent validity studies (Startup, Jackson & Bendix, 2002; Amodei & Lamb, 2003; Mirrett, Bailey, Robert & Hatton, 2004). Following the traditional trend, a correlational design was used in this study to investigate the relationships between the test to be validated (The Assessment, Evaluatoin, and Programming System, 2nd Edition) and the measure serving as the criterion (The Battelle Developmental Inventory, 2nd Edition).

In order to examine the concurrent validity of Assessment, Evaluation, and Programming System, 2nd Edition (AEPS®), scores generated from the test were correlated with scores generated from the criterion measure. In this study, the Battelle Developmental Inventory, 2nd Edition (BDI-2), a widely recognized and validated standardized measure, was chosen as the criterion measure. The correlational design examined whether scores generated from the AEPS® correlated with scores generated from the BDI-2. If there were statistically significant correlations between these scores, evidence would support the hypotheses that the AEPS® is a valid measure in assessing young children's competence in cognitive and communication domain.

Measures

Assessment, Evaluation and Programming System 2nd Edition (*AEPS*®). The AEPS® (Bricker, Pretti-Frontczak, Johnson, & Straka, 2002) is a curriculum-based assessment for infants and young children. It is designed for 1) determining a child's present level of functioning, 2) developing developmentally appropriate goals for individual child, 2) planning intervention, and 4) evaluating a child's performance overtime. The AEPS® has two separate sets of tests designed for children of different ages, one for children from birth-to-three, and another for children from three to six. In this study, only the three to six set was used. The AEPS® assesses young children's competency in fine motor, gross motor, adaptive, cognitive, social-communication and social areas based on their performance on everyday activities. For the purpose of this study, only the cognitive area and social-communication area in the AEPS® three to six set were used.

The history of the AEPS® can be dated back to 1974 when a group of concerned professionals decided to develop an alternative measurement for young children who ranged from birth to 2 and that would yield educationally relevant outcomes. In the early 1980s, the first complete and usable assessment/evaluation tool became available for comprehensive field testing. The tool was called *Adaptive Performance Instrument* (API). Between 1983 and 1984, the API was modified and renamed twice. The name was changed to the *Evaluation and Programming System: For Infants and Young Children* (EPS). In 1993, the name of the mearsure was changed again to AEPS for Birth to Three Years. With the pressure of expanding the AEPS to cover the developmental range from 3 to 6 years, work has been done on the development of a test and associated curriculum to address the developmental range from 3 to 6 years. In 1996, the test was titled the *Assessment, Evaluation, and Programming System Test for Three to Six Years*. In 2002, the birth to three and three to six AEPS tests have been combined and revised as the AEPS second Edition.

In the AEPS®, there are 54 items divided into 8 strands in the cognitive area. The eight strands are: concepts, categorizing, sequencing, recalling events, problem solving, play, pre-math, phonological awareness, and emergent reading. For each item one of the 3 scores was assigned. A score of 0 indicated the child is not able to perform the corresponding skill, a score of 1 indicated the child can perform the skill inconsistently or with some assistance, and a score of 2 indicated the child is able to perform the skill without assistance and consistently. After each item was scored, a composite score was obtained by adding item scores together. There are 49 items divided into 2 strands in the social-communication area. The strands were social-communicative interactions, and production of words, phrases, and sentences. The scoring rule for the social-communication area was the same as that of the cognitive area.

Multiple studies have been done to examine the psychometric properties of the AEPS® test and its revised version. The original version of the AEPS® test- the Evaluation and Programming Sytem For Infants and Young Children (EPS) was tested for its reliability, validity and utility (Bricker, Bailey, & Slentz, 1990; Notari & Bricker, 1990; Bailey and Bricker, 1986; Slentz, 1983). It has been suggested to be reliable, valid and useful in generating appropriate information for education programming for children with disabilities. For example, the test-retest reliability coefficients ranged from adequate to good for all areas except for gross motor and adaptive areas. Concurrent validity of EPS was examined with McCarthy Scales of Children's Abilities (McCarthy, 1972) and the Uniform Performance Assessment System (Haring, White, Edgar, Affleck, & Hayden, 1981). The correlation coefficients ranged from .37 for the Adaptive

Area to .97 for the cognitive area, all of which were significant. All these figures lay out the evidence of the reliability and validity of the AEPS® measure.

When the AEPS® was further developed, its reliability and validity were re-examined. The Assessment, Evaluation and Programming System (AEPS®) was valid in assessing functional skills and corroborating eligibility decisions (Macy, Bricker, & Squires, 2005; Bricker, Yovanoff, Capt, & Allen, 2003; Kim, 1997; Straka, 1994), sensitive to the differences among children of varying ages and children with and without disabilities (Hsia, 1993), and provided higher quality IEP goals and objectives for children with disabilities than other assessment system (Pretti-Frontczak & Bricker, 2000; Hamilton, 1995). The AEPS® also has been examined for its psychometric properties. It was found to have satisfactory interrater reliability and internal consistency. It also was sensitive to differences demonstrated by children of different ages and disability status (Noh, 2005). However, few studies has been done on the validity of AEPS® , and this study intended to provide one type of validity evidences of the AEPS®.

The Battelle Development Inventory, 2nd Edition (BDI-2). The BDI-2 is a standardized measure designed for the purpose of screening, diagnosis and evaluation of young children's early development. It was developed based on the concept of milestones. The conceptual foundation of the original BDI was that a child attain skills in certain sequence, and the acquisition of each skill depends on the acquisition of the preceding skills. Based on this underlying concept of development, the BDI development team analyzed large number of items from different measurement instruments and clustered items together based on the behaviros these item measures. From these clusters, a sequence of behaviors was derived to describe the functioning of typically developing children at various stages of development. Theses behaviors were later analyzed and categorized into five major areas of development and smaller subdomains. After identifying these behaviors, items were developed to assess these behaviors. The second version of the BDI has blended many of the important features of the earlier edition with the improvements in psychometric design, changes in the life-experiences of children and the availability of user-friendly materials and technology (Newborg, 2004).

The BDI-2 is ideal for several uses: identification of children with special needs, evaluation of children with special needs in early education programs, assessment of typically developing children, screening for school readiness, and program evaluation for accountability. This measure is appropriate for ages from birth to seven. The assessment is appropriate for both typically developing children and children with special needs. The BDI-2 assesses children's development in five domains: personal-social, adaptive, motor, communication, and cognitive domain. Each domain includes several sub-domains. For the purpose of this study, only the cognitive and communication domain were used. There are three sub-domains in the cognitive domain: attention and memory, reasoning and academic, and perception and concepts. There are two sub-domains in the communication domain: receptive communication and expressive communication. In each sub-domain, items are listed according to chronological order. All BDI-2 items are presented in a standard format that specifies the behavior to be assessed, the material needed, and the recommended procedures for administering the specific item. Each item was scored 0, 1 or 2. A score of 0 indicated that the child has not mastered the skill at all; a score of 1 indicated that the skill is emerging; and a score of 2 indicated the child has already mastered the skill. Based on the child's chronological age, there are different starting points for each age level. From the first item being administered, if the child gets three consecutive 2s, the basal is established. When calculating the raw score, any item before the basal is scored 2. When the child gets three consecutive 0s, a ceiling is obtained. Any item after the ceiling is scored as 0.

When adding all the scores together a raw score of the domain is obtained and can be converted to a scaled score. According to the age level, the sum of scaled scores in one domain can be transferred to the developmental quotient (DQ) score.

The internal consistency of the BDI-2 was examined using the split-half method. The split-half method splits a single administration of the test into two halves for analysis and the scores from these two halves of test will be correlated. According to Bracken (1987), Nunnally (1978) and Salvia & Ysseldyke (2001), in order for test score to be considered reliable, the reliability coefficient for the two halves should be higher than .80 for the subdomain score and higher than .90 for domain score and total scores. The reliability coefficient for the total BDI-2 DQ scores is average at .99 across all 16 age groups, and internal consistencies of the 13 subdomains report a range of .89-.93, which indicates that the measure is sufficiently reliable. Also, inter-rater reliabilities on 17 subjective items were examined, and 94% to 99% agreements were achieved.

Besides being validated with the original BDI, the BDI-2 was also validated with Bayley Scales of Infant Development-II (Bayley, 1993), and moderate relationships between the two tests were found A correlation of .61 was reported between BDI-2 cognitive domain and BSID-II mental index, and .75 was reported as the correlation coefficient between BDI-2 communication domain and BSID-II mental index. The BDI-2 was also validated with the Denver Developmental Screening Test-II (Frankenburg et al.., 1992) High agreements on identifying potential problems (range from 83% to 89%) were found. Among all domains, the agreement on identifying potential problem in communication domain was the highest, at 89%. The DQ scores from the BDI-2 communication domain and the scaled scores from the two subscales in the domain were correlated with the Preschool Language Scale, Fourth Edition (PLS-4)

(Zimmerman, Steiner, & Pond, 2002). Moderate to high correlations were found between BDI-2 communication scores and PLS-4 scores. These evidences indicate that BDI-2 is a valid measure in measuring young children's cognitive and communication ability.

Recruitment and participants

All children enrolled in five preschool classrooms in an elementary school in Fayette County were recruited as participants. Recruitment lasted one month, from January 2006 to February 2007. The researcher contacted all five preschool teachers working in these five preschool classrooms. Teachers were asked if they were willing to distribute parental consents to the parents of children enrolled in their classroom. Once all teachers agreed to distribute the consents, the researcher furthered the recruitment procedure by confirming with teachers the number of children enrolled in their classrooms. The researcher then made enough copies of informed consents [IRB Stamped] for each parent and distributed them to each classroom.

The informed consent form listed the research purpose, the measures for the research, and the procedure of testing and explained the potential risk and benefits in sufficient details. Contact information of the researcher and the faculty supervisor also were listed. By signing the consent form, parents agreed to: 1) give permission to the researcher to conduct the BDI-2 test on the child, 2) give permission of disclosing the child's demographic information and AEPS® data to the researcher.

Parents who were interested in letting their children participate in the study signed and returned the informed consent form to the child's classroom teacher. After collecting the signed informed consents, the researcher started to conduct BDI-2 assessment on children with signed consent.

There were 100 children enrolled in these 5 classrooms during the 2005-2006 school year. Ninety percent of the population was African American and the rest of them were Caucasians and Hispanics. The age range of enrolled children was from 3 to 5 years old. Based on the power analysis formula, (Cohen, 1988) in order to predict a statistically significant correlation at 0.5 level, the sample size had to be at least 28 to reach a conventional 0.8 statistical power. For the purpose of this study, the researcher attempted to recruit 50 English speaking children. However, only 34 parents returned signed consent forms. Based on the teachers' report, all of these 34 children were fluent in English.

Procedures

The researcher started to collect parental consents two weeks after they were distributed by teachers to each parent. Once an informed consent was collected, the researcher went into the classroom and tested the child using the BDI-2. The test of BDI-2 occurred while children's AEPS® data were collected by teachers in the classroom. Both the BDI-2 data collection and AEPS® data collection occurred between the second week of February 2006 and second week of March 2006. By arranging the BDI-2 test concurrently with the AEPS® data collection, the concurrence of scores generated from these measures was ensured.

Before the researcher collected BDI-2 data, the researcher had participated in two BDI trainings. The trainings ensured that she was familiar with the measure and procedures of administering the test, as well as the scoring procedures. The first BDI-2 training was provided by the publisher, focusing on the historical and policy issues regarding the BDI-2 test development. The second training was provided by someone who had been trained by the publisher, focusing on the administration and scoring issues of the BDI-2 test. During the second training, the researcher was required to administer items in BDI-2 in front of the trainer so that the procedural reliability was checked. These two trainings prepared the researcher with both theoretical background and administering experience of the BDI-2. During the data collection period, the researcher went into the classroom with BDI-2 test books and manipulative kits, called children according to the participants list, and took children to the designated testing areas which included both the staff conference room and an area which consisted of one round table and two small chairs outside of a preschool classroom. There were three ways the researcher could administer the BDI-2 test items. The researcher could obtain information through structured test format, observation and interview with caregivers. The recommended ways of administering the specific item was listed for each item, and the researcher chose the one that was most suitable for the situation based on her previous knowledge. Most of the items were conducted using the structured test format because it was the most efficient way to collect data.

According to the child's chronological age, the researcher chose the appropriate starting item to begin the assessment. If the child scored 2 on the first item, the researcher continued to the next item. When the child scored 2 on three consecutive items, a basal was established. Once the basal was established, the researcher continued testing the child according to the order of the items until the child scored 0s on three consecutive items, meaning a ceiling was established. However, if the child failed to score 2 on one of the first three items, the researcher administered items in the reverse order until the child got 2s on three consecutive items. The researcher stopped testing when both basal and ceiling were established. The approximate time for conducting the BDI-2 cognitive and communication subdomains was about 45 minutes. On average, the researcher tested 3 children per day during the days she went into the classroom and collected the data. The data collection period lasted about a month in February. Some children were tested in the month of March.

Between the middle of February and middle of March, teachers in these five classrooms collected children's AEPS® data. The AEPS® data collection was mandatory by the school. The teachers in these 5 classrooms collected data using a set of activities during which they observed children and gave scores on AEPS® items. All five teachers had received technical assistance on how to collect children's AEPS® data. Five research staff from University of Kentucky went into classrooms and helped teachers in administering the AEPS® test by modeling how to conduct activity-based assessment, answering questions about the scoring of the AEPS®, and assisting data entry into the online AEPS® data system. An AEPS® certified trainer calculated a reliability session with all five teachers to ensure the scoring accuracy of teachers. Eighty percent agreement was reached. The AEPS® data were collected three times across the 2005-2006 school year. The fall semester data were collected between September 2005 and October 2005, the mid-point data were collected between February 2006 and March 2006, and the spring semester data were collected between March 2006 and April 2006. For the purpose of this study, only the AEPS® data collected between February 2006 and March 2006 were used. That time period was chosen to ensure the concurrency between AEPS® data and BDI-2 data.

Analyses

For this study, data on two areas were analyzed. They were children's cognitive area and communication area. The cognitive areas of both tests describe children's understanding of numbers, letters, consequences, logical relationships, spatial relationships and print. The communication areas of both tests describe children's ability to use verbal or non-verbal language to communicate with his or her environment.

SPSS 11.5 for Windows was used for child data analysis. Descriptive statistics including mean and standard deviation were run for the AEPS® cognitive and social-communication areas.

They were also run for the BDI-2 cognitive and communication domains. The evidence of concurrent validity was demonstrated by the correlation between the test score and criterion score (Carmines & Zeller, 1983; Messick, 1983; Nunnally, 1978). Therefore, Pearson's product moment correlations were run between the AEPS® score and the BDI-2 scores. Scores on AEPS® cognitive area were correlated to scores on BDI-2 cognitive domain, and scores on AEPS® social-communication domain were correlated to scores on BDI communication domain.

Raw scores were used for analyses for all the AEPS® areas and strands. In order to be consistent with technical adequacy studies on the BDI-2 where the developmental quotient (referred as DQ in the following text) scores were used, the DQ scores were used for BDI-2 cognitive and communication domains (Newborg, 2004). The DQ score is the normalized standard scores with mean of 100 and standard deviations of 15. It depicts the child's relative standing among the population. Based on the raw score, the DQ score of each domain can be obtained by using Appendix C of the examiner's manual. Because DQ score was not available for BDI subdomains, raw scores were used for the subdomains.

Before correlating scores generated from both tests, some preliminary data analyses were conducted. Data analyses included producing descriptive statistics for domain scores and sub-domain scores. In the descriptive statistics, mean, standard deviation and range of scores were reported. Due to the small sample size, efforts were made to avoid missing data. Two cases with missing data were eliminated from the analysis.

After conducting preliminary analysis, scores of both tests were entered into SPSS version 11.5. Correlation analyses were run by this statistic software. Correlation coefficients were generated by the correlation analysis. The statistical significance of the correlation was calculated by the correlation analysis as well.

After correlating the AEPS® cognitive and social-communication area scores with BDI-2 cognitive and communication domain scores, correlation analysis were conducted to explore the relationship between each strand under AEPS® cognitive and social-communication area and each subdomain under BDI-2 cognitive and communication domains. Further analyses also included correlations between social communicative strand and words, phrases, and sentences strand from AEPS® communication area and expressive communication subdomain and receptive communication subdomain from BDI-2 communication domain, as well as correlations between concept, category, sequence, recall, problem-solving, play, premath, phonological awareness, and emergent reading strands from AEPS® cognitive area and attention and memory, reasoning and academic, and perception and concepts subdomain from BDI-2 cognitive domain. As the DQ scores were not available for BDI-2 subdomains, raw scores were used in the exploratory analysis, and raw score were used for all the AEPS® areas as well. Correlating subdomain scores from these two measures provided detailed information on whether and how children score similarly on these two measures.

Qualitative. A qualitative approach was used to answer research question three: what are teachers' perceptions on using authentic assessment to obtain information on children? The qualitative approach was used because: 1) research questions start with the word, *what*, 2) research questions need to be explored. When variables are not easily identified and theories have not been established to explain the behaviors and perceptions of the participants, a qualitative approach is appropriate (Creswell, 1998). Theories on how teachers perceive both traditional and authentic assessments have not been systemically developed yet, therefore, the qualitative approach is needed for exploring this topic.

In this study, the social validity of the proposed measure relied on teachers' perceptions on the assessment practice. In order to obtain the in-depth information on teachers' perceptions of the assessment practices and the factors that influence teachers' perceptions, a focus group session was conducted. A focus group was selected for this study because it: 1) has been proven to be effective in social science research for exploring perceptions of participants (Brotherson, 1994; Wesley, Buysse, & Tyndall, 1997; Kern, 2007), 2) elicits multiple perspectives from participant, and 3) addresses questions that inform or assess practice (Brotherson, 1994). The focus group was facilitated with a 11-question structured open-ended interview. Seven of the questions were asked individually and four of them were asked together.

Recruitment and participants

Lead teachers in the five preschool classrooms where children were recruited for the child assessment part of this study were asked to participate in the social validity component of this study by attending a focus group. These teachers' perceptions on both standardized and authentic assessment were discussed because all of them had experiences with both types of assessment. They had been collecting child data on standardized measures for accountability purposes. They were also trained to collect child data on authentic assessments.

After getting oral agreement from teachers, the researcher brought a copy of the teacher consent to each teacher. In the teacher consent, the research purpose, teacher's responsibility, research procedures, and potential benefits and risks for participation were explained. Contact information of the researcher and the faculty advisor was also listed. Teachers were encourage to contact either of them if they had any question or concern.

Focus Group Protocol

The focus group took place in a reserved conference room in the elementary school where all participants (preschool children and classroom teachers) were recruited. Permission for use of the school facility was obtained from the school principle.

The focus group lasted approximately 45 minutes including the set up and group interview. The room was set up using a conference table and several conference chairs. The chairs were placed around the table. Sandwiches, vegetable plate and drinks, provided by the researcher, were placed in the middle of the table, together with the tape recorder the researcher used to record the group interview.

The researcher started the structured group interview by introducing herself, followed by the explanation of the research purpose for this study, and the goal for the focus group. The researcher also explained to the participants that participation was voluntary and they could request to stop the recording any time they felt uncomfortable.

The guideline and norms of the focus group were then read by the researcher. The researcher ensured the participants understood the guidelines and norms by allowing them to ask any questions. The participants indicated no questions.

After the introduction and explanation of the process, the tape recorder was then turned on. The researcher initiated the interview process by starting with the first component that discussed the experience teachers had using standardized and authentic assessment. After introducing the components, specific question related to the component were asked.

All participants were asked to comment on the question introduced by the researcher. If they failed to comment on the question on the first round, the researcher reminded them by giving them a choice of either stating “my opinion has been stated by others” or commenting on the question.

When all the questions were answered and commented, the researcher stopped the tape recorder and thanked the participants. The researcher then delivered a package of classroom materials worth $25 to each teacher as a thank you gift for participation.

Analyses

The focus group data analyses were accomplished by the researcher to answer research question three. There are two approaches to analyze focus group data in the literature, ethnographic summary and content analysis (Morgan, 1988). The ethnographic approach usually uses more direct quotation of the group discussion, while content analysis typically produces numerical description of the data (Morgan, 1988).

In this study, focus group data were analyzed using both approaches. Coding schemas were used, combined with the direct quotations from the discussion. Each question asked by the researcher represented one coding category so that responses to each question could be organized in an orderly fashion (Bogdan & Biklen, 2003). The categories were developed exclusively so that a code can only be placed under one category.

The group discussion was audio-taped and lasted about 30 minutes. The audio-taped material was transcribed verbatim by using the transcribing machine. Responses to all questions were open-ended. After reviewing the transcripts twice, the researcher assigned a code to each response, and listed the codes beside the text. For example, the first question explored teachers' perception of how children responded to a standardized test. Two codes (*not responded* and *response doesn't reflect what they know*) were assigned, with at least one code assigned to each participant's response. After the original codes were assigned to each response, transcripts were reviewed again and necessary modifications were made to make them consistent across the board. After all the codes were assigned and modified, a content analysis approach was employed. The researcher listed all the different codes occurring in one coding category and indicated the number of times this specific code had been assigned to all responses for the theme by putting tallies beside the code.

Each code was then examined to see if it fit into a different category other than the one under which it was originally assigned. For example, a code "unfair" was originally assigned under category *use of standardized test*, but after reviewing the transcripts carefully, the researcher believed it fit better under the question/category *disadvantages of standardized assessment*. After the researcher decided a specific code fit under a different category instead of the original one, a tally was placed besides the corresponding code in the new category. The process repeated itself until all responses were coded and placed under the appropriate coding category.

Finally, the list of codes and coding categories were analyzed and compared. This process allowed for larger theme to emerge from the data. For example, all codes, regardless which category they were under, related to the administration of assessment were grouped together as *administrative issues*, and all codes related to parents were grouped as *parent involvement*.

Inter-Rater Reliability

Inter-rater reliability check was achieved by using a code by code comparison method between the researcher and an outside coder. The dissertation co-chair served as the outside coder. After the data were transcribed each coder received a copy of transcript. Each coder then read the transcript independently and analyzed the transcribed data as previously described. Both coders listed codes and made notes on the margin of the transcripts so the comparison could be made. Code by code comparison was made in the form of discussion. The researcher first read the categories and codes under the category. The outside coder checked off the codes from her list. When the researcher missed codes that the outside coder had, discussion occurred to determine whether the codes should be added, and vice versa. For example, *tester familiarity* was listed as a code under *child response* category for the outside coder but not for the researcher. The researcher explained that she coded it as rigid setting under category *disadvantages of standardized test*. However, the outside coder considered it as a factor that impact child response. After carefully reviewing of the transcript again and the discussion, an agreement was reached that it should be listed as *tester familiarity* under *child response*. This process was repeated until 100% agreements were reached on each category and code.

CHAPTER FIVE

Results

Concurrent Validity

There were a total of thirty four English speaking children recruited in the study. These thirty four children were enrolled in five preschool classrooms during the 2005-2006 academic year. Two of the participating children were eliminated from the final analysis because their AEPS® data were not available. Therefore, a total of 32 were children included in the analyses. All children were from 3 to 5 years old. The average age for all 32 children was 57.81 months. The oldest participant was 64 months and the youngest participant was 40 months. Among the 32 children, 18 of them were female and 14 of them were male. Gender difference was examined using the independent sample T-test. The results indicated no significant difference between male and female in terms of their scores on either of the measures. Seven out of 32 of them were white, 19 were black, 5 were Hispanic and 1 was biracial. None of the 32 children had a disability. Table 5.1 presents the descriptive statistics for children's characteristics.

Table 5.2 presents the descriptive statistics for AEPS® cognitive and social-communication areas. The descriptive statistics of AEPS® included mean scores, standard deviations and the range of scores for each strand as well as the whole areas. Table 5.3 presents the descriptive statistics for BDI-2 cognitive and communication domains. The descriptive statistics of BDI-2 also included means scores, standard deviations and range of scores for each subdomain as well as whole domains.

The total score for the AEPS® cognitive area ranges from 0 to 108. In this study, the highest score was 106 and the lowest score was 47. The average score for the 32 participants was

87.53. The total score for the AEPS® social-communication area ranges from 0 to 98. In this study the highest score was 98 and the lowest score was 50. The average social-communication score for all 32 participants was 88.88.

Table 5.1

Descriptive Statistics for Children's Characteristics

Characteristics	N=32	Percent%
Gender		
Male	18	56.3
Female	14	43.7
Ethnicity		
White	7	21.9
Africa American	19	59.3
Hispanic	5	15.6
Biracial	1	3.1

The DQ scores for both BDI-2 cognitive and communication domains ranged from 40 to 160. In this sample, the highest cognitive DQ score was 111 and the lowest was 55. The average cognitive DQ score for all 32 children was 84.16. The highest communication DQ score was 119 and the lowest communication DQ score was 55. The average communication DQ score was 88.38.

Table 5.4 presents the Pearson's correlation coefficient between AEPS® social-communication score and BDI-2 communication scores. The result indicated that a positive correlation existed between the AEPS® social-communication area score and BDI communication domain score. The correlation was statistically significant. The correlation coefficient was .60 (p<.001).

Table 5.2

AEPS® Descriptive Statistics (N=32)

	Mean	SD	Minimum	Maximum
Cognitive Area	87.53	2.44	47	106
Concept Strand	16.53	4.30	4	20
Category Strand	7.75	12.35	4	8
Sequence Strand	11.19	4.32	7	12
Recall Strand	5.50	8.97	2	6
Problem-Solving Strand	12.03	2.44	4	14
Play Strand	12.81	4.30	7	14
Premath Strand	9.25	12.35	3	12
Phonological Awareness Strand	11.78	4.32	4	22
Social-Communication	88.88	8.97	50	6
Social-Communication Strand	32.69	2.44	21	6
Words, Phrases and Sentences Strand	55.38	4.30	29	6

Strands in AEPS® social-communication area also were positively correlated to BDI communication domain as well as its subdomains. The two strands in AEPS® social-communication area were a) social-communicative interaction, and b) word, phrases, and sentences. And the two subdomains in BDI communication domain were a) receptive communication and b) expressive communication. The results of Pearson's correlation indicated that both strands in AEPS® social-communication area were significantly correlated to BDI receptive communication and expressive communication. Scores of the AEPS®

Table 5.3

BDI-2 Descriptive Statistics (N=32)

	Mean	SD	Minimum	Maximum
Cognitive DQ	84.16	17.35	55	111
Attention and Memory	46.94	5.96	32	57
Reasoning and Academic	31.75	8.20	14	45
Perception and Concepts	43.22	12.43	20	67
Communication DQ	88.38	16.83	55	119
Receptive Communication	52.59	8.87	24	64
Perceptive Communication	57.19	11.95	22	75

Table 5.4

Pearson's Correlation Coefficient Between AEPS® Social-communication Area and BDI-2 Communication Domain

	BDI Communication		
AEPS® Social-Communication	Expressive	Receptive	**Communication**
Social-Communicative Interaction	.63 ***	.67 ***	.50
Words, Phrase, and Sentences	.64 ***	.72 ***	.58 ***
Social-Communication Area	.68 ***	.76 ***	.60 ***

Note. ***: significant at .001 level (2 tailed)

**: significant at .05 level (2 tailed)

*: significant at .01 level (2 tailed)

social-communicative interaction strand was significantly correlated to scores of the BDI receptive communication subdomain (r (32) = .67, p<.001) and the BDI expressive communication subdomain (r (32) =.63 p<.001). It meant that higher scores of the AEPS® social-communicative interaction strand were associated with higher scores of BDI-2 receptive communication strand and BDI expressive communication.

Scores of The AEPS® word, phrase, and sentences strand were significantly correlated to both BDI receptive communication (r (32) =.72, p<.001) and expressive communication (r (32) = .64, p<.001). Higher scores of the AEPS® word, phrase, and sentences strand were associated with higher scores of the BDI-2 receptive communication and expressive communication.

Table 5.5 shows the Pearson's correlation between AEPS® cognitive domain and BDI cognitive domain. The results indicate that a positive correlation existed between AEPS® cognitive score and BDI cognitive scores, and the correlation was statistically significant (r (32) = .57, p<.001). The results indicated that higher AEPS® cognitive score was associated with higher BDI cognitive score.

Eight strands in the AEPS® cognitive domain were correlated to the BDI cognitive domain as well as the three sub-domains. The eight AEPS® strands were concept, category, sequence, recall, problem-solving, play, premath, and phonological awareness and emergent reading. The three sub-domains in BDI cognitive domain were attention and memory, reasoning and academic, and perception and concepts. Among the eight AEPS® strands, seven of them (concept, category, sequence, recall, problem-solving, permath, and phonological awareness and emergent reading) were significantly correlated to all three sub-domains in BDI cognitive domain. The play strand was significantly correlated to one of the sub-domains (reasoning and academic), but not the other two. All but one strand (category) were significantly correlated to

Table 5.5

Person's Correlation Coefficient between AEPS® Cognitive and BDI-2 Cognitive Domain

	BDI Cognitive Domain			
AEPS® Cognitive Area	Attention and Memory	Reasoning and Academic	Perception and Concepts	Cognitive DQ
Concept	.39*	.62 ***	.52 **	.47 *
Category	.50 **	.36 *	.37 *	.15
Sequence	.56 **	.64 ***	.49 *	.37 *
Recall	.53 **	.65 ***	.51 **	.55 **
Problem-Solving	.62 ***	.70 ***	.47 *	.45 *
Play	.25	.37 *	.34	.40 *
Premath	.41 *	.49 *	.42 *	.37 *
Phonological Awareness and Emergent Reading	.52 **	.69 ***	.46 *	.53 **
Cognitive Domain	.64 ***	.78 ***	.61 ***	.57 **

Note. ***: significant at .001 level (2 tailed)

**: significant at .05 level (2 tailed)

*: significant at .01 level (2 tailed)

BDI-2 cognitive score.

The results from the correlational analyses indicated that the scores from the concept strand in AEPS® cognitive area were correlated to scores from the attention and memory (r (32) =.39, p<.05) subdomain, reasoning and academic subdomain(r (32) =.62, p<.001), and perception and concepts (r (32) =.52, p<.01) subdomain in BDI-2 cognitive domain. The higher AEPS® cognitive score was associated with higher attention and memory score, reasoning and academic score, and perception and concepts scores. The concept strand score was also correlated to BDI-2 cognitive score (r (32) = .47, p<.05), which means that higher concept scores on AEPS® was associated with higher BDI cognitive score.

Results also indicated that the scores of AEPS® category strand were correlated to scores on the BDI-2 attention and memory subdomain (r (32) =.50, p<.01), reasoning and academic subdomain (r (32) = .36, p<.05), and perception and concepts subdomain (r (32) =.37, p<.05). The higher AEPS® category score indicated higher attention and memory score, reasoning and academic score, and perception and concept score in BDI cognitive domain.

Scores of the AEPS® sequence strand were found to be correlated to scores from BDI-2 attention and memory subdomain (r (32) = .56, p<.01), reasoning and academic subdomain (r (32) = .64, p<.001), perception and concepts subdomain (r (32) = .49, p<.05), as well as BDI-2 cognitive domain scores (r (32) = .55, p<.01). Higher scores of sequence strand were associated with higher scores of attention and memory subdomain, reasoning and academic subdomain, perception and concepts subdomain, and BDI-2 cognitive domain.

Scores of the AEPS® strand were correlated to BDI-2 attention and memory subdomain (r (32) =.53, p<.01), reasoning and academic subdomain (r (32) = .65, p<.001), and perception and concepts subdomain (r (32) = .51, p<.01). They were also correlated to scores of BDI-2 cognitive domain (r (32) =.55 (p<.01). Higher AEPS® recall scores were associated with higher scores of the attention and memory subdomain, reasoning and academic subdomain, perception and concepts subdomain, as well as BDI cognitive domain.

The results also indicated correlations between scores of AEPS® problem solving strand and scores of the BDI-2 attention and memory subdomain (r (32) = .63, p<.001), reasoning and academic subdomain (r (32) =.70, p<.001), perception and concepts subdomain (r (32) = .47, p<.05), as well as cognitive domain (r (32) = .45, p<.05). Higher AEPS® problem solving scores were associated with higher scores of BDI-2 attention and memory subdomain, reasoning and academic subdomain, perception and concepts subdomain, and cognitive domain.

Scores of the AEPS® play strand were found to be correlated to scores of BDI-2 reasoning and academic subdomain (r (32) =.37, p<.05), and scores of BDI cognitive domain (r (32) =.40, p<0.5). Higher play scores were associated with higher scores of BDI-2 reasoning and academic subdomain as well as BDI-2 cognitive domain.

Scores of the AEPS® premath strand were correlated to scores of BDI-2 attention and memory subdomain (r (32) =.41, p<.05), reasoning and academic subdomain (r (32) = .49, p<.05), perception and concepts subdomain (r (32)= .42, p<.05), and cognitive domain (r (32)= .37, p<.05).

Last, scores of AEPS® phonological awareness strand and emergent reading strand were correlated to BDI-2 attention and memory subdomain (r (32) = .52, p<.01), reasoning and academic subdomain (r (32) =.69, p<.001), perception and concepts subdomain (r (32) =.46, p<.05). The scores from this strand were also correlated to scores of BDI-2 cognitive domain (r (32) = .53, p<.01).

Social Validity

After analyzing, comparing and discussing the transcribed data, both coders agreed that the four majors themes emerged from the focus group: 1) *administrative issues of assessment*; 2) *the use of assessment and its results*, 3) *parent involvement in the assessment*, and 4) *teacher preference*. In each of the four themes, different codes were assigned.

Administrative Issues

For standardized assessment, three categories were included under the theme of administrative issues: *child response*, *advantages of standardized test*, and *disadvantages of the standardized test*. Teachers commented that in standardized tests children either do not respond to the questions or their responses do not reflect what they know. Three teachers indicated that children do not respond to standardized tests, and two teachers indicated that children's response do not reflect what they know. Examples of some of the responses are reflected below:

> "…they don't want to do anything like they will just quit or confused so that really doesn't make sense."
>
> "…even though they may know the correct answer they may not respond."
>
> "I know that their perception did not reflect what they knew."

In terms of what impacted the child's response in standardized tests, one teacher mentioned that it was because children were aware of the fact that they are being tested. Two teachers mentioned that tester familiarity and test settings could impact children's responses. For example, one teacher said:

> "…it depends on who the person is to test the child and if it's that child is not familiar with that person then they may not respond to the question that you are asking them, even

> though they may know the correct answer they may not respond because they are with someone who was not familiar to them."

When discussing the advantages of standardized test, three factors were listed as advantages: 1) it was easier to administer; 2) it cost less time; and 3) it was fun for some children. When talking about the disadvantages of standardized test, three teachers indicated it was not fair for children. Two teachers indicated the rigid setting was a disadvantage for standardized test. Examples of responses are listed below:

> "Even though you feel like if you worded differently maybe that child will be able to answer or do that, with standardized you cannot."
>
> "…it was kind of unfair because some of my kids didn't even, they then looked at some of the pictures and they called it different name than what they are probably called."

For authentic assessment, two categories were included under the administrative theme: *child response* and *advantages and disadvantages of authentic assessment*. Two teachers indicated that authentic assessment elicited natural responses from children. All teachers indicated that compared to standardized tests children were less aware of the fact that they are being tested and that authentic assessment was easier for children because items in authentic assessment are embedded in the natural environment. Examples of some responses are as follow:

> "…because it was not set up the way that they are directly tested and then it can convey the information from them."
>
> "I think it's a little bit easier like I said before because they don't necessary know that they are being tested…"

Even though *time-consuming* was listed by all teachers among one of the disadvantages of authentic assessment, two teachers mentioned that it could be less time-consuming once teachers know their children. For example, one teacher stated:

> "I think that's not true and like I can think of a handful of my students that turn 5, like in 2005, like they've been missing the deadline for kindergarten, and by the second round I could have just gone through the AEPS® and ask them the questions, and they, you know, and that would have been easier and taking less time than trying to complete the activity protocol that we have for the AEPS®. Then, you know, I just think it would have been, it would be a lot easier if we can just ask them questions."

Use of the assessment results

For standardized test, teachers had different approaches regarding their uses of the test. Two teachers indicated *they did not use the test results* at all. One teacher used the assessment as a *screening tool* to inform the need of further assessment, and used items on standardized test for instructional purposes so that children will perform better on "*post-test*". Two teachers mentioned the test results made them *aware of children's needs*.

For authentic assessment, all teachers indicated that they use the results to *inform their classroom instruction*. Teachers *trusted the test results*. They also believe that skills reflected in the authentic assessment are *linked to daily life and curriculum*. As for why teachers trusted the results, one teacher indicated:

> "because I mean their, their responses are authentic. Umm, you know because it was not set up the way that they are directly tested and then it can convey the information from them"

Parent involvement

According to the teachers, some parents *don't care* about the standardized test because they don't understand it, while others are *curious* about how their children are doing based on the standardized scores. Meanwhile, with the authentic assessment, some parents need more information on the tool, some parents appreciate the fact that authentic assessment *monitors the progress*, some parents think *it's easier to read*, and others just think it is *really cool*.

Teacher preference

Teachers had different opinions when asked about their preferences on the tests. First of all, all of them indicated that their preference depended on the child or situation. The factors mentioned that influence their preferences were: 1) *child's characteristics*, 2) *how many times the assessment had to be conducted*, 3) *how many children to be assessed*, and 4) *how well the materials were prepared*. Meanwhile, two teachers indicated both tests had its benefits, and two of them indicated authentic assessment would be more beneficial.

CHAPTER SIX

Discussion, Implications, Limitation, and Future Research

Discussion of the Results

Results from this study supported the hypothesis that the AEPS® is a concurrently valid measure for reporting children's accountability data on language, literacy, and pre-math area and it is also perceived as a useful measure by teachers. Correlations were run between the AEPS® social-communication area and BDI-2 communication domain. According to the results, scores from the AEPS® social-communication area were significantly correlated to scores from BDI-2 communication domain. Based on Cohen (1988)'s definition of strength of correlation, it is moderately correlated (r=.60, p<.001). Compared to the validity coefficiency of .75 when the BDI-2 communication domain was correlated with BSID-II mental index, a validity coefficient of .60 is lower. However, the BDI-2 cognitive domain only reached the validity coefficient of .61 with the BSID-II mental index and was still considered valid, therefore, the correlation coefficient of .60 is still acceptable.

Upton further examinations of the strands, strands under AEPS® social-communication area were all significantly correlated to both receptive subdomain and expressive subdomain under the BDI-2 communication domain, with validity coefficients range from .63 to .72. When comparing these numbers to other validity studies (Newborg, 2004) of the BDI-2, it is confident to claim that scores of the AEPS® social-communication area were similar enough with the scores of the BDI-2 communication domain to claim its concurrent validiy. The possible explanations of the correlations are listed as follow:

- The BDI-2 receptive communication subdomain intended to measure a child's ability to understand information received through verbal or nonverbal communication (Newborg, 2004).
- The BDI-2 expressive communication subdomain intended to measure a child's production and use of sounds, words, or gesture to express information to others (Newborg, 2004). These skills are necessary for children to convey information correctly and develop grammatical understanding of words, phrase, and sentences.
- The AEPS® social-communicative strand intended to measure child's ability to convey information by using words, phrases, and sentences. In order to convey information correctly, one not only need to have the ability of producing sounds, words, and sentences, but also need to have the abilitiy of receiving information correctly. Therefore, scores of the AEPS® social-communicative strand were similar to scores of the BDI-2 receptive and expressive subdomain (Bricker, et al..).
- The AEPS® production of words, phrases, and sentences strand intended to examine a child's grammatical understanding of words, phrase, and sentences. Only if a child understands the communicative information he or she received, he or she could develop the correct grammatical understanding of words, phrase, and sentences. Therefore, scores of this strand were similar to scores of the BDI-2 receptive and expressive communication (Bricker, et al..).

Based on the above explanations, the correlations between the AEPS® social-communication area and the BDI-2 communication domain supported the idea that the AEPS® can be used as alternative measure to report children's communication scores.

Correlations also were run between the AEPS® cognitive area and the BDI-2 cognitive domain. According to the correlation anaylsis, scores from AEPS® cognitive area were significantly correlated to scores from BDI-2 cognitive domain. Based on Cohen (1988)'s definition of strength of correlation, they are moderately correlated. The validity coefficient between the AEPS® cognitive area and the BDI-2 cognitive domain is .57. Compared to the .61 validity coefficient of the BDI-2 when it was validated with the BSID-II mental index, .57 is an acceptable figure.

Upton further examinations of the AEPS® cognitive area, scores from most strands under that area were significantly correlated to scores from all three of the subdomains under the BDI-2 cognitive domain. Scores from seven out of eight strands under AEPS® cognitive area were significantly correlated to scores from the BDI-2 attention and memory subdomain. Five of them (category, sequence, recall, problem-solving, and phonological awareness and emergent reading) had moderate correlations with the attention and memory subdomain.

Also, scores from all strands under the AEPS® cognitive area were significantly correlated to BDI-2 reasoning and academic subdomain. Among these strands, five of them (concept, sequence, recall, problem-solving, and phonological awareness and emerging reading) had moderate correlations with the reasoning and academic subdomain.

In addition, scores from seven out of eight strands under the AEPS® cognitive area were significantly correlated to BDI-2 perception and concepts subdomain. However, most of the correlations were weak based on Cohen (1988)'s definition of strength of correlation. Only two strands (concept and recall) had moderate correlations with the BDI-2 perception and concepts subdomain. The possible explanations for the similiarities of scores from both measures are listed as follow:

- BDI-2 attention and memory subdomain intended to measure a child's ability to attend to environmental stimuli for certain length of time and to retrieve information (Newborg, 2004).
- BDI-2 reasoning and academic subdomain intended to measure children's critical thinking skills. These skills are necessary for reading, writing and mathematics (Newborg).
- BDI-2 perception and concepts subdomain intended to measure young children's interaction with the immediate environment. Most items in this subdomain were social in nature and focused on self-concept and interactions. That explains why most of the AEPS® cognitive strands only had weak correlation with this subdomain (Newborg; Bricker et al.., 2002).
- The AEPS® category strand intended to measure children's ability in grouping objects or people based on their characters. In order to group and compare, it requires children to attend to the physical or functional attributes and later retrieve the information on these attributes. It also requires the critical thinking skills. Therefore, the performance on AEPS® category strand reflected their ability to attend to environmental stimuli and retrieve information. That explains why scores from this strand were similar to scores from the BDI attention and memory subdomain (Bricker et al..).
- The AEPS® sequence and recall strands intended to measure children's ability to understand the sequences of verbal orders, objects, and stories, as well as to retrieve information later. In order to understand the concept of sequence and to recall events, children also need to attend to stimuli and retrieve information, as well as use their

critical thinking skills. Therefore, children's scores on the AEPS® sequence and recall strands were similar to their scores on all three subdomains under BDI-2 cognitive domain (Newborg; Bricker et al..).

- The AEPS® problem-solving strand intended to measure children's ability of evaluating the problem, understand cause and effect, and find solutions for the problem. In order to solve a problem, a child has to be able to attend to a task and retrieve previous information such as causal relations. Therefore, children's performance on AEPS® problem-solving strand were similar to their scores on the BDI-2 attention and memory subdomain (Bricker et al..).
- The AEPS® phonological awareness and emergent reading strand intended to measure children's ability to match sounds and letters. In order to demonstrate skills such as letter-sound association and rhyming skills, a child has to use his or her attention and memory ability, therefore, children's performance on AEPS® phonological awareness and emergent reading strand was also similar to their performance on BDI-2 attention and memory subdomain (Bricker et al..).

The correlations between scores from the AEPS® cognitive area and its strands and scores from the BDI-2 cognitive domain and its subdomains reflected the link between these two measures.

Based on the above results, the idea that the AEPS® can be used as an alternative measure to report children's cognitive score which is now measured by standardized tests like the BDI-2 is supported by the data.

Based on all the above findings, scores from the AEPS® cognitive and social-communication area as well as their strands are reflective of children's performance on the BDI-

2 cognitive and communication domains and their subdomains. Therefore, instead of conducting the standardized test, the AEPS® can be used as an alternative measure to report children's scores in cognitive and communication areas. Since the cognitive and communication areas included items that measure children's language, literacy and pre-math abilities, the AEPS® can be used as a valid measure to report children's accountability data on these areas.

Social validity

According to the results from the focus group, teachers indicated that authentic assessment such as the AEPS® truly reflected what a child knows regardless of his or her personality, and they all used it to gain information from children. Teachers also indicated that they did not use standardized tests much because they did not reflect a child's ability as accurate as the authentic assessment, especially when the child is not familiar with the test administrator and the test environment. This is consistent with the notion mentioned by other researchers that young children's behaviors in the testing situation may affect the accuracy of testing results (Nagle, 2000). However, some teachers did use items from the standardized tests to "teach the test". This finding is consistent with some of the concerns around standardized test that when using standardized test to measure outcomes for accountability, teachers are alternating their instructions and the curriculum is narrowed to a focus on skills that are on the tests (Kohn, 2001; Hess & Brigham, 2000; Shepard, Taylor, & Kagan, 1996; James & Tanner, 1993). Teachers did indicate the benefits standardized tests in terms of time. Using standardized test is efficient in a sense.

Among the advantages of authentic assessment, two factors were listed as the most influential reasons for teachers to use it: 1) authentic assessment elicits natural response from children, and 2) authentic assessment is easier for children. All teachers indicated that authentic assessment put less pressure on children because most of the time the child was not aware that he or she was being tested. Teachers' opinions on authentic assessment are consistent with the other literatures that pointed out preschool school children as different as their school-age counterparts. Literatures (Nagle, 2000) indicated that preschool children approach the test with a different motivational style than older children. Unlike older school age children, younger children tend not to place importance on answering questions correctly, persisting on test items, pleasing the examiners, and responding to social reinforcement. Younger children also have lower tolerance and higher level of frustration than older children (Nagle).

Even though teachers indicated that authentic assessment was naturally embedded and it reflected natural responses from children, when asking about their preferences, teachers did mention that both standardized test and authentic assessment have their benefits. Four factors were mentioned as the conditions or barriers which they have to consider in choosing authentic assessment: 1) child's character, 2) frequency of conducting assessment, 3) number of children, and 4) preparation of the materials. Time-consuming is one of the concerns that has negative impact on teacher's preference over authentic assessment. When the assessment has to be repeated three times a year it could add pressures to teachers. Also, when there are many children to assess, the time-consuming nature of authentic assessment may keep teachers away from it. However, these barriers could be removed by getting familiar with children. As two of the teachers mentioned, when they know their children better it does not cost much time to check off the items from the authentic assessment. Since the items in the authentic assessment are naturally embedded and linked to daily life skills, teachers have enough opportunities to see them on a daily base. Therefore, when it comes to the "assessment time", teachers should know them already if they know their kids well enough.

According to teachers' discussion, parents appreciated the fact that the authentic assessment shows the progress over the time. Because items on an authentic assessment are linked to the daily life, parents can read understand authentic assessment more easily. This finding is consistent with the researches indicating that when results from alternative assessment were used for reporting to parents, the performance indictors were detailed and concrete enough for parents to understand what curriculum expectations were being addressed (Shepard, Taylor, & Kagan, 1996).

Based on the perceptions of teachers, even though standardized tests have the benefits of being efficient and fun for some children, the authentic assessment is an appealing alternative for traditional standardized test for its authenticity and naturally embedded characters. It is also easier for children and parents. As long as its time-consuming issue is address by better preparing the materials and getting to know children better, it can be served as an efficient method to assess young children.

Implications

Implication for research

Different professional organizations such as the Division of Early Childhood (DEC) and the National Association for the Education of Young Children (NAEYC) (DEC, 2007; NAEYC and NAECS/SDE, 2003) have recommended desirable assessment practices. According to their statements, assessment should be 1) conducted in a naturalistic environment, 2) reflecting functional skills, 3) involving families, and 4) linked to the curriculum and individual goal development. In this particular study, the results indicated that the AEPS® fits all these described characteristics because: first, it was conducted in the classroom by teachers during the regular activities so that it ensured naturalistic environment; second, according to teachers' perceptions, the items in the AEPS® are functional items that assess children's real life skills instead of abstract forms of knowledge; third, the AEPS® also required some parent input while teachers scored children on the measure, and it was easier for parents to understand; fourth, the AEPS® was developed in the way that items can be directly linked with curriculum. Therefore, the AEPS®, as well as many other similar curriculum-based assessments, is consistent with the recommended assessment practice for its appropriate use in the classroom. However, when using for the accountability purpose, the reliability and validity of such assessment have been questioned (Harbin et al.., 2005; Neisworth & Bagnato, 2004; Stewart & Kaminski, 2002).

This study answered part of the psychometric questions raised from the researchers. The results indicated that scores from the AEPS® are reflective of scores generated from the traditionally used BDI-2. The results demonstrated the technical adequacy of the measure and convinced the researchers, educators and other customers that the AEPS® meets the traditional technical adequacy criteria to be used as an alternative measure to report children's accountability data in early language, literacy and pre-math areas.

Implications for practices

Also, the issue was addressed regarding teachers' ability to implement the measure. There have been debates about whether human judgment can be relied on to provide reliable data to be used for accountability purpose (Shavelson, Baxter, & Gao, 1993). For the purpose of this study, an AEPS® certified trainer conducted reliability check with each teacher and reached at least 80% agreement with all of them. Therefore, the reliability indicator tells the public that teachers are trustworthy in terms of using authentic assessment measure to provide accountability data. In addition, based on the teachers' perceptions on the AEPS®, using the

AEPS® is appealing for several reasons: First, the efficiency issue of the curriculum-based assessment can be addressed. Teacher indicated that the AEPS® cost less time than standardized tests when the teacher is familiar with children. One of the advantages of teacher perferring standardized test over authentic assessment is the efficiency issue. When it is as efficient to administer authentic assessment in classroom as using standardized test, authentic assessment is preferred. Second, using the AEPS® can obtain more natural and authentic responses from children than the traditional standardized test. Third, the results from the AEPS® can be directly linked with classroom instructions and individualized planning. Finally, the AEPS® is more family-friendly because parents found it easier to read and follow their children's progress. Based on the reasons stated above, teachers in this study felt more comfortable to use the AEPS® to record and report child outcome. It is preferable for teachers to use the measure that they are comfortable with. And since it is suggested that teachers can implement the measure reliably, the AEPS® can serve as a desirable alternative for assessing children and report their accountability data. Furthermore, the AEPS® Test has also been developed as a measure that also provides cut-off scores for determining eligibility for special service for infants and young children (Macy et al., 2005; Bricker, Yovanoff et al., 2003). When combining all these benefits together, the AEPS® Test serves all purposes for an assessment. It can be used as a tool for determining eligibility for special services; it can be used to provide information for classroom instruction; it can be used as a measure that records child's developmental progress; and it can provide the accountability data for programs. Now that the technical adequacy issue of the test has been examined, public's concerns on the psychometric properties of the measure were eliminated. And the social consequence of using the AEPS® Test has also been examined in this study.

Many states are now in the process of implementing curriculum-based assessment for accountability (Early Childhood Outcomes Center, 2007). One of the obstacles for fully implementing non-standardized measurement is the technical adequacy issue. Now that the evidences of technical adequacy of the AEPS® Test has been collected, it sheds light on other curriculum-based measurement. It convinces the public that non-standardized measurement can be as reliable and valid as standardized tests. Therefore, using curriculum-based assessment or authentic assessment for accountability is a feasible alternative.

Limitations and Future Research

This preliminary study has several limitations. First, although the sample size met the minimum requirement for demonstrating statistic significance (if there was any), the similar language, socioeconomic and disability status of the children enrolled in the study raised the question of whether the results from this study can be generalized to a larger and more diversified population. The results would be more convincing if future research on this topic can include larger sample size including children from different regions, language, and socio-economic backgrounds.

Second, besides the homogenous nature of the children enrolled in this study, all five classrooms were in the same public preschool. The lack of variety in program type limited the generalization of the study results as well. Future research should include different types of programs (e.g. private preschools and day care centers). Holding educational institution accountable should not be the responsibility only for state-funded public preschools but also for all educational programs. Using an assessment measure that can both accurately report child outcomes and provide information for curriculum development should be encouraged and disseminated to all types of schools. Research data on those programs which do not have the pressure of reporting child outcomes may give more insights on what difference it can make by switching traditionally used standardized test to the authentic, curriculum-based assessment when there is no accountability pressure. With the limitation of homogeneousness, this study did

Third, due to the limited resources, data were only collected on children's early-language, literacy, and pre-math abilities. That means, only part of the authentic assessment measure was examined for its validity. Even though these are the areas that received most attention (Strickland & Riley-Ayers, 2006; Harbin, Rous & McLean, 2005), there are still puzzle pieces that are missing. The AEPS® fine motor, gross motor, adaptive and social areas have not been examined closely for its technical adequacy in this study. Future research may include more areas of the test. Other criterion-referenced measures may be used so that it can include a broader range of criteria.

The fourth limitation was the small number of teachers that participated in the focus group. The five teachers participated in the focus group possessed same level of education, similar educational background, and similar length of years of teaching experiences. Future research should investigate perceptions of teachers from different backgrounds and with different experiences to portrait a diversified picture of people's opinions on the use of assessment.

Even with these limitations, findings from this study are still meaningful for research and practice for several reasons: 1) the sample size reached the minimum number for statistic power of .80 at the .05 significance level (Cohen, 1988). The correlations presented in this study demonstrated the relations "scientifically" between the AEPS® and BDI-2 regardless of the small sample size; 2) Children who participated in this study were almost equally divided by gender. As there was no significant difference found by gender differences, the results can be replicated on both male and female; 3) Scores in language, literacy and pre-math areas are critical indicators for school-readiness and school-readiness is the “major theme” for accountability requirement, therefore, any information on how to generate reliable and valid scores on school readiness is of critical use for both researchers and practitioners.

Conclusion

This study addressed the issue of technical adequacy for authentic assessment. This study also explored teachers’ perceptions on using different assessment measure to record child outcome data. To provide meaningful information on authentic assessment, this study examined the statistic correlations between the authentic assessment and a traditional standardized test. This study evidenced the close relations on scores generated from the authentic assessment and the traditional standardized test in the areas of language, literacy and pre-math. The close relations between the two measures solved the “validity mystery” of authentic assessment. This study also found the barriers for teachers to use the authentic assessment. To minimize these obstacles, several factors should be taken into consideration as the time comes to decide on assessment measure for young children: 1) assessment should be appropriate for children with different characters, 2) assessment can be used in a way that no extensive “extra time” involved for teachers, and 3) assessment should be able to conducted with materials that are used frequently in classroom daily activities.

With the technical adequacy questions being answered, and the practical concerns being addressed, authentic assessment is presented as a prospective alternative for documenting young children’s accountability data. With the accountability movement evolving further into the field of early childhood education, authentic assessment should be given equal opportunity as standardized test to serve as a tool for high-stake decision making.

APPENDICES

APPENDIX A

ASSESSMENT, EVALUATION, AND PROGRAMMING SYSTEM, 2ND EDITION (AEPS®)

COGNITIVE AND SOCIAL COMMUNICATION RECORDING FORMS

APPENDIX B

BATTELLE DEVELOPMENTAL INVENTORY, 2ND EDITION (BDI-2)

COGNITIVE AND COMMUNICATION SCORE SHEETS

APPENDICE C

FOCUS GROUP DISSCUSSION GUIDE

APPENDIX D

RESEARCH APPROVAL LETTER FROM FAYETTE COUNTY PUBLIC SCHOOLS

APPENDIX E

INFORMED CONSENT FORMS

References

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

Anderson, R.C., Hiebert, E.H., Scott, J.A., & Wilkinson, I.A.G. (1985). Becoming a nation of readers: The report of the Commission of reading. Washington, DC: National Institute of Education.

Anderson, S. (1998). The trouble with testing. *Young Children, 53(4),* 25-29.

Appl, D. J. (2000). Clarifying the preschool assessment process: traditional practices and alternative approaches. *Early Childhood Education Journal, 27(4),* 219-225.

Archbald, D.A., & Newmann, F.M. (1988). *Beyond standardized testing: Assessing authentic achievement in the secondary school.* Reston, VA: National Association of Secondary Principals.

Authentic Assessment (2004). Retrieved July 24, 2004, from http://www.tki.org.nz/r/gifted/handbook/related/authentic_e.php

Barnett, D. W., & Macmann, G. M. (1992). Early intervention and the assessment of developmental skills: Challenges and directions. *Topics in Early Childhood Special Education, 12*(1), 21-42.

Bailey, E., & Bricker, D. (1986). A psychometric study of a criterion-referenced assessment designed for infants and young children. Journal of the Division of Early Childhood, 10(2), 124-134.

Baroody, A.J. (1992). The development of preschoolers' counting skills and principles. In J. Bideau, C. Meljac, & J.P. Fischer (Eds.), Pathway to number (pp.99-126). Hillsdale, NJ: Erlbaum.

Beck, S.S., & Pierson, C.A. (1993). Performance assessment: the realities that will influence the rewards. Childhood Education, 70, 99-102.

Bell, S. H., & Barnett, D. W. (1999). Peer micronorms in the assessment of young children: Methodological review and examples. *Topics in Early Childhood Special Education, 19* (2), 112-123.

Bereiter, C., & Engelman, S. (1996). *Teaching disadvantaged children in preschool.* Englewood Cliffs, NJ: Prentice-Hall.

Bergen, D (1994). Authentic Performance Assessments. *Childhood Education 70, 99-102.*

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56 (1),* 137-172.

Belak, H. (1992). The need for a new science of assessment. In H. Belak, F.M. Newmann, E. Adams, D.A. Archbald, T. Burgess, J. Raven & T. A. Romberg (Eds.). *Toward a New Science of Educational Testing and Assessment (pp. 1-21).* Albany, NY: SUNY.

Biggar, H. (2005). NAEYC recommendations on screening and assessment of young English-Language learners. Young Children, 60(6), 44-47.

Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectual des anormaux. L'Année psychologique, 11, 191–336

Biondi, L. A. (2001). Authentic assessment strategies in fourth grade. (ERIC Document Reproduction Service No. ED460165).

Black, P. & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80(2),* 139-148.

Bobbitt, F. (1992). The elimination of waste in education. *The Elementary School Teacher, 12,* 259-271.

Brandt, R. (1992). On performance assessment: A conversation with Grant Wiggins. *Educational Leadership, 49(8),* 35-37.

Bredekamp, S., & Rosegrant, T. (1992). *Reaching potentials: appropriate curriculum and assessment for young children.* Washington, DC: NAEYC.

Brigance, N.A. (1985). Brigance Screens. Curriculum Associates, Inc. North Billerica, MA.

Bricker, D., Bailey, E., & Slentz, K. (1990). Reliability, validity, and utility of the Evaluation and Programming System: For infants and young children (EPS-I). *Journal of Early Intervention, 14*(2), 147-160.

Bricker, D., Pretti-Frontczak, K., *Assessment, Evaluation and Programming System for Children and Infants, Second Edition*. Baltimore, MD: P.H. Bookers.

Bricker, D., Yovanoff, P., Capt, B., & Allen, D. (2003). Use of a curriculum-based measure to corroborate eligibility decisions. Journal of Early Intervention, 26, 20-30.

Brown, D. F. (1993). The political influence of state testing reform through the eyes of principals and teachers. (Report No. EA-025-190). Atlanta, GA: Conference Paper. (ERIC Document Reproduction Service No. ED360737).

Bruder, M.B. (1994). Working with members of other disciplines: Collaboration for success. In M. Wolery & J.S. Wilbers (Eds.), *Including children with special needs in early childhood programs* (pp. 45-70). Washington, DC: National Association for the Education of Young Children.

Bufkin, L. J., & Bryde, S. M. (1996). Young children at their best: Linking play to assessment and intervention. *Teaching Exceptional Children, 29* (2), 50-53.

Clements, D.H., Swaminathan, S., Hannibal, M.A.Z., & Sarama, J. (1999). Young children's concept of shape. *Journal for Research in Mathematics Education, 30,* 192-212.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Craig, C. L., & McCormick, E. P. (2002). Improving student learning through authentic assessment. (ERIC Document Reproduction Service No. ED468053)

Cresswell, J. W. (1998). *Qualitative inquiry and research design: choosing among five traditions*. Thousand Oaks, CA: SAGE.

Cronbach, L.J. (1988). Five perspective on validity argument. In H.Wainer (Ed.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Crawford, J. (2005, May/June). Test Driven. *NABE News, 28,* 1.

Daniels, V. I. (1999). The assessment maze: Making instructional decisions about alternative assessment for students with disabilities. *Preventing school failure, 43*(4), 171-178.

Division of Early Childhood (DEC). (2005). *Division for Early Childhood companion to the NAEYC and NAECS/SDE Early Childhood Curriculum, Assessment, and Program Evaluation: Building an effective, accountable system in programs for children birth through age 8.*

Dorr-Bremme, D., & Herman, J. (1983). *Assessing student achievement: a profile of classroom practices.* Los Angeles, UCLA: Center for the Study of Evaluation.

Drawing Value (1992). Retrieved July 29, 2004, from http://www.ssta.sk.ca/research/evaluation_and_reporting/92-09.htm

DuBose, R.F. (1979). Working with sensorily impaired children. In S.G. Garwood (Ed.), *Educating young handicapped children*. Rockville, MD: Aspen Systems Corp.

DuBose, R.F. (1981). Assessment of severe impaired young children: Problems and recommendations. *Topics in Early Childhood Special Education,* 1(2), 9-22.

Dunst, C.J., & Rheingrover, R. (1981). An analysis of the efficacy of infant intervention programs with organically handicapped children. *Evaluation & Program Planning, 4*, 287-323.

Early Childhood Outcome Center (ECO) (2007). *National Research Council Report on Developmental Outcomes and Assessments for Children Aged Zero to Five.*

Elliott, J., Ysseldyke, J., Thurlow, M., & Erickson, R. (1998). What about assessment and accountability? *Teaching Exceptional Children, 30(1),* 20-27.

Farr, R., & Greene, B. (1993). Improving reading assessment: Understanding the social and political agenda for testing. *Educational Horizons(72),* 20-27.

Fuchs, L. S., & Fuchs, D. (1999). Fair and unfair testing accommodations. *The School Administrator, 56(10),* 24-29.

Gagne, R. M. (1965). *The conditions of learning.* New York: Rinehard & Winston.

Geocaris, C. & Ross, M. (1999). A test worth taking. Educational Leadership, 57 (1), 29-33.

Greenspan, S. I., Meiseld, S.J., & the Zero to Three Work Group on Developmental Assessment. (1996). Toward a new vision for the developmental assessment of infants and young children. In S.J. Meisels & E.Fenichel (Eds.), *New visions for the developmental assessment of infants and young children* (pp. 11-26). Washington, DC: Zero to Three: National Center for Infants, Toddlers, and Families.

Goodwin, W. L. & Goodwin, L. D. (1982). Young children and measurement: standardized and nonstandardized instruments in early childhood education. *Handbook of Research in Early Childhood Education, 28,* 441-456.

Grisham-Brown, J. L. (2000). Transdisciplinary activity-based assessment for young children with multiple disabilities: A program planning approach. *Young Exceptional Children, 3*, 3-10.

Grisham-Brown, J., Hallam, R., & Brookeshire, R. (2006). Using authentic assessment to evidence children's progress toward early learning standards. *Early Childhood Education Journal, 34* (1), 45-51.

Grisham Brown, J. L., Hemmeter, M. L., & Pretti-Frontczak, K. L. (2005). *Blended practices for teaching young children in inclusive settings*. Baltimore, MD: Paul Brookes Publishing Company.

Government Accountability Office (2005, May). Further development could allow results of new test to be used for decision making. Retrieved March 12, 2008, from http://www.gao.gov/mew.items/d05343.pdf

Gilbert, J. C. (1990). Performance-based assessment resource guide. Denver, CO: Colorado Department of Education (ERIC Document Reproduction Service No. ED327304).

Hatch, J. A. (2002). Accountability shovedown: resisting the standards movement in early childhood education. *Phi Delta Kappan, 83(6)*, 457-462.

Hatch, J. A. & Grieshaber, S. (2002). Child observation in Australia and the USA: A cross-national analysis. *Early Child Development and Care, 169*, 39-56.

Hallam, R., Grisham-Brown, J. L., Gao, X., & Brookshire, R. (2007). The effects of outcomes-driven authentic assessment on classroom quality. *Early Childhood Research and Practice, 9*(2), 1-9.

Hamilton, D. A. (1995). The utility of the assessment evaluation programming system in the development of quality IEP goals and objectives for young children, birth to three, with visual impairments. Unpublished doctoral dissertation, University of Oregon, Eugene, Oregon.

Harbin, G. (1977). Educational assessment. In L. Cross & K. Goin (Eds.), *Identifying handicapped children: A guide to casefinding, screening, diagnosis, assessment, and evaluation*. New York: Walker.

Harbin, G., Rous, B., & McLean, M. (2005). Issues in designing state accountability systems. *Journal of Early Intervention, 27 (3),* 137-164.

Herman, J., & Golan, S. (1991). Effects of standardized testing on teachers and learning-another look. *CSSE Technical Report # 334*, Los Angeles: Center for the Study of Evaluation.

Herman, J. L. (1992). What research tells us about assessment. *Educational Leadership, 49(8).*

Hsia, T. (1993). Evaluating the psychometric properties of the Assessment, Evaluation, and Programming System for Three to Six Years: AEPS Test. Unpublished doctoral dissertation, University of Oregon, Eugene, Oregon.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory.* New York: Appleton-Century.

Hynd, G. W., & Semrud-Clikeman, M. (1993). Assessment of learning and cognitive dysfunction in young children. In D.J. Willis, & J.L. Culbertson (Eds.), *Testing young children: A reference guide for developmental, psychoeducational, and psychosocial assessments* (pp.167-191). Austin, TX: PRO-ED.

James, J.C., & Tanner, C.K. (1993). Standardized testing of young children. *Journal of Research and Development in Education, 26(3),* 143-152.

Janesick, V.J. (2001). *The Assessment Debate.* Santa Barbara, CA: ABC-CLIO.

Johnson-Martin, N. (1985). Sources of difficulty. In J. Danaher (Ed.), *Assessment of child progress* (TADS Monograph No.2). Chapel Hill, NC: Technical Assistance Development System.

Kamii, C. & Kamii, M. (1990). Why achievement testing should stop. In C. Kamii (Ed.), *Achievement Testing In the Early Grades: The Games Grown-Ups Play* (pp.15-38). Washington, DC: National Association for the Education of Young Children.

Kaufman, A.S., & Kaufman, N.L. (1983). *Interpretive manual for the Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Kern, T. T. (2007). *Program availability and quality of child care in center-based programs for young children with disabilities in Kentucky: an exploration of conditions and parental perceptions.* Unpublished doctoral dissertation, University of Kentucky, Lexington.

Kim, Y. (1997). Activity-Based Assessment: A functional approach to determining the eligibility of young children for special education services. Unpublished doctoral dissertation, University of Oregon, Eugene, Oregon.

Kellagham, T. & Madaus, G. (1991). National Testing: lessons for American from Europe. *Educational Leadership, 49(3),* 87-93.

Klein, A. S. & Estes, J. S. (2004). Using observation for performance assessment. *Early Childhood News, 23,* 32-39.

Kohn, A. (2001). Fighting the tests: a practical guide to rescuing our schools. *Phi Delta Kappan, 82(5),* 348-357.

Kornhaber, M. L. (2004). Appropriate and inappropriate forms of testing, assessment, and accountability. *Educational Policy, 18(1)*, 45-70.

Langer, J.A. (1984). *Literacy instruction in America's schools.* New York: Crown.

Letendre, L. (2001) An emerging paradigm of testing. In L. Letendre (Ed.), *Assessment: issues and challenges for the millennium* (pp.29-40). ERIC Document Reproduction Service No. ED457426.

Macy, M. G., Bricker, D. D., & Squires, J. K. (2005). Validity and reliability of a curriculum-based assessment approach to determine eligibility for part C services. *Journal of Early Intervention, 28*, 1-16.

McCarthy, D. (1972). *McCarthy Scales of Children's Ability.* San Antonio, TX: Psychological Corporation.

Mahoney, M. J. (1995). *Constructivism in psychotherapy.* Washington, DC: American Psychological Association.

Manset-Williamson, G., John, E. Hu, S., & Gordon, D. (2002). Early literacy practices as predictors of reading related outcomes: Test scores, test passing rates, retention, and special education referral. *Exceptionality, 10(1)*, 11-28.

McAfee, A., Leong, D. J., & Bodrova, E. (2004). *Basics of assessment. A primer for early childhood education.* Washington, DC: National Association for the Education of Young Children.

Meadow, S., & Karr-Kidwell, P. J. (2001). The role of standardized tests as a means of assessment of young children: a review of related literature and recommendations of alternative assessment for administrators and teachers (ERIC Document Reproduction Service No. ED456134).

Meisel, S.J. (1987). Uses and Abuses of Developmental Screening and School Readiness Tests. *Young Children, 42(2),* 4-6, 68-73.

Meisels, S.J., Steele, D.M., & Quinn-Leering, K. (1993). Testing, tracking, and retaining young children: An analysis of research and social policy. In B.Spodek (Ed.), *Handbook of research on the education of young children* (pp.279-292). New York: Macmillan.

Meisels, S. J., Xue, Y., Bickel, D.D., Nicholson, J. Atkins-Burnett, S. (2001). Parental reactions to authentic performance assessment. *Educational Assessment, 7(1),* 61-85.

Meisels, S. J., & Fenichel, E. (Eds.). (1996). *New visions for the developmental assessment of infants and young children.* Washington, DC: Zero to Three: National Center for Infants, Toddlers, and Families.

Miller, P. H. (1992). *Theories of developmental psychology* (3rd ed.). New York, NY: W H Freeman/Times Books/ Henry Holt & Co.

Mitchell, R. (1995). *The promise of performance assessment: how to use backlash constructively.* Paper presented at AERA annual conference. San Francisco, CA.

Moorcroft, T. A., Desmarais, K. H., Hogan, K., & Berkowitz, A. R. (2000). Authentic assessment in the informal setting: how it can work for you. *Journal of Environmental Education,* 31(3), 20-24.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62(3),* 229-258.

Nagle, R. J. (2000). Issues in preschool assessment. In B.A. Bracken (Eds.), The *Psychoeducational assessment of preschool children* (pp. 19-32). Needham Heights, MA: Pearson Education.

Nagle, R. J. & Paget, K. D. (1986). A conceptual model of preschool assessment. *School Psychology Review, 15(2),* 154-165.

National Association for the Education of Young Children (NAEYC) & National Association of Early Childhood Specialist in State Departments of Education (NAECS/SDE). 1991. Position Statement. Guidelines for appropriate curriculum content and assessment in programs serving children ages 3 through 8. *Young Children 46 (3)*, 21-38.

National Association for the Education of Young Children (NAEYC) & National Association of Early Childhood Specialist in State Departments of Education (NAECS/SDE). 2003. Position Statement. Guidelines for appropriate curriculum content and assessment in programs serving children ages 3 through 8. *Young Children*.

National Research Counsil (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Neisworth, J. T. & Bagnato, S. J. (2004). The mismeasure of young children: The authentic assessment alternative. *Infants and Young Children, 17*(3), 198-212.

Newborg, J., Stock, J., Wnek, L., Guidubaldi, J., & Svinicki, J. (1984). *Battelle Developmental Inventory.* Dallas: DLM/Teacher Resources**.**

Newborg, J. (2005) *Battelle Development Inventory (2nd Edition).* Rolling Meadow, IL: Riverside Publishing

Newcombe, N.S., & Huttenlocher, J. (2000). *Making space: The development of spatial representation and reasoning.* Cambridge, MA: MIT Press.

Newmann, F. & Archbald, D.A. (1992). The nature of authentic academic achievement. In H. Belak, F.M. Newmann, E. Adams, D.A. Archbald, T. Burgess, J. Raven & T. A. Romberg (Eds.). *Toward a New Science of Educational Testing and Assessment (pp. 71-83).* Albany, NY: SUNY.

Newmann, F., Mark, H., & Gamoran, A. (1996). Authentic pedagogy and student performance. *American Journal of Education, 104(4),* 280-312.

Noh, J. (2005). Examining the psychometric properties of the second edition of the Assessment, Evaluation, and Programming System for Three to Six Years: AEPS Test 2nd Edition (3-6), Unpublished doctoral dissertation, Eugene: University of Oregon.

Notari, A., & Bricker, D. (1990). The utility of a curriculum-based assessment instrument in the development of individualized education plans for infants and young children. *Journal of Early Intervention, 14*(2), 117-132.

Nunnally, J. C. (1978). Psychometric Theory. New York: McGraw-Hill.

Parette, H.P. Jr, Bryde, S. Hoge, D.R., & Hogan, A. (1995). Pragmatic issues regarding arena assessment in early intervention. *Infant and Toddler Intervention, 5*(3), 243-253.

Perrone, V. (1990). How did we get here? In C. Kamii (Ed.), *Achievement Testing in The Early Grades: The Games Grown-Ups Play* (pp.1-14). Washington, DC: National Association for the Education of Young Children.

Piaget, J. (1954). *The Construction of Reality in the Child, translated by Cook M*. New York: Basic Books.

Piaget, J. (1954). *The Child's Conception of Number.* London: Routledge & Kegan Paul Ltd.

Piaget, J. (1978). *The development of Thought: Equilibration of Cognitive Structures, translated by Rosin A.* Oxford: England: Viking.

Pierson, C. A., & Beck, S. S. (1993). Performance assessment: the realities that will influence the rewards. *Childhood Education, 80(1),* 29-32.

Pintner, R. (1923). *Intelligence Testing.* Oxford: Holt.

Popham, W.J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56(6),* 8-15.

Pretti-Frontczak, P., & Bricker, D. (2000). Enhancing the quality of Individualized Education Plan (IEP) goals and objectives. Journal of Early Intervention, 23(2), 92-105.

Purves, A.C. (1984). The potential and real achievement of U.S. students in school reading. *British Journal of Education, 93,* 82-106.

Ratcliff, N.J. (1995). The need for alternative techniques for assessing young children's emerging literacy skills, *Contemporary Education, 66(3),* 169-171.

Ratcliff, N.J. (2002). Using authentic assessment to document the emerging literacy skills of young children. *Childhood Education, 78(2),* 66-69.

Roe, M., & Vukelich, C. (1994). Portfolio implementation: What about R for realistic? *Journal of Research in Childhood Education, 9 (1),* 5-14.

Ross, M., & Ceocaris, C. (1999). A test worth taking. *Educational Leadership, 57(1),* 29-33.

Salvia, J., & Ysseldyke, J.E. (1995). *Assessment, Sixth Edition.* Boston, MA: Houghton Mifflin.

Schweinhart, L. J. (1993). Observing young children in action: the key to early childhood assessment. *Young Children* (July 2003).

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessment. *Journal of Educational Measurement, 30(3),* 215-232.

Shepard, L. (1991). Will national tests improve student learning? *Phi Delta Kappan, 73(3),* 232-238.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher, 29 (7),* 4-14.

Shepard, L., Kagan, S. L., Lynn, S., & Wurtz (1998). *Principle and recommendations for early childhood assessments.* Washington, DC: National Education Goals Panel.

Shepard, L., Tylor, G. A., & Kagan, S. L. (1996). *Trends in early childhood assessment policies and practices.* Washington, DC: National Education Goals Panel.

Skinner, B. F. (1954). The science of learning and the art of teaching, *Harvard Educational Review, 24,* 86-97.

Slentz, K. (1986). Evaluating the instructional needs of young children with handicaps: Psychometric adequacy of the Evaluation and Programming System-Assessment Level II. Dissertation Abstracts International, 47(11), 4072A.

Smith, L. & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10(4),* 7-11.

Smith, S. S. (1999). Reforming the kindergarten round-up. *Educational Leadership, 56,* 39-44.

Starkey, P., & Cooper, R.G. (1995). The development of subitizing in young children. *British Journal of Developmental Psychology,* 13, 399-420.

Stiggins, R. J. & Bridgeford, N. J. (1986). The ecology of classroom assessment. *Journal of Educational Measurement, 22(4),* 271-286.

Suen, H.K., Lu, C.H., Neiworth, J. & Bagnato, S. (1993). Measurement of team decision-making through generalizability theory. *Journal of psychoeductaional assessment, 11(2),* 120-132.

The Gesell Institute (1985). *Gesell Developmental Observation.* New Haven, CT: The Gesell Institute.

Thompson, S. (2001). The authentic standards movement and its evil twin. *Phi Delta Kappan, 82(5),* 358-362.

Thorndike, E. L. (1921). Intelligence and its measurement. *Journal of Educational Research, 12,* 124-127.

Tudge, J.R.H., & Doucet, F. (2004). Early mathematical experiences: observing young Black and White children's everyday activities. *Early Childhood Research Quarterly, 19,* 21-39.

Vacc, N. A., & Ritter, S. H. (1995). Assessment of preschool children. (ERIC Document Reproduction Service No. ED389964)

Valencia, S.W. (1997). Authentic classroom assessment of early reading: Alternatives to standardized tests. *Preventing School Failure, 41(2),* 63-70.

Vygotsky, L.S. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on language production and socioeconomic factors. *Child Development, 65,* 606-621.

Wesley, P. W., Buysse, V., Tyndall, S. (1997). Family and professional perspectives on early intervention: *An exploration using focus groups. Topics in Early Childhood Special Education, 17,* 435-457.

Wesson, K.A. (2001). The "Volvo effect"- questioning standardized tests. *Young Children, 56(2),* 16-18.

Wiersma, W., & Jurs, S.G. (1985). *Educational measurement and testing.* Boston: Allyn & Bacon.

Whitehurst, G.J. & Lonigan, C.J. (1998). Child development and emergent literacy. *Child Development,* 68, 848-872.

Whitehurst, G.J. & Lonigan, C.J. (2001) Emergent literacy: development from prereaders to readers. In Neuman, S.B. & Dickinson, D.K. (Ed.), *Handbook of early literacy research.* (pp.11). New York, NY: Guilford.

Wiggins, G. (1989). Teaching to the authentic test. *Educational Leadership, 46(7),* 41-47.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49,* 26-33.

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavioral Analysis, 11(2),* 203-214.

Woodcock, R.W., & Johnson, M.B. (1977). *Woodcock-Johnson Test of Achievement.* Rolling Meadows, IL: Riverside Publishing.

Wortham, S.C. (1996). *Measurement and evaluation in early childhood education.* New Jersey: Englewood Cliffs.

Wortham, S.C. (2008). *Assessment in early childhood education, fifth Edition.* New Jersey: Pearson Merrill Prentice Hall.

Xue, Y., Meisels, S.J., Bickel, D.D., & Atkins, B.S. (2000). An Analysis of Parents' Attitudes towards Authentic Performance Assessment. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA.

Zatta, M. & Pullin, D. (2004). Education and alternative assessment for students with significant cognitive disabilities: implications for educators. *Educational Policy Analysis Archives, 12(16).*