The Pennsylvania State University

The Graduate School

Department of Educational Psychology, School Psychology, and Special Education

EXAMINING THE DIMENSIONALITY OF EARLI NUMERACY SKILL MEASURES

A Thesis in

Educational Psychology

by

Weiyi Cheng

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2011

The thesis of Weiyi Cheng was reviewed and approved* by the following:

Pui-Wa Lei
Associate Professor of Education
Thesis Advisor

Hoi K. Suen
Distinguished Professor of Educational Psychology
Professor in Charge

Spencer G. Niles
Head of the Department of Educational Psychology

*Signatures are on file in the Graduate School

ABSTRACT

Test validity is a critical issue in test construction and evaluation. Evidence for construct validity can be gathered by examining the factor structure of a test (Messick, 1995). One assumption of many assessments is unidimensionality, which means that the test is constructed to represent a single attribute (Banerji, Smith, & Dedrick, 1997).

The Early Arithmetic, Reading, and Learning Indicators (EARLI, Diperna, Morgan, & Lei, 2006) is a recently developed screening instrument that aims at providing practitioners with a psychometrically sound assessment to assess key academic skills and monitor growth of children at preschool age. The current study examined the dimensionality of the EARLI numeracy measures, which target key skills such as number identification, counting, and basic arithmetic.

Considering the small sample size of the study, three programs, NOHARM, DIMTEST and DETECT were selected to provide a comprehensive picture of the underlying structure of the EARLI numeracy measures. Initial evidence of unidimensionality was obtained for some of the measures (Counting Aloud, Counting Object, and Measurement), which contributed to the support of their use in preschools. However, the dimensionality for two measures (Number Naming and Pattern Recognition), remained inconclusive due to discrepancy in results between programs. Further evidence concerning dimensionality is needed before any substantive conclusion can be made.

Our literature review revealed the lack of empirical studies on dimensionality evaluation with small sample sizes. Some inconsistent results found in this study suggested that more research is needed to provide guidelines for dimensionality assessment with small samples.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

This work would not have been possible without the support and encouragement of my advisor, Dr. Pui-Wa Lei. I want to express my gratitude for the guidance and inspiration she has provided me along my academic career. I also would like to thank my second reader, Dr. Hoi Suen for insightful comments.

I am also grateful for all my friends who have been my source of inspiration and motivation. I cannot end without thanking my family, on whose constant encouragement and love I have relied throughout my time at the Academy.

Chapter 1

Introduction

There is a consensus that academic development in early years plays a critical role in children's later school success (e.g., Duncan, Dowsett, & Claessens, 2007). Academic development is a cumulative process beginning with the acquisition of many basic knowledge and concepts that provide the basis for subsequent advanced learning (Feinstein & Duckworth, 2006; Fox & Diezmann, 2007).

Particularly, researchers suggest that children's early mathematical experiences play an important role in the development of their understanding of advanced mathematics and predict later school success (Wolfgang, Stannard & Jones, 2001). Mathematical proficiency is of increasing importance in children's education and career, given that studying advanced mathematics is a key factor for success in college science (Sadler & Tai, 2007) and that high math achievement is essential for a wide range of vocations such as engineering, technology and science of the 21st century (Mazzocco & Thompson, 2005).

Early academic assessment is an effective way to identify children at risk at early age and prevent disparity among children at school entry. It is suggested by researchers (Floyd, Hojnoski, & Key, 2006) that the assessment of mathematics and number skills should begin as early as age 3 to promote early identification of learning problems. Therefore, more psychometrically sound assessments are needed to prevent children from lagging behind in their early years.

The Early Arithmetic, Reading, and Learning Indicators (EARLI; DiPerna, Morgan, & Lei, 2006) is a recently developed screening instrument that aims at providing practitioners with a psychometrically sound tool to assess key academic skills and monitor growth of children at preschool age. Test validation is an essential process to provide supporting evidence for the valid

use of EARLI probes. Some encouraging findings for EARLI probes concerning reliability and validity evidence were reported in previous studies (Reid, Morgan, DiPerna, & Lei, 2006).

The dimensional structure of a test is usually used to provide one type of validity evidence concerning construct validity (Messick, 1995). Unidimensionality is one of the important assumptions of most of the academic assessments, the violation of which suggests potential existence of construct-irrelevant factors or biased items. Though unidimensionality is a critical component of validity evidence, it is often overlooked during test evaluation (Wang, 2006).

Dimensionality of the EARLI literacy subtests was investigated (Hochsted, Lei, DiPerna, & Morgan, 2010), and preliminary evidence of unidimensionality was found for most of the EARLI literacy skill measures except for Sound Deletion. To provide further validity evidence for test development and test use, the current study intends to examine the dimensionality of EARLI numeracy skill scores. Indications of unidimensionality will provide more evidence to support the use of the EARLI numeracy skill measures in preschools. Indications of multidimensionality may suggest a need for further investigation of construct-irrelevant factors or modification of test items and score interpretations.

Literature Review

Early education and emergent numerical skills

Early childhood education has been a focus of attention because developments in early childhood years "are remarkable for their speed, comprehensiveness, and complexity" (Perry & Dockett, 2002, p. 83). Cumulative studies focusing on children's early learning and capacities in the first six years of life reveal that early experiences and development have long-lasting outcomes, which are essential determinants for later achievement (e.g., Duncan et al., 2007; Aubrey, Dahl, & Godfrey, 2006; Bowman, Donovan, & Burns, 2001).

Particularly, it is recognized that the acquisition of early academic skills is a vital part of a developmental continuum (Feinstein & Duckworth, 2006) in that early acquisitions of basic concepts and skills during preschool years serve as the foundation on which later advanced knowledge can be built and developed during formal schooling (Entwisle & Alexander, 1990; Pungello, Kupersmidt, Burchinal, & Patterson, 1996; Whitehurst & Lonigan, 1998). Researchers who examined data from six studies of close to 36,000 preschoolers in the United States, Canada and England reported that children entering kindergarten with elementary math and reading skills are the most likely to do well in school later, and that the mastery of early math concepts on school entry was the strongest predictor of future academic success (Duncan et al., 2007).

Study results indicate that children's mathematical concepts and competences start to form and develop through informal experience during preschool years. For example, toddlers are reported to have a variety of mathematical competencies (Hughes, 1986) and that even infants are able to represent numbers in a nonverbal manner (Jordan & Kaplan, 2009).

Researchers and educators realize that early mathematics concepts go beyond learning to count and acquiring a few number facts (Ginsburg, Klein, & Starkey, 1998). Tasks such as "quantity discrimination (magnitude comparison), counting objects, counting aloud, number identification, basic computation, estimation, understanding measurement concepts, number production, and identifying a missing number" (Lago & Diperna 2010, p. 166) exemplify important early numerical skills.

Many empirical studies suggest that the acquisition of basic numerical concepts during early childhood serves as a foundation for the acquisition of later higher order mathematical concepts (Ginsburg & Allardice, 1984), Children who enter school with strong numerical competences are more likely to be advantaged throughout primary years (Aubrey, Dahl, & Godfrey, 2006), whereas the children at school entry with weak number competencies, especially with respect to operational knowledge and skills, have to make up for basic skills and may never catch up to their advanced peers (Jordan & Kaplan, 2009). Therefore, it is of particular importance to indentify children who are at risk for future math failure and offer them appropriate intervention as early as possible. Early identification of at risk children requires sound academic assessments.

Need for early academic assessment

Early assessment and intervention are receiving increasing attention because there is a need to identify children at risk in early age, promote early childhood development and prevent discrepancy at school entry (Methe, Hintze, & Floyd, 2008; Clarke & Shinn, 2004; VanDerHeyden, Witt, Naquin, & Noell, 2001). Preschool assessment can help to prevent children from lagging behind in their early years. Previous research indicated that approximately 10% of children have some form of learning disability (Kenny & Culbertson, 1993). By having

appropriate assessment in place, children who have potential development delays and need further evaluation and intervention can be identified.

Early assessment is also essential for effective instructions (e.g., Aubrey, Dahl, & Godfrey, 2006; Spodek & Saracho, 1997). By identifying children's diverse learning styles, unique strengths and special needs, teachers can provide more appropriate and intentional instructions with explicit purpose (Brenneman, Stevenson-Boyd, & Frede, 2009).

The increasing number of children enrolled in a variety of preschool education programs (e.g., Child Care, Head Start, Early Head Start) and the large amount of money invested in these programs also demand appropriate assessment to evaluate their effectiveness (Spodek & Saracho, 1997; Bracken & Nagle, 2007). Sound early assessments assist educators in program planning, help policymakers decide whether publicly funded early education programs are actually benefiting children and deserve continuous funding, and inform parents of their children's progress (Appl, 2000; Kochanoff et. al., 2003).

The importance of preschool academic assessment calls for the development of more appropriate instruments to monitor academic growth or evaluate instruction outcomes. Fox and Diezmann (2007) found that only 3% of the research on mathematics education and early childhood between 2000 and 2005 concerned preschool assessment. This phenomenon of lacking preschool academic assessment is especially noticeable with respect to early math assessments (Floyd, Hojnoski, & Key, 2006). Floyd and his colleagues (2006) reviewed 25 published measurements that focus on mathematics and number skills, only 1 study included children as young as age 3 (i.e., Magliocca, Rinaldi, Crew, & Kunzelmann, 1977).

Test validation and dimensionality

In addition to the need for preschool academic assessments, more extensive investigations of instruments' technical adequacy are also needed in the area of early education (Hojnoski & Missall, 2006). Psychometric characteristics such as reliability and validity are primary considerations in the selection of an appropriate test (Bagnato, Neisworth, & Munson, 1993). Reviews of several early mathematic assessments (Methe, Hintze, & Floyd, 2008; Foegen, Jiban & Deno, 2007; Floyd, Hojnoski, & Key, 2006) show that all but two (Chard et. al., 2005; Joyce & Wolking, 1987) have examined score reliability such as test-retest reliability (e.g., Clarke & Shinn, 2004; Daly, Wright, Kelly, & Martens, 1997; VanDerHeyden, Broussard, Fabre, Stanley, LeGrendre, & Creppell, 2004; Floyd, Hojnoski, & Key, 2006) and alternate form reliability (e.g., Clarke & Shinn, 2004; VanDerHeyden et al., 2004; VanDerHeyden et. al., 2001), and that moderate to strong reliability were found in most studies.

However, score validity has been less adequately investigated in existing early mathematic assessments. Validity is a concept that has evolved over time. As defined by 1999 *Test Standards*, validity refers to "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests…" (AERA, APA, & NCME, 1999, p. 9), which requires accumulative evidences based on test content, response processes, internal structure, relations to other variables, and consequence of the testing. Though many early mathematic measures (e.g., VanDerHeyden, Broussard, & Cooley, 2006; Daly et al., 1997; Chard, Clarke, Baker, Otterstedt, Braun, & Katz, 2005; Clarke & Shinn, 2004) provided concurrent and predictive correlations as validity evidence, few studies, except for the one conducted by Floyd and his colleagues (2006), provided information concerning internal structure of the test. Floyd, Hojnoski, and Key (2006) examined the factor structure of a screening instrument, Preschool Numeracy Indicators (PNI). The results of confirmatory factor analysis

conducted by program Amos 5.0 suggested that a single factor (Number Sense) affected all four PNI measures (count objects aloud, count orally from 1, name printed numbers, and compare and name larger quantity).

It appears that construct validity (internal structure of a test) fails to receive enough attention in the validation process of many early mathematic measures. Construct validation requires evidence of internal structure that examines the degree to which the test items and components match the defined construct (AERA/APA/NCME, 1999). It is a process of gathering evidence from a number of perspectives to support the use of a test as a measure of a construct (Shore, Thornton, & Shore, 1990). Frederiksen (1986) considered construct validity to be "the basis not only for better tests but also better understanding of what our tests measure" (p. 4). Construct validity is particularly important, but is also the most difficult and complex property to obtain (Benson & Clark, 1982). Investigating the dimensionality of a measure is one way to produce construct validity evidence (Jasper, 2010). Dimensionality of a test is defined as the number of factors that is necessary to account for the inter-relationship among test items (McDonald, 1986). Many researchers agree that investigation of the dimensional structure of a measure is a "requisite part of a comprehensive validation process" (Jang & Roussos, 2007, p. 2).

A test usually assumes unidimensionality in test construction, which "impinges directly upon construct validity" (Banerji, Smith, & Dedrick, 1997, p. 57). Unidimensionality means that the test is constructed to represent a single attribute (Banerji, Smith, & Dedrick, 1997). The assumption of unidimensionality is satisfied only if a single ability is required to respond to the items (Nandakumar, 1994). The importance of unidimensionality is stated by Hattie (1985, p. 139) that: "one of the most critical and basic assumption of measurement theory is that a set of items forming an instrument all measure just one thing in common". A number of studies indicated that test items often measure other abilities in addition to the intended trait, in which case task-irrelevant abilities or characteristics of examinees may affect dimensionality of the

measures (e.g., Stout, 1987; Reckase, 1985). Therefore, dimensionality assessment is an indispensible process to provide empirical support for the content and construct aspects of validity and to ensure that appropriate interpretation of test results can be made.

From the perspective of test fairness, unidimensionality is also an important consideration in test development since "fairness requires that all examinees be given a comparable opportunity to demonstrate their standing on the construct(s) that the test is intended to measure" (AERA/APA/NCME, 1999, p. 74). Test fairness cannot be achieved if the test does not measure the same construct in targeted populations. The presence of bias items, which are usually derived from multidimensional test structure with construct-irrelevant abilities, threatens test fairness (Tindal & Haldyna, 2002). Stout (1987) also suggested that a test that is supposed to measure individual differences must measure a unified "trait". Otherwise it will be difficult and unreliable to make fair comparison across subgroups.

In addition, psychometric models that are commonly applied to describe item functions assume that the items measure a single trait (Reckase, 1985). For example, Item Response Theory (IRT) makes an explicit assumption of unidimensionality for unidimensional models (e.g., Hambleton, Swaminathan & Rogers, 1991; Henning, 1988). Unidimensional IRT models are widely used in scaling, equating, and computer adaptive testing. If the assumption of unidimensionality is met, IRT methodology could be applied legitimately (e.g., Stout, 1987). Violating this assumption could seriously bias parameter estimation (Wainer & Wang, 2001; Ackerman, 1989). Although classic test theory does not have an explicit assumption of test dimensionality, it assumes "homogeneous" items on the test, which is essentially an assumption of unidimensionality.

In general, it is more difficult to develop measures with adequate validity and reliability evidences for young children (Goodwin & Driscoll, 1980) because they are usually not good at taking tests, which may result in additional errors in their responses. Given the expectation of

more construct-irrelevant factors in early assessment, it is of more importance to empirically confirm predefined test constructs through dimensionality examination.

Methods of dimensionality examination

Given the importance of verifying unidimensionality of a test, it is necessary to have appropriate and reliable methods to statistically test unidimensionality. To date, numerous statistical methods have been developed and applied to check the dimensionality of tests. Several researchers (e.g., Hattie, 1984, 1985; Tate, 2003) gave comprehensive reviews of methods that have been developed at different times.

Classical linear factor analysis (FA), which is an approach usually associated with construct validity (Thompson & Daniel, 1996), represents a observed response variable as a linear function of multiple latent traits. However, it should be noted that in addition to the general concern about overestimating the number of underlying dimensions (e.g., De Ayala & Hertzog, 1989), linear FA method assumes continuous observed variables. Though analysis on tetrachoric correlation matrix is applied in an attempt to solve the problem, the analysis on this matrix is inappropriate when the distribution of latent trait is not normal (e.g., Jones, Sabera, & Trosset, 1987). As alternatives to linear FA, parametric nonlinear FA methods have been proposed. Examples include the Normal-Ogive Harmonic Analysis Robust Method (NOHARM, McDonald, 1967) and Full-Information Item Factor Analysis (TESTFACT, Bock, Gibbons & Muraki, 1988).

Factor-analytic methods assess the "traditional unidimensionality" which is also known as *strict unidimensionality*, without distinguishing between major and minor dimensions (e.g., Nandakumar, 1991; Tindal & Haladyna, 2002). However, it has long been argued that the restrictive assumption of unidimensionality can never be strictly met in reality (e.g., Stout, 1990; Nandakumar, 1991; Humpherys, 1984) since many factors related to examinees, environment and

their interactions with the test always have an effect on test performance to some extent (Blais & Laurier, 1995). In other words, it is more realistic to look for a "dominant" dimension that is responsible for examinees' performance. Stout (1990) proposed to replace traditional IRT assumptions of unidimensionality and local independence with the weaker assumptions of essential unidimensionality and essential independence, respectively. As opposed to the factor-analytic approach, several conditional-covariances-based nonparametric methods have been developed to assess the assumption of "essential unidimensionality". Examples include the Dimensionality Evaluation to Enumerate Contributing Traits (DETECT) procedure by Zhang and Stout (1999), cluster analysis of items (HCA/CCPROX) by Roussos, Stout, and Marden (1998), and the DIMTEST program (Stout, Douglas, Junker, & Roussos, 1993; Stout, Froelich, & Gao, 2001).

However, no dimensionality analysis procedures have been found to be completely satisfactory under all conditions (De Ayala & Hertzog, 1992; Stout, 1987). The test statistics may be inaccurate when there is a mismatch between test structure and that assumed by the method (Tate, 2003), or under conditions with small samples and short test lengths (De Champlain & Gessaroli, 1996, 1998; Monahan, Stump, Finch, & Hambleton, 2007).

The major challenge of the current study is small sample size given that the largest sample size for an EARLI numeracy measure is in the mid-200s, with which many of the dimensionality indices may not work well. Few studies, especially empirical ones, have examined the behavior of dimensionality assessment procedures with sample size less than 500 (Monahan et al., 2007; De Champlain & Gessaroli, 1998). Therefore, it is desirable to rely on multiple methods to determine dimensionality. Literature review suggested that the following procedures might work with small sample sizes.

NOHARM is a widely used procedure that is found to be powerful in detecting multidimensionality in most cases (Tate, 2003). This program uses a nonlinear factor analytic

approach to estimate item parameters in either an exploratory or confirmatory mode. The normal orgive multidimensional IRT model can be represented as (McDonald, 1999)

$$P\{U_i=1 \mid \Theta=(\theta_1,\ldots, \theta_k)\}=N\{\beta_{i0}+ \beta_{i1}\theta_1,+\ldots+ \beta_{ik}\theta_k\}.$$

Here the probability of a correct response to item i given the latent traits is a function of normal orgive N{}, item difficulty parameters β_{i0}, and item discrimination parameters β_{ik}. An approximate chi-square test of model fit is provided by the CHIDIM program (De Champlain & Tang, 1997).

With sample size of 2000, NOHARM was reported to have "generally good to excellent" performance in confirming unidimensionality for unidimensional data and recovering the generating model for multidimensional data (Tate, 2003). De Champlain and Gessaroli (1996, 1998) examined NOHARM performance when small sample sizes were applied and found that the NOHARM approximate chi-square statistic maintained Type I error rate close to the nominal alpha level (.05) when the sample size was as small as 250 and it was able to correctly reject unidimensionality for two-dimensional data sets. Finch and Habing (2007) recommended that "when it is known that guessing is not present in the data, one of the NOHARM-based test statistics should be used" (p.304). The exceptions to the good NOHARM performance were found when there were extremely large discrimination parameters or when guessing was present in the data such that NOHARM tended to generate a difficulty factor in the former case (Tate, 2003) and failed to maintain the nominal Type I error rate in the latter (Finch & Habing, 2007).

DIMTEST (Stout, Douglas, Junker, & Roussos, 1993) is a nonparametric statistical procedure based on Stout's (1987) concept of "essential unidimensionality". The purpose of DIMTEST is to conduct a hypothesis test on two sets of items, taken by the same examinees, called AT and PT. AT stands for the "assessment subtest," and PT stands for the "partitioning subtest." The formation of two subsets can be conducted using either a confirmatory or an exploratory approach. In a confirmatory approach, the AT items are identified using expert

opinion based on test content. For an exploratory approach, AT items can be selected through a statistical procedure (e.g., by using the items with the highest factor loadings from an unrotated principal factor analysis of the item tetrachoric correlations). DIMTEST is based on the idea that if a test is unidimensional then the conditional covariance between any two items on the AT subset should be zero after conditioning on PT.

Gessaroli and De Champlain (1996) found that DIMTEST had good Type I error control for sample size from 500 to 1000 and test length from 15 to 45, and high power in detecting departure from unidimensionality with sample size of 1000 and more than 30 items on the test. Similar findings were reported by Nandakumar (1991) and Tate (2003) with sample sizes of 750 to 2000. DIMTEST has been applied to assess dimensionality for real data sets using sample size smaller than 250 (Hsieh, 2010). Yet the author did not discuss the accuracy of DIMTEST performance with such small sample size. DIMTEST was also recommended when guessing was known to be present, in which case, NOHARM had inflated Type I error rate (Finch & Habing, 2007). However, it has been reported that in some cases, DIMTEST has higher Type I error rate than NOHARM, especially for short tests (number of items<15) (Gessaroli & De Champlain; 1996; Folske, Gessaroli, & De Champlain, 1998).

DIMTEST is not able to identify the exact number of dimensions of the data when essential unidimensionality is rejected. DETECT (Kim, 1994; Zhang & Stout, 1999) is often used as a complementary procedure to DIMTEST to indicate the degree of multidimensionality and assign items to different dimensions. The DETECT procedure is based on the idea that after conditioning on the test composite Θ_{TT}, items measuring the same dimension on a test should exhibit positive conditional covariances, and items measuring different dimensions should exhibit negative conditional covariances. DETECT searches for a partition (P*) of the items into a set of non-overlapping clusters that maximize the DETECT index D(P), which is given by the equation

$$D(P) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \delta_{ij} E[Cov(X_i, X_j) | \Theta_{TT}],$$

where n is the number of items on the test, P denotes the partitioning of n items into k clusters, Θ_{TT} is the test composite, X_i and X_j are observed scores on items i and j, and δ_{ij} is equal to 1 when item i and j are from the same cluster, and -1 otherwise. Each of the possible partitions of items would be examined, and the one that results in the largest D(P) index would be selected. Theoretically, the D(P) index is expected to be zero under the condition of unidimensionality (Zhang & Stout, 1999). Larger D(P) values indicate more departure from unidimensionality. It has been suggested that D(P)< 0.2 represents essential unidimensionality, between 0.2 and 0.4 indicates weak to moderate multidimensionality, between 0.4 and 1.0 indicates moderate to large multidimensionality, and >1.0 indicates strong multidimensionality (Roussos & Ozbek, 2006).

Another index, r, is usually reported with DETECT index as an indicator of approximate simple structure. Simple structure is present when in a multidimensional data structure, each item measures only a single latent ability (McDonald, 1999). However, simple structure rarely exists in real data, the term approximate simple structure is often used instead. The data will be said to have "approximate simple structure" when it consists of dimensionally homogenous clusters that can be *sufficiently* separated (Zhang & Stout, 1999). The approximate simple structure index r can be computed using the following ratio

$$r_{\max} = \frac{D(P^*)}{\hat{D}(P^*)},$$

where $\hat{D}(P^*)$ is the maximum possible value that can be obtained by summing the absolute values of the estimated conditional covariances across all item pairs. If the partition, P*, results in a strictly simple structure solution, $D(P^*)$ will be equal to $\hat{D}(P^*)$, and the ratio r will be one. If a strict simple structure solution is not obtained, then $D(P^*)$ will be smaller than $\hat{D}(P^*)$ and ratio r will be smaller than 1. When complex structure is present, some of the conditional between-

cluster item covariances will be positive and result in a $D(\mathrm{P}^*)$ value smaller than $\widehat{D}(\mathrm{P}^*)$. According to Kim (1994), r index equal to or larger than 0.80 indicated approximate simple structure. DETECT can accurately identify the correct partition and confirm the dimensional structure with presence of approximate simple structure when the test has 20 or more items and relative large sample size (400 to 2000) (e.g., Roussos & Ozbek, 2005; Zhang & Stout, 1999).

Item discrimination, correlations between dimensions and degree of complexity of the data were found to have noticeable influence on DETECT performance (e.g., Tate, 2003; Tan & Gierl, 2006; Gierl, Leighton, & Tan, 2006; Zhang & Stout, 1999). Gierl, Leighton, and Tan (2006) found that when data displayed complex structure (30% of the items measured more than one dimension), DETECT was able to classify items accurately if the correlation between dimensions was no larger than 0.60 and sample size exceeded 500; if 50% of the items displayed complex structure, more than 1000 sample size was needed; when the correlation was 0.90, DETECT did not work well regardless of sample size. Similarly, Tan and Gierl (2006) reported that when items with low discrimination power were involved, DETECT worked properly only under limited conditions with approximate simple structure (i.e., correlation between dimensions ≤ 0.70).

The DETECT procedure can operate in either the exploratory mode or in the cross-validation mode. In the exploratory model, the DETECT index obtained by the program is the largest DETECT index over all searched partitions. The cross-validation index is intended to avoid mistaking sampling error as multidimensionality (Monahan et al., 2007). In the cross-validation model, the DETECT index is calculated on the validation data set given the partition determined in the training data set. In a simulation study conducted by Monahan and his colleagues (2007), the performance of the two DETECT indices were examined under varied test lengths and sample sizes (the smallest sample size condition was 100). Their results based on bias, standard error, and root mean square error suggested that the cross-validated index

consistently performed better than the exploratory index with respect to bias and RMSE, although standard error of the exploratory index was less affected by sample size than that of the cross-validated index. When sample size dropped from 500 examinees to 100 examinees, the increase in bias and RMSE were greater for exploratory index than for cross-validated index with any test length. The better performance of cross-validated index compared to exploratory index, with regard to bias and RMSE, was most pronounced when sample size was 100.

EARLI probes and purpose of current study

The Early Arithmetic, Reading, and Learning Indicators (EARLI, DiPerna, Morgan, & Lei, 2006) is a recently developed screening instrument that aims at providing practitioners with a psychometrically sound assessment to assess key academic skills and monitor growth of children at preschool age. EARLI includes two distinct sets of measures. The literacy measures target key early skills such as phonological awareness, vocabulary, and print knowledge; the numeracy measures target key skills such as number identification, counting, and basic arithmetic skills.

Preliminary findings with respect to psychometric properties of the EARLI measures were reported in a previous study (Reid et al., 2006). Specifically, all item-discrimination indices of the numeracy measures were positive and high, and Cronbach's alphas for the different subtests ranged from .82 to .98. Evidence of concurrent correlation was obtained by computing correlations of EARLI subtest scores with the Woodcock_Johnson III Test of Achievement of early academic skills (2011). The EARLI literacy skill measures exhibited small to large (.28-.89) correlation with the WJ- III, among which the Letter Naming measure demonstrated the strongest (.70 to .89) correlation with WJ- III subtests. The EARLI numeracy skill measures exhibited moderate to large correlation (.30-.88) with the WJ- III subtests (see Reid et al., 2006 for details).

Dimensionality of the EARLI literacy subtests has been investigated (Hochsted et al., 2010), and preliminary evidence of unidimensionality was found for most of the EARLI literacy skill measures except for Sound Deletion. The potential problem of the study was small sample size ($N<300$) as mentioned by the authors. Although the authors showed that it was possible to have high power ($>.95$) to reject unidimensionality at such small sample sizes under simple structure and low inter-factor correlation ($\leq.5$), actual performance of the dimensionality assessment methods (NOHARM and DIMTEST) used in the study was not examined under these conditions. It is still inconclusive as to what methods may work best in determining dimensionality under small sample conditions.

The purpose of the current study is to investigate the dimensionality of the EARLI numeracy skill measures. Considering the small sample size of the study, a single dimensionality assessment method may not yield an accurate and reliable conclusion. It is desirable to compare the results from different procedures before making any conclusion with confidence. DIMTEST and NOHARM are two of the most popular methods for assessing dimensionality (Gessaroli, 1994). These two procedures test theoretically different hypotheses, either strict unidimensionality or essential unidimensionality, and they have unique strengths under different conditions. Because DIMTEST does not indicate the number of dimensions when essential unidimensionality is rejected, DETECT has been used as a complementary procedure to identify potential multidimensional structures. Results from these programs would give us a more comprehensive picture of the dimensional structure of the EARLI numeracy measures.

Chapter 2

Methods

Measures

The EARLI include six numeracy measures that assess early mathematics skills (e.g., measurement, number, identification, counting aloud, pattern recognition, subitizing, counting objects) (Reid et al., 2006).

Counting Aloud requires children to count aloud, staring with 1, in the correct sequence. Testing is discontinued if the child hesitates for more than 10 seconds.

Number Naming requires children to provide the name of each number or shape in isolation. No corrective feedback is provided on the test items.

Counting Objects requires children to count different sets of objects as quickly and accurately as they can. No corrective feedback is provided on the test items.

Grouping requires children to immediately identify the number of objects in small groups of six or fewer. No corrective feedback is provided on the test items.

Measurement requires children to identify fundamental measurement concepts (e.g., taller, shorter, higher, lower) using basic shapes. No corrective feedback is provided on the test items.

Pattern Recognition requires children to recognize and complete patterns using combination of three shapes (i.e., circle, square, triangle). Children is presented with an incomplete pattern (e.g., triangle, triangle, circle, [black space]) and asked to name which shape finishes the pattern. Corrective feed back only is provided if the child fails any of the first 3 items.

Two alternate forms (A and B) of Counting Objects, Grouping, Pattern recognition, and Measurement were used in the study, while Counting Aloud and Number Naming have one form. The alternative forms of each measure were equivalent in length (number of items) and shared a subset of common items.

For each measures, except for Grouping, responses are scored dichotomously (correct=1, incorrect=0). Non-responses are scored as incorrect. Dimensionality assessments were conducted on the five dichotomously scored measures. Grouping items were scored polytomously and was not examined in current study.

Data Source

Data were collected from a Head Start program in the Northeastern United States (Lei et al., 2009). Two cohorts of children (ages 3 and 4, defined at the beginning of the school year and remained the same despite some turning 4 or 5 during the academic year) completed the EARLI numeracy measures at three different time points during the year (October, January, and April). Test length and sample size for each form at three time points were summarized in Table 1.

Table 1. Test length and sample size for each form and time points

Measures	Items	Examinees		
		Time 1	Time 2	Time 3
Counting Aloud	60	244	279	254
Number Naming	38	244	279	254
Counting Objects (Form A)	20	112	139	125
Counting Objects (Form B)	20	132	140	129
Measurement (Form A)	20	112	139	125
Measurement (Form B)	20	132	140	129
Pattern Recognition (Form A)	20	112	139	125
Pattern Recognition (Form B)	20	132	140	129

Procedures

A one-dimensional exploratory factor analysis (EFA) was first applied to the five measures by form and time point using the NOHARM program. NOHARM provides root mean squared residual (RMSR) to measure the lack of fit of the model. A RMSR value that is equal to or less than $4/\sqrt{N}$, where N is sample size, indicates a good model fit (Fraser & McDonald, 1988). An approximate chi-square test (CHIDIM; De Champlain & Tang, 1997) was used to further assess the test dimensionality. If the approximate chi-square test was not significant and that percentages of large standardized residuals were small (<7%), unidimensionality would not be rejected (De Champlain & Tang, 1997). For those measures with significant approximate chi-square results, a two-factor EFA was analyzed by NOHARM. If the two-factor model resulted in more than 10% decrease in RMSR (Tate, 2003), unidimensionality would be rejected in favor of the two-factor model. A model with one more factor would be estimated until the model was not rejected by the approximate chi-square test and the decrease in RMSR from the previous model was less than 10%. In addition, the pattern of the varimax-rotated factor loadings was explored to help explain the rejection of unidimensionality and determine the number of interpretable dimensions for the measure.

DIMTEST and DETECT were applied as a comparison against NOHARM results. The most recent version of DIMTEST (Froelich, 2000) was used to test the hypothesis H_0: d=1 versus H_a: d>1, where d is the number of dominant dimensions. In current study, the program was allowed to automatically select the optimum number of items for the AT subset based on a principal-axis factor analysis of a tetrachoric correlation matrix. When DIMTEST rejected the null hypothesis at the .05 alpha level at any time point, DETECT was further applied for all three time points to examine the degree of multidimensionality and pattern of item classification. The cross-validated DETECT index, D(P), was calculated by specifying a 50%/50% split of total

sample size randomly selected by the program into training and cross-validation subsamples. This study followed the guidelines for interpreting the DETECT index by Roussos and Ozbek (2006): D(P)< 0.2 represents essential unidimensionality, 0.2 to 0.4 indicates weak to moderate multidimensionality, 0.4 to 1.0 indicates moderate to large multidimensionality, and above 1.0 indicates strong multidimensionality. When the D(P) index indicated at least weak multidimensionality (D(P)> 0.2), the r index was examined to find out if approximate simple structure existed (r>.8). The number of clusters identified by DETECT is the partition used to maximize the D(P) value. When D(P) index indicates the presence of essential unidimensionality, the number of clusters can be ignored. When multidimensionality and approximate simple structure are indicated, the number of sizable clusters in the optimal partition equal to the number of dominant dimensions present in the test (Kim, 1994; Zhang & Stout, 1999).

Reasonable conclusions about the dimensionality of the five EARLI numeracy measures would be made by comparing the results from the three programs and examining the consistency of the results across time points.

Chapter 3

Results

Summaries of results from NOHARM, DIMTEST and DETECT at three time points were provided in Table 2. It is noticeable that when NOHARM results were available, the values of RMSR were smaller than $4/\sqrt{N}$ for all the measures across all three time points. This was not surprising because small sample size leads to large value of $4/\sqrt{N}$. However, RMSR values for the unidimensional model from NOHARM were generally small (RMSR<.03). This made the % decrease in RMSR uninformative because a small denominator would make the ratio look large even for a very small amount of decrease. Therefore, we relied on other statistics for indications of dimensionality.

Table 2. Model fit statistics for NOHARM, DIMTEST and DETECT

	Time 1	Time 2	Time 3
Counting Aloud			
NOHARM 1-dimensional model			
RMSR ($4/\sqrt{N}$)	NR	0.0106734 (0.239)	0.0146459(0.251)
% decrease in RMSR		-	-
Approximate χ^2		15876.744 (p=0)	8527.755 (p=0)
% SE>|2|		22.94	13.90
DIMTEST	NR	NR	p=0.2145
DETECT			
D(P)	0.006	0.006	0.005
r index	0.91	0.82	0.95
number of clusters	2	2	2
Number Naming			
NOHARM 1-dimensional model			
RMSR ($4/\sqrt{N}$)	0.0061092 (0.256)	0.0085726 (0.239)	0.0094615 (0.251)
%decrease in RMSR	28.27	22.44	18.85
Approximate χ^2	2036.540 (p=0)	1345.029 (p=0)	1028.5879 (p=0)
% SE>|2|	14.51	12.38	8.96
NOHARM 2-dimensional model			
RMSR	0.0043819	0.0066488	0.0076782
% decrease in RMSR	22.51	18.40	10.09
Approximate χ^2	1202.633 (p=0)	676.165 (p=0.089)	680.385 (p=0.072)
% SE>|2|	9.957	4.836	4.552

(continued)

Table 2. (continued)

	Time 1	Time 2	Time 3
NOHARM 3-dimensional model			
RMSR	0.0033954		
%decrease in RMSR	12.06		
Approximate χ^2	972.961 (p=0)		
% SE>|2|	8.108		
NOHARM 4-dimensional model			
RMSR	0.0029858		
% decrease in RMSR	-		
Approximate χ^2	927.276 (p=0)		
% SE>|2|	7.112		
DIMTEST	p=0.0715	p=0.0017	p=0.0108
DETECT			
D(P)	0.036	0.082	0.151
r index	0.2	0.24	0.37
number of clusters	3	3	2
Counting Objects (Form A)			
NOHARM 1-dimensional model			
RMSR (4/√N)	NR	0.0054549 (0.339)	0.0092464 (0.358)
%decrease in RMSR		25.42	36.20
Approximate χ^2		251.934 (p<0.001)	189.116 (p=0.1501)
% SE>|2|		8.95	5.26
NOHARM 2-dimensional model			
RMSR		0.004068	
%decrease in RMSR		17.9	
Approximate χ^2		198.691 (p=0.006)	
%SE>|2|		7.895	
NOHARM 3-dimensional model			
RMSR		0.0033398	
% decrease in RMSR		17.6	
Approximate χ^2		162.508 (p=.042)	
% SE>|2|		5.263	
DIMTEST	p=0.0956	p=0.1444	p=0.0754
Counting Objects (Form B)			
NOHARM 1-dimensional model			
RMSR (4/√N)	NR	0.0094734(0.338)	0.0087895 (0.352)
% decrease in RMSR		32.11	23.76
Approximate χ^2		180.207 (p=0.2813)	98.9644 (p=0.9999)
% SE>|2|		4.74	2.11
DIMTEST	p=0.0685	p=0.1619	p=0.3187
Measurement (Form A)			
NOHARM 1-dimensional model			
RMSR (4/√N)	0.0158670 (0.378)	0.0130430 (0.339)	0.0149766 (0.358)
% decrease in RMSR	17.01	14.32	13.36
Approximate χ^2	136.942 (p=0.9705)	133.943 (p=0.9811)	141.529 (p=0.9455)
% SE>|2|	2.11	2.11	2.11
DIMTEST	p=0.0806	p=0.0934	p=0.0438
Measurement (Form B)			
NOHARM 1-dimensional model			
RMSR (4/√N)	0.0145869 (0.348)	0.0143302 (0.338)	0.0119335 (0.352)
% decrease in RMSR	16.35	15.29	12.58
Approximate χ^2	119.921 (p=0.9987)	141.734 (p=0.9441)	99.924 (p=0.9999)
% SE>|2|	1.05	3.16	1.05
DIMTEST	p=0.0674	p=0.1691	p=0.1763

(continued)

Table 2. (continued)

	Time 1	Time 2	Time 3
Pattern Recognition (Form A)			
NOHARM 1-dimensional model			
RMSR ($4/\sqrt{N}$)	0.0178256 (0.378)	0.0171160 (0.339)	0.0188511 (0.358)
% decrease in RMSR	32.42	14.95	28.26
Approximate χ^2	134.379 (p=0.9798)	129.710 (p=0.9906)	144.757 (p=0.9202)
% SE>|2|	2.11	2.11	2.11
DIMTEST	p=0.0387	p=0.0174	p=0.0064
DETECT			
D(P)	0.278	0.569	1.430
r index	0.22	0.29	0.61
number of clusters	3	5	3
Pattern Recognition (Form B)			
NOHARM 1-dimensional model			
RMSR ($4/\sqrt{N}$)	0.0213749 (0.348)	0.0170728 (0.338)	0.0170728 (0.352)
% decrease in RMSR	30.43	25.72	25.72
Approximate χ^2	199.760 (p=0.0589)	131.236,(p=0.9878)	120.699 (p=0.9984)
% SE>|2|	3.16	1.05	0.53
DIMTEST	p=0.0513	p=0.0086	p=0.0533
DETECT			
D(P)	0.262	1.187	0.466
r index	0.12	0.57	0.34
number of clusters	3	3	3

Note: NR= no result (program failed to generate results); "-" indicates no decrease in RMSR.

Counting Aloud

NOHARM failed to generate results for Counting Aloud in Time 1 due to the existence of items with no variations in item responses, and rejected the one-dimensional model in Time 2 and Time 3 when results were available. However, RMSR for the two-dimensional model was larger than that for the unidimensional model, supporting the assumption of unidimensionality. The varimax-rotated factor loadings (Table 3) for the two-factor model in Time 2 and Time 3 showed that almost all the items loaded on one factor, indicating the existence of a dominant dimension.

Table 3. Two-factor varimax-rotated loadings and item Discrimination and Difficulty estimates for Counting Aloud

	Time 2				Time 3			
Item	p-value	pbis	2-Factor loading		p-value	pbis	2-Factor loading	
1	0.95	0.27	**0.938**	0.348	0.98	0.20	**0.924**	0.383
2	0.94	0.29	0.657	**0.754**	0.96	0.27	0.405	**0.914**
3	0.91	0.33	0.540	**0.842**	0.95	0.29	0.401	**0.916**
4	0.84	0.43	0.454	**0.891**	0.94	0.32	0.397	**0.918**
5	0.79	0.49	0.437	**0.899**	0.91	0.37	0.392	**0.920**
6	0.75	0.53	0.431	**0.902**	0.85	0.45	0.388	**0.922**
7	0.71	0.56	0.427	**0.904**	0.81	0.50	0.387	**0.922**
8	0.68	0.58	0.426	**0.905**	0.81	0.50	0.387	**0.922**
9	0.66	0.59	0.425	**0.905**	0.80	0.52	0.387	**0.922**
10	0.59	0.62	0.422	**0.906**	0.76	0.54	0.386	**0.923**
11	0.54	0.64	0.421	**0.907**	0.72	0.57	0.385	**0.923**
12	0.54	0.64	0.421	**0.907**	0.70	0.58	0.385	**0.923**
13	0.43	0.67	0.419	**0.908**	0.61	0.63	0.384	**0.923**
14	0.37	0.69	0.418	**0.908**	0.51	0.70	0.383	**0.924**
15	0.29	0.73	0.418	**0.909**	0.44	0.75	0.383	**0.924**
16	0.22	0.77	0.417	**0.909**	0.39	0.80	0.383	**0.924**
17	0.21	0.78	0.417	**0.909**	0.37	0.81	0.383	**0.924**
18	0.20	0.79	0.417	**0.909**	0.37	0.82	0.383	**0.924**
19	0.19	0.79	0.417	**0.909**	0.35	0.83	0.383	**0.924**
20	0.18	0.79	0.417	**0.909**	0.33	0.84	0.383	**0.924**
21	0.17	0.80	0.417	**0.909**	0.32	0.84	0.383	**0.924**
22	0.16	0.80	0.417	**0.909**	0.32	0.84	0.383	**0.924**
23	0.15	0.79	0.417	**0.909**	0.31	0.85	0.383	**0.924**
24	0.14	0.79	0.417	**0.909**	0.30	0.85	0.383	**0.924**
25	0.14	0.79	0.417	**0.909**	0.30	0.85	0.383	**0.924**
26	0.14	0.79	0.417	**0.909**	0.29	0.85	0.383	**0.924**
27	0.13	0.79	0.417	**0.909**	0.28	0.84	0.383	**0.924**
28	0.13	0.79	0.417	**0.909**	0.28	0.84	0.383	**0.924**
29	0.11	0.76	0.417	**0.909**	0.26	0.83	0.383	**0.924**
30	0.05	0.69	0.417	**0.909**	0.18	0.79	0.383	**0.924**
31	0.05	0.69	0.417	**0.909**	0.14	0.78	0.383	**0.924**
32	0.05	0.69	0.417	**0.909**	0.13	0.78	0.383	**0.924**
33	0.05	0.69	0.417	**0.909**	0.13	0.78	0.383	**0.924**
34	0.05	0.69	0.417	**0.909**	0.13	0.78	0.383	**0.924**
35	0.05	0.69	0.417	**0.909**	0.13	0.78	0.383	**0.924**
36	0.05	0.69	0.417	**0.909**	0.13	0.78	0.383	**0.924**
37	0.05	0.69	0.417	**0.909**	0.13	0.78	0.383	**0.924**
38	0.05	0.69	0.417	**0.909**	0.13	0.77	0.383	**0.924**
39	0.05	0.69	0.417	**0.909**	0.13	0.77	0.383	**0.924**
40	0.03	0.61	0.416	**0.910**	0.07	0.66	0.383	**0.924**
41	0.03	0.61	0.416	**0.910**	0.06	0.65	0.383	**0.924**
42	0.03	0.61	0.416	**0.910**	0.06	0.65	0.383	**0.924**
43	0.03	0.61	0.416	**0.910**	0.06	0.65	0.383	**0.924**
44	0.03	0.61	0.416	**0.910**	0.06	0.64	0.383	**0.924**
45	0.03	0.61	0.416	**0.910**	0.06	0.64	0.383	**0.924**
46	0.03	0.61	0.416	**0.910**	0.06	0.64	0.383	**0.924**
47	0.03	0.61	0.416	**0.910**	0.06	0.64	0.383	**0.924**
48	0.03	0.61	0.416	**0.910**	0.06	0.64	0.383	**0.924**
49	0.03	0.61	0.416	**0.910**	0.06	0.62	0.383	**0.924**
50	0.01	0.51	0.411	**0.912**	0.03	0.49	0.383	**0.924**
51	0.01	0.51	0.411	**0.912**	0.03	0.49	0.383	**0.924**
52	0.01	0.51	0.411	**0.912**	0.03	0.49	0.383	**0.924**
53	0.01	0.51	0.411	**0.912**	0.03	0.49	0.383	**0.924**
54	0.01	0.51	0.411	**0.912**	0.02	0.47	0.383	**0.924**
55	0.01	0.51	0.411	**0.912**	0.02	0.47	0.383	**0.924**
56	0.01	0.51	0.411	**0.912**	0.02	0.47	0.383	**0.924**
57	0.01	0.51	0.411	**0.912**	0.02	0.47	0.383	**0.924**
58	0.01	0.51	0.411	**0.912**	0.02	0.47	0.383	**0.924**
59	0.01	0.51	0.411	**0.912**	0.02	0.47	0.383	**0.924**
60	0.01	0.51	0.411	**0.912**	0.02	0.47	0.383	**0.924**

Note: pbis = point biserial correlation. Values in boldface indicate dominate rotated factor loadings

DIMTEST only generated result for Time 3, which failed to reject null hypothesis of essential unidimensionality. Since DIMTEST failed to give a solution at Time 1 and Time 2, DETECT was applied instead. Small DETECT index indicated little amount of departure from unidimensionality. DETECT index for Counting Aloud ranged from .005 to .006, much smaller than the .2 criterion recommended by Roussos and Ozbek (2005). Therefore, unidimensionality was not rejected for Counting Aloud.

Number Naming

Approximate chi-square statistics consistently rejected the one-dimensional model across three time points with small p-values and large percentage of SE >|2|. Approximate chi-square statistics stopped rejecting the two-factor model for Time 2 and Time 3 with non-significant p-values and small percentage of SE >|2| (<5%). For Time 1, however, the chi-square test kept rejecting the model with small p-values even when the number of factors was increased to four. Yet the value of the chi-square statistic did not change much from the three-factor model (χ^2_{592} =972.96) to the four-factor model (χ^2_{557} =927.28) and that the percentage of SE >|2| was not large for the 3-factor model. So it should be reasonable to stop at 3-factor model since additional factor added did not bring much improvement in the model fit. Factor loadings for the two- and three-factor model at Time 1 and two-factor model at Time 2 and Time 3 were then explored to find a consistent solution for factorial structure of Number Naming from NOHARM (Table 4). For Time 1, the two- and three-factor loading patterns were quite similar (only 4 items loaded on different factors in two- and three-factor models), implying that item responses could be modeled in a two-factor structure generated by NOHARM. NOHARM also suggested two-factor models for Time 2 and Time 3, but the two-factor loadings across three time points were inconsistent and difficult to interpret. The items did not seem to be separated by item difficulty or discrimination.

Though NOHARM suggested the existence of a second latent factor, it could not be identified based on factor loadings.

Table 4. Two- and three-factor varimax-rotated loadings and item Discrimination and Difficulty estimates for Number Naming

	Time 1							Time 2				Time 3			
Item	p-value	pbis	2-Factor loading		3-Factor loading			p-value	pbis	2-Factor loading		p-value	pbis	2-Factor loading	
1	0.17	0.65	**0.775**	0.419	**0.685**	0.318	0.468	0.31	0.64	**0.726**	0.422	0.41	0.57	**0.670**	0.403
2	0.47	0.58	**0.746**	0.427	**0.666**	0.360	0.413	0.68	0.50	**0.819**	0.216	0.72	0.51	**0.808**	0.390
3	0.34	0.67	**0.868**	0.372	**0.782**	0.287	0.450	0.54	0.62	**0.917**	0.206	0.66	0.59	**0.813**	0.481
4	0.35	0.66	**0.814**	0.417	**0.734**	0.342	0.433	0.54	0.62	**0.914**	0.240	0.63	0.64	**0.873**	0.487
5	0.32	0.68	**0.955**	0.294	**0.833**	0.157	0.531	0.52	0.65	**0.899**	0.299	0.63	0.60	**0.750**	0.552
6	0.28	0.73	**0.921**	0.389	**0.788**	0.244	0.565	0.49	0.65	**0.914**	0.275	0.59	0.63	**0.787**	0.532
7	0.18	0.65	**0.842**	0.304	**0.787**	0.283	0.341	0.33	0.68	**0.733**	0.512	0.41	0.70	0.597	**0.706**
8	0.18	0.75	**0.915**	0.402	**0.824**	0.364	0.433	0.29	0.73	**0.787**	0.540	0.46	0.72	0.660	**0.751**
9	0.19	0.76	**0.904**	0.427	**0.797**	0.347	0.495	0.33	0.75	**0.861**	0.508	0.43	0.72	0.560	**0.770**
10	0.11	0.68	**0.810**	0.450	**0.748**	0.465	0.333	0.23	0.73	**0.773**	0.557	0.29	0.75	0.524	**0.805**
11	0.07	0.62	**0.655**	0.603	0.635	**0.643**	0.217	0.18	0.72	**0.733**	0.591	0.27	0.73	0.420	**0.852**
12	0.07	0.62	**0.837**	0.443	**0.855**	0.500	0.135	0.16	0.69	**0.701**	0.608	0.26	0.74	0.434	**0.865**
13	0.04	0.50	**0.692**	0.477	**0.662**	0.474	0.256	0.08	0.60	**0.707**	0.575	0.16	0.71	0.275	**0.961**
14	0.03	0.58	0.619	**0.784**	0.599	**0.772**	0.210	0.05	0.58	**0.809**	0.588	0.12	0.67	0.297	**0.955**
15	0.05	0.59	**0.751**	0.551	**0.799**	0.593	0.098	0.09	0.65	**0.765**	0.584	0.17	0.73	0.288	**0.958**
16	0.02	0.53	0.576	**0.817**	0.610	**0.785**	0.107	0.05	0.55	**0.938**	0.347	0.11	0.69	0.339	**0.941**
17	0.04	0.54	**0.782**	0.499	**0.846**	0.530	0.051	0.08	0.62	**0.726**	0.602	0.15	0.71	0.259	**0.966**
18	0.04	0.62	0.643	**0.765**	0.668	**0.733**	0.129	0.08	0.66	**0.841**	0.542	0.13	0.72	0.308	**0.951**
19	0.04	0.58	**0.887**	0.457	**0.841**	0.530	0.103	0.08	0.63	**0.870**	0.492	0.12	0.65	0.216	**0.976**
20	0.04	0.54	**0.705**	0.573	**0.769**	0.638	0.023	0.06	0.58	**0.912**	0.409	0.11	0.66	0.271	**0.962**
21	0.03	0.54	0.534	**0.844**	0.563	**0.817**	0.125	0.07	0.60	**0.756**	0.574	0.09	0.68	0.422	**0.907**
22	0.02	0.49	0.401	**0.915**	0.339	**0.871**	0.355	0.05	0.55	0.501	**0.815**	0.09	0.61	0.586	**0.728**
23	0.04	0.56	0.545	**0.763**	0.399	**0.793**	0.459	0.12	0.68	0.492	**0.871**	0.13	0.71	0.635	**0.773**
24	0.02	0.56	0.503	**0.864**	0.464	**0.841**	0.277	0.09	0.67	0.469	**0.883**	0.08	0.64	**0.766**	0.643
25	0.02	0.49	0.442	**0.896**	0.401	**0.872**	0.280	0.08	0.63	0.407	**0.914**	0.09	0.68	**0.734**	0.679
26	0.02	0.56	0.503	**0.864**	0.464	**0.841**	0.277	0.08	0.64	0.410	**0.912**	0.10	0.70	**0.718**	0.696
27	0.02	0.49	0.413	**0.910**	0.369	**0.887**	0.276	0.08	0.67	0.493	**0.870**	0.09	0.65	**0.757**	0.654
28	0.01	0.44	0.372	**0.928**	0.314	**0.893**	0.323	0.08	0.63	0.449	**0.894**	0.10	0.70	0.629	**0.777**
29	0.02	0.56	0.503	**0.864**	0.464	**0.841**	0.277	0.08	0.65	0.444	**0.896**	0.10	0.69	**0.713**	0.701
30	0.02	0.55	0.462	**0.886**	0.418	**0.855**	0.306	0.05	0.51	0.297	**0.955**	0.07	0.59	**0.796**	0.605
31	0.01	0.37	0.343	**0.938**	0.309	**0.919**	0.244	0.04	0.53	**0.770**	0.563	0.07	0.62	0.541	**0.841**
32	0.43	0.52	0.426	**0.676**	0.322	**0.558**	0.465	0.62	0.42	0.376	**0.561**	0.70	0.36	**0.459**	0.265
33	0.28	0.59	0.515	**0.680**	0.246	0.352	**0.903**	0.43	0.51	**0.645**	0.267	0.52	0.54	**0.568**	0.462
34	0.64	0.39	0.175	**0.833**	0.096	**0.775**	0.368	0.83	0.35	0.309	**0.951**	0.87	0.22	**0.603**	-0.022
35	0.61	0.41	0.316	**0.725**	0.245	**0.669**	0.365	0.75	0.21	0.048	**0.555**	0.75	0.30	**0.375**	0.235
36	0.26	0.54	0.516	**0.578**	0.339	0.326	**0.681**	0.42	0.51	**0.629**	0.273	0.50	0.54	**0.518**	0.477
37	0.21	0.59	0.497	**0.666**	0.303	0.424	**0.719**	0.46	0.41	**0.567**	0.205	0.40	0.40	0.370	**0.388**
38	0.82	0.33	0.422	**0.621**	0.383	**0.579**	0.311	0.88	0.26	0.291	**0.655**	0.91	0.17	**0.629**	-0.077

Note: pbis = point biserial correlation. Values in boldface indicate dominate rotated factor loadings

DIMTEST also rejected essential unidimensionality at Time 2 and Time 3 at the .05 alpha level, but failed to reject it at Time 1. However, the DIMTEST p values at Time 2 and Time 3 were close to be nonsignificant had Bonferroni adjustment been applied to account for the

number of tests conducted. DETECT indices were also not large (< .2), indicating essential unidimensionality.

NOHARM suggested that strict unidimensionality did not hold for Number Naming. However, the pattern of factor loadings for the two-factor model was inconsistent across time points and the factors were uninterpretable. DIMTEST and DECTEC indicated that the assumption of essential unidimensionality was not violated for Number Naming. Thus, further verification is needed before a definitive conclusion of dimensionality for Number Naming can be reached.

Counting Objects

NOHARM failed to generated results for both forms at Time 1. When solutions were available, NOHARM supported unidimensionality for Form A at Time 2 and for Form B at Time 2 and Time 3, but rejected it for Form A at Time 2. At Time 2, p-value of approximate chi-square statistic was still significant when three factors were included in the model, but percentage of large standardized residuals was small (5.3%), implying that no additional factor was needed. The varimax-rotated factor loadings for Form A (Table 5) were examined to find the cause of the rejection of unidimensionality at Time 2. The pattern of 3-factor loadings at Time 2 was not informative, but the 2-factor loading patterns was quite similar to that at Time 3, with the first six items loading on the first factor, and remaining items loading on the second factor. The item difficulty index showed that items loading on the first factor were easier than those loading on the second factor, which suggested that the factors were separated by difficulty. As the approximate chi-square test failed to reject the unidimensional model for Form B and for Form A at Time 3 and that NOHARM is known to have a tendency to produce difficulty factors, we tentatively conclude that the unidimensional model is plausible for Counting Objects.

Table 5. Two- and three-factor varimax-rotated loadings and item Discrimination and Difficulty estimates for Counting Objects Form A

Form A	Time 2							Time 3			
Item	p-value	pbis	2-Factor loading		3-Factor loading			p-value	pbis	2-Factor loading	
1	0.96	0.25	**0.968**	0.251	**0.873**	0.487	-0.004	0.97	0.23	**0.735**	0.463
2	0.93	0.31	**0.999**	0.037	**0.751**	0.633	0.184	0.96	0.20	**0.713**	0.049
3	0.81	0.38	**0.706**	0.191	0.504	**0.531**	0.167	0.88	0.31	**0.672**	0.125
4	0.58	0.47	**0.835**	0.130	0.135	**0.977**	0.159	0.71	0.43	**0.820**	0.101
5	0.50	0.52	**0.676**	0.362	0.229	**0.549**	0.462	0.62	0.45	**0.952**	0.004
6	0.37	0.53	**0.614**	0.480	0.380	**0.479**	0.468	0.51	0.57	**0.718**	0.413
7	0.32	0.62	0.545	**0.688**	0.164	0.404	**0.865**	0.49	0.54	0.523	**0.586**
8	0.35	0.61	0.423	**0.816**	0.434	0.263	**0.743**	0.44	0.68	0.579	**0.762**
9	0.28	0.57	0.469	**0.648**	**0.750**	0.288	0.439	0.35	0.67	0.587	**0.752**
10	0.24	0.61	0.606	**0.626**	**0.547**	0.451	0.530	0.34	0.63	0.446	**0.778**
11	0.04	0.54	0.291	**0.957**	0.642	0.181	**0.744**	0.11	0.64	0.450	**0.872**
12	0.10	0.64	0.468	**0.884**	0.440	0.284	**0.852**	0.21	0.73	0.576	**0.817**
13	0.04	0.55	0.448	**0.894**	0.395	0.268	**0.878**	0.13	0.64	0.379	**0.904**
14	0.05	0.50	0.494	**0.782**	0.166	0.290	**0.942**	0.09	0.57	0.293	**0.906**
15	0.02	0.43	0.089	**0.996**	**0.800**	0.075	0.594	0.03	0.46	0.022	**1.000**
16	0.02	0.41	0.145	**0.989**	0.657	0.010	**0.752**	0.07	0.57	0.309	**0.920**
17	0.01	0.29	0.261	**0.757**	0.558	0.158	**0.569**	0.04	0.48	-0.002	**1.000**
18	0.01	0.33	0.295	**0.806**	0.120	0.029	**0.990**	0.02	0.45	0.104	**0.995**
19	0.01	0.24	0.095	**0.825**	**0.917**	0.256	0.290	0.02	0.41	0.297	**0.850**
20	0.01	0.38	0.107	**0.994**	**0.789**	0.098	0.603	0.03	0.40	-0.112	**0.994**

Note: pbis = point biserial correlation. Values in boldface indicate dominate rotated factor loadings

DIMTEST provided more evidence for unidimensionality for Counting Objects with consistent non-significant test statistics for both forms across all three time points. Thus, it was concluded that unidimensionality of responses to Counting Objects was not violated.

Measurement

NOHARM failed to reject one-dimensional model for both forms across three time points with non-significant approximate chi-square statistics and small percentage of large standardized residuals.

DIMTEST also supported the null hypothesis of essential unidimensionality except for Form A at Time 3. As the p value for Form A at Time 3 was close to .05, the test would fail to reject undimensionality had Bonferroni adjustment been applied to account for the number of tests conducted. Therefore, unidimensionality for Measurement appears to hold.

Pattern Recognition

Discrepancy among programs was found in the results for Pattern Recognition. The non-significant approximate chi-square statistics and small percentage of large standardized residuals from NOHARM were in favor of one-dimensional model for both forms across all three time points. DIMTEST rejected essential unidimensionality for Form A across all three time points and for Form B at Time 2. However, p-values of DIMTEST were non-significant had Bonferroni adjustment been applied to account for the number of tests conducted. DETECT indexes suggested existence of multidimensionality for both forms across all three time point. The D(P) values fluctuated considerably (0.26 – 1.43) across time points, indicating weak to strong multidimensionality. Yet, the r values were not large enough to indicate approximate simple structure for the data, in which case it was not surprising to see the inconsistent item classifications across time points (Table 6 & Table 7). Though DETECT indexes indicated some degree of multidimensionality, the unstable item classification was not informative enough to tell us the dimensional structure of the data. Considering the support from NOHARM, further evidence is needed before rejecting unidimensionality for Pattern Recognition.

Table 6. Cluster Classification for Pattern Recognition Form A

Clusters	Item Classification			Classification Consistency			
	Time 1	Time 2	Time 3	Time 1&2	Time 1&3	Time 2&3	Overall
Cluster 1	1,7,8,916,18	1,2,5,8,9,15,16	1,2,3,4,8,16,17,20				
Cluster 2	2,3,4,5,611,14,15,19,20	3,19,17	5,6,9,10,11,12,13,14,18, 19	30.0%	40.0%	30.0%	15.0%
Cluster 3	10,12,13,17	4,7,12,20	7,15				
Cluster 4		6,13,14,18					
Cluster 5		19					
Cluster corr.	0.77 – 0.80	0.44 – 0.73	0.54 – 0.67				

Note: The consistency was calculated as the percentage of items consistently partitioned into the same dimension across the samples.
Cluster corr. was range of the correlation between every two clusters identified for each time point.

Table 7. Cluster Classification for Pattern Recognition Form B

Clusters	Item Classification			Classification Consistency			
	Time 1	Time 2	Time 3	Time 1&2	Time 1&3	Time 2&3	Overall
Cluster 1	1,2,4,7,8,911,14,15,18	1,2,4,6,7,8,913	1,4,11,14,18				
Cluster 2	3,5,6,10,12,16,17,19,20	3,5,10,12,15,16,17,19,20	2,5,6,7,810,12,20	70.0%	55.0%	30.0%	30.0%
Cluster 3	13	11,14,18	3,9,13,15,16,1719				
Cluster corr.	0.41 – 0.78	0.66 – 0.69	0.78 – 0.83				

Note: The consistency was calculated as the percentage of items consistently partitioned into the same dimension across the samples.
Cluster corr. was range of the correlation between every two clusters identified for each time point.

Chapter 4

Discussion and Implication

Early academic assessment is an effective way to identify children at risk at early age and prevent disparity among children at school entry. To achieve these purposes, it is essential that the assessment has sound psychometric characteristics with respect to reliability and validity. Dimensionality examination is a critical but often overlooked process in test evaluation. The current study attempts to gather evidence for construct validity and appropriateness of using EARLI numeracy skill measures.

Though NOHARM rejected unidimensionality of Counting Aloud across all three times, it was reasonable to believe that assumption of unidimensionality was not violated as suggested by DIMTEST and DETECT results. Counting Aloud was scored in a way that item responses followed a perfect Guttman pattern. Children need to count the numbers in order without skipping any number or counting out of order, which established a one-dimensional continuum for the measure. Therefore, the conclusion of unidimesnionality for Counting Aloud is plausible.

Counting Objects received quite consistent supporting evidence for unidimensionality from all three assessment procedures selected in this study. Though NOHARM rejected one-dimensional model for Form A at Time 2, the pattern of factor loadings revealed that the items were separated by difficulty. Difficulty factor usually results from the factorization of a correlation matrix, in which case "a set of items or tests which differ 'widely' in difficulty level may yield a reduced covariance matrix of rank greater than unity, even when the set is 'really' unidimensional" (McDonal, 1967, p.55). As difficulty factor is often a methodological artifact, it should not be counted as an additional dimensionality.

For Number Naming, NOHARM chi-square test favored a two-factor model for Time 2 and Time 3 but suggested that additional factors were needed for Time 1. But a close look at the factor-loadings for Time 1 suggested that a two-factor model was enough to represent the underlying structure. However, results from DIMTEST and DETECT did not quite agree with that of NOHARM. DIMETEST p-values were close to be non-significant had Bonferroni adjustment been applied and that small DETECT indexes across all three time points suggested essential unidimensionality. NOHARM factor loadings did not indicated difficulty factors. There are a few shape recognition items in Number Naming. However, inconsistent NOHARM factor loadings did not indicate that the items were separated by numbers and shapes. Even though NOHARM suggested the presence of a second factor, the proportion of the items that require the additional factor to get the correct response was not large, since DIMTEST and DETECT did not reject essential unidimensionality. DIMTEST did not indicate the presence of a second dimension until the items measuring the second dimension exceeded one-third of the number of items measuring the dominant dimension (Elias, Hattie, & Douglas, 1998).

Considerable discrepancy in results was found between programs for Pattern Recognition. DETECT indexes suggested weak to strong degree of multidimensionality for both forms across time points, whereas NOHARM consistently supported unidimensionality. Though NOHARM usually has good performance without a pseudo-guessing parameter and its results in this case were consistent across test forms and time points, the considerable departure from unidimensionality suggested by DETECT should not be neglected. Tate (2003) reported that DETECT was able to make correct decisions of essential unidimensionality for all of the unidimensional cases (the cases varied in item difficulty and item discrimination) with sample size 2000 and 60 test items. It was not certain if DETECT could keep up the good performance with such small sample size in our case though. No study was found discussing the conditions under which DETECT might generate large index (>.2) for actually unidimensional data.

Therefore, dimensionality for Pattern Recognition remains inconclusive before further evidence is collected.

Disagreement between NOHARM and DETECT results has been noted by Finch and Habing (2005). The authors compared the performance of NOHARM and DETECT in identifying the number of dimensions and in allocating items into different dimensions. They concluded that NOHARM was generally more likely to group together items that were supposed to be separated, while DETECT tended to separate items that should be kept together. We also observed similar disagreement between NOHARM and DETECT results in this study. When the DETECT D(P) index supported essential unidimensionality (e.g., Number Naming), NOHARM factor loadings suggested difficulty factors. On the other hand, when NOHARM approximate chi-square tests failed to reject unidimensionality (e.g., Pattern Recognition), the DETECT indexes suggested weak to strong multidimensionality with inconsistently classified clusters of items. Yet in both cases, there was not enough evidence to choose one program over the other, suggesting the need for future study.

However, unstable DETECT performance across time points was noticeable in the current study. D(P) values and item classifications fluctuated considerably across time points, while the r index was unresponsive with pretty low value (except for Counting Aloud). As mentioned in literature review, DETECT did not work well under certain conditions such that the classification accuracy and consistency of DETECT decrease when data displays complex structure with high inter-dimension correlations and when item with low discrimination powers are involved. Item discrimination power should not be a concern of this study because EARLI numeracy items showed moderate to high item discrimination. The correlations between clusters identified by DETECT for Pattern Recognition are displayed in Table 6 and Table 7. Though some cluster correlations were a little high (i.e., Form A at Time 1 and Form B at Time 3), no extremely high correlation (>.90) was found. Therefore, item discrimination and inter-dimension

correlation do not seem to explain the inconsistent DETECT performance. Small r index was supposed to indicate complex structure of the data. However, unlike simulation studies, the true underlying dimensional structure in this study was unknown, so it was difficult to tell whether the small r index in this study was true representation of complex structure or the result of small sample size. Tan and Gierl (2006) found that the magnitude of r index was also influenced by sample size such that as sample size increased, the r index increased. Perhaps the small r indexes observed in this study are due to small sample size.

Apparently, small sample size is the primary limitation of this study, which is a possible cause of the discrepancies among programs and inconsistencies across time points. It is expected that the power and accuracy of these statistical indices decrease as sample size decreases. Though the programs selected in this study are recommended in a few studies for their comparatively desirable performance with small sample size, these findings are not enough to tell if the performance of these dimensionality analysis procedures are accurate or reliable for real data set with sample size as small as 100. Particularly, it is important to note that there could be many unexpected "noise" in real data, which makes the results less interpretable. Therefore, more empirical study is needed to provide guidelines for dimensionality assessment with small samples. It would also be desirable to increase sample size in future research.

Besides, the data in this study were collected with children enrolled in Head Start, which may not be representative of preschool population. Ackerman (1994) contended that the dimensionality of an assessment was the result of the interaction of the examinees with the items, which means certain property of the items may affect the performance of some examinees to a greater extent than others given individuals' unique characteristics. Children from other settings could be different from children enrolled in Head Start with respect to their abilities and other characteristics, which may results in different interaction with the test items. Therefore, results of the current study should not be generalized to other populations and it would be advantageous to

replicate the study with a representative preschool sample for more generalizable results concerning dimensionality.

Test validation is an on-going process requiring cumulative evidence from different aspects. Preliminary evidence of unidimensionality was obtained for some of the EARLI numeracy measures (Counting Aloud, Counting Object, and Measurement), which can be viewed as contributing evidence for construct validity of the EARLI numeracy measures. However, the dimensionality of Number Naming and Pattern Recognition remained inconclusive. Further studies with larger sample sizes, using different assessment procedures or on different groups are needed to collect more evidence of unidimensionality.

References

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items are measuring. Applied Measurement in Education, 7, 255-278.

AERA, APA, & NCME (1999). Standards for educational and psychological testing. Washington, D. C.: Author.

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

Aubrey, C., Dahl, S., & Godfret, R. (2006). Early mathematical development and later achievement: Further evidence. Mathematics Education Research Journal, 18(1), 27-46.

Bagnato, S. J., Neisworth, J. T., & Munson, S. M. (1993). Sensible strategies for assessment in early intervention. In D. M. Bryant & M. A. Graham (Eds.), Implementing early intervention. New York: Guilford Press.

Banerji, M., Smith, R.M. & Dedrick, R.F. (1997). Dimensionality of an early childhood scale using Rasch analysis and confirmatory factor analysis. Journal of Outcome Measurement, 1(1), 56-85.

Benson, J., & Clark, F. (1982). A guide for instrument development and validation. The American Journal of Occupational Theory, 36, 789-800.

Blais, J. & Laurier, M. (1995). Methodological considerations in using DIMTEST to assess unidimensionality. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April.

Bock, R. D., Gibbons, R., &Muraki, E. (1988). Full-information item factor analysis. Applied Psychological Measurement, 12, 261–280.

Bowman, B. T., Donovan, M. S., & Burns, M. S. (Eds.). (2001). Eager to learn: Educating our preschoolers. Washington, DC: National Academy Press.

Bracken, B. A., & Nagle, R. (2007). Psychoeducational assessment of preschool children

(4th ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers

Brenneman, K., Stevenson-Boyd, J.S., & Frede, E. (2009). Math and science in preschool: Policies and practice. Preschool Policy Matters, Issue 19. New Brunswick, NJ: National Institute for Early Education Research.

Chard, D., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number-sense to screen for difficulties in mathematics: Preliminary findings. Assessment for Effective Intervention, 30, 3-14.

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. School Psychology Review, 33(2), 234-248.

Daly, E. J., Wright, J. A., Kelly S. Q., & Martens, B. K. (1997). Measures of early academic skills: Reliability and validity with a first grade sample. School Psychology Quarterly, 12, 268-280.

De Ayala, R. J., & Hertzog, M.. A. (1980). A comparison of methods for assessing dimensionality for use in item response theory. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

De Champlain, A., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. Applied Measurement in Education, 11, 231-253.

De Champlain, A. F.,&Tang, K. L. (1997). CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model Educational and Psychological Measurement, 57, 174-178.

De Champlain, A. F., & Gessaroli, M. E. (1996). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

DiPerna, J. C., Morgan, P. L., & Lei, P. (2006). Development of Early Arithmetic, Reading, and Learning Indicators for Head Start (The EARLI Project). Semi-annual performance report to the U.S. Department of Health and Human Services Administration for Children and Families. University Park: The Pennsylvania State University, College of Education.

Duncan, G.J., Dowsett, C. J., & Claessens, A. (2007). School readiness and later achievement. Developmental Psychology, 43(6), 1428-1446.

Elias, S., Hattie, J., & Douglas, G. (1998). An assessment of various item response model and structural equation model fit indices to detect unidimensionality. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Entwisle, D.R., & Alexander, K.L. (1990). Beginning school math competence: minority and majority comparisons. Child Development, 61, 454-471.

Feinstein, L., & Duckworth, K., (2006). Development in the early years: its importance for school performance and adult outcomes. Wider Benefits of Learning Research Report No 20, Institute of education, London.

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality. Applied Psychological Measurement, 31, 292-307.

Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in Item Cluster Recovery: Counting dimensions and allocating items. Journal of Educational Measurement, 42(2), 149-169.

Floyd, R. G., Hojnoski, R. L., & Key, J. (2006). Preliminary evidence of technical adequacy of the Preschool Numeracy Indicators. School Psychology Review, 35, 627-644.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. Journal of Special Education, 41, 121-139.

Folske, J. C., Gessaroli,M. E., & De Champlain, A. F. (1998). Comparing a likelihood-ratio chi-

square statistic and DIMTEST in conditions of correlated proficiencies and pseudo-guessing. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Fox, J. L. & Diezmann, C. M. (2007). What counts in research? A survey of early years' mathematical research 2000-2005. Contemporary Issues in Early Childhood, 8(4), 301-312.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 23, 267-269.

Frederiksen, N. (1986). Construct validity and construct similarity: methods for use in test development and test validation. Multivariate Behavioral Research, 21, 3-28.

Froelich, A. G. (2000). Assessing unidimensionality of test items and some asymptotics of parametric item response theory. Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.

Gessaroli, M. E. (1994). The assessment of dimensionality via local and essential independence: a comparison in theory and practice. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), Modern Theories in Measurement: Problems and Issues. Ottawa: University of Ottawa.

Gierl, M. J., Leighton, J. P., Tan, X. (2006). Evaluating DETECT classification accuracy and consistency when data display complex structure. Journal of Educational Measurement, 43, 265-289.

Ginsburg, H. P. , & Allardice, B. S. (1984). Children's difficulties with school mathematics. In J. Lave & B. Rogoff (Eds.), Everyday cognition: Its development in social context (pp. 194-219). Cambridge, MA: Harvard University Press.

Ginsburg, H. P., Klein, A., & Starkey, P. (1998). The development of children's mathematical knowledge: Connecting research with practice. In I. E. Sigel & K. A.

Renninger (Eds.), Handbook of child psychology: Vol. 4. Child psychology in practice (5th Ed., pp. 401–476). New York: Wiley & Sons.

Godwin, W. L., & Driscoll, L. A. (1980). Handbook for measurement and evaluation in early childhood education: Issues, measures, and methods. San Francisco: Jossey-Bass.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139 –1 64.

Henning, G. 1988: The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. Language Testing 5, 83-99.

Hochstedt, K. S., Lei, P. W., DiPerna, J. C., & Morgan, P. L. (2010). Examining the dimensionality of EARLI literacy skill scores using nonlinear factor analysis. Journal of Psychoeducational Assessment.

Hojnoski, R., & Missall, K. (2006). Addressing school readiness: Expanding school psychology in early education. School Psychology Review, 35, 602-614.

Hsieh, M. (2010). Assess unidimensionality of computerized reading comprehension and math tests. International journal of intelligent technology and applied statistics, 3(1), 93-105.

Hughes, M. (1986). Children and Number: difficulties in learning mathematics. New York: Basil Blackwell.

Humphreys, L. (1984). A theoretical and empirical study of the psychometric assessment of psychological test dimensionality and bias (ONR Research Proposal). Washington, DC: Office of Naval Research.

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using

conditional covariance-based nonparametric approach. Journal of Educational Measurement, 44, 1-22.

Jasper, F. (2010). Applied dimensionality and test structure assessment with the START-M Mathematics Test. The International Journal of Educational and Psychological Assessment, 6(1), 104-125.

Jordan, N., Kaplan, D. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. Developmental Psychology, 45(3), 850-867.

Joyce, B. G., & Wolking, W. D. (1987). Standardized tests and timed curriculum-based assessments: A comparison of two methods for screening high-risk students. Journal of Psychoeducational Assessment, 5, 185-193.

Kenny, T.J., & Culbertson, J.L. (1993). Developmental screening for preschoolers. In J.L. Culbertson & D.J.Willis (Eds.), Testing young children.Areference guide for developmental, psychoeducational, and psychosocial assessments. Austin, TX: Pro-Ed.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data. (Doctoral dissertation, University of Illinois at Urbana-Champaign). Dissertation Abstracts International, 55-12B, 5598.

Kim, H. R., Zhang, J., & Stout, W. (1995). Anew index of dimensionality—DETECT. Unpublished manuscript.

Kochanoff, A. T., Hirsh-Pasek, K., Newcombe, N. & Weinraub, M. (2003). Using science to inform preschool assessment: A summary report of the Temple University Forum on Preschool Assessment. Temple University's Department of Psychology.

Lago, R. M., & DiPerna, J. C. (2010). Number sense in Kindergarten: A factor-analytic study of the construct. School Psychology Review, 39, 164-180.

Lei , P. W., Wu, Q., DiPerna, J. C., & Morgan. P. L. (2009). Developing short forms of the EARLI Numeracy Measures : Comparison of item selection methods. Educational and

Psychological Measurement. 69, 825.

McDonald, R. P. (1967). Nonlinear factor analysis (PsychometricMonographs,No. 15). The Psychometric Society.

McDonald, R. P. (1986). Describing the elephant: Structure and function in multivariate data. Psychometrika, 4, 513-534.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum.

Magliocca, L. A., Rinaldi, R. T., Crew, J. L., & Kunzelmann, H. P. (1977). Early identification of handicapped children through a frequency sampling technique. Exceptional Children, 43, 414—420.

Mazzocco, M. M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. Learning Disabilities Research & Practice, 20, 142–155.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into scoring meaning. American Psychologist, 9, 741-749.

Methe, S. A., Hintze, J. M., & Floyd, R. G. (2008). Validation and decision accuracy of early numeracy skill indicators. School Psychology Review, 37, 359-373.

Monahan, P., Stump, T. E., Finch, H., & Hambleton, R. K. (2007). Bias of exploratory and cross-validated DETECT index under unidimensionality. Applied Psychological Measurement. 31, 483.

Nandakumar, R. (1991). Traditional dimensionality vs. essentialdimensionality. Journal of Educational Measurement, 28, 99-117.

Nandakumar, R. (1994). Assessing the dimensionality of a set of item responses—Comparison of different approaches. Journal of Educational Measurement, 31, 17-35.

Perry, B. & Dockett, S. (2002). Young children’s access to powerful mathematical ideas. In L. English (Ed.) Handbook of International Research in Mathematics Education. Mahwah,

NJ: Lawrence Erlbaum.

Pungello, E.P., Kupersmidt, J.B., Burchinal, M.R., & Patterson, C.J. (1996). Environmental risk factors and children's achievement from middle childhood to early adolescence. Developmental Psychology, 32(4), 755-767.

Pyo, K. H. (2000). Assessing dimensionality of a set of language test data. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Reckase, M. (1985). The di!culty of items that measure more than one ability. Applied Psychological Measurement, 9 (5), 401–412.

Reid, E. E., Morgan, P. L., DiPerna, J. C., & Lei, P.W. (2006). Development of measures to assess young children's early academic skills: Preliminary findings from a Head Start-university partnership. Insights on Learning Disabilities, 3(2), 25-38.

Roussos, L. A., Sout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. Journal of Educational Measurement. 35(1), 1-30.

Roussos, L. A., & Ozbek, O. (2005). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. Manuscript under review.

Roussos, L. A., & Ozbek, O. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. Journal of Educational Measurement, 43(30), 215-243.

Sadler, P. M., & Tai, R. H. (2007). The two high-school pillars supporting college science. Science, 317, 457–458.

Shore, T. H. and Thornton, G. C., & Shore, L. (1990). Construct validity of two categories of assessment center dimension ratings. Personnel Psychology, 43, 1-12.

Spodek, B., & Saracho, O. N. (1997). Issues in early childhood educational assessment and

evaluation. New York: Teachers College, Columbia University.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Stout, W. F., Douglas, J., Junker, B., & Roussos, L.A. (1993). DIMTEST manual. Unpublished manuscript available from W. F. Stout, University of Illinois at Urbana-Champaign, Champaign.

Stout, W., Froelich, A., & Gao, F. (2001). Using resampling to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijin, & T. A. B. Snijders (Eds), Essays on item response theory. New York: Springer-Verlag.

Tan, X., & Gierl, M. J. (2006). Evaluating the consistency of DETECT indices and item clusters using simulated data that display both simple and complex structure. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Tindal, G., & Haladyna, T. M. (2002). Large-scale assessment programs for all students: Validity, technical adequacy, and implementation. Mahwah, N.J: L. Erlbaum.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. Applied Psychological Measurement, 27, 159-203.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. Educational and Psychological Measurement, 56, 197–208.

VanDerHeyden, A. M., Broussard, C., & Cooley, A. (2006). Further development of measures of early math performance for preschoolers. Journal of School Psychology, 44(6), 533-553.

VanDerHeyden, A. M., Broussard, C., Fabre, M., Stanley, J., LeGrendre, J., & Creppell, R.

(2004). Development and validation of curriculum-based measures of math performance for preschool children. Journal of Early Intervention, 27, 27-41.

VanDerHeyden, A.M.,Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. School Psychology Review, 30, 363−382.

Wainer, H., & Wang, X. (2001). Using a new statistical model for testlets to score TOEFL (TOEFL Technical Report No. 16). Princeton, NJ: Educational Testing Service.

Wang, S. (2006). Validation and Invariance of Factor Structure of the ECPE and MELAB across Gender, Spaan Fellow Working Papers in Second or Foreign Language Assessment, 4, 41-56.

Whitehurst, G.J. & Lonigan, C.J. (1998) 'Child development and emergent literacy', Child Development, 69, 848-872.

Wolfgang, C. H., Stannard, L. L., & Jones, I. (2001). Block play performance among preschoolers as a predictor of later school achievement in mathematics. Journal of Research in Childhood Education, 15(2), 173-180.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. Psychometrika, 64, 213-249.

Appendix 1. DIMTEST AT & PT selection for Counting Aloud

	Time 1	Time 2	Time 3
AT list	1,2,3,4,5,6,7,8,9,10,11,12,13,14, 15,16	1,2,3,4,5,6,7,8,9,10,11,12,13,14, 15,16,17,18,19,20,21,22,23,24, 25,26	23,24,25,26,27,28,29,30,31
PT list	17,18,19,20,21,22,23,24,25,26, 27,28,29,30,31,32,33,34,35,36, 37,38,39,40,41,42,43,44,45,46, 47,48,49,50,51,52,53,54,55,56, 57,58,59,60	27,28,29,30,31,32,33,34,35,36, 37,38,39,40,41,42,43,44,45,46, 47,48,49,50,51,52,53,54,55,56, 57,58,59,60	1,2,3,4,5,6,7,8,9,10,11,12,13,14 15,16,17,18,19,20,21,22,32,33, 34,35,36,37,38,39,40,41,42,43, 44,45,46,47,48,49,50,51,52,53, 54,55,56,57,58,59,60

Appendix 2. DIMTEST AT & PT selection for Number Naming

	Time 1	Time 2	Time 3
AT list	1,2,3,4,5,6,7,8,9,10,11,12,14,15, 17,19,22,23,27	22,23,24,25,26,27,28,29,30	1,22,23,24,25,26,27,28,29,30,31, 32,34,35,38
PT list	11,13,16,18,20,21,24,25,26,28,29, 30,31,32,33,34,35,36,37,38	1,2,3,4,5,6,7,8,9,10,11,12,13,14, 15,16,17,18,19,20,21,31,32,33, 34,35,36,37,38	2,3,4,5,6,7,8,9, 10,11,12,13,14, 15,16,17,18,19,20,21,33,36,37

Appendix 3. DIMTEST AT & PT selection for Counting Object Form A

	Time 1	Time 2	Time 3
AT list	3,4,5,8	6,7,8,9,10	3,4,5,6,
PT list	1,2,6,7,9,10,11,12,13,14,15,16, 17,18,19,20	1,2,3,4,5,11,12,13,14,15,16,17, 18,19,20	1,2,7,8,9,10,11,12,13,14,15,16, 17,18,19,20

Appendix 4. DIMTEST AT & PT selection for Counting Object Form B

	Time 1	Time 2	Time 3
AT list	1,2,7,9,10	3,4,5,7	12,15,17,18
PT list	3,4,5,6,8,11,12,13,14,15,16,17, 18,19,20	1,2,6,8,9,10,11,12,13,14,15,16,17, 18,19,20	1,2,3,4,5,6,7,8,9,10,11,13,14,16, 19,20

Appendix 5. DIMTEST AT & PT selection for Measurement Form A

	Time 1	Time 2	Time 3
AT list	4,5,6,10	9,11,15,17	4,6,10,16,17
PT list	1,2,3,7,8,9,11,12,13,14,15,16, 17,18,19,20	1,2,3,4,5,6,7,8,10,12,13,14,16, 18,19,20	1,2,3,5,7,8,9,11,12,13,14,15,18, 19,20

Appendix 6. DIMTEST AT & PT selection for Measurement Form B

	Time 1	Time 2	Time 3
AT list	3,8,10,13,16	7,8,12,16,18	1,3,6,13,19
PT list	1,2,4,5,6,7,9,11,12,14,15,17,18, 19,20	1,2,3,4,5,6,9,10,11,13,14,15,17, 19,20	2,4,5,7,8,9,10,11,12,14,15,16,17, 18,20

Appendix 7. DIMTEST AT & PT selection for Pattern Recognition Form A

	Time 1	Time 2	Time 3
AT list	1,2,3,8,17	4,7,19,20	2,3,15,16
PT list	4,5,6,7,9,10,11,12,13,14,15,16, 18,19,20	1,2,3,5,6,8,9,10,11,12,13,14,15, 16,17,18	1,4,5,6,7,8,9,10,11,12,13,14,17, 18,19,20

Appendix 8. DIMTEST AT & PT selection for Pattern Recognition Form B

	Time 1	Time 2	Time 3
AT list	2,10,16,20	3,10,12,15,17	4,8,9,11,18
PT list	1,3,4,5,6,7,8,9,11,12,13,14,15, 17,18,19	1,2,4,5,6,7,8,9,11,13,14,16,18, 19,20	1,2,3,5,6,7,10,12,13,14,15,16,17, 19,20