

The Use of Authentic Assessment to Report Accountability Data on Young Children's Language, Literacy and Pre-math Competency

Xin Gao

University of Kentucky

E-mail: kitty_gao@hotmail.com

Jennifer Grisham-Brown

University of Kentucky

E-mail: jgleat@uky.edu

Received: February 11, 2011 Accepted: February 28, 2011 doi:10.5539/ies.v4n2p41

Abstract

This validity study examined the validity of Assessment, Evaluation, and Programming System, 2nd Edition (AEPS®), a curriculum-based, authentic assessment for infants and young children. The primary purposes were to: a) examine whether the AEPS® is a concurrently valid tool for measuring young children's language, literacy and pre-math skills for accountability purpose and b) explore teachers' perceptions on using authentic assessment and standardized tests. This was accomplished through implementing both quantitative and qualitative methods. Findings from the study indicated (a) the AEPS® is a concurrently valid (b) there were both advantages and disadvantages of using authentic assessment such as the AEPS® and using standardized tests based on teachers' perceptions, however, the practical issues of using the authentic measure can be addressed by providing in-depth trainings to teachers and increasing teachers' familiarity with their children; and (c) families preferred authentic assessment such as the AEPS® because it is easier.

Keywords: Authentic assessment, Concurrent validity, Social validity

1. Introduction

Since the beginning of the 21st century, the public's push for using standardized tests to hold k-12 schools accountable has become a noticeable phenomenon (Dorn, 1998). Test-based accountability models have received widespread support from the business sector (Kornhaber, 2004) in k-12 programs. The No Child Left Behind legislation signed by President Bush in 2002 requires all students from grades 3 to 8 to take annual tests on literacy and mathematics in order to hold students, teachers, and schools accountable, which also leads to some punishment for students, teachers, and schools who fail to demonstrate the adequate yearly progress as shown by scores on standardized tests. Even though the formal testing is not required for younger children, the atmosphere also has influenced the practice in early childhood education. Some states required programs for young children to take accountability testing, others tried to link the testing in kindergarten to the performance of state-funded pre-K program in the previous year (Meisels, 2006). Studies have reported the frequent use of "testing" for young children (Meisels, 2006; Harbin, Rous & McLean, 2004; Hatch & Grieshaber, 2002; Hatch, 2002). One of the famous accountability tests in early childhood was the National Reporting System [NRS] implemented in Head Start programs in 2004. Just like the NRS, most child assessment tools for the accountability purpose are standardized tests.

Standardized tests are commercially published tests that contain a set of items and have uniform procedures for administration and scoring (Anderson, Hiebert, Scott & Wilkinson, 1985; Popham, 1999). These tests provide a quick and simple answer to the question that concerns policy makers most: are children learning with the money invested in their programs? If public funded programs are able to show improvement in test scores, the public will be satisfied with the benefits of their investment.

However, many claim disadvantages of using standardized test for accountability purposes, especially for those that are high-stakes. Experts have criticized norm-referenced tools as inappropriate for use with children from different ethnic, racial and socio-economics backgrounds (Meisels & Atkins-Burnet, 2000). Standardized tests are usually normative and summative in nature and do not have strong curricular links (Militello, Schweid, & Sireci, 2010).

Because young children have rapid and uneven development, they are developmentally unreliable test takers. Researchers often concerned about the reliability and validity of standardized tests when high-stakes decisions were made based on performance on such test (Appl, 2000; Brown, Kohn, 2001; Linn, 2000, McLean, Meisels, 2006; 1996; Nagle, 2000; Ratcliff, 1995). Also, in terms of the connection between pre-k readiness scores and child's later academic performance, studies found out standardized test possess inadequate predictive value for later school performance (Laparo & Pianta, 2000; Miseils, 2009; Kim & Suen, 2003). The concerns made sense because if the measure itself is flawed, how could we use the flawed judgment to deprive children of their chance to the appropriate education and intervention, and to punish the agencies that actually made effort to improve their services. As a result of such concern, the technical adequacy aspect of the NRS was examined, and the results indicated that it was not valid and reliable overtime (Government Accountability Office [GAO], 2005). The use of the NRS was then suspended.

In addition to the technical adequacy concerns, the social consequence of using standardized tests as high-stakes accountability tool is also problematic. Standardized assessments impact children's self-esteem because they may cause unnecessary stress and unrealistic expectations for children, especially for young children (Andersen, 1998; Hatch, 2002; Meisels, 2006). When using standardized tests to report accountability data, pressures force 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 the real world (Hatch, 2002; Mintrop, 2004). This practice narrowed the curriculum by overemphasizing basic skills and neglecting higher-order thinking (Hatch, 2002; Shepard et al, 2001; Kellaghan and Madaus, 2001; Shepard, 1991). Furthermore, such narrowing is likely to be greatest in programs that serve at-risk and disadvantaged children where there is the most pressure to improve outcomes (Shepard et al., 2001; Herman and Golan, 1991). The narrowing of curriculum distorts the real goal for the education (Parini, 2005).

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 thus called for a stop to the use of standardized tests (Shepard, Tylor & Kagan, 1996; NAEYC, 2009; NAECS/SDE, 2003; NASP, 2002) and suggest alternatives to evaluate and observe young children as they are learning. Recommended practices in early childhood field have suggested the use of authentic assessment that is aligned with developmentally appropriate assessment principles (NAEYC, 2009; Neisworth & Bagnato, 2004, 2005).

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), 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 for accountability purpose.

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 AEPS[®] is considered as an authentic assessment because 1) data collection occurred in children's familiar environment; 2) data is collected through ongoing observation of children; 3) it assesses child's functional skills that are closely related to their real life experiences; and 4) information is gathered from people who are familiar with the child. The AEPS[®] included 54 items divided into 8 strands in the cognitive area and 49 items divided into 2 strands in the social-communication area. 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. Teachers who work with the child every day score these items in an ongoing and informal format. A combination of data collection method such as observation of daily activities, collection of child's work samples, and parent interview are used to score all the items.

Even though authentic assessment is advocated for its relevance to young children's learning processes and fits the characteristics for developmentally appropriate practice, reliability and validity of this type of assessment remains unclear. As mentioned above, not until the reliability and validity evidences have been collected, tests may not be

appropriate for reporting and accountability purposes. For this purpose, there is no published data on the validity of the AEPS®. It is critical to provide such evidence before it can be used for accountability purpose. In this study, the AEPS® was validated by comparing its scores with a criterion measure for the criterion validity, and teachers' perceptions of the authentic assessment were explored to provide evidence of the social validity.

Three research questions were examined in this study: 1) Is the AEPS® a concurrently valid measure for assessing young children's competence in the cognitive domain for accountability purposes? 2) Is the AEPS® a concurrently valid measure for assessing young children's competence in the communication domain for accountability purposes, 3) What are teachers' perceptions on using authentic assessment and standardized tests?

2. Methods

2.1 Concurrent Validity

2.1.1 Recruitment and participants

Children enrolled in five public pre-k classrooms in a southeast state were recruited as participants. Recruitment lasted one month. Preschool teachers distributed parental consents to the parents of children enrolled in their classrooms. Parents signed and returned the informed consent form to the teacher if they are interested. After collecting the signed informed consents, the researcher started to conduct Battelle Developmental Inventory, 2nd Edition (referred as BDI-2) on children with signed consent.

There were a total of thirty four (34) English speaking children recruited for the study. Two of the participating children were eliminated from the final analysis because their AEPS® data were not available. Therefore, a total of 32 children were included in the analyses. The average age for the 32 children was 58 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 the 32 of them were white, 19 were black, 5 were Hispanic and 1 was biracial. None of the 32 children had a disability. Based on the power analysis formula (Cohen, 1988), in order to predict a statistically significant correlation at the 0.5 level, the sample size had to be at least 28 to reach a conventional 0.8 statistical power. For this purpose, our sample size of 32 meets the criteria.

2.1.2 Measures

The AEPS® (Bricker, 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 levels designed for children of different ages, one for children from birth-to-three years old (level I), and another for children from three to six years old (level II). In this study, only level II was used because all the participants were older than 3. The AEPS® assesses young children's competency in fine motor, gross motor, adaptive, cognitive, social-communication and social areas based on their performance in everyday activities. For the purpose of this study, only the cognitive and social-communication areas of the AEPS® three to six set were used because these are the two domains that reflect the learning in early language, literacy and pre-math areas which policy makers concerns the most.

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.

The Battelle Developmental Inventory, 2nd Edition (BDI-2) is a norm-referenced measure designed for the purpose of screening, diagnosis and evaluation of young children's early development. It 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 was developed based on the concept of milestones. The conceptual foundation of the original BDI was that a child attains 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 numbers of items from different instruments and clustered items together based on the behaviors of these item measure. From these clusters, a sequence of behaviors was derived to describe the functioning of typically developing children at various stages of development. 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 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.

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.

2.1.3 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. 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.

2.1.4 Analyse

Children's performance in the cognitive and communication areas was recorded and analyzed. The cognitive areas of both tests describe children's understanding of numbers, letters, consequences, logical relationships, spatial relationships and print, which reflects children's learning in early literacy and pre-math areas. The communication areas of both tests describe children's ability to use verbal or non-verbal language to communicate with his or her environment, which reflects children's learning in early language area.

SPSS was used for data analysis. Raw scores were used for analyses for all the AEPS® areas and strands. In order to be consistent with other technical adequacy studies on the BDI-2, the developmental quotient (referred as DQ in the following text) scores were used for its cognitive and communication domains (Newborg, 2004). When the DQ score was not available for the subdomains, raw scores were substituted. 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 performed between the AEPS® score and the BDI-2 scores. Scores on the AEPS® cognitive area were correlated to scores on the BDI-2 cognitive domain, and scores on the AEPS® social-communication domain were correlated to scores on the BDI communication domain. Before correlating scores generated from both tests, some preliminary data analyses were conducted. Descriptive analyses were performed on both AEPS® and BDI-2 domain and subdomain scores. 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. Correlation analyses were performed 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, relationship between each strand under AEPS® cognitive and social-communication area and each subdomain under BDI-2 cognitive and communication domains was also explored. Correlating subdomain scores from these two measures provided detailed information on whether and how children score similarly on these two measures.

2.2 Social Validity

2.2.1 Recruitment and Participant

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. Teachers who are interested in participation returned signed consent. All five teachers agreed to participation.

2.2.2 Analyses

The focus group data analysis used ethnographic summary and content analysis (Morgan, 1988). The ethnographic approach usually uses more direct quotations of the group discussion, while content analysis typically produces numerical description of the data (Morgan, 1988). Coding schemas were used, combined with the direct quotations from the discussion. Each open-ended 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 could only be placed under one category.

The group discussion was audio-taped and lasted about 30 minutes and then transcribed verbatim. After reviewing the transcripts twice, the researcher developed a list of codes according to the responses, and assigned a code to each response. 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 their actual knowledge*) 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 ensure the consistency 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 fits 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 fits 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 broader themes 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*.

An inter-rater reliability check was achieved by using a code by code comparison method between the researcher and an outside coder. After the data were transcribed each coder received a copy of the transcript. Each coder then read the transcript independently and analyzed the transcribed data as previously described. Both coders listed codes and made notes in the margin of the transcripts so the comparison could be made. Code by code comparisons occurred 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 the *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 impacted child response. After carefully review of the transcript again, an agreement was reached that it should be listed as *tester familiarity* under *child response*. This process was repeated until 100% consensus was reached on each category and code.

3. Results

3.1 Concurrent Validity

Table 1 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, with the correlation coefficient of .60 ($p<.001$).

Strands in the AEPS® social-communication area also were positively correlated to BDI communication domain as well as its subdomains. The two strands in the 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 these strands and subdomains were significantly correlated. Higher scores of the AEPS® social-communicative interaction strand were associated with higher scores of BDI-2 receptive communication and expressive communication. Higher scores of the AEPS® word, phrase, and sentences strand were also associated with higher scores of the BDI-2 receptive communication and expressive communication.

Table 2 presents the Pearson's correlation between the AEPS® cognitive domain and the BDI cognitive domain. The results indicate that a positive correlation existed between the AEPS® cognitive score and the BDI cognitive scores, and the correlation was statistically significant ($r (32) = .57, p <.001$). The results indicated that higher AEPS® cognitive scores were associated with higher BDI cognitive scores.

Eight strands in the AEPS® cognitive domain were also positively 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 under BDI-2 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, premath, and phonological awareness and emergent reading) were significantly correlated to all three sub-domains in the BDI-2 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) was significantly correlated to the BDI-2 cognitive DQ score.

3.2 Social Validity

After analyzing, comparing and discussing the transcribed data, both coders agreed that four major 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*.

Child response. 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 responses 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.”

Advantages/disadvantages of standardized tests. Three factors emerged as advantages of standardized tests: 1) they are easier to administer; 2) they take less time; and 3) they are 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.”

Advantages/disadvantages of authentic assessments. 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 necessarily 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 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 do not care about the standardized test because they do not 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 is 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.

Scores generated from the instrument are valid for the purposes for which they are being used (National Research Council, 2008). Assessors should be trained to meet a clearly specified level of expertise in administering assessments, should be monitored systemically, and should be reevaluated occasionally because implementing assessment For assessment that can have significant consequences for children, teachers, or programs, following the best possible assessment practice is crucial. Curriculum-based assessment can be used for accountability under the condition that objectivity of such assessment is ensured by checking for reliability and consistency.

4. Discussion

4.1 Concurrent Validity

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 skills and it also is perceived as a useful measure by teachers. Correlations were run between the AEPS® social-communication area and the BDI-2 communication domain. According to the results, scores from the AEPS® social-communication area were significantly correlated to scores from the BDI-2 communication domain. Based on Cohen’s (1988) definition of strength of correlation, it is moderately correlated ($r=.60$, $p<.001$). Compared to the validity coefficient 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 acceptable.

Upon further examination of the strands, strands under AEPS® social-communication area were all significantly correlated to both subdomains under the BDI-2 communication domain. The validity coefficients of these correlations range from .63 to .72. When comparing these numbers to other validity studies (Newborg, 2004) of the BDI-2, it is appropriate 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 validity.

Correlations also were run between the AEPS® cognitive area and the BDI-2 cognitive domain. According to the correlation analysis, scores from the AEPS® cognitive area were significantly correlated to scores from the BDI-2

cognitive domain. Based on Cohen's (1988) definition of strength of correlation, they were 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, the coefficient of .57 found in this study is an acceptable figure.

Based on the further examination 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 the 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 the 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 the 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 concept. 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 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.

4.2 Social Validity

According to the results from the focus group, teachers indicated that authentic assessments such as the AEPS® truly reflect what a child knows regardless of his or her personality, and they all used it to gain information from children about their skills. Teachers also indicated that they did not use standardized tests much because these tools did not reflect a child's ability as accurately 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 admitted using 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 instruction 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 puts 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 literature that pointed out preschool school children as different as their school-age counterparts. Literature (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 levels and higher levels of frustration than older children (Nagle, 2000).

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. The fact that alternative assessment was viewed as a time-consuming task impacted teacher's preferences. When the authentic assessment has to be repeated three times a year it could add pressure 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 take 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 understand authentic assessment more easily. This is consistent with findings that when results from alternative assessments were used for reporting to parents, the performance indicators 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 because of its authenticity and naturally embedded characteristics. It also is easier for children and parents. As long as the time-consuming issue is addressed by better preparing the materials and getting to know children better, it can serve as an efficient method to assess young children.

5. Implications 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: 1) it was conducted in the classroom by teachers during regular activities so that it ensured naturalistic environment; 2) 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; 3) the AEPS® required some parent input, and was easier for parents to understand; and 4) 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 assessments have been questioned (Harbin et al., 2005; Neisworth & Bagnato, 2004; Stewart & Kaminski, 2002).

This study answered one aspect of psychometric questions raised from the researchers. The results indicated that scores from the AEPS® are similar as 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 the area of early language, literacy and pre-math areas.

6. Implication for Practices

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 a 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 an authentic assessment measure to provide accountability data. In addition, based on the teachers’ perceptions of the AEPS®, using the AEPS® is appealing for several reasons: First, the efficiency issue of the curriculum-based assessment can be addressed. Teachers indicated that the AEPS® cost less time than standardized tests when the teacher is familiar with children. Before the time-consuming issue of the AEPS® being addressed, one of the advantages of teacher’s preference standardized test over authentic assessment was 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 result in more natural and authentic responses from children than those collected for 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 using 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 reporting their accountability data. Furthermore, the AEPS® Test also has been developed as a measure that provides cut-off scores for determining eligibility for special service for infants and young children (Bricker et al., 2008; 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. And 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 partially answered. 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 have 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.

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, S. (1998). The trouble with testing. *Young Children*, 53(4), 25-29
- 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.
- Appl, D. J. (2000). Clarifying the preschool assessment process: traditional practices and alternative approaches. *Early Childhood Education Journal*, 27(4), 219-225
- Black, P. & William, D. (1998). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan*, 80, 139-148
- 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
- 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).
- Black, P. & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-148
- Bogdan, R.C., & Biklen, S.K. (2003). Qualitative research for education: an introduction to theories and methods. Boston, MA: Allyn & Bacon.
- Bricker, D., Pretti-Frontczak, K. (2002). *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
- 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
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Crawford, J. (2005, May/June). Test Driven. *NABE News*, 28, 1.
- Division of Early Childhood (DEC). (2007). *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*.
- Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6(1), Available from [Online] Available: <http://epaa.asu.edu/epaa/v6n1.html>
- 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
- Hatch, J. A. (2002). Accountability shovels: 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
- 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.
- Hess, F., & Brigham, F.J. (2000). The promises and pitfalls of high stakes testing. *American School Board Journal, 187 (1)*, 26-29
- 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.
- Kaplan, R.M., & Saccuzzo, D.P. (2001). *Psychological testing: principles, applications, and issues* (5th ed.). Belmont, CA: Wadsworth/Thomson Learning.
- Kellaghan, T. & Madaus, G. (1991). National Testing: lessons for American from Europe. *Educational Leadership, 49(3)*, 87-93
- Kim, J., & Suen, H.K. (2003). Predicting children's academic achievement from early assessment scores: a validity generalization study. *Early Childhood Research Quarterly, 18 (4)*, 547-566
- 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
- La Paro, K.M., & Pianta, R.C. (2000). Predicting children's competence in the early school years: a meta-analytic review. *Review of Educational Research, 70(4)*, 443-484
- 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
- Matthew, M., Schweid, J., & Sireci, S.G. (2010). Formative assessment systems: evaluating the fit between school districts' needs and assessment systems' characteristics. *Educational Assessment, Evaluation and Accountability, 22(1)*, 29-52
- McLean, M. Bailey, D., & Wolery, M. (1996). *Assessment of infants and preschoolers with special needs*. Columbus, OH: Merrill/Prentice-Hall.
- Meisels, S.J. (2006), Accountability in early childhood: No easy answers.
- Meisels, S. J., & Atkins-Burnett, S. (2000). The elements of early childhood assessment. In J. P. Shonkoff, & S. J. Meisels (Eds.), *Handbook of early childhood intervention* (2nd ed., pp. 231– 257). New York: Cambridge University Press.
- Klein, A. S. & Estes, J. S. (2004). Using observation for performance assessment. *Early Childhood News, 23*, 32-39
- Mitchell, R. (1995). *The promise of performance assessment: how to use backlash constructively*. Paper presented at AERA annual conference. San Francisco, CA.
- Mintrop, H. (2004). High-stakes accountability, state oversight and educational equity. *Teacher College Record, 106 (11)*, 2128-2145
- 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
- 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*.
- 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. (2000). Issues in preschool assessment. In B.A. Bracken (Eds.), *The Psychoeducational assessment of Neisworth, J. T. & Bagnato, S. J. (2004). The mismeasure of young children: The authentic assessment alternative. Infants and Young Children, 17(3)*, 198-212
- Neuman, S.B., & Roskos, K. (2005). The state of state pre-kindergarten standards. *Early Childhood Research Quarterly, 20*, 125-145

- Newborg, J. (2005) *Battelle Development Inventory (2nd Edition)*. Rolling Meadow, IL: Riverside Publishing.
- Nunnally, J. C. (1978). Psychometric Theory. New York: McGraw-Hill.
- preschool children* (pp. 19-32). Needham Heights, MA: Pearson Education.
- Ratcliff, N.J. (1995). The need for alternative techniques for assessing young children's emerging literacy skills, *Contemporary Education*, 66(3), 169-171
- Shepard, L. (1991). Will national tests improve student learning? *Phi Delta Kappan*, 73(3), 232-238
- Smith, L. & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, 10(4), 7-11
- Shepard, L., Tylor, G. A., & Kagan, S. L. (1996). *Trends in early childhood assessment policies and practices*. Washington, DC: National Education Goals Panel.
- Scott-Little, C., Lesko, J., Martella, J., & Milburn, P. (2007). *Early learning standards: Results from a national survey to document trends in state-level policies and practices*. Retrieved April 1, 2008, [Online] Available: <http://ecrp.uiuc.edu/v9n1/little.html>
- Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessment. *Journal of Educational Measurement*, 30(3), 215-232

Table 1. 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 .01 level (2 tailed)

*: significant at .05 level (2 tailed)

Table 2. Pearson'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 .01 level (2 tailed)

*: significant at .05 level (2 tailed)