

**A COMPARISON OF EARLY CHILDHOOD ASSESSMENTS  
AND A STANDARDIZED MEASURE  
FOR PROGRAM EVALUATION**

by

Stephanie Hildegard Zadro Jacobson

Dissertation submitted to the Faculty of the  
Virginia Polytechnic Institute and State University  
in partial fulfillment of the requirements for the degree of

**DOCTOR OF PHILOSOPHY**

in

Educational Research and Evaluation

APPROVED:

Marvin G. Cline, Chairperson  
Victoria R. Fu  
Barbara A. Hutson  
Javaid Kaiser  
Ronald L. McKeen

April 17, 1997  
Blacksburg, Virginia

# **A COMPARISON OF EARLY CHILDHOOD ASSESSMENTS AND A STANDARDIZED MEASURE FOR PROGRAM EVALUATION**

by

Stephanie Hildegard Zadro Jacobson

Marvin G. Cline, Chairman  
Educational Research and Evaluation  
(ABSTRACT)

Traditionally, standardized achievement tests have been used to monitor program effectiveness. Recently, however, educators have questioned the appropriateness of standardized tests for this purpose, especially for programs designed for young children. Early childhood advocates suggest using developmentally appropriate assessments instead of standardized achievement tests for making classroom-level decisions about children and for program evaluation. Proponents, however, have not fully identified the psychometric properties of the assessments, certainly not for the purposes of program evaluation.

Although developmentally appropriate assessments have been implemented in a number of classrooms across the country, few studies have verified their ability to discriminate among developmental levels. In addition, even fewer studies have addressed their use for evaluating program effectiveness.

Using the records of 293 students from the local site of a National Transition Project and both classical test theory (CTT) and item response theory (IRT) procedures, three assessment instruments and a standardized test were examined. It was shown that the Concepts about Print portion of the Early Childhood Assessment Package, the Language Arts component of the kindergarten developmental progress reports, and the first grade Early Literacy Scale tasks are, in fact, developmental assessments. Additionally, IRT procedures located students on the developmental continuum underlying the assessments.

Although classical ANCOVAs were unable to identify Treatment or Head Start program effects beyond the kindergarten year, IRT procedures showed that the expected proportion of students at the highest latent ability levels tended to be greater for students in Demonstration schools and Head Start graduates than their counterparts throughout kindergarten and first grade. A standardized reading achievement measure administered to the students in second grade, was unable to differentiate program effects through either classical or IRT procedures. This suggests that the concepts underlying standardized tests differ from those underlying developmentally appropriate assessments. As a result, the key issue to be resolved is which type of measure is more valid, that is, more appropriate, for evaluating early childhood programs.

## **Dedication**

to my parents  
who have always believed that I could do anything I wanted to do  
and have been certain that my accomplishments would far exceed their expectations

## Acknowledgments

A great many people contributed to helping me attain my life long dream of successfully completing this degree in Educational Research and Evaluation and earning the title, Doctor of Philosophy. I am fortunate to have had the support and encouragement of my committee, my family, and the people involved with the Virginia site of the National Head Start Transition Demonstration Project throughout this process.

I would like to thank my committee chairperson, Dr. Marvin G. Cline, for his wisdom and patience. Without his professional expertise, assistance, and reassurance, I would never have been able to complete my program and this dissertation in such a timely manner. I am especially indebted to him for opening up the world of educational research to me and creating unique learning experiences that have been educational as well as personally meaningful.

The members of my committee helped me to probe deeper and clarify my thinking on a number of issues. I want to thank Dr. Victoria Fu for her expertise in observation and assessment in child development. I am grateful to Dr. Barbara Hutson and her provocative questions which helped me to focus my research within the area of literacy development. Dr. Javaid Kaiser provided the theoretical and statistical background for classical measurement theory, and for that I am most appreciative. In addition, I wish to recognize Dr. Ronald McKeen for helping me to view issues from multiple perspectives and to see the value of them all.

Most importantly, I am grateful to my husband, Sherwin, and my children, Jennifer, Jonathan, and Joshua, who have continually supported my quest for knowledge. I thank them for preparing numerous dinners while I attended evening classes, for maintaining our home in Fairfax while I completed my residency in Blacksburg, and for providing emotional support and assurance throughout this process.

# TABLE OF CONTENTS

	Page
Abstract .....	ii
Dedication.....	iii
Acknowledgments .....	iv
Table of Contents .....	v
List of Tables .....	viii
List of Figures .....	xi
CHAPTER I. INTRODUCTION.....	1
A Context for Developmentally Appropriate Assessment Issues .....	1
Purpose of the Research.....	4
Research Questions.....	5
Significance of the Study.....	5
Limitations.....	5
Definition of Terms.....	6
CHAPTER II. REVIEW OF RELATED LITERATURE .....	8
Inappropriateness of Standardized Tests for Monitoring Development .....	8
Narrow definition of emergent literacy.....	9
Lack consideration for literacy development and characteristics of young learners .....	10
Tasks unrelated to classroom practices .....	12
Limited use for instruction.....	13
Developmentally Appropriate Literacy Assessments .....	15
Concepts About Print .....	17
Running Records.....	18
Writing Assessments.....	19
Story Retellings.....	20
Developmental Checklists and Rating Scales.....	20
Psychometric Considerations.....	21

Problems with Classical Test Theory .....	21
Advantages of Item Response Theory.....	22
IRT in Relation to Alternative Assessments .....	23
Conclusions .....	23
Statement of the Problem.....	24
 CHAPTER III. METHODS .....	 26
Overview .....	26
Sample.....	26
Instruments .....	27
Early Childhood Assessment Package.....	27
Developmental Progress Reports .....	29
Early Literacy Scale.....	29
Reading Comprehension Subtest of Metropolitan Achievement Tests ....	31
Data Collection Method.....	31
Data Analysis.....	32
Descriptive Statistics .....	32
Developmental Nature of the Assessments.....	32
Literacy Measures for Program Evaluation .....	33
 CHAPTER IV. RESULTS .....	 35
Sample.....	35
Descriptive Statistics.....	37
Early Childhood Assessment Package (Concepts About Print).....	37
Kindergarten Progress Reports (Language Arts).....	38
Early Literacy Scale.....	40
Metropolitan Achievement Tests (Reading Comprehension) .....	42
Correlations and Reliability Estimates .....	45
Developmental Nature of the Assessments .....	47
Early Childhood Assessment Package (Concepts About Print).....	48
Kindergarten Progress Reports (Language Arts).....	51
Early Literacy Scale.....	53

Metropolitan Achievement Tests (Reading Comprehension) .....	54
Location of Students on the Underlying Developmental Continuum .....	57
Literacy Measures for Program Evaluation.....	57
Assessments .....	57
IRT Procedures for the Assessments.....	59
Reading Comprehension Subtest of the Metropolitan Achievement Tests .....	63
IRT Procedures for the MAT6 Reading Comprehension Subtest.....	64
CHAPTER V. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS .....	70
Summary and Conclusions .....	70
Recommendations.....	73
REFERENCES .....	75
APPENDIX .....	81
VITA .....	108

## LIST OF TABLES

TABLE	Page
3.1 Ethnicity .....	26
4.1 Project participants by Demonstration v. Comparison schools and Head Start v. non-Head Start participation .....	35
4.2 Students receiving ESL and/or special education services in second grade.....	36
4.3 Descriptive statistics for the Concepts About Print component of the Early Childhood Assessment Package using classical test theory.....	37
4.4 ECAP Factor Matrix showing factor loadings for each item .....	38
4.5 Mean teacher ratings for the Language Arts component of the Kindergarten Developmental Progress Reports by quarter.....	39
4.6 Factor loadings for the Language Arts ratings on the kindergarten progress reports .....	40
4.7 Mean teacher ratings of student performance on the four literacy tasks in the Early Literacy Scale.....	41
4.8 Correlation coefficients for Early Literacy Scale data .....	41
4.9 Factor loadings for the Early Literacy Scale .....	42
4.10 Descriptive statistics for the Reading Comprehension subtest of the MAT6 using classical test theory .....	43
4.11 Factor loadings for the MAT6 Reading Comprehension items .....	45
4.12 Correlation coefficients for total scores for the literacy assessments .....	46
4.13 Correlation coefficients for literacy assessment factor scores .....	46
4.14 Reliability estimates for the literacy measures.....	47

4.15	Rasch statistics for the Concepts About Print component of the Early Childhood Assessment Package .....	48
4.16	Mean ability levels by teacher rating of the use of literacy behaviors as documented in the kindergarten developmental progress reports .....	52
4.17	IRT difficulty indices for teacher ratings of student responses to the Early Literacy Scale tasks .....	53
4.18	Rasch statistics for the Reading Comprehension subtest of the Metropolitan Achievement Tests.....	55
4.19	ANCOVA on the factor scores for the 3 <sup>rd</sup> quarter kindergarten Language Arts items by experimental status with ESL and special education enrollment as covariates....	58
4.20	ANCOVA on the first grade ELS factor scores by experimental status with ESL and special education enrollment as covariates.....	59
4.21	Expected proportion of ratings for degree of use of 3 <sup>rd</sup> quarter kindergarten L.A. behaviors by experimental status .....	61
4.22	Expected proportion of ratings for ELS literacy levels by experimental status.....	62
4.23	ANCOVA on MAT6 Reading Comprehension total raw scores by Head Start and school status with ESL and special education as covariates.....	64
4.24	MAT6 (Reading Comprehension) IRT parameter estimates and expected proportion of correct responses by experimental groups .....	66
A.1	t-tests for independent samples of Head Start and non-Head Start mean factor scores by program enrollment .....	82
A.2	Correlation coefficients for Concepts About Print items in the Early Childhood Assessment Package .....	83
A.3	Correlation coefficients for the 3 <sup>rd</sup> quarter Language Arts scores on the kindergarten progress reports.....	84
A.4	Inter-item correlations for Reading Comprehension items in the MAT6 .....	85
A.5	ECAP (Concepts About Print) Rasch ability estimates .....	90

A.6	Estimated ability levels for third quarter kindergarten Language Arts behaviors ....	91
A.7	Estimated ability levels for Early Literacy Scale scores .....	97
A.8	MAT6 (Reading Comprehension) Rasch ability estimates.....	103
A.9	Factor loadings for the 3 <sup>rd</sup> quarter kindergarten Language Arts ratings.....	106
A.10	Factor loadings for the first grade ELS ratings .....	107

## LIST OF FIGURES

FIGURE	Page
4.1 Item Characteristic Curves for the Concepts About Print items in the Early Childhood Assessment Package generated through the one-parameter Rasch analysis .....	50

## **CHAPTER 1 INTRODUCTION**

During the 1986-87 school year 39.6 million school children took 206 million standardized tests. Two million of those standardized tests were administered to kindergarten children (Moore, 1992). The results of these outcome achievement measures were used to determine readiness for schooling and to support retention practices. No measurable change in the use of standardized tests has been noted over the last decade. In fact, there has been an increase in using standardized test results to make high-stakes decisions affecting the placement of children in instructional programs and in evaluating educational programs in general. Critical decisions are made based on the results of these standardized tests -- tests which measure student achievement at one point in time.

Children's growth and development, however, progress at different rates for each child, yet the developmental process remains consistent for normal, healthy children (Gullo, 1994). Observation and assessment in child development support these emotional, social, physical, and cognitive developmental growth patterns. By their nature, observation and assessment -- unobtrusive, on-going assessments within the context of the learning environment -- document the varying rates of development. These assessments consider the child's developmental stage and response characteristics. They provide the foundations for instructional decisions and inform parents and children of progress toward curriculum goals. Standardized tests do not.

Assessments and standardized tests have different purposes. Traditionally, assessments are used by teachers to make informed decisions about the instructional program in the classroom. Standardized tests, on the other hand, offer reliable measures for evaluating programs. The question among educators and program evaluators today is: Can assessments which have been useful for teachers at the classroom level, also be used to evaluate programs?

### **A Context for Developmentally Appropriate Literacy Assessment Issues**

Instructional practices for beginning readers and writers have changed dramatically in the last fifteen years. The teaching of skills in isolation and practice drills are no longer seen in the classroom. Real literature and authentic writing experiences replace the myriad of worksheets that were used two decades ago. Now big books and shared reading experiences abound in the early childhood classroom. Students practice "real" reading behaviors. In addition, writing instruction is no longer limited to handwriting exercises and weekly spelling tests. Teachers encourage students to write authentic stories based on personal experiences and knowledge of a topic. Today, teachers reinforce and augment approximations in reading and writing to help children grow and develop into literate members of society. Teachers plan new learning experiences to foster literacy development by building on what children already know about reading and writing and what teachers know about the development of literacy.

Along with these changes in instructional practices, assessment of literacy development in the early childhood classroom has also undergone a transformation. Standardized tests are no longer valid for measuring individual growth and evaluating developmentally appropriate programs in emergent literacy classrooms. These norm-referenced, standardized tests do not match the curriculum or objectives inherent in developmentally appropriate programs, nor do they adequately provide information to teachers for future instruction.

In addition, early childhood education advocates question the use of standardized tests for young children. "The Southern Association on Children Under Six recommends a ban of the routine, mass use of standardized intelligence, achievement, readiness and developmental screening tests for children through the age of eight" (SACUS, 1990, p.12). The National Association for the Education of Young Children (NAEYC) and the National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE) (1991) advocate the use "of an array of tools and a variety of processes" that document performance during "real" activities as opposed to "only skills testing" (p.32).

As a result of these recommendations, several states began to move away from wholesale standardized tests, especially for young children. The state of Maine opposes group standardized testing for young children at the preoperational stage of development (Maine State Department of Education and Cultural Services, 1988). Mississippi, North Carolina, Texas, and Arizona also object to standardized tests for young children (SACUS, 1990). Illinois, Maryland, Michigan, Rhode Island, and Vermont incorporate more authentic assessments in their statewide testing programs for children in kindergarten through grade eight (Valencia, Hiebert, & Afflerbach, 1994; Valencia, Pearson, Peters, & Wixson, 1989).

It is generally agreed that developmentally appropriate programs need developmentally appropriate assessments and that these assessments differ from traditional standardized tests and teacher made classroom tests of old. In 1991 the National Association for the Education of Young Children (NAEYC) and the National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE) proposed guidelines for developmentally appropriate assessment in programs serving children between the ages of three and eight. These organizations established the guidelines on the belief that "curriculum and assessment should be based on the best knowledge of theory and research about how children develop and learn with attention given to individual children's needs and interests in relation to program goals" (p.21).

Teachers who use developmentally appropriate assessment, observe individual children regularly and systematically in a variety of natural classroom interactions that encompass the scope of the program. Teachers document what the children know on their own and what children can do with the assistance of others. Teachers use this information as well as that gathered from students and parents to provide experiences that meet individual learner needs and correspond to the goals and objectives of the program.

Cunningham (1992) claimed teacher observational assessments possess validity. In addition, she noted that these assessments have high reliabilities although not numerically

derived, that is, quantified with statistically determined reliability estimates. Cunningham based her comments on reviews of numerous teacher observational assessments used in elementary schools.

Moore (1992) agreed that there is a need for informal classroom based assessments. Teachers need a variety of on-going informal measures to provide guidance for the daily instructional program. Moore contends, however, that there is also a need for formal, standardized tests. Administrators, curriculum specialists, and program evaluators need formal, standardized measures to monitor district-wide programs and to provide accountability to state and local governments and the public in general.

Historically, monitoring programs and addressing accountability issues have been accomplished through program evaluations, which focus on change. Differences between pre and post standardized test results have typically been used to document that change. Recent instructional practices, however, focus on the learning process and the growth and development of learning. Consequently, standardized tests no longer match the content and methods of current instructional programs. Therefore, standardized tests are not valid measures of developmental changes and program implementation.

Wiggins (1993) proposed that authentic assessments could be viable alternatives to formal, standardized tests to monitor change. By maintaining validity and reliability, authentic assessments could provide administrators and policy makers with indicators for addressing accountability and program evaluation. These assessments would incorporate the validity and on-going nature of teacher observational assessments as well as the reliability of more formal measures.

Engel (1991) concurred with Wiggins on the use of developmentally appropriate assessments for program evaluation. She offered suggestions for using narratives and graphical representations of children's literacy growth to inform administrators of program effectiveness. Additionally, several school systems have proposed the use of portfolios containing samples of developmentally appropriate literacy assessments to monitor growth in individual students and "common measures" for curriculum specific learning outcomes to allow comparisons across individuals and districts and for accountability purposes (Valencia, Hiebert, & Afflerbach, 1994).

Although developmentally appropriate assessments are currently being used by a number of school systems across the country and are increasingly becoming a part of state assessment packages, few of these assessments have been investigated from a psychometric perspective. First, the developmental nature of the assessments have not been confirmed, that is, whether, in fact, the assessments represent a developmental continuum. Second, little is known about the capabilities of the assessments to accurately identify the location of students on the developmental continuum underlying the assessments. Third, to date, few school systems or state agencies have validated the reliability of the assessments and the appropriateness for using these assessments as accountability measures or for program evaluation. Vermont has attempted to do just that but has so far been unsuccessful.

Attempts to establish the reliability and validity of outcome achievement measures used for accountability purposes or program evaluation, such as standardized, norm-

referenced tests, have typically relied on classical test theory (CTT) analyses. Alternative assessments, and in particular developmentally appropriate assessments, need analyses that are more suitable to their natures than CTT methods. As a result, measurement specialists are now investigating the psychometric properties of assessments through methods based on item response theory.

Mislevy (1994) claims that IRT procedures are more appropriate than CTT methods for determining the reliability of alternative assessments: “[t]he relationship between observable, and by implication between observable and hypothesized unobservable hypotheses, are laid out more explicitly than in CTT” (p.12). When alternative assessments are tailored to the individual or involve observations that differ from individual to individual, IRT offers a way to compare individuals independently of the items to which they responded and to compare items regardless of the sample of respondents. Item response theory also offers a theoretical foundation and procedures for confirming the developmental aspect of the assessments by equating response behavior to underlying ability on the construct of interest. IRT identifies variations in responses across latent ability levels and places them on a continuum. By comparing individuals’ responses within an assessment one can identify an individual’s location on that developmental continuum.

In addition, item response theory procedures offer a method for identifying differences among groups of respondents. When groups of students exhibit significant differences in their location on the developmental continuum underlying the concept of interest, it can be logically referred to differences in programs and their implementations. These differences can be used for evaluating program effectiveness.

### **Purpose of the Research**

The purpose of the research was fourfold. First, the study investigated the psychometric properties of developmentally appropriate assessments to verify the developmental nature of the assessment instruments. Specifically, the research examined emergent literacy assessment instruments: the Early Childhood Assessment Package (1994), the Early Literacy Scale (1993), and the Language Arts portion of developmental progress reports, for a developmental sequence of skills underlying each assessment. Second, analyses were conducted using item response theory procedures to locate respondents on that developmental sequence. Third, the study explored the use of the assessment instruments for program evaluation by identifying their ability to capture differences among groups of students. Fourth, the research investigated the use of item response theory procedures on a standardized measure for program evaluation. Specifically, the research used data from the local site of a national project to compare responses of Head Start and non-Head Start graduates in Treatment and Comparison schools.

## **Research Questions**

The following research questions drove this study to investigate the psychometric properties of developmentally appropriate emergent literacy assessments and their utility for program evaluation:

- Are the assessment instruments: the Early Childhood Assessment Package, the Early Literacy Scale, and the literacy portion of developmental progress reports developmentally appropriate, that is, do they represent a developmental continuum?
- To what extent do item response theory procedures facilitate the location of students on the developmental continuum underlying the assessments?
- Can these assessment instruments be used to discriminate among groups as required for program evaluation?
- Do item response theory procedures distinguish among groups on a standardized measure?

## **Significance of the Study**

The Early Childhood Assessment Package and the Early Literacy Scale are grounded in theory on the observation and assessment of young children and emergent literacy development. The psychometric properties of the assessment instruments have not been identified. To date, no statistical analyses have been performed on the assessments. Neither statistical confirmation of a true developmental assessment nor reliability estimates for the Early Childhood Assessment Package have been reported. Although reliability estimates for running records have been reported in the literature (Clay, 1993), no statistical analyses have been conducted on the developmental rating assigned to this component nor to any other component of the Early Literacy Scale. There are no reported reliability estimates nor validation procedures documenting the developmental nature of the ratings for the Early Literacy Scale. Similarly, no statistical analyses nor reliability estimates have been provided for the literacy portion of the developmental progress reports. To date, no studies have analyzed a comprehensive package of developmentally appropriate assessments to verify their appropriateness as developmental assessments and to investigate their usefulness for program evaluation.

## **Limitations**

The Early Childhood Assessment Package (ECAP) was implemented for the first time during the 1993-94 school year. During that time, teachers were learning how to document student literacy behaviors in the new format. Therefore, documentation was not systematic across all teachers and all schools within the system. As a result, not all student assessments were complete enough to be used for this research resulting in a smaller sample than anticipated.

In addition, the ECAP was revised for following school year. The new instrument includes a modified Concepts About Print assessment for emergent reading behaviors. For this study, the researcher coded teacher comments from related objectives in the original ECAP into the revised format. Because responses to the original ECAP did not address

all of the emergent reading behaviors included in the revised ECAP, the analyses in this study will be performed only on those literacy behaviors for which the data are available.

The students in the research study are part of a larger national Head Start Transition Demonstration Project sample. Students were selected for the national study based on the fact that they participated in a Head Start preschool program or were from a matched sample with no Head Start experience. Subjects selected for the current research sample may be overrepresentative of lower ability levels in literacy acquisition. There is evidence to suggest that measurement errors are greater for individuals with low ability on the concept being measured than individuals with average ability. Consequently, the reliability estimates calculated from this study may be lower than expected.

Further, a longitudinal study would be necessary to ascertain the developmental nature of the assessments over time. The data for such a study was not available for this research. As a result, a cross-sectional study was carried out. Because of this design, the results can only suggest the developmental nature of the early childhood assessment package. However, the assumption that a developmental sequence is present in a cross-sectional sample is highly supportable in this instance. Nevertheless, a longitudinal sample would be necessary for a definitive demonstration of a developmental sequence.

In addition, in order to verify the ability of the measures to identify differences among groups for program evaluation, groups with known differences must be used. Because the experimental groups in this study have not been identified to be different, it is impossible to compare the assessments with the standardized measure in their ability to capture differences among the groups. As a result, this study examined the classical reliability of the assessments in order to determine their potential use in discriminating among groups. The assessments were then used to note the consistency across assessments in making group comparisons. However, a definitive test of the capacity of these assessments to discriminate among groups can be accomplished only with groups that are known to be different. This will be accomplished in future research.

Finally, the size of the sample limited the interpretation of the results of some of the analytic procedures. For example, a sample size ten times greater than the number of items in a measure is needed to conduct a factor analysis. Because the  $n$  of 170 in this study did not meet that criteria, the results of the factor analysis of the standardized measure must be interpreted with caution. Clearly, this study needs to be replicated with a larger sample.

### **Definition of Terms**

***Alternative Assessment:*** Alternative assessment refers to an evaluation of student behavior that is an alternative to a standardized measure such as a traditional standardized achievement test. The term alternative assessment is often used interchangeably with authentic or performance assessment. For the purpose of this research, an alternative assessment is an authentic, performance assessment. Meyer (1992) distinguishes an authentic from performance assessment by the context in which the behavior is elicited. “In a performance assessment, the student completes or demonstrates the same behavior that the assessor desires to

measure. . . . In an authentic assessment, the student not only completes or demonstrates the desired behavior, but also does it in a real-life context” (p.40).

***Developmentally Appropriate Assessment:*** The NAEYC and NAECS/SDE (1991) define developmentally appropriate assessment as: the process of observing, recording, and otherwise documenting the work children do and how they do it, as a basis for a variety of educational decisions that affect the child, including planning for groups and individual children, and communicating with parents. Assessment encompasses the many forms of evaluation available to educational decision makers. Assessment in the service of curriculum and learning requires teachers to observe and analyze regularly what the children are doing in light of the content goals and the learning processes (pp.21-22).

Since literacy learning is a developmental process, assessment must focus on that development. The most cogent assessments document literacy behaviors in meaningful contexts with a variety of textual materials. They note these behaviors over time. These assessments may include work samples, anecdotal and narrative records, interviews, and developmental checklists. The assessments focus on which behaviors the child controls, which ones he or she is working on controlling, and what intervention(s) would enhance the development of literacy.

***Standardized Tests:*** Standardized tests are commercially prepared measures that “provide methods of obtaining samples of behavior under uniform procedures” (Mehrens & Lehmann, 1969, p.3), i.e., the same set of directions and time limitations apply for all administrations of a common set of items. Responses are scored using the same scoring procedure. Standardized tests may be norm-referenced or criterion-referenced measures of achievement, ability, or attitudes. For the purpose of this research, standardized tests will refer to norm-referenced, multiple-choice measures of achievement.

***Early Childhood Assessment Package:*** The kindergarten Early Childhood Assessment Package (ECAP) includes a variety of assessment instruments and data collection forms documenting student growth and learning across the physical, social, emotional, and cognitive domains. The Developmental Profile Assessment of the ECAP provides guided tasks for assessing student growth and development in reading, writing, math/science concepts, and fine- and gross- motor skills. The literacy portion documents the child’s development in reading, story retelling, and writing through the observation and assessment of significant learning behaviors.

***Early Literacy Scale:*** The Early Literacy Scale (ELS) is “an end-of-the-year literacy ‘snapshot’ of first-grade students.” The ELS assesses learning behaviors in the areas of reading, writing, and oral language. For each task, teachers rate students in one of five stages: Emergent, Novice, Apprentice, Developing, and Independent.

## **CHAPTER 2**

### **REVIEW OF RELATED LITERATURE**

To develop the theoretical foundations for the use of developmentally appropriate emergent literacy assessments for program evaluation, the review of related literature will focus on three overriding themes. First, an overview of the inappropriateness of the use of standardized achievement tests for monitoring literacy development will be presented. Next, the literature on selected developmentally appropriate emergent literacy assessments will be surveyed. Assessments similar to those used in this study will be detailed including reliability estimates and validation procedures cited in the literature. Finally, an overview of the psychometric considerations for measurements and assessments will be described. Validation procedures will be outlined, and methods for estimating reliability will be presented from two theoretical perspectives: classical test theory and item response theory. Conclusions will focus on the design of the present research in the context of the theoretical issues presented.

#### **Inappropriateness of Standardized Tests for Monitoring Development**

Traditionally, standardized tests have been used to measure student ability and achievement (Cronbach, Linn, Brennan, & Haertel, 1995; Stiggins, 1995). The information from standardized achievement test results has been used to make decisions affecting student placement in instructional settings, to evaluate programs, allocate funding, and to hold teachers, individual schools, and school divisions accountable (Office of Technology Assessment, 1992; Madaus & Tan, 1993; Pearson & Stallman, 1993; Stiggins, 1995). Proponents for standardized tests argue that these scores are a viable tool for measuring change and evaluating educational programs because the tests are carefully constructed, highly reliable, and yield objective results (Ebel, 1977; Green, 1991; Madaus & Tan, 1993). Casteen (1982) claims that tests validate what people do and, therefore, are proper for documenting learning and progress. Additionally, Ravitch (1982) notes that “the uses of standardized tests are clear and limited,” she contends that their major value is “to check up on how well our children are learning” (p.67). Others maintain that alternative assessments are more appropriate for documenting learning and evaluating programs (Teale, 1988; National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education, 1991; Moore, 1992; Wiggins, 1993; Stiggins, 1995).

In response to the supporters of standardized testing, some educators and assessment specialists provide the rationale for the inappropriateness of using standardized tests to monitor literacy development in young children. Typically, four reasons for not monitoring early literacy development with standardized tests emerged from the review of the literature. First, standardized tests offer a narrow definition of emergent literacy. Second, the tests fail to consider literacy development and the characteristics of young learners. Third, the tasks in standardized tests are unrelated to developmentally

appropriate practices in the classroom. Fourth, standardized test results are limited for use in the instructional program. These four reasons are explained next.

#### *Narrow definition of emergent literacy*

Standardized achievement tests and readiness measures offer a limited view of reading and writing (Chittenden & Courtney, 1989; Tierney, Carter, & Desai, 1991). Tierney, Carter, & Desai (1991) contend “[f]ormal tests focus upon a narrow band of acquired understandings – namely, whether students can produce according to established expectations” (p.30). In other words, correct responses to test items reflect the test constructor’s interpretation of text. They add that by doing so, standardized tests ignore the current research in reading and writing which cites reading comprehension and writing are influenced by the students’ background experiences and purposes. Chittenden and Courtney (1989) argue that standardized tests narrow the definition of early reading by “accentuating knowledge of letters and words to the exclusion of other indicators of literacy” (p.108). As a result, according to Tierney, Carter, & Desai (1991), the tests “tend to focus upon verbatim recall and judge reading comprehension by a predetermined standard of response” (p.30).

Experts in the field of emergent literacy education point out that reading and writing are complex behaviors (Teale & Sulzby, 1989; Strickland, 1990; Clay, 1991). Standardized achievement tests, on the other hand, focus on behaviors that are easily measured, often resulting in the measurement of only lower-order thinking skills (Moore, 1992). In addition, a standardized test summarizes ability into a global score which maximizes reliability estimates (Tierney et al., 1991; Wiggins, 1993). Wiggins (1993) claimed that although the elimination of “bias, shift, and other forms of judgment errors” is desirable, the use of multiple choice scoring procedures results in “taking complex performances and dividing them into discrete, independent tasks that minimize the ambiguity of the results” (p.15). Combining these small, discrete task scores into a single, global score, however, yields scores which have different meanings despite their numerical similarity.

Traditional readiness measures have been an attempt to assess the student’s cognitive ability on the constructs of interest. Measures such as the Metropolitan Readiness Tests (MRT) and the Boehm Test of Basic Concepts, purport to assess skills necessary for school success. Use of these tests, however, has decreased in recent years because the concepts assessed by the MRT: auditory memory, beginning consonants, letter recognition, visual matching, school language and listening, quantitative language, listening (Nurss & McGauvran, 1986) and those measured by the Boehm: direction, amount, and time (Sattler, 1992) do not match current research on emergent literacy tasks (Stallman & Pearson, 1990).

Further, reliability estimates of these readiness measures have been less stable than estimations of reliability for other achievement measures. Sattler (1992) reports split-half reliability coefficients for Form C of the Boehm Test of Basic Concepts - Revised range from 0.55 to 0.85 and from 0.57 to 0.87 for Form D. The variability of these coefficients reflects the lack of parallelism between the halves of the tests. Further, test-retest reliability estimates ranged from 0.73 to 0.88 for samples of 63 to 111 children with a one

week retest interval (Sattler, 1992). These relatively low estimates may be due to developmental changes in the respondents or to response behavior characteristic of young children.

*Lack consideration for literacy development and characteristics of young learners*

Standardized achievement tests fail to consider characteristics of young learners and the developmental nature of literacy acquisition. Moore (1992) expressed concern with the reliability and validity claims for tests used with young children: “Available research shows how students develop in different ways at different rates and how individuals will give varying responses that are all ‘correct.’ This seems to demonstrate the inability of standardized tests to take these variations into account” (p.125). Smith (cited in Moore, 1992) also doubts the reliability and validity data reported for the tests given the unpredictable nature and rapid development of young children. Further, in regard to emergent literacy development, Chittenden and Courtney (1989) stated that “norm-referenced tests allow little room for the natural variation among children in the rate at which they become relatively proficient as readers” (p.108). Because standardized tests reflect responses at one point in time they are unable to capture the variability of development and learning. Learning does not progress in a strict linear fashion. As a result, standardized tests may not reflect a young child’s true location on a developmental continuum.

Gullo (1994) identified four developmental issues that should be considered in any type of formal or informal testing program designed for young learners:

1. *Developmental constraints affecting children’s responses.* “[T]he child’s level of social, language, cognitive, and motor development often affects how he or she will interpret and respond, during both formal and informal assessment situations” (Gullo, 1994, p.28). Some standardized tests for young children require “controlled fine motor movements,” such as, filling in a bubble on a sheet with many bubbles. In addition, the peculiar language of the test may be unfamiliar to the child or inconsistent with the child’s own level of language development. Further, impulsive behavior, typical of young children, may impact response behavior. Consequently, it may be difficult to determine whether incorrect responses to standardized test items are indicative of a lack of knowledge or are the result of a mismatch between the tasks and the child’s level of development.
2. *Motivational differences.* Many young children do not understand the implications of their performance on a standardized test. “Many times, the reinforcement or incentive to perform is simply to complete the task, so that they can go on to a more comfortable or enjoyable circumstance” (Gullo, 1994, p.28). As a result, response behavior may not be characteristic of what the child really knows and can do. In addition, Gullo (1994) suggested that some groups of children may be more motivated to perform well on standardized tests because of their background experiences with “assessment-like” situations.
3. *Exaggerated perception of performance.* Gullo (1994) stated that children do not perceive their performance on tasks in the same way adults view that performance. As a result, children often have an exaggerated perception of their own performance.

- Gullo (1994) attributed this belief to the egocentric nature of preoperational children. “When they receive feedback from the environment, both positive and negative, they may focus on only the positive, thus getting a false sense of competence” (p.30).
4. *Contextual differences*. “It is not appropriate to assume that because children’s performance in academic settings indicates that they possess knowledge or skills within one particular context, that they will be able to generalize this knowledge or skill and demonstrate it in all contexts” (Gullo, 1994, p.30). Gullo (1994) concludes that preoperational children at the concrete level of cognitive development may not be able to transfer the knowledge and skills that they demonstrate in authentic situations to the “contrived” tasks in formal tests either because of the decontextualized nature of the tasks or the semi-abstract format of the tests.

In researching how young children learn to read and write, Teale and Sulzby (1989) highlighted the developmental nature of literacy acquisition. Based on a belief that “literacy is not only a cognitive skill; it is a complex activity with social, linguistic, and psychological aspects” (Strickland, 1990, p.19), they identified four themes in emergent literacy development: learning to read and write begins very early in life; oral language, reading, and writing are interrelated; literacy development requires active engagement in activities that are meaningful to the child; and adult-child interactions with print enhance literacy development.

First, “[l]earning to read and write begins early in life and is on-going” (Strickland, 1990, p.19). Parents surround children with language from the moment they are born. In addition to oral language, they read to their children and provide experiences with print. Research shows that by the time many children reach preschool, they can identify environmental print, that is “signs, labels, and logos in their homes and community” (Teale & Sulzby, 1989, p.3).

Second, listening, speaking, reading, and writing are integrated, not separate developmental strands (Teale & Sulzby, 1989; Strickland, 1990). Teale & Sulzby (1990) described the interrelatedness of speaking, reading, and writing as follows:

Educators have long seen that a strong oral language base facilitates literacy learning. Furthermore, it is clear that children’s developing reading abilities influence their writing. However, we must also recognize that reading experiences influence oral language (e.g., reading books to children enhances vocabulary), and writing actually improves children’s reading skills (e.g., allowing kindergartners to write builds decoding skills). (p.4)

Third, literacy development requires active engagement in activities that are meaningful to the child. Teale and Sulzby (1989) noted young literacy learners discover the functions of literacy and construct knowledge of how language works. “Literacy develops from real life settings in which reading and writing are used to accomplish goals” (p.3). Young children observe adults reading and writing for various purposes. As young children begin to develop their own literacy, they engage in reading and writing activities that help them understand how language works. Young children “read” books by telling the story through the pictures. In addition, they construct knowledge through writing development.

Ferreiro (1990) described three developmental levels for “the interpretation systems children build in order to understand the alphabetic representation of language” (p.12). During the first level children consider strings of letters as representations of objects and develop a distinction between pictures and writing . Having made those discoveries, children struggle with the number and type of symbols, or letters, needed to represent meaning during level two. At the third level, children begin to develop phonemic awareness and alphabetic sense.

Although Ferreiro’s theories evolved from her work with Spanish-speaking children, she reported that researchers working in other countries found similar patterns of development. This lead her to conclude “differences in language did not constitute a barrier to the application of the basic ideas in a field so language dependent as literacy” (p.12).

Fourth, adult-child interactions with print enhance literacy development. Strickland (1990) noted “[s]haring books with young children has long been recognized as a crucial aid to their language and literacy development and as a socializing process within families” (p.20). Teale and Sulzby (1989) added that children learn much about literacy through parent-child interactions “because parents work with them to jointly achieve the goal of the activity” (p.5) whether reading a book or preparing a shopping list.

#### *Tasks unrelated to classroom practices*

Instructional practices have changed since standardized achievement tests were first developed. From 1908 to 1916 Thorndike and his students at Columbia University created achievement scales for arithmetic, handwriting, spelling, drawing, reading, and language ability. By the 1920s, schools across the nation had adopted these tests for measuring student achievement (Wigdor & Garner, 1982; Pearson & Stallman, 1993; Cronbach, Linn, Brennan, & Haertel, 1995). In the last 75 years standardized achievement tests have changed very little. Today’s tests closely resemble those of the 1930s when the invention of the scanner made scoring quick and inexpensive (Pearson & Stallman, 1993).

Wiggins (1993) stated that “[f]or a variety of policy-related and historical reasons, testing in this country has become generic . . . in the sense of being linked neither to a particular curriculum nor to realistically complex problems and their natural settings” (p.15). He reported that this is due to an effort to establish common indicators for comparing schools and school districts and for accountability purposes. Generic tests have high reliability estimates which are required for accountability.

Tierney et al. (1991), however, documented the problems associated with this mismatch between reading and writing practices in the classroom and standardized testing:

1. Reading and writing instruction are integrated in the classroom, yet they are tested separately with standardized tests.
2. In the classroom students are encouraged to self-select reading material and topics for writing, establish purposes for reading and writing, formulate questions as they read, construct meaning, collaborate with others to develop understanding, and revise their thinking. Standardized tests, however, present texts and topics selected by others

- without regard to student interests; questions for reading and purposes for writing are also determined by others. In addition, students are expected to work alone.
3. Today's classrooms encourage students to read trade books with extended discourse, yet reading tests focus on abbreviated texts that limit the story line and character development. In addition, these texts often focus on obscure subjects to counter the effects of prior knowledge.
  4. Reading and writing programs support synthesis of ideas from multiple sources; whereas, tests rely on a single text.
  5. Instructional programs support diverse responses depending upon the texts and purposes for reading or writing. Standardized tests, on the other hand, assume student performance is the same across contexts.
  6. Today's instructional programs strive to empower students and encourage independent learning, reading for enjoyment, and self-assessment; yet standardized tests do not address these issues (p.29).

Moore (1992), Pearson and Stallman (1993), and others pointed out that because current instructional practices differ from what is measured by standardized tests, teachers incorporate testing skills in the curriculum to prepare their students for formal testing. Heibert and Calfee (1992) claimed that teaching test content and format invalidates the results from standardized testing. Nevertheless, great amounts of time are spent on test preparation to the detriment of the identified goals of the program (Perrone cited in Moore, 1992).

Wiggins (1990) stated that we come to value what is tested. Therefore, if standardized tests consist of "decontextualized, unambiguous items" to evaluate learning, then classroom instruction will focus on mastery of specific skills in isolation. This presentiment has been documented in the literature. Teale (1988) noted that increased emphases on testing has resulted in curriculums driven by the tests. Pearson and Stallman (1993) reported this often occurs in states that mandate standardized tests where the belief is if it is tested it should be taught. Shephard and Dougherty (1991, cited in Pearson & Stallman, 1993) also concluded that widespread use of standardized tests caused teachers to focus the curriculum on discrete, measurable tasks. In support of that belief, Hiebert and Calfee (1992) reported that teachers devoted instructional time to teaching these identified basic skills as well as encouraged students to memorize vocabulary and equations for the tests. Further, Koretz, Linn, Dunbar, & Shephard (1991, cited in Pearson & Stallman, 1993) have documented additional evidence of "teaching to the test" in an effort to raise test scores.

#### *Limited use for instruction*

Standardized test results are limited for use in the instructional program. Readence and Martin (1988) noted that standardized comprehension tests "were developed for the sake of convenience in test construction and scoring rather than for diagnostic information" (p.67). As a result, they lack useful information for instructional program planning. Tierney, Carter, & Desai (1991), on the other hand, believe assessments should be "useful to teachers and students, better inform teachers and students about achievement, progress, and effort, and fuel teacher and student

instructional collaboration” (p.34). They contend that standardized tests “for reporting literacy achievement are oftentimes scores or grades that do not translate into practice” (p.35).

Traditionally, standardized achievement test results report the information generated by the measure as a total raw score, or the number of items the examinee answered correctly. In order to interpret this score it is then converted to a percentile, a stanine, or other standard score which indicates the position of the examinee relative to the group upon which the test was normed. The percentile, stanine, or standard score does little to inform the teacher about the instructional needs of the student. These scores do not indicate what the child can and can not do. They are unable to provide information on the level of ability the examinee possesses on the underlying trait tapped by the measure nor what instructional practices are most appropriate for furthering the child’s development in that area.

In her review of the literature, Green (1991) found “a review of past practice suggests minimal use by teachers of the results of standardized tests in making instructional decisions” (p.200). Williams (1991) concurred: “There are many who contend that standardized tests are of limited use for teachers for daily instructional planning, especially with low-achieving students” (p.109).

Whereas, Moore (1992) and others cited previously, contended that standardized tests “narrow curriculum and inhibit good instructional practices” (p.126), Durkin (1987) (cited in Moore) found in her research that the results of standardized tests had no affect on the instructional program. Teachers continued to follow the plans they had made before the test results became available.

Additionally, standardized tests measure behavior at one point in time only. As noted previously, this may not be representative of young children’s literacy behavior in the classroom. Chittenden and Courtney (1989) stated that “tests tend to be one shot attempts at assessment; by definition they ignore evidence of learning that children exhibit in other, nontesting contexts. Children’s interest in books, and their ability to listen to stories or to make sense of meaningful words (such as classmates’ names or signs around the room) are not measured by tests” (p.108).

Standardized tests do not match today’s instructional practices in emergent literacy classrooms nor do they capture the processes involved in learning the concepts. Tierney Carter, and Desai (1991) stated “[w]hat is measured on these tests fails to approach what we view as literacy and the changes that are occurring in classrooms” (pp.25-26). The process students go through in learning is considered today to be as important as outcome measures of achievement. As a result, Tierney et al. (1991), among others, question the utility of standardized test scores for documenting reading ability and for predictive purposes. Readence and Martin (1988) stated that standardized tests with the traditional multiple choice format offer one way to measure reading comprehension, but “[s]tudents can show they have understood a text by producing miscues, retelling passages, or dramatizing stories” (p.68). In addition, teachers can use the information from these assessments to design an instructional program to meet specific student needs, whereas, standardized tests offer no diagnostic information.

The research supports the inappropriateness of using standardized tests with young children in light of the available information on child development and the nature of today's classroom experiences. In addition, it questions the validity and reliability of standardized tests for monitoring emergent literacy development. As a result, early childhood educators advocate the use of developmentally appropriate assessments instead.

### **Developmentally Appropriate Literacy Assessments**

Wiggins (1993) described the difference between a test and an assessment. A test is an instrument or measuring device. "It is an evaluation procedure in which responsiveness to individual test-takers and contexts and the role of human judgment are deliberately minimized" (p.15). Assessment, however, involves observing and recording what respondents know and can do. "An assessment is a comprehensive, multifaceted analysis of performance; it must be judgment-based and personal" (p.13). In addition, Archbald and Newmann (1988) stated that "a valid assessment system provides information about the particular tasks on which students succeed or fail, but more important, it also presents tasks that are worthwhile, significant, and meaningful – in short, *authentic*" (p.1).

The National Association for the Education of Young Children (NAEYC) and the National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE) (1991) support a multifaceted documentation of performance during ongoing classroom activities. The NAEYC and NAECS/SDE (1991) define developmentally appropriate assessment as:

the process of observing, recording, and otherwise documenting the work children do and how they do it, as a basis for a variety of educational decisions that affect the child.... Assessment in the service of curriculum and learning requires teachers to observe and analyze regularly what the children are doing in light of the content goals and the learning processes (pp.21-22).

NAEYC and NAECS/SDE (1991) identified three uses for assessment: the first is for planning the instructional program and for communicating with parents; the second is for identifying students with special needs; and the third is for program evaluation. The primary use of assessment is to guide instruction and inform parents of student progress. Based on the premise that curriculum and assessment are interrelated, the NAEYC and the NAECS/SDE proposed guidelines for assessment when it is used as a tool to guide teachers in planning instruction and as a method for communicating with parents (Hills, 1992).

Teachers who use developmentally appropriate assessment should, then, observe individual children regularly and systematically in a variety of natural classroom interactions that encompass the scope of the program. Johns (1982) termed this form of observation the "inner-ocular technique" hoping to legitimize teacher observation with a pseudo-scientific name. Cunningham (1982) called it "diagnosis by observation." Regardless of the terminology for the technique, good teachers have been using regular

and systematic observation to document what the children know on their own and what they can do with the assistance of others (Cunningham, 1982; Moore, 1992). Teachers incorporate this information with that gathered from students and parents to provide experiences that meet individual learner needs and correspond to the goals and objectives of the program (Hills, 1992).

Developmentally appropriate literacy assessments include ongoing samples of students' work, checklists, anecdotal records, logs, journals, and interviews (e.g., Chittenden & Courtney, 1989; Hills, 1992; Tierney et al., 1991). These assessments consider the developmental stage and characteristics of the child (Gullo, 1994).

Bergan (1988) provided the rationale for the use of developmental assessments in describing the development of cognitive measures in language, mathematics, nature and science, perception, reading, and social development for the Head Start program. The assessments were designed "to plan learning experiences based on the child's position in a developmental sequence" (p.234). He noted "[n]orm-referenced assessment was ruled out for describing children's performance because it describes ability in terms of position in a norm group rather than position in a developmental sequence. Moreover, it does not link ability to the performance of specific skills, which limits its use in educational planning" (p.235). He also dismissed the use of a criterion-referenced measure because "[d]escribing a child's performance in terms of the proportion of items mastered may restrict the amount of information provided by specific skills. In addition, it offers no information about the sequencing of skills" (p.235). Further, Bergan (1988) added that the theory underlying criterion referenced measures is not based on a latent trait, or ability, affecting performance.

Although developed for instructional planning, information gathered through developmentally appropriate assessments can be summarized for administrators to document student growth and to monitor programs (Engel, 1991). Pearson and Stallman (1993) reported that a suburban Chicago school district has already "translated student portfolio entries into data for wide-scale evaluation" (p.9). Valencia, Hiebert, and Afflerbach (1994) cited several schools that are monitoring literacy development with individual portfolios and noted that some use responses to common measures to document learning objectives for program evaluation.

Maeroff (1991) stated that individual schools and school systems are at various stages in developing alternative assessments for wide-spread use. In addition, "40 states are planning to use some form of alternative assessment at the state level, with writing samples as the most common alternative" (p.274). California, Connecticut, Rhode Island, and Vermont have begun to use alternative assessments on a large scale. Hiebert and Calfee (1992) reported "California has opted for gathering writing samples and assessments that integrate reading and writing" (p.72). In 1989, Rhode Island piloted reading and writing assessments in its elementary schools (Maeroff, 1991). Since 1990, Vermont has been developing their state-wide assessment program based on portfolios of classroom-generated student writing samples (Hiebert & Calfee, 1992; Koretz, Klein, McCaffrey, & Stecher, 1993).

The school system for the local site of the National Head Start Transition Project bases its documentation of emergent literacy development on current research in the field of literacy acquisition (Early Childhood Assessment Package, 1993, 1994; Early Literacy Scale, 1994). The kindergarten assessment, known as the Early Childhood Assessment Package (1994), consists of a portfolio of student work and a Developmental Profile Assessment of common indicators for literacy acquisition. Among these indicators is an assessment of students' knowledge of concepts about print. This assessment is founded on the Concepts About Print component of Marie Clay's (1993) Observation Survey of Early Literacy Achievement. The modifications allow for the use of standard texts commonly used in emergent literacy classrooms rather than the formal texts developed by Clay for Concepts About Print.

The first grade assessment, the Early Literacy Scale (1994), documents students' literacy behaviors in four areas: running records, oral language in writing conferences, writing samples, and story retellings. The assessments are completed by classroom teachers over the course of four weeks in the spring of the first-grade year. First, running records are taken on "bench marked" texts appropriate for the grade level. Teachers assess accuracy and miscues as well as analyze reading strategies used by the students. Second, classroom teachers document students' oral language behaviors during writing conferences. Third, student writing samples are gathered and assessed according to a rubric designed by the school system. Fourth, teachers evaluate listening comprehension through story retellings.

### ***Concepts About Print***

Marie Clay (1991) developed Concepts About Print to assess what young children know about print and book handling skills. Concepts About Print (CAP) is one component of the Observation Survey (Clay, 1985, 1993), a structured, multifaceted observational assessment designed to provide information for emergent literacy instruction. The CAP has 24 concepts, or items, administered individually. The child demonstrates knowledge of the concepts as the teacher reads a story designed for the assessment. Some of the concepts include identifying the front of the book, that the print not the picture contains the message, left to right progression, what a letter is, what a word is, and punctuation marks.

The assessment is designed to be used with children entering school and non-readers. Clay (1991) stated "It can show individual differences and how well-prepared children are for a particular program. It points the way to instruction for particular children, and it is a way of recording how the child behaves toward print before, during, and after the first year of instruction" (p.147).

Clay (1985) identified developmental age norms for the Concepts About Print items. The norms are based on a 1978 sample of 48 European children from New Zealand ranging in age from 5 to 7 years old. The developmental age for a task was set at the age at which 50% of the children in the sample correctly identified the concept. For example, 50% of the children 5 years 6 months of age were able to correctly identify the front of the book, left to right progression, top to bottom, return sweep, and a letter.

According to the Clay (1991), Concepts About Print (CAP) “was shown in research studies to have the qualities of a good test. Firstly, it captured shifts in book behavior that change rather rapidly the first year or two at school. Secondly, the items used discriminated quite well. Thirdly, it had a good range for use from preschool children through to high progress readers after a year at school” (pp.150-151). Day and Day’s (1978) research with kindergartners in Texas supported the sensitivity of the CAP for documenting rapid behavioral changes. Johns (1980) upheld the validity of the CAP with his study that distinguished low, average, and high ability readers.

Reliability coefficients for Concepts About Print have ranged from 0.73 to 0.95. Clay (1970) reported a Kuder-Richardson reliability coefficient of 0.95 in her normalization study with 40 urban New Zealand children ages 5 to 7 in 1968. To date no other researchers have been able to match that reliability estimate. Day and Day (1978) reported test-retest reliability coefficients from 0.73-0.89 in their study of 56 kindergarten students in Texas. They also reported odd-even split-half correlations corrected by the Spearman-Brown method that ranged from 0.84-0.88 and Kuder-Richardson coefficients from 0.83-0.92. Pinnell, McCarrier, and Button (cited in Clay, 1993), however, reported an internal consistency reliability estimate (Cronbach Alpha) of 0.78 (n=107) for urban Ohio children in August, 1990.

In addition, criterion-related validity coefficients have been reported for the Concepts About Print. Clay (1985, 1993) reported validity for the Concepts About Print assessment by correlating the scores with scores on Word Reading – a list of 15 words commonly found in beginning reading material. She derived the correlation of 0.79 from the responses of 100 6 year-old children who were administered both instruments as part of the Observation Survey (Clay, 1985). Day and Day (1978) also reported concurrent validity coefficients of 0.65 to 0.72, that is, correlations between responses for three administrations of Concepts About Print and the Record of Oral Language. Predictive validity coefficients, or correlations between three Concepts About Print administrations in kindergarten with the pre-reading composite of the first-grade Metropolitan Readiness Test, ranged from 0.69 to 0.72 (Day & Day, 1978). Although Clay (1991) stated that in her research “[c]orrelations with reading progress have ranged from 0.63 to 0.69” (p.151), she noted that Concepts About Print is not intended to be used as a predictor for reading achievement.

### ***Running Records***

Clay (1985, 1993) developed the running record as another component in the Observation Survey. The running record is a modification of Goodman and Burke’s (1972) miscue analysis. Taking a running record, however, does not require a prepared script of the text, a tape recorder, nor a “technical knowledge of linguistic concepts” (Clay, 1993, p.22) to interpret the record.

The running record is an account of a child’s behavior in reading text. The teacher records the child’s responses during the reading. After the reading, the teacher analyses the record and identifies the student’s accuracy rate (percent of words correctly identified in the text) and self-correction rate (ratio of self-corrections to miscues). In addition, the teacher records what cues the child appeared to attend to during the reading and with any

self-corrections. Clay (1985) categorized the cues attended to as meaning, structure, and visual.

Few reliability estimates have been reported in the literature for running records. Clay (1985), however, reported 0.90 for accuracy and error reliabilities on running records with a self-correction reliability of 0.70. Inter-rater correlations were higher for identifying word accuracy and miscues than for categorizing self-corrective behaviors. This appeared to be due to the assessors' disagreement over cues used by the reader when self-correcting (Clay, 1985). Clay (1966, cited in Clay, 1985) also reported reliability estimates for tape recorded running records for four children taken over a one year period. The estimates ranged from 0.98 for error scoring and 0.68 for self-corrections. Reliability estimates vary according to the teacher's experience taking and analyzing running records and with the difficulty of the text for the reader (Clay, 1985).

### ***Writing Assessments***

The literature on the use of authentic, or alternative, assessments as indicators of writing ability is prolific. Recent attempts to assess writing ability have concentrated on pieces written in the classroom during the course of the school year and those written to a given prompt under standardized conditions. Some large-scale writing assessments require students to write to a given prompt; others, such as the writing portion of the state of Vermont's portfolio assessment, require students to submit samples of their best writing for the year. In both cases, readers rate the written pieces according to a rubric and assign scores.

A study comparing the two types of writing assessments, that is, writing samples generated in the classroom and those developed under more formal conditions, found significant differences in the quality of the pieces. Simmons (1990) described his study comparing judges' ratings of student-selected pieces from classroom portfolios with a timed-test writing sample. Results indicated that the timed-tests underrated writing ability overall, especially for lower ability students. In other words, judges' ratings of portfolio pieces for these lower ability students were higher than scores for the same students' timed writing samples. In addition, Simmons (1990) noted that the holistically-scored timed writing assessments "failed to describe how writers of differing ability actually behave, and offered no instructional strategies tailored to specific achievement groups" (p.28).

Studies on the reliability of large-scale writing assessments have focused primarily on the comparison of readers' ratings. For example, interrater correlation coefficients for the writing assessment component of the Vermont Project ranged from 0.35 for within dimension correlations averaged across dimensions to 0.63 for correlations combining dimensions prompting researchers to question the use of writing assessments for large-scale accountability purposes (Koretz, Klein, McCaffrey, & Stecher, 1993). Resnick, Resnick, & DeStefano (1993) reported comparable interrater correlation coefficients (ranging from 0.53 to 0.56) from the New Standards Project Big Sky Scoring Conference. They suggested using a combination of interrater correlation coefficients, exact scorer agreement, and adjacent agreement to enhance the reliability of ratings.

### ***Story Retellings***

After reading the story or hearing one read to them, children may be asked to retell the story to demonstrate comprehension of text. Johnston (1983) declares that “retelling is the most straightforward assessment . . . of the result of text-reader interaction” (p.54). Morrow (1988) points out that the technique has been used for assessing reading comprehension in research studies. “Because retelling can indicate a reader’s or listener’s assimilation and reconstruction of text information, it can reflect comprehension” (Morrow, 1988, p.128). Morrow also notes that retelling has an advantage over traditional questioning for assessing comprehension in that it “allows a reader or listener to structure response according to personal and individual interpretations of the text” (p.128).

Story retelling is not a new concept for assessing comprehension of text. Early achievement tests relied on retelling for measuring reading comprehension (Johnston, 1984; VanLeirsburg, 1991). Today, the technique is widely used for informal diagnosis of both listening and reading comprehension through individual reading inventories and as an instructional strategy to develop comprehension ability (Morrow, 1988).

Scoring procedures for story retellings vary according to their theoretical foundations. Early reading assessments relied on verbatim recalls, that is, counting words (Starch in 1915 – cited in Johnston, 1984) and idea units (Courtis in 1914, Brown in 1914, Kintsch & van Dijk in 1978 – cited in Johnston, 1984). Johnston (1984) reported that for the most part scoring procedures based on “reproduction” were discontinued because recall ability is not the same as the ability to gain meaning from text. Recall, however, remained part of the silent reading portion of the Durrell Analysis of Reading Difficulty. In 1970 Goodman & Burke revived the use of recall for measuring reading ability; this scoring procedure relies not on reproduction but on the recall of literary elements (Goodman & Burke, 1970; Johnston, 1984).

Retelling scoring procedures that focus on the recall of literary elements, i.e., characters, setting, events, main idea or theme and ideas, are rarely used for reading assessments because they are time consuming (Johnston, 1984); however, this scoring procedure is commonly used in today’s classrooms through informal reading inventories. Typically, reading comprehension is defined by the number and quality of responses. Goodman & Burke’s (1970) scoring procedure for narrative text awards points for recalling the literary elements: 15 each for characters and character development, 20 each for theme and plot, and 30 for a list of events in the story (Allen & Watson, 1976).

### ***Developmental Checklists and Rating Scales***

Emergent literacy developmental checklists and rating scales document children’s behaviors in the areas of oral language and beginning reading and writing. The checklists record the existence or nonexistence of behaviors related to developmental goals. Rating scales, on the other hand, describe the degree to which those behaviors or traits are thought to be present in an individual.

Checklists can be used to prepare individual progress reports or as a guide to understanding children’s development. In addition, checklists can be of assistance in developing curriculum. When appropriately designed and implemented, checklists can be

used to assess individual children, inform parents of children's progress, and guide the development of the instructional program (Gullo, 1994).

Wortham (1990 cited in Gullo, 1994), however, cautioned against viewing checklists as assessment instruments. Checklists are designed as "organizing mechanisms" to describe curriculum or developmental sequences. Although Gullo (1994) questioned the objectivity of rating scales as well as the ambiguity of the descriptive terms used in rating scales, he noted that they are "often used to measure those traits not easily described using other assessment procedures" (p.78). Gullo (1994) suggested that both checklists and rating scales can be "used in conjunction with other observational forms to provide a more comprehensive assessment picture" (p.79).

Developmentally appropriate assessments for early literacy, such as concepts about print indicators, running records, writing samples, story retellings, and developmental checklists and rating scales have been an integral part of the classroom teacher's instructional program although not traditionally used as large-scale assessments nor for program evaluation. There is a movement, however, to incorporate these assessments in the evaluation of instructional programs and to use the results of the assessments to compare individual schools and school districts.

### **Psychometric Considerations**

Traditionally, classical test theory (CTT) analyses have been used to identify the psychometric properties of standardized tests and validate their use for program evaluation. Likewise, these same CTT analyses have been employed for alternative assessments. However, a review of the literature shows that classical test theory analyses are problematic for use with developmentally appropriate assessments. Modern test theory procedures, particularly those of item response theory (IRT), offer advantages over CTT for addressing the psychometric issues of assessments. In addition, IRT is especially suited for analyzing the psychometric properties of developmentally appropriate assessments and their utility for program evaluation.

#### ***Problems with Classical Test Theory***

Hambleton, Swaminathan, & Rogers (1991) list three main disadvantages with the use of classical test theory (CTT) methods for identifying the psychometric properties of measures:

First, characteristics of the examinees and those of the test can not be separated. These characteristics are group-dependent and test-dependent. CTT item difficulty and discrimination indices are calculated from the specific set of items included in the measure for a given group of respondents. Variations in either the set of items or the respondent group affect the item statistics of the instrument. As a result, these statistics as reported by the testing company for the norming group, may not be accurate for another sample of students.

Second, traditional (CTT) reliability estimates are calculated as the correlation between examinees scores on parallel measures. Mislevy (1994) elaborated: "This definition reflects the classic norm-referenced usage of tests: locating people along a single dimension, for selection and placement decisions. A high reliability coefficient indicates

that a different sample of tasks of the same kind would order the examinees similarly, leading to the same decision about them” (p.9).

Third, classical test theory is test-oriented rather than item-oriented. In the classical true score model there is no regard for an examinee’s response to a given item. As a result, CTT does not allow predictions to be made about how an individual or group of examinees will perform on a given item. According to Hambleton et al. (1991) this is necessary “to predict test score characteristics for one or more populations of examinees or to design tests with particular characteristics for certain populations of examinees” (p.5).

### ***Advantages of Item Response Theory***

Item response theory (IRT) is a modern test theory designed to address the shortcomings inherent in classical test theory methods for designing, constructing, and evaluating educational and psychological tests (Hambleton, Swaminathan, & Rogers, 1991). IRT is based on the concept that responses to items vary according to an individual’s ability level for the latent trait, or underlying construct, assessed by the test (Crocker & Algina, 1986; Hambleton et al., 1991; Baker, 1992). Hambleton & Swaminathan (1985) summarize the foundation of item response theory presented by Lord and Novick in 1968:

Any theory of item responses supposes that, in testing situations, examinee performance on a test can be predicted (or explained) by defining examinee characteristics, referred to as *traits*, or *abilities*; estimating scores for examinees on these traits (called “ability scores”); and using the scores to predict or explain item and test performance. Since traits are not directly measurable, they are referred to as *latent traits* or *abilities*. An item response model specifies the relationship between *observable* examinee test performance and the *unobservable* traits or abilities assumed to underlie performance on the test. (p.9)

An item response model that fits the data explains the complete latent space specified by the assessment. As a result, there are a number of advantages to item response theory (IRT) over classical test theory (CTT) methods (e.g., Hambleton & Swaminathan, 1985, Hambleton et al., 1991).

First, item response theory methods are invariant because they are item oriented rather than test oriented, that is, the procedures calculate statistics for each of the items in the measure or assessment in relation to the ability level of respondents rather than for the items as components of the total test for a particular group of respondents.

Second, an item characteristic curve, or function, reflects the true relationship among the unobserved variables (abilities) and the observed responses. In other words, the logistic function plots respondent behavior across the continuum of abilities underlying the concept of interest.

Third, the information generated from IRT lends itself to modern test formats in which individuals respond to selected items based on their ability levels and not to all items in the test. As a result, measurement specialists now use item response theory procedures over classical test theory methods for developing and evaluating tests and assessments.

### ***IRT in Relation to Alternative Assessments***

To evaluate the developmental nature of alternative assessments two psychometric issues must be addressed. First, the psychometric properties of the developmentally appropriate assessments must be identified, that is, the extent to which the instruments represent a developmental continuum needs to be determined. Second, the ability of the assessments to identify the location of students on the developmental continuum must be ascertained.

Item response theory techniques offer an advantage over classical test theory methods for identifying the developmental aspects of an assessment. For example, cognitive measures for the Head Start program were constructed and their developmental nature validated using item response theory, also known as latent trait, techniques. Bergan & Smith (1984, cited in Bergan, 1988) described the development of the measures as follows:

The Head Start Measures Battery is based on a latent ability model of development, which links latent trait constructs to psychological constructs that describe cognitive growth. The constructs of developmental level and developmental path are particularly important in the model. Developmental level indicates degree of growth. It is an individual characteristic that can be represented by the latent ability parameter in a latent trait model. A developmental path is a set of ordered competencies that comprise a developmental sequence. Developmental paths are defined by task characteristics. Paths can be described in terms of the item parameters in a latent trait model (pp.249-250).

Bergan & Smith (1984, cited in Bergan, 1988) concluded that unlike CTT, IRT afforded methods “to validate hypothesized developmental sequences, . . . to estimate ability represented by a developmental score, . . . [and] to relate developmental level to position in a developmental sequence” (p.236). They determined developmental sequences by holding the discrimination parameter constant across items in an instrument and ordering the items by the values of the difficulty parameter. In addition, IRT procedures generate an ability estimate, which is on a common scale, for each item in the developmental sequence. Bergan and Smith, therefore, were able to use individual estimates of ability, that is, developmental level, to locate respondents on the developmental sequence.

With the exception of Bergan and Smith’s (1984) cognitive measures for Head Start, few assessments have been evaluated psychometrically in relation to an underlying developmental sequence. In addition, to date, little is known about the ability of developmentally appropriate assessments to evaluate the effectiveness of educational programs.

### **Conclusions**

The uses and misuses of standardized tests have been well documented in the literature. The trend now is to supplement, if not replace, them with alternative assessments. Several states and school systems are doing just that. These states and school systems, however, are faced with problems as they move to implement alternative

assessments on a large-scale. Major issues involve identifying the psychometric properties, that is, the validity and reliability, of the alternative assessments before they can be used for documenting learning and evaluating programs (Pearson & Stallman, 1993). Farr & Carey (1986) claim informal classroom assessments offer the most potential for valid literacy assessments because they gather samples of behavior in a variety of contexts, include a greater sample of behaviors than traditional standardized tests, and are useful for instructional decisions; however, few studies have investigated the reliability and validity of these informal classroom assessments.

So far, the few studies to determine the reliability of large-scale alternative assessments have not been promising. Reliability estimates have been low – too low for making high-stakes decisions (Koretz, Klein, McCaffrey, & Stecher, 1993). Pearson and Stallman (1993) point out that most of these reliability studies have depended upon classical test theory methods for generating the estimates. Hiebert and Calfee (1992) noted that the technical aspects of alternative assessments may best be determined through modern test theory methods. Mislevy (1994) explained the rationale behind this: “The IRT formulation lends itself well to the machinery of statistical inference. The relationships among observed variables, and by implication between observed and hypothesized unobservable variables, are laid out more explicitly than in CTT” (p.12).

Item response theory (IRT) also appears promising for assessing latent abilities from a developmental perspective. Binet and Simon’s work with cognitive development made use of the basic principle underlying IRT, the item characteristic curve (Hambleton & Swaminathan, 1985; Baker, 1992). Rather than represent the components of an assessment as mere numbers, as in classical test theory item analyses, IRT also considers the amount of the underlying ability that the respondents possess in determining the difficulty and discriminatory power of items. Additionally, Bergan (1988) offered support for using IRT in monitoring development through his work with cognitive measures for the Head Start program.

### **Statement of the Problem**

Traditionally, standardized achievement tests have been used to monitor program effectiveness. Recently, however, educators have questioned the appropriateness of standardized tests for this purpose, especially for programs designed for young children. Early childhood advocates suggest the use of developmentally appropriate assessments instead of traditional standardized achievement tests for program evaluation. None of these proponents, however, have identified the psychometric properties of the assessments.

Although developmentally appropriate assessments have been implemented in a number of classrooms across the country, few studies have verified their ability to discriminate among developmental levels. In addition, even fewer studies have addressed the utility of these assessments for evaluating program effectiveness. Before administrators and policy makers will consider replacing standardized tests with authentic assessments for program evaluation, these two issues must be resolved. First, the developmental nature of the assessments must be confirmed and the ability of these

assessments to document development along a continuum must be verified. Second, the feasibility of developmentally appropriate assessments for evaluating program effectiveness must be explored.

For the present research, then, identifying the psychometric properties of emergent literacy assessments required analyzing a common data set through classical test theory (CTT) analyses and item response theory (IRT) procedures. The data consisted of responses to three literacy assessments designed by a local school system and a second grade standardized reading achievement test. Respondents were Cohort II participants from the local site of the National Transition Project. The instruments were: the Concepts About Print portion of the kindergarten Early Childhood Assessment Package, the Language Arts component of the kindergarten developmental progress reports, the first grade Early Literacy Scale, and the second grade Reading Comprehension subtest of the Metropolitan Achievement Tests. CTT analyses were performed to generate descriptive statistics and reliability estimates using both total and factor scores. Item response theory (IRT) procedures were used to investigate the developmental nature of the assessments, that is, the extent to which the assessments represented a developmental continuum, as well as to identify the location of students in relation to an underlying emergent literacy abilities' continuum for each of the assessments. In addition, the data from an early childhood literacy package consisting of a kindergarten and a first grade measure were submitted to both classical and IRT procedures to investigate the feasibility of using the instruments for program evaluation. These results were compared to the results of a second grade standardized achievement measure also submitted to both classical and IRT procedures.

## CHAPTER 3 METHODS

### Overview

Using a set of three instruments devised by a local school system and designed to be assessments of early literacy and a standardized reading achievement test, both classical test theory (CTT) and item response theory (IRT) procedures were applied to a common set of data. The instruments were: Concepts About Print portion of the Early Childhood Assessment Package, Language Arts component of the kindergarten developmental progress reports, first grade Early Literacy Scale, and second grade Reading Comprehension subtest of the Metropolitan Achievement Tests (MAT6). A comparison of the properties of this data set generated by CTT and IRT procedures allow an approximation of the general strengths and weaknesses of both procedures.

### Sample

The sample for the research consisted of 293 students from the local site of a national research project. These students comprised the second cohort (Cohort II) of the National Transition Project, a longitudinal study which followed the students from kindergarten through third grade. The students were enrolled in kindergarten during the 1993-94 school year and are currently in third grade.

The participants represented the diversity of the student population in the Virginia school system selected to participate in the national project. Approximately 10% of the participants in this study were Caucasian, 29% African American, 44% Latino, and 16% Asian. Table 3.1 shows the ethnic make-up of the participants in the study.

**Table 3.1.** Ethnicity (n=293).

<b>Ethnic Group</b>	<b>Frequency</b>	<b>Percent</b>
White / Caucasian	30	10.2
Black / African American	85	29.0
Hispanic / Latino	129	44.0
Asian or Pacific Islander	49	16.7

It should be noted that the ethnic make up of the participants in this research was not representative of that of the school system as a whole. In the 1994-95 school year, 67% of the students in the school system reported being Caucasian, 11% African American, 9% Latino, and 13% Asian (Cohn, 1995 p.B6).

All of the participants in this research attended fifteen special needs schools. These schools were randomly drawn from the population of special needs, low income schools in the County. The sampling strata and sampling frame were designed to yield a sample of schools in which approximately 300 Head Start graduates were enrolled in kindergarten and which reflected the proportional distribution of these graduates in the population of low income schools. Further, eight of the schools were randomly assigned to Treatment status (see below) and seven assigned to Comparison status. These special needs schools had been identified to receive additional support and resources because the majority of their students lived in families from lower socioeconomic groups. One hundred eleven participants (42.4%) in this study lived in households that reported receiving some type of public assistance during the 1994-95 school year. One of the additional supports afforded to special needs schools in the County is a fifteen-to-one student teacher ratio in all of the first-grade classrooms. However, the Treatment schools were part of the National Transition Project and received a wide range of social and educational support services. It was expected as a consequence of these services and other experimental treatments that there would be differential development of children participating in the treatment programs. Should such differences occur they would be useful in testing the capacity of assessment instruments to identify such differences. Consequently, children from the Treatment and Comparison groups were used in this study.

## **Instruments**

Because this school system's policy limits the testing of young children with standardized tests, there were no formal kindergarten and first grade achievement measures available for this population. Students participating in the National Transition Project, however, are administered a standardized reading comprehension test in both second and third grades. Literacy assessments for the participants in the national project were selected for this study because the records were readily available to the researcher. In addition, there was an interest to document the cognitive development of this particular group of students for the national research project.

### *Early Childhood Assessment Package*

The Early Childhood Assessment Package (1993) used in the Virginia site's kindergarten program is a developmental assessment which documents student growth and learning across the physical, social, emotional, and cognitive domains. The assessment is grounded in theory on developmentally appropriate practices and assessment of young children. The literacy portion is a record of the child's progress on the six outcomes and learning indicators synthesized from the Integrated Language Arts curriculum used in the public school system.

The Early Childhood Assessment Package (ECAP) (1993) identified six outcomes for K-12 as follows:

- Children read and write in a variety of forms.
- Children use strategies to construct meaning.

- Children adapt their language to communicate.
- Children respond critically to ideas in written and spoken language.
- Children use language processes to acquire, organize, and communicate information.
- Children enjoy and appreciate language and literature (p.5).

Throughout the school year, classroom teachers observe their students “during regular learning activities within the regular classroom setting” and record the students’ progress on each of the learning indicators which “identify observable behaviors that support a child’s development toward each of the six outcomes” (ECAP, 1993 p.5). Teachers assess literacy learning throughout the year but typically document a child’s progress on each learning indicator twice a year through the Early Childhood Assessment Package. The assessment allows the teacher to record the context of the observation, the social setting in which the behavior occurred, the date the behavior was observed, and any additional comments appropriate for each learning indicator.

The original Early Childhood Assessment Package was revised for use in the school system beginning with the 1994-95 school year. The revised format of the Early Childhood Assessment Package (1994) sought to remove variability in teachers’ narrative records by replacing these open-ended narratives with a more structured set of scales for recording student behaviors. This new instrument, known as the Developmental Profile Assessment of the Early Childhood Assessment Package, records student behaviors on common indicators for assessing emergent literacy, math/science concepts, and fine- and gross-motor ability. The emergent literacy portion includes a modified Concepts About Print (Clay, 1985) assessment, a story retelling task, and a drawing/writing sample.

For each of the literacy tasks, the teacher records the student’s responses, comments, and goals. In addition, for the first and fourth quarters, the teacher records the child’s literacy level for each task. Literacy levels are defined as: Early Emergent, Developing Emergent, Emergent, and Novice. Guidelines for determining each literacy level for each of the tasks are included in the Developmental Profile Assessment of the ECAP. Typically, children demonstrate behaviors that are representative of more than one literacy level. The teacher uses professional judgment in deciding the child’s literacy level (ECAP, 1994).

The Concepts About Print assessment provides information on the student’s understanding of printed language through observable behaviors that support reading acquisition (ECAP, 1994). As the teacher reads a story, the child demonstrates book handling skills. During the activity, the teacher observes and records the child’s ability to:

- identify the front of the book,
- predict text using pictures,
- demonstrate print conveys meaning,
- move left to right across a line of text,
- move top to bottom of the page,
- use return sweep (following two lines of text),

- match voice to print (matching word by word for both one-syllable and multi-syllabic words),
- demonstrate page sequence by turning the page for the teacher to continue reading the story, and
- demonstrate concept of letter/word by framing a letter and a word in the text.

(Early Childhood Assessment Package, 1994, pp.45-46)

The literacy component of the revised Early Childhood Assessment Package (1994) “reflects current research about the developmental growth and learning of young children.” (p.1) It offers a variety of authentic assessment instruments and data collection forms to record on-going observations of children’s literacy behaviors. The Developmental Profile Assessment component of the Early Childhood Assessment Package provides common indicators for assessing literacy development during the kindergarten year. To date, no reliability estimates have been identified for the instrument. The validity of this instrument is contained in the fact that committees of teachers developed the items to be consistent with the curriculum guiding their work. These items were reviewed for their face validity with respect to the curriculum by several other groups of teachers, and the assessment went through several iterations before the teachers were confident in its validity.

#### *Developmental Progress Reports*

The kindergarten developmental progress reports document achievement in the areas of language arts, social and emotional development, mathematics and science, and art and movement. The progress reports are completed three times per year: mid-year, or the second quarter, the third quarter, and the end of the year.

The classroom teacher documents students’ demonstrated behaviors for the Language Arts component of the kindergarten developmental progress reports in eight areas. These include: listens with understanding, communicates ideas verbally, explores language through rhyme, poetry, and/or dramatics, shares in group reading and writing experiences, understands ideas from literature, uses beginning reading strategies, communicates ideas with drawings and/or written symbols, chooses reading and writing as independent activities. The teacher rates student behaviors for each area on a 4-point scale: ‘not at this time,’ ‘sometimes,’ ‘usually,’ and ‘consistently.’

The intent of the progress reports is to represent a developmental growth pattern. To date, the developmental nature of the pattern has not been confirmed. In addition, no reliability estimates have been generated for the kindergarten developmental progress reports, and teacher rating behavior has not been studied.

#### *Early Literacy Scale*

The Early Literacy Scale (1993) is an end-of-the-year “snapshot” of a child’s literacy development in reading, writing, and oral language. The assessment takes place in the classroom over a four-week period in the spring of the first grade year. For each task, the classroom teacher rates students in one of five stages of literacy development: Emergent, Novice, Apprentice, Developing, and Independent. Although the children typically exhibit behaviors indicative of more than one stage, the teacher is asked to use

professional judgment in assigning a single stage of literacy development for each task for each child.

The reading portion of the Early Literacy Scale consists of a running record taken on a benchmark text, that is, one identified at a first-grade reading level. For children unable to read at that level, the teacher selects an appropriate level text to take a running record. After introducing the text, the teacher records the student's literacy behaviors while reading. The teacher notes any errors and self-corrections as well as cues the student used while reading. At the end of the assessment, the teacher records the text difficulty level based on the student's word accuracy rate: Independent (95-100%), Instructional (90-94%), or Frustration (below 90%). In addition, the teacher notes the student's self-correction rate, such as 1:5, or one self-correction for every five errors. From this information, the teacher assigns a developmental rating to the student based on guidelines developed by the local school system.

Students demonstrate their writing ability on the Early Literacy Scale by writing a piece on a self-selected topic. The teacher encourages the student to write about an area of interest or an event. The teacher instructs the students to plan prior to writing by drawing a picture, talking with fellow students, creating a list or web, or using other planning techniques. The student independently writes a draft and revises it the following day. Any changes in the writing are student initiated, although the teacher "reminds students to reread their writing to see if it makes sense" and "asks if there is anything they want to add or change" (Early Literacy Scale, 1993 p.19). As each student writes, the teacher highlights observed behaviors on a checklist developed by the school system. Information from this form guides the teacher in determining the student's stage of writing literacy development.

Oral language in the Early Literacy Scale is assessed with two tasks. One is participation in a writing conference; the other is retelling a story. Each student shares a piece of writing in a small group writing conference, which consists of three to five heterogeneously grouped students and the teacher. Students read their stories to the group and invite comments or questions about the writing from the other group members. The teacher notes individual student responses and observed behaviors on checklists developed by the local school system. The teacher then uses this documentation to determine the student's stage of oral literacy development.

Retelling a story demonstrates a student's comprehension of text and provides an opportunity for teachers to observe "a student's ability to use oral language, organize and sequence events, integrate story structure, and engage in a complete literacy event" (Early Literacy Scale, 1993 p.29). For the retelling task on the Early Literacy Scale, students listen to a taped story while following along with the text. After hearing the story, the teacher removes the text and elicits unassisted responses by asking the student to tell everything remembered about the story. When the student offers no additional information, the teacher prompts with, "Can you tell me anything else?" The teacher records unassisted responses and comments on a checklist developed by the school system to document retelling behavior. Without giving new information, the teacher elicits assisted responses by asking: "Can you remember anything else in the story?" "What else

happened?” “What was the problem in the story?” and/or “Where did the story happen?” Assisted responses and comments are also recorded on the checklist. The teacher highlights observed behaviors on another form developed by the school system and determines the stage at which most characteristics of literacy development are observed.

As with the Early Childhood Assessment Package, the Early Literacy Scale reflects current research on literacy acquisition and developmentally appropriate assessments for young children. Although there are reliability estimates for running records in the literature (Clay, 1993), there are no reliability estimates for the various components of the Early Literacy Scale. There was no attempt on the part of the school system to use multiple raters or to examine rating patterns used by individual teachers. Although teachers were trained in the use of these scales, no measures of the response of teachers to the training was attempted.

#### *Reading Comprehension subtest of the Metropolitan Achievement Tests*

The Reading Comprehension subtest of the Metropolitan Achievement Tests, sixth edition (MAT6), is a standardized norm-referenced test. The instrument consists of a series of short passages accompanied by questions with multiple choice responses. The Kuder-Richardson #20 reliability estimate for second grade students in the national standardization sample is 0.93 ( $SE_m=2.8$ ) (Prescott, Balow, Hogan, & Farr, 1989).

Although the local school system limits standardized testing of children in the early grades, students in the schools participating in the National Transition Project are administered the Reading Comprehension subtest of the MAT6 at the end of second and third grade. Scores from the subtest are used as an outcome measure for the local site of the national research project. To date, no information has been generated from these data for Cohort II students in the project.

### **Data Collection Method**

Early Childhood Assessment Packages (ECAP) for participants in the Head Start Transition Demonstration Project were collected at the end of the 1993-94 school year. In developing the ECAP, a team of teachers working with the Office of Early Childhood at the local educational agency (LEA) identified eleven Concepts About Print items as representative of emergent reading behaviors in kindergarten and included them in the revised ECAP. Members of the evaluation team for the Head Start Transition Demonstration Project selected these Concepts About Print items from the Early Childhood Assessment Packages for participants. Student responses were coded dichotomously for each item, that is, 1 for exhibiting the behavior and 0 for not exhibiting the behavior. These data were made available to the researcher for use in the present study.

Kindergarten developmental progress reports were collected by the local National Transition Project Evaluation Team. These reports were made available for use in the present study.

Scores from the first-grade Early Literacy Scale (ELS) were obtained through the LEA's Office of Planning, Testing, and Evaluation (OPTE). The OPTE created a data file of the spring, 1995, ELS scores for the 1,839 first-grade students enrolled in a sample of

schools with a high proportion of at-risk students. This was the same population from which the Transition sample was drawn. Scores for the Transition sample were taken from this data set.

Scores from the Reading Comprehension subtest of the Metropolitan Achievement Tests were collected by the Evaluation Team at the local site of the National Transition Project. These scores were made available for use in this study.

## **Data Analysis**

Several analyses were performed on the data from the literacy portion of the Early Childhood Assessment Package, the Language Arts component of the kindergarten developmental progress reports, the Early Literacy Scale, and the Reading Comprehension subtest of the Metropolitan Achievement Tests. Both classical test theory (CTT) and item response theory (IRT) techniques were applied to the scores from the instruments. CTT analyses were performed in SPSS for Windows, Release 6.1 (1994). IRT procedures were completed through MULTILOG, Release 6.30 (Thissen, 1991) and the TESTAT module of SYSTAT (1989).

### *Descriptive Statistics*

Descriptive statistics were generated for individual components and scale scores for all instruments in the data set. Within each instrument, item by item correlations were also generated where appropriate. In addition, total assessment scores were intercorrelated. These correlations were examined for inconsistencies, outliers, and other anomalies.

There are clearly issues of dimensionality and measurement error contained in such a complex and largely untested data set. In order to bring some order out of these scores by searching for underlying structures and clearing out components with high measurement error, exploratory factor analyses (EFAs) were performed on the scores within each of the instruments. For the present study, EFAs consisted of principal component analyses with varimax rotations.

Classical test theory techniques for estimating reliability were performed on all constructed factor scores. A correlation between the factor scores for each pair of assessments was generated. In addition, Cronbach's  $\alpha$  for internal-consistency reliability was calculated for each of the assessments to determine whether the performance indicators, or items, within each assessment measured the same phenomenon.

### *Developmental Nature of the Assessments*

To investigate the developmental nature of the assessments, responses from the Concepts About Print component of the Early Childhood Assessment Package, ratings from the Language Arts component of the developmental progress reports and the Early Literacy Scale, and responses from the Reading Comprehension subtest of the MAT6 were submitted to item response theory (IRT) procedures. IRT seeks to link an individual's responses to the instrument with the unobservable factor, or latent variable, underlying that measure. "The point of IRT is to use observed behaviors -- usually responses to many items -- to estimate an individual's standing on the latent characteristic.

The individual's standing on the characteristic can then be used to predict other behaviors" (Hulin, Drasgow, & Parsons, 1983 p.15).

Whereas, classical test theory analyses provide item difficulty indices and determine to what extent items distinguish between high and low scoring individuals, IRT procedures identify the order of the observable tasks, or the developmental scale of an assessment. This developmental scale is represented by the item parameters of the model. In addition, IRT identifies an individual's developmental level represented by the latent ability parameter (Bergan, 1988).

IRT procedures to determine item characteristics and to support the construct validity of the measures were conducted through MULTILOG, Version 6.30 (Thissen, 1991) and the TESTAT module of SYSTAT (1989). Dichotomously scored items in the Early Childhood Assessment Package and the Metropolitan Achievement Tests were fitted to a one-parameter logistic model in MULTILOG (Thissen, 1991) and the Rasch model, a special one-parameter model, in SYSTAT (1989). The polytomous scored items in the developmental progress reports and the Early Literacy Scale were submitted to graded model procedures.

#### *Literacy Measures for Program Evaluation*

In order to explore the use of developmentally appropriate assessments for program evaluation, scores from the Language Arts component of the kindergarten developmental progress reports and the Early Literacy Scale were submitted to both classical and IRT procedures. First, two analyses of covariance (ANCOVAs) by experimental groups, that is, Treatment/Head Start, Treatment/non-Head Start, Comparison/Head Start, and Comparison /non-Head Start were performed. Covariates included enrollment in English as second language (ESL) and special education programs. These covariates were selected because students enrolled in ESL and special education classes are clearly different from those students in the regular educational program. The use of ANCOVAs, then, results in removing the contribution of the variance of non-valid scores. The dependent variable for the initial ANCOVA was the third quarter kindergarten Language Arts factor scores. A Least Significant Difference (LSD) post hoc test was performed to determine which groups differed. The factor scores for the first grade Early Literacy Scale ratings were the dependent variable for the second ANCOVA.

Second, the ratings from the two assessments were fitted to a graded IRT model. Next, using the ability parameter estimates generated in that procedure, respondents were located on the developmental continuum, or scale, underlying the instruments. Finally, groups by experimental status were compared by their location on the developmental scale.

In addition, Cohort II responses to the second grade Reading Comprehension subtest of the MAT6 were submitted to both classical and IRT procedures. First, an ANCOVA by experimental group, that is, Treatment/Head Start, Treatment/non-Head Start, Comparison/Head Start, and Comparison /non-Head Start was performed using the total raw scores from the MAT6 as the dependent variable. Raw scores were selected over factor scores for two reasons: raw scores, traditionally, form the basis for calculating differences among groups of students in program evaluation and the measure did not

produce a clean, interpretable single factor. Covariates included enrollment in ESL and special education programs. The rationale for the use of these covariates is explained above.

Next, the data from the individual item responses to the MAT6 were fitted to a three-parameter IRT model. Groups by experimental status were compared. Results of the classical and IRT procedures on the MAT6 data were then compared to the results obtained from the classical and IRT procedures performed on the data set consisting of kindergarten and first grade assessment ratings.

As a result of the analyses, conclusions were drawn to address each of the purposes of the research. The purposes were: to investigate the developmental nature of the assessments, to locate students on the developmental continuum underlying the assessments, to explore the usefulness of the assessments for program evaluation, and to investigate the utility of IRT procedures on the MAT6 data for program evaluation. In addition, recommendations for further research were offered.

## CHAPTER 4 RESULTS

Using a set of three early literacy assessment instruments devised by a local school system and a standardized reading comprehension measure, both classical test theory (CTT) methods and item response theory (IRT) procedures were applied to a common set of data. The instruments were: Concepts About Print portion of the Early Childhood Assessment Package, Language Arts component of the kindergarten developmental progress reports, first grade Early Literacy Scale, and second grade Reading Comprehension subtest of the Metropolitan Achievement Tests (MAT6).

To address the purposes of this study, the research focused on a data set consisting of responses to the four literacy instruments by local participants in a National Transition Project. The purposes of the research were: to confirm the developmental nature of the assessments, to verify the ability of the assessments to locate students on the developmental continuum underlying the assessments, to investigate the usefulness of an early literacy assessment package for program evaluation, and to explore the use of IRT procedures on a standardized measure for determining program effectiveness.

### Sample

A local sample of students participating in the National Head Start Transition Demonstration Project comprised the sample for the present study. The table below shows the distribution of the 293 Cohort II participants by Demonstration and Comparison school as well as Head Start and non-Head Start participation.

**Table 4.1.** Project participants by Demonstration v. Comparison schools and Head Start v. non-Head Start participation.

	Head Start	Non-Head Start
Demonstration	96	49
Comparison	101	47

The National Transition Project is a longitudinal study designed to follow students from kindergarten through their third grade year. During the course of the study, students have transferred from school to school within the system as well as outside the system. In addition, some students have moved from Treatment to Comparison schools and vice versa. As a result, data on one or more of the four literacy measures were available for 235 of the 293 project participants.

After cleaning the data, 31 Early Childhood Assessment Package, 190 kindergarten Language Arts, 192 Early Literacy Scale, and 170 MAT6 Reading Comprehension subtest records remained for use in this study. The criteria for cleaning

the data set were: complete records for the three locally developed literacy assessments and MAT6 records in which students attempted at least half of the items in the test, that is, 27 or more of the 53 items. The latter criterion was chosen to obtain data representative of the reading behavior of the sample and to offset potential confounding of the results by including outliers.

Of the 235 students with literacy records, 167 had Head Start experience, 68 did not. Those Head Start graduates attended either a federally sponsored program located in a Head Start Center or a Family and Early Childhood Educational Program (FECEP) housed in a County school and operated by the local educational agency (LEA). Fifty (50) participated in a Head Start Center program; 117 attended FECEP.

Further, of the students with literacy records, 210 remained in the same school from kindergarten through second grade. One hundred-five (105) of these students were in Treatment schools; 105 were in Comparison schools. An additional five (5) students transferred from one Treatment school to another, or one Comparison school to another, within the first three years of schooling. Sixteen (16) students moved from a Treatment to a Comparison school, or vice versa, within the same time period, and four (4) students transferred to another County school that was not part of the National Transition Project.

Additionally, by the spring of the second grade year, ninety (90) of the students had been identified to receive additional services through one of two programs. The programs were: English as a Second Language (ESL) and special education. Table 4.2 shows the number of students in each of the programs.

**Table 4.2.** Students receiving ESL and/or special education services in second grade (n=235).

		ESL	
		none	services
Special Education	none	145	69
	services	16	5

To address the potential confounding of the study by including students who receive ESL and/or special education services, independent t-tests were performed on the factor scores for each of the literacy assessments (Table A.1, Appendix). Results showed that there were no significant differences in mean factor scores between students enrolled in ESL or special education and those who received no additional services. As a result, it was decided to include scores on the literacy instruments for all 235 project participants.

## Descriptive Statistics

### *Early Childhood Assessment Package (Concepts About Print)*

Initially, descriptive statistics were generated for the Cohort II data from the Concepts About Print component of the Early Childhood Assessment Package (ECAP), a kindergarten literacy assessment. Thirty-one (31) records were available for the analysis which identified item means, or proportion of respondents recording a correct response for each item in the assessment. These item means represent the level of difficulty for the dichotomously scored items. Items with higher means signify easier items, that is, most respondents correctly identified the concept about print. Table 4.3 shows the ECAP items ranged from the most difficult item ( $m=0.52$ ,  $sd=0.51$ ) for ‘uses pictures to predict the text,’ to the easiest items ( $m=0.97$ ,  $sd=0.18$ ) for ‘front-to-back page sequencing’ and ‘left-to-right progression across a line of print.’ In other words, 52% of the respondents were able to use picture cues to predict the text; whereas, 97% demonstrated understanding of the left to right sequencing items.

The corresponding standard error of the mean for each item indicated the precision of the mean value, that is, given another sample from the population of interest, 95% of the time the mean for an item would fall within  $\bar{n}$  1.96 standard errors from the given item mean score. For example, 95% of the time the mean score for ‘identifies a letter’ would range from 0.61 to 0.93. An inspection of the standard errors of the mean in Table 4.3 indicates that the ECAP means were acceptably precise for the sample.

**Table 4.3.** Descriptive statistics for the Concepts About Print component of the Early Childhood Assessment Package using classical test theory ( $n=31$ ).

Variable	Mean	S.E. of the Mean	Std.Dev.
Front-to-Back	.97	.03	.18
Left-to-Right	.97	.03	.18
Letter	.77	.08	.43
Top-to-Bottom	.71	.08	.46
Letter Sound	.68	.09	.48
Return Sweep	.68	.09	.48
Word	.68	.09	.48
One-to-One	.63	.09	.49
Predicts	.61	.09	.50
Pictures	.52	.09	.51

Inter-item correlations were also generated from the Concepts About Print responses (Table A.2, Appendix). Because the ECAP revealed moderate to strong intercorrelations among the majority of items, it was suspected that the variance would also be shared rather than unique. As a result, an exploratory factor analysis (EFA), that is, a principal components analysis (PCA) with varimax rotation, was performed using SPSS for Windows, Release 6.1 (1994). The EFA revealed a dominant first factor. This factor explained 61.4% of the variance in responses to the 10 items in the assessment. The second factor extracted by the analysis explained 17.6% of the variance.

**Table 4.4.** ECAP Factor Matrix showing factor loadings for each item (n=31).

	Factor 1 Loadings	Factor 2 Loadings
Word	.92905	-.14905
Letter Sound	.92905	-.14905
Top-to-Bottom	.89100	-.07623
One-to-One	.88418	-.18546
Letter	.87216	-.00660
Return Sweep	.82830	-.06504
Predicts	.75308	-.14273
Pictures	.68300	-.13060
Front-to-Back	.42348	.90344
Left-to-Right	.42348	.90344
Percent of Variance	61.4	17.6

An inspection of the rotated factor matrix (Table 4.4) for the items in the Early Childhood Assessment Package (ECAP) show eight items loaded on the first factor: ‘identifies a word,’ ‘letter sound relationships,’ ‘identifies a letter,’ ‘one-to-one correspondence,’ ‘top-to-bottom,’ ‘return sweep,’ ‘predicts, or anticipates, text,’ and ‘uses pictures to predict text.’ This factor could be called ‘use of cueing systems’ because the items reflected the use of graphophonic, contextual, and visual cues. Two items loaded on both the first and second factor: ‘front-to-back page sequencing’ and ‘left-to-right progression.’ This factor could be called ‘left-to-right sequencing.’

*Kindergarten Progress Reports (Language Arts)*

Next, descriptive statistics were generated for the kindergarten progress reports available for the Cohort II participants in the National Transition Project. Developmental progress reports were completed quarterly for kindergarten students. The first quarter reports were narrative and shared with parents through parent-teacher conferences. Therefore, only the teacher ratings for the students’ mid-year, 3<sup>rd</sup> quarter, and end-of-year assessments were submitted to analyses.

**Table 4.5.** Mean teacher ratings for the Language Arts component of the Kindergarten Developmental Progress Reports by quarter (n=190).

	Mid-Year		3 <sup>rd</sup> Quarter		End-of-Year	
	Mean	SD	Mean	SD	Mean	SD
Listens with understanding	2.80	0.66	3.11	0.64	3.34	0.58
Communicates ideas verbally	2.68	0.74	2.99	0.71	3.26	0.61
Explores language through rhyme, poetry, and movement	2.72	0.73	2.98	0.63	3.26	0.67
Shares in group reading & writing activities	2.61	0.84	2.92	0.75	3.17	0.74
Understands ideas from literature	2.62	0.74	2.91	0.73	3.14	0.72
Uses beginning reading strategies	2.22	0.85	2.65	0.77	2.95	0.76
Communicates ideas with drawings and words	2.87	0.70	3.13	0.66	3.43	0.60
Chooses reading and writing as independent activities	2.34	0.78	2.69	0.74	2.92	0.80

Table 4.5 shows the results of the descriptive statistics for individual items in the Language Arts component of the kindergarten developmental progress reports by quarter. Results of the analysis showed that the literacy behaviors exhibited by the children developed over time. At the mid-year evaluation teachers rated student use of literacy behaviors between 2.0 and 3.0 on the rating scale, that is between ‘sometimes’ and ‘usually.’ By the end-of-the-year, students ‘usually’ demonstrated use of these literacy behaviors.

Because the Early Childhood Assessment Package and the Early Literacy Scale data were collected in the spring, only the third quarter language arts ratings from the kindergarten developmental progress reports were used for the inter-item correlations. A correlation matrix for these teacher ratings from the developmental progress reports is shown in the Appendix (Table A.3).

As shown in the matrix, correlations among the third quarter language arts ratings ranged from 0.35 to 0.80. The weakest correlation was between ‘communicates ideas verbally’ and ‘chooses reading and writing as independent activities.’ The strongest relationship was between ‘shares in group reading and writing activities’ and ‘understands ideas from literature.’ Overall, the relationships between variables were moderate.

An exploratory factor analysis (PCA with varimax rotation) was performed using SPSS for Windows, Release 6.1 (1994). The analysis revealed a single factor for the third quarter language arts behaviors.

**Table 4.6.** Factor loadings for the Language Arts ratings on the kindergarten progress reports (n=190).

	Factor Loadings
Understands ideas from literature	.89662
Shares in group reading & writing activities	.85150
Uses beginning reading strategies	.84064
Listens with understanding	.83739
Explores language through rhyme, poetry, and movement	.81128
Communicates ideas verbally	.77027
Communicates ideas with drawings and words	.71175
Chooses reading and writing as independent activities	.68206
Percent of Variance	64.5

Table 4.6 shows the factor loadings for the literacy variables. The correlations between the variables and the single factor were moderate to strong, ranging from 0.68 for ‘chooses reading and writing as independent activities’ to 0.90 for ‘understands ideas from literature.’ Overall, the factor explained 64.5% of the variance in third quarter teacher ratings of kindergarten student literacy behaviors.

*Early Literacy Scale*

Descriptive statistics using SPSS for Windows, Release 6.1 (1994) were generated for the first grade Early Literacy Scale (ELS) data. One hundred ninety-two (192) available records were submitted to the analyses. These data represent Cohort II participants from the seven demonstration schools and six comparison schools that administer the ELS.

On a five-point scale from 1=‘emergent’ to 5=‘independent,’ mean teacher ratings showed students at the apprentice level (3) for all four literacy tasks in the Early Literacy Scale. Mean ratings depicted in Table 4.7 ranged from 3.37 (sd=1.05) for oral behavior during a ‘writing conference’ to 2.99 (sd=0.93) for stage of writing development as shown through the ‘writing sample.’

**Table 4.7.** Mean teacher ratings of student performance on the four literacy tasks in the Early Literacy Scale (n=192).

	Mean	SE of the Mean	Std.Dev.
Writing Conference	3.37	0.08	1.05
Retell	3.37	0.09	1.18
Running Record	3.26	0.11	1.46
Writing Sample	2.99	0.07	0.93

The means for all four tasks were reasonably consistent as were the standard errors of the mean. A review of the standard deviations in Table 4.7 shows greater variability in ratings for the ‘running record’ than any other task. This was not unexpected as the students read text appropriate for their instructional levels rather than one story for all respondents. As a result, teacher ratings were more variable.

The relationships among the teacher ratings of the emergent literacy behaviors assessed by the Early Literacy Scale are depicted in Table 4.86 below. Bivariate correlation coefficients were included for the overall ratings of the literacy behaviors exhibited through the ‘writing samples,’ during ‘writing conferences,’ in ‘retelling’ story elements, and during oral reading, that is, ‘running records.’ These correlations ranged from 0.26 for ‘running records’ and ‘writing conference’ behaviors to 0.59 for ‘running records’ and ‘writing sample’ behaviors indicating the strength of the relationship between each pair of variables was weak to moderate.

**Table 4.8.** Correlation coefficients for Early Literacy Scale data (n=192).

	Running Record	Writing Sample	Writing Conference	Retell
Running Record	1.0000			
Writing Sample	.5879	1.0000		
Writing Conference	.2560	.3197	1.0000	
Retell	.4096	.5261	.3828	1.0000

An exploratory factor analysis of the Early Literacy Scale (ELS), that is, a principal components analysis with varimax rotation, revealed one factor. This factor

explained 56.5% of the variance for the teacher ratings of the four emergent literacy assessments that comprise the Early Literacy Scale.

**Table 4.9.** Factor loadings for the Early Literacy Scale (n=192).

	Factor Loadings
Writing Sample	.83610
Retell	.78099
Running Record	.76607
Writing Conference	.60448
Percent of Variance	56.5

Table 4.9 shows the factor loadings for each of the ELS items. The loadings, ranged from 0.60 for ‘writing conference’ to 0.84 for ‘writing sample’ indicating a moderate to strong relationship between each variable, or item, and the factor.

*Metropolitan Achievement Tests (Reading Comprehension)*

Descriptive statistics were also generated from the second grade Cohort II data for the Reading Comprehension subtest of the Metropolitan Achievement Tests (MAT6). Item means and standard deviations are presented in Table 4.10 (n=170). The item mean, or proportion of respondents identifying the correct response, represents the level of difficulty for the dichotomously scored items in the MAT 6 subtest. Items with higher means signify easier items, that is, most respondents selected the keyed answers.

**Table 4.10.** Descriptive Statistics for the Reading Comprehension subtest of the MAT6 using classical test theory (n=170).

Item #	Mean	S.E. of the Mean	Std. Dev.
3	.98	.01	.13
1	.98	.01	.15
5	.96	.02	.20
2	.95	.02	.21
6	.95	.02	.21
8	.93	.02	.25
4	.88	.02	.32
24	.85	.03	.36
29	.82	.03	.38
17	.81	.03	.40
7	.80	.03	.40
9	.80	.03	.40
13	.80	.03	.40
26	.80	.03	.40
39	.80	.03	.40
30	.78	.03	.41
27	.78	.03	.41
36	.78	.03	.42
37	.77	.03	.42
11	.76	.03	.43
10	.75	.03	.43
23	.75	.03	.43
19	.74	.03	.44
25	.74	.03	.44
21	.73	.03	.44
22	.72	.03	.45
32	.72	.04	.45
18	.71	.04	.45
28	.69	.04	.46
44	.69	.04	.46
31	.67	.04	.47
14	.67	.04	.47
33	.67	.04	.47
43	.63	.04	.48
15	.62	.04	.49
20	.61	.04	.49
34	.61	.04	.49
40	.61	.04	.49
41	.60	.04	.49
47	.58	.04	.49
42	.57	.04	.50
35	.54	.04	.50
12	.54	.04	.50
16	.53	.04	.50
45	.53	.04	.50
48	.46	.04	.50
52	.45	.04	.50
46	.45	.04	.50
53	.40	.04	.49
50	.38	.04	.49
38	.38	.04	.49
49	.36	.04	.48
51	.35	.04	.48

MAT6 items ranged in difficulty from the easiest item, # 3 (m=0.98, sd=0.13) to the most difficult item, # 51, (m=0.35, sd=0.48). For 45 out of 53 items in the subtest, over fifty percent (50%) of the respondents attempting an item were able to identify the correct answer.

Inter-item correlations are presented in Table A.4 in the Appendix. They ranged from the highest correlation between items 1 and 3 ( $r=0.7032$ ) to no correlation between items 34 and 4 ( $r=0.0009$ ). Item 6 is not included in the matrix because the item had no variance. An inspection of the table shows that, in general, the items have either no relationship or a weak relationship at best.

Because the items in the instrument appeared to share little variance, it was expected that a principal components analysis (PCA) with varimax rotation would reveal several factors. In fact, the initial analysis revealed 16 factors with eigenvalues in excess of 1.0. Due to the fact that the PCA showed a dominant first factor for the MAT6 Reading Comprehension items (eigenvalue=11.14), a single factor solution was retained for use in further analyses. The factor explained 23.2% of the variance. Although the  $n$  of 170 is rather low for a factor analysis using 53 variables, or items, a single factor solution is suggestive of the underlying structure of this instrument. Table 4.11 shows the item loadings for this single factor. Items 1, 2, 3, 5, and 6 are not included in the table because these items had no variability.

**Table 4.11.** Factor loadings\* for the MAT6 Reading Comprehension items (n=170).

Item #	Factor Loadings
22	.68149
19	.65475
26	.65456
41	.65113
32	.64355
20	.63778
30	.61320
40	.59156
17	.58597
47	.56817
27	.55903
21	.55657
48	.55307
15	.54596
25	.53626
43	.52322

Item #	Factor Loadings
46	.52003
23	.51863
12	.51000
31	.50445
44	.49822
18	.49769
11	.49662
38	.47582
16	.47382
14	.47108
35	.44517
29	.44399
33	.43674
13	.42850
51	.41328
9	.41004

Item #	Factor Loadings
24	.40824
37	.40792
10	.39028
42	.37638
34	.37023
39	.37011
28	.36998
49	.36770
7	.36004
4	.33226
8	.29012
36	.28335
53	.14320
50	.11705
52	.03977
%ofVar	23.2

\*Only a single, dominant factor emerged from this procedure

Correlations between each item and the factor ranged from 0.68 for Item # 22 to 0.04 for Item # 52. Generally, the correlations were moderate. In fact, the single factor solution surprisingly revealed few items which would not be included in a table of factor loadings using a traditional criterion of exclusion, that is, an item-factor correlation less than 0.40.

*Correlations and Reliability Estimates*

In addition to descriptive statistics for the individual items within each assessment, total scores for the assessments were also generated and correlated to determine relationships among the early childhood literacy assessments. Further, correlation coefficients for the assessments' factor scores were also generated. These correlation coefficients, excluding cases pairwise, are presented in Tables 4.12 and 4.13.

**Table 4.12.** Correlation coefficients for total scores for the literacy assessments.

	ECAP	Progress Reports (Kgn)	ELS	MAT6
ECAP	1.00 (n=31)			
Progress Reports (Kgn)	.26 (n=24)	1.00 (n=190)		
ELS	.32 (n=20)	.39 (n=155)	1.00 (n=192)	
MAT6	.48 (n=23)	.43 (n=139)	.63 (n=146)	1.00 (n=170)

**Table 4.13.** Correlation coefficients for literacy assessment factor scores.

	ECAP 'use of cueing systems'	ECAP 'left-to-right sequencing'	Progress Reports (Kgn)	ELS	MAT6
ECAP 'use of cueing systems'	1.00 (n=31)				
ECAP 'left-to-right sequencing'	.00 (n=31)	1.00 (n=31)			
Progress Reports (Kgn)	.25 (n=24)	.09 (n=24)	1.00 (n=190)		
ELS	.28 (n=20)	.27 (n=20)	.38 (n=155)	1.00 (n=192)	
MAT6	.41 (n=16)	-.27 (n=16)	.47 (n=78)	.62 (n=77)	1.00 (n=92)

In general, correlation coefficients for the total scores were moderate to weak. They ranged from 0.63 for the ELS and MAT6 scores to 0.26 for the ECAP and third

quarter kindergarten Language Arts scores. Correlation coefficients between factor scores ranged from a moderate relationship (0.62) between the ELS and MAT6 factor scores to no relationship for the two ECAP factor scores. The latter was expected because factor scores are required to be orthogonal. Because there was limited information available to identify students who responded to the Early Childhood Assessment Package (ECAP), a smaller number of cases than expected were submitted to the correlation analyses. As a result, the correlation coefficients depicted in Tables 4.12 and 4.13 must be interpreted with caution.

The internal consistency reliability coefficient (Cronbach's  $\alpha$ ) was also generated for each of the assessments and the standardized comprehension measure. Cronbach's  $\alpha$  could be considered an estimate of the lower bound of a theoretical reliability coefficient, or the coefficient of precision. Table 4.14 shows the  $\alpha$  coefficient for the Early Childhood Assessment Package (Concepts About Print), the third quarter Language Arts ratings from the kindergarten developmental progress reports, the Early Literacy Scale, and the MAT6 (Reading Comprehension). The third quarter progress report ratings were selected to maintain consistency with the spring administration of the other instruments.

**Table 4.14.** Reliability estimates for the literacy measures.

	Cronbach's $\alpha$	
	estimate	SE <sub>m</sub>
Early Childhood Assessment Package (n=30)	0.93	0.09
Kindergarten Developmental Progress Report (Language Arts component) (n=190)	0.92	0.16
Early Literacy Scale (n=192)	0.72	0.45
MAT6 (Reading Comprehension) (n=91)	0.92	0.05

Internal consistency reliability estimates (Cronbach's  $\alpha$ ) were strong. They ranged from 0.72 for the four tasks in the Early Literacy Scale to 0.93 for the ten Concepts About Print items in the Early Childhood Assessment Package. The reliability estimates for the kindergarten progress reports (Language Arts component) and the MAT6 (Reading Comprehension subtest) were the same; however, the standard errors of measurement indicated that the MAT6 estimation of reliability was more precise.

### **Developmental Nature of the Assessments**

To investigate the developmental nature of the early literacy assessments, item response theory (IRT) procedures were performed using the TESTAT module of

SYSTAT (1989) for dichotomous data and MULTILOG, Release 6.30 (Thissen, 1991) for both dichotomous and polytomous. Because IRT procedures provide estimates of the ability levels at which 50% of the respondents are expected to correctly identify items, these procedures can be used to confirm the developmental sequence of the items within assessments.

*Early Childhood Assessment Package (Concepts About Print)*

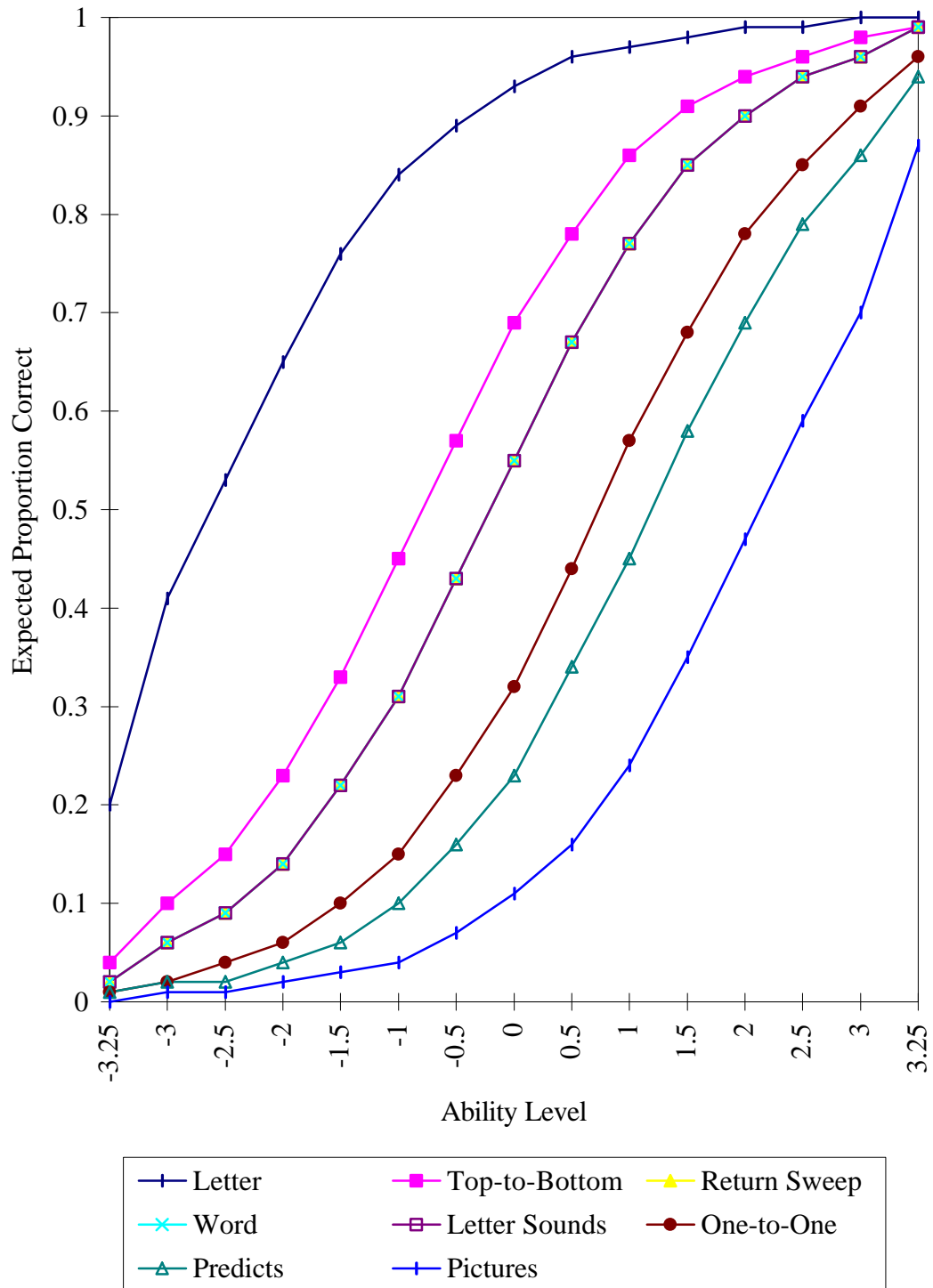
The ECAP data was fitted to a Rasch (IRT) model using the TESTAT module of SYSTAT (1989). Of the 31 ECAP records, 13 were retained for the analysis. The analysis ignored 18 records because the respondents had either incorrectly identified all 10 items or correctly responded to all 10 items, conditions that invalidated a record for the analysis. In addition, two items were deleted from the analysis because of perfect scores from all thirteen respondents. The Rasch one-parameter model procedure provided an item P value, that is, the probability of obtaining a correct response to an item at a given ability level ( $\theta$ ). The lower the ability level the easier the item. Item P values ranged from 0.31 at  $\theta=2.1$  for ‘uses pictures to predict the text’ to 0.92 at  $\theta=-2.6$  for ‘identifies a letter.’ In other words, the probability of a respondent with an underlying ability of 2.1, using ‘pictures to predict text’ was 0.31; whereas the probability of a respondent with an ability level of -2.6 correctly identifying a ‘letter’ was 0.92. An inspection of the Table 4.13 shows that the Rasch procedure ranked the ECAP items in the same order of difficulty as the CTT descriptive statistics.

**Table 4.15.** Rasch statistics for the Concepts About Print component of the Early Childhood Assessment Package (n=30).

Variable	Item P	Rasch Difficulty	Std.Error
Front-to-Back	unusable item: P=1.00		
Left-to-Right	unusable item: P=1.00		
Letter	.92	-2.63	1.25
Top-to-Bottom	.77	-0.78	.79
Letter Sound	.69	-0.21	.72
Return Sweep	.69	-0.21	.72
Word	.69	-0.21	.72
One-to-One	.54	0.74	.67
Predicts	.46	1.18	.67
Pictures	.31	2.13	.72

In addition to item P values, the Rasch procedure identified the difficulty parameter for each item, that is, the point on the ability scale ( $\theta$ ) at which 50% of respondents correctly identified the item. The Rasch one-parameter IRT procedure assumed that the difficulty parameter was the sole factor affecting respondent performance. ECAP item difficulties ranged from -2.6 to 2.1. The easiest usable item, 'identifies a letter' was at a low level of difficulty because 50% of the respondents with an underlying ability of -2.6 were able to correctly identify the concept. On the other hand, the item with the highest difficulty was 'predict text from pictures.' Fifty percent (50%) of the respondents with an underlying ability of 2.1 were able to correctly identify the concept.

The Rasch procedure supplied the mean latent ability level of respondents for each item, thereby, supporting the developmental nature of the Concepts About Print portion of the Early Childhood Assessment Package. Respondents with lower ability levels were able to correctly answer easy items. More difficult items required a higher underlying ability to produce a correct response. Items ranged from very easy to somewhat difficult for the kindergarten students in the sample. Prediction concepts, that is, 'predicts, or anticipates, text' and 'predicts text from pictures,' appeared to develop later than letter identification.



**Figure 4.1:** Item Characteristic Curves for Early Childhood Assessment Package items generated through Rasch (IRT) procedure (n=12).

Figure 4.1 depicts the item characteristic curves (ICCs) for the ECAP items generated through the one-parameter Rasch model. A visual inspection of the figure clearly indicates the varying item difficulty levels. Fifty-three percent (53%) of the respondents with a -2.5 ability ( $\theta$ ) level would be expected to correctly identify the easiest item in the assessment, whereas 99% of the respondents with a 2.5 ability level would be expected to correctly identify a 'letter.'

*Kindergarten Progress Reports (Language Arts)*

To verify the developmental nature of kindergarten literacy behaviors, teacher ratings of these behaviors from the developmental progress reports were fitted to a graded model using MULTILOG, Release 6.30 (Thissen, 1991). For each of the behaviors in the assessment, the procedure identified the mean level of underlying ability ( $\theta$ ) for the teacher ratings of students' use of literacy behaviors. These ratings included: 'not at this time,' 'sometimes,' 'usually,' and 'consistently.' The mean ability for students at each level of the rating scale across the three measures in the assessment is presented in Table 4.16.

**Table 4.16.** Mean ability levels by teacher rating of the use of literacy behaviors as documented in the kindergarten developmental progress reports (n=190). 1=not at this time 2=sometimes 3=usually 4=consistently

	Mid-Year				3 <sup>rd</sup> Quarter				End-of-Year			
	1	2	3	4	1	2	3	4	1	2	3	4
Listens with understanding	-6.36	-3.15	-0.15	2.47	-6.37	-3.00	-1.00	1.57	-1.39	-14.06	-1.80	0.98
Communicates ideas verbally	-6.74	-2.48	0.24	2.60	-6.56	-3.31	-0.53	1.72	-1.39	-16.98	-1.60	1.24
Explores language through rhyme, poetry, and movement	-6.60	-1.98	-0.07	2.60	-6.18	-2.91	-0.68	1.94	-1.39	-16.06	-1.08	1.02
Shares in group reading & writing activities	-6.77	-1.30	0.32	2.23	-6.05	-2.26	-0.19	1.62	-5.06	-3.39	-0.64	1.06
Understands ideas from literature	-6.61	-1.63	0.32	2.50	-4.78	-1.95	-0.27	1.70	-5.09	-3.41	-0.73	1.21
Uses beginning reading strategies	-7.60	-0.76	1.29	2.75	-7.05	-1.80	0.36	2.23	-6.20	-3.24	0.42	1.77
Communicates ideas with drawings and words	-6.99	-3.64	0.34	2.20	-6.99	-3.61	-1.23	1.60	-1.39	-13.98	-2.30	0.66
Chooses reading and writing as independent activities	-7.48	-1.92	1.31	3.27	-7.43	-3.34	0.27	2.50	-8.86	-6.10	-0.32	1.89

Results of the IRT procedure supported the developmental nature of the progress reports. Not only was there a progression in mean ability levels across item ratings within an observation period, but there was also a clear progression of mean ability levels within rating levels across quarters. For example, within the third quarter mean ability levels for ‘shares in group reading and writing activities’ included -6.05 (not at this time), -2.26 (sometimes), -0.19 (usually), and 1.62 (consistently). This showed that for ‘not at this time’ 50% of the students at the -6.05 ability level were able to demonstrate use of the task. Fifty percent (50%) of the students with an ability level of -2.26 were ‘sometimes’ able to perform the task, and 50% of the students with an ability level of -0.19 ‘usually’ ‘shared in reading and writing activities.’ Additionally, 50% of the students with a mean ability level of 1.62 ‘consistently’ demonstrated the behavior. From mid-year to end-of-year observations, the mean ability level, or the ability level at which 50% of the students achieved the rating of ‘usually,’ for ‘shared in reading and writing activities’ developed from 0.32 to -0.19 and finally to -0.64. Results of this progression showed that the mean ability level of students rated ‘usually’ in degree of use for the behavior decreased as the year progressed.

*Early Literacy Scale*

Teacher ratings of tasks in the Early Literacy Scale (ELS) were also fitted to a graded model IRT procedure in MULTILOG, Release 6.30 (Thissen, 1991). The data fit the model ( $\chi^2=2.4$ ,  $p<0.99$ ). Results of the procedure showed that there was a progression of teacher ratings for the tasks according to the latent ability level of the students (Table 4.17). For example, the procedure identified the mean ability level of students rated at the ‘emergent,’ ‘novice,’ ‘apprentice,’ ‘developing,’ and ‘independent’ levels for the ‘running record’ task. Students rated ‘emergent’ had on average an underlying ability of -2.45 (se=0.39); whereas, those rated ‘independent’ had an underlying ability of 1.32 (se=0.20).

**Table 4.17.** IRT difficulty indices for teacher ratings of student responses to the Early Literacy Scale tasks (n=192).

	Difficulty Indices of Teacher Rating by Task									
	Emergent		Novice		Apprentice		Developing		Independent	
	Diff	SE	Diff	SE	Diff	SE	Diff	SE	Diff	SE
Writing Conference	-8.60	****	-3.08	1.08	-2.29	0.70	0.20	0.24	2.38	0.58
Retell	-6.54	****	-1.95	0.33	-1.32	0.26	0.09	0.16	1.43	0.21
Writing Sample	-5.10	****	-1.65	0.15	-0.67	0.09	0.57	0.08	2.17	0.24
Running Record	-2.45	0.40	-1.53	0.23	-0.77	0.17	-0.37	0.15	1.32	0.20

\*\*\* indicates no standard error calculated due to estimated difficulty index

Table 4.17 shows the each of the ability levels at which 50% of the students were rated on the four tasks of the literacy scales. Clearly, the underlying, or latent, ability of students across the ratings of each task increased. In other words, for the 'retell' task students rated 'emergent' had an estimated underlying ability of -6.54, on average; whereas those students rated 'independent' on the 'retell' task averaged a latent ability of 1.43. In all instances, the average estimated underlying ability of the students showed a marked increase as rating levels increased. This, too, was not unexpected since the assessment sought to rate student literacy ability in the four areas, and it would be expected that the latent ability of 'independent' readers would be greater than that of 'emergent' readers.

*Metropolitan Achievement Tests (Reading Comprehension)*

The dichotomous MAT6 data were fitted to a Rasch (IRT) one-parameter model using the TESTAT module of SYSTAT (1989). The purpose of the procedure was to identify an ability continuum underlying the measure. Of the 170 MAT6 records, 90 were retained for the analysis. The analysis ignored 80 records because the respondents had either responded incorrectly or correctly to all 53 items or had not attempted a response to one or more items, conditions that invalidated a record for the procedure. In addition, one item (#6) was deleted from the analysis because all respondents correctly identified the keyed answer. Table 4.18 shows the item P values, that is, the probability of obtaining a correct response to an item at a given ability level ( $\theta$ ), the Rasch difficulty indices, and their corresponding standard errors of measurement.

**Table 4.18.** Rasch statistics for the Reading Comprehension subtest of the Metropolitan Achievement Tests (n=90).

Item #	Item P	Rasch Difficulty	Std.Error
2	.989	-4.534	1.162
3	.989	-4.534	1.162
1	.978	-3.593	.826
5	.956	-2.637	.591
8	.933	-2.064	.488
4	.911	-1.647	.428
24	.900	-1.473	.406
29	.844	-.798	.336
39	.844	-.798	.336
7	.833	-.688	.327
9	.833	-.688	.327
27	.833	-.688	.327
13	.822	-.584	.319
17	.822	-.584	.319
30	.811	-.485	.311
26	.800	-.390	.305
10	.789	-.299	.299
11	.778	-.212	.293
18	.778	-.212	.293
36	.778	-.212	.293
37	.778	-.212	.293
22	.756	-.046	.283
23	.756	-.046	.283
19	.744	.033	.279
21	.744	.033	.279
25	.744	.033	.279
28	.744	.033	.279
32	.744	.033	.279
14	.700	.330	.266
33	.700	.330	.266
15	.667	.536	.258
20	.667	.536	.258
31	.667	.536	.258
41	.667	.536	.258

**Table 4.18 (continued).** Rasch statistics for the Reading Comprehension subtest of the Metropolitan Achievement Tests 6 (n=90).

Item #	Item P	Rasch Difficulty	Std.Error
44	.667	.536	.258
43	.656	.602	.256
47	.656	.602	.256
42	.644	.667	.255
34	.622	.795	.251
40	.611	.858	.250
12	.589	.981	.248
45	.578	1.043	.247
16	.556	1.163	.245
35	.533	1.283	.244
52	.489	1.521	.244
46	.478	1.580	.244
48	.433	1.820	.246
50	.422	1.880	.247
53	.422	1.880	.247
38	.378	2.128	.251
49	.311	2.523	.263
51	.300	2.593	.266

MAT6 items 2 and 3 were the easiest in the subtest as shown by the Item P values. The higher the Item P value, the easier the item. Given a latent ability level of -4.5, the probability of correctly identifying items 2 and 3 was .99. Item 51 was the most difficult item. The probability of a correct response was .30 with an underlying ability of 2.6.

Both the IRT Item P and the CTT item mean values distinguished among the easy, medium, and hard items in the instrument. Although the rank order of items differed between the two outputs, this may have been due more to the differences in the data sets than to the procedures themselves. The Rasch procedure eliminated 80 of the 170 records used for the CTT descriptive statistics.

The difficulty parameter for each item, that is, the point on the ability scale ( $\theta$ ) at which 50% of respondents selected the correct answer, was also identified by the Rasch procedure. MAT6 item difficulties ranged from -4.5 to 2.6. The easiest usable items, 2 and 3, were at a low level of difficulty because 50% of the respondents with an underlying ability of -4.5 were able to correctly identify the answer. On the other hand, for the item with the highest difficulty, 51, fifty percent (50%) of the respondents with an underlying ability of 2.6 correctly identified the answer.

Traditional descriptive statistics, that is, item means and standard deviations, provided information about the items within each of the instruments. It was unclear from these classical test theory results, however, whether the items within an assessment were truly developmental or merely reflective of item difficulty levels. As a result, the data were fitted to item response theory models. Results of these procedures supported the developmental nature of the items, behaviors, or tasks, within each of the literacy assessments. The probabilities of respondents with lower underlying ability levels correctly identifying the Concepts About Print assessed by the Early Childhood Assessment Package were lower than respondents with higher latent abilities. Likewise, the probabilities of respondents with lower latent ability levels achieving high ratings on both the Language Arts items within the kindergarten developmental progress reports and Early Literacy Scale tasks were less than respondents with higher underlying abilities.

### **Location of Students on the Underlying Developmental Continuum**

In addition to item parameters, an item response theory (IRT) procedure provides an ability parameter, that is, an estimate of a respondent's latent ability on the concept of interest. Ability estimates ( $\hat{\theta}$ ) are reported as a z-scale score and range from -4.0 to 4.0 with a mean of 0 and a standard deviation of 1.

Of the thirty (31) Early Childhood Assessment Package records submitted to the Rasch IRT procedure, ability estimates ( $\hat{\theta}$ ) were generated for thirteen (13) students (Appendix). These ability estimates ranged from -2.47 ( $SE_E=1.20$ ) to 2.47 ( $SE_E=1.20$ ). For example, a student who responded correctly to one out of the eight usable items, would be located at -2.47 on the developmental continuum underlying the assessment.

Ability estimates ( $\hat{\theta}$ ) were also generated for the third quarter kindergarten Language Arts scores ( $n=190$ ), the first grade Early Literacy Scale (ELS) ( $n=192$ ), and the second grade MAT6 Rasch scores ( $n=90$ ). These estimates are presented in the Appendix. Ability estimates for the kindergarten Language Arts behaviors ranged from -3.56 ( $SE_E=0.40$ ) to 2.14 ( $SE_E=0.44$ ). Those for the ELS ratings ranged from -2.34 ( $SE_E=0.54$ ) to 2.09 ( $SE_E=0.41$ ). In addition, ability estimates ( $\hat{\theta}$ ) for the MAT6 scores ranged from -1.08 ( $SE_E=0.35$ ) to 5.23 ( $SE_E=1.17$ ).

### **Literacy Measures for Program Evaluation**

#### *Assessments*

To explore the use of an early childhood literacy assessment package for program evaluation a data set consisting of the teacher ratings from the Language Arts component of the 3<sup>rd</sup> quarter kindergarten developmental progress reports and the Early Literacy Scale (ELS) tasks for the Cohort II sample were submitted to both classical and item response theory (IRT) procedures. Kindergarten progress reports and first grade ELS ratings were available for 155 students. Because seven of the 155 students had moved from a Treatment to a Comparison school, or vice versa, between kindergarten and first grade, 148 records were submitted to the analyses. Responses from the Concepts About

Print portion of the Early Childhood Assessment Package were omitted from the procedures because of the limited data available.

Analyses of covariance (ANCOVAs) were performed on the kindergarten and first grade assessment data using SPSS for Windows, Release 6.1 (1994). Mean scores for the four experimental groups were compared. The groups were: Treatment/Head Start, Treatment/non-Head Start, Comparison/Head Start, and Comparison/non-Head Start. Enrollment in English as a second language (ESL) and special education programs were the covariates to remove the effects of non-valid scores. The dependent variable for the first ANCOVA was the factor score for the third quarter Language Arts items in the kindergarten progress reports generated through a principal components analysis (Appendix A). The dependent variable for the second ANCOVA was the factor scores of the combined ELS ratings (Appendix A). Tables 4.19 and 4.20 show the results of the two ANCOVA procedures.

**Table 4.19.** ANCOVA on the factor scores for the 3<sup>rd</sup> quarter kindergarten Language Arts items by experimental status with ESL and special education enrollment as covariates (n=148).

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Covariates	12.795	2	6.397	7.449	.001
ESL	9.912	1	9.912	11.540	.001
Special Education	2.883	1	2.883	3.357	.069
Main Effects	12.246	3	4.082	4.753	.003
Groups	12.246	3	4.082	4.753	.003
Explained	25.041	5	5.008	5.831	.000
Residual	121.959	142	.859		
Total	147.000	147	1.000		

**Table 4.20.** ANCOVA on the first grade ELS factor scores by experimental status with ESL and special education enrollment as covariates (n=148).

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Covariates	23.503	2	11.751	13.760	.000
ESL	13.235	1	13.235	15.498	.000
Special Education	10.267	1	10.267	12.023	.001
Main Effects	2.230	3	.743	.870	.458
Groups	2.230	3	.743	.870	.458
Explained	25.732	5	5.146	6.026	.000
Residual	121.268	142	.854		
Total	147.000	147	1.000		

Results of the ANCOVA using kindergarten factor scores as the dependent variable show a main effect for the experimental groups [F(3)= 4.753, p=0.003]. A significant difference existed between the means for the Treatment/Head Start (m=0.29), Treatment/non-Head Start (m=0.05), Comparison/Head Start (m=-0.14), and Comparison/non-Head Start (m=-0.45) groups after accounting for the effects of the covariates, that is, enrollment in special education or English as a second language (ESL) programs. A Least Significant Differences (LSD) post hoc test identified significant differences between the Treatment/Head Start group mean and the means of both the Comparison/Head Start and Comparison/non-Head Start groups. These results were surprising as schools were randomly assigned to Treatment or Comparison status and one would expect a non-significant difference in mean scores across the groups.

Using the ELS factor scores as the dependent variable, results of an ANCOVA showed no significant main effects across experimental groups [F(3)=0.870, p=0.458]. There were no differences in mean factor scores among the Treatment/Head Start (m=-0.15), Treatment/non-Head Start (m=0.05), Comparison/Head Start (m=0.17), and Comparison/non-Head Start (m=-0.21) groups after accounting for the effects of the covariates.

#### *IRT Procedures for the Assessments*

The kindergarten and first grade assessment data were also fitted to an IRT graded model using MULTILOG, Release 6.30 (Thissen, 1991). The IRT procedure was performed using data grouped by experimental status. The four groups were: Treatment/Head Start, Treatment/non-Head Start, Comparison/Head Start, and Comparison/non-Head Start. The parameters estimated in the graded model included: item difficulty (*b*) and item discrimination (*a*). The difficulty parameter (*b*) is the point on the underlying ability continuum at which 50% of the respondents are expected to receive

that rating. The item discrimination parameter ( $a$ ) represents the slope of the item characteristic curve (ICC) at point  $b$ . “Items with steeper slopes are more useful for separating examinees into different ability levels than are items with less steep slopes” (Hambleton, Swaminathan, & Rogers, 1991, p.15). Tables 4.21 and 4.22 show the results of the IRT procedure.

**Table 4.21.** Expected proportion of ratings for degree of use of 3<sup>rd</sup> quarter kindergarten L.A. behaviors by experimental status (n=148).

Variable	Treatment/Head Start				Treatment/non-Head Start				Comparison/Head Start				Comparison/non-Head Start			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Listens with understanding	.00	.06	.56	.38	.00	.07	.60	.33	.00	.12	.64	.23	.01	.20	.65	.14
Communicates ideas verbally	.00	.14	.53	.33	.00	.17	.55	.28	.01	.23	.55	.21	.01	.33	.52	.13
Explores language through rhyme, poetry, and movement	.00	.11	.61	.28	.00	.14	.63	.23	.00	.20	.64	.16	.01	.30	.59	.09
Shares in group reading & writing activities	.00	.14	.51	.34	.01	.18	.53	.29	.01	.25	.53	.20	.03	.36	.49	.12
Understands ideas from literature	.01	.14	.56	.29	.01	.17	.57	.24	.02	.25	.56	.16	.05	.36	.50	.09
Uses beginning reading strategies	.02	.29	.46	.23	.03	.33	.45	.19	.05	.41	.41	.13	.10	.51	.33	.07
Communicates ideas with drawings and words	.00	.09	.52	.39	.00	.11	.54	.34	.01	.16	.57	.27	.01	.23	.57	.18
Chooses reading and writing as independent activities	.01	.29	.46	.24	.01	.33	.45	.21	.01	.40	.43	.16	.02	.49	.38	.10

Degree of use for kindergarten items: not at all (1), sometimes (2), usually (3), consistently (4)

**Table 4.22.** Expected proportion of ratings for ELS literacy levels by experimental status (n=148).

Variable	Treatment/Head Start					Treatment/non-Head Start					Comparison/Head Start					Comparison/non-Head Start				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Running Record	.07	.12	.06	.44	.27	.08	.13	.07	.43	.24	.10	.14	.07	.42	.19	.13	.17	.08	.39	.14
Writing Sample	.05	.14	.40	.37	.03	.06	.16	.41	.34	.03	.08	.18	.42	.30	.02	.10	.22	.42	.25	.02
Writing Conference	.07	.06	.33	.38	.16	.07	.06	.34	.38	.15	.08	.07	.36	.36	.13	.10	.08	.38	.33	.11
Retell	.05	.07	.31	.36	.20	.06	.08	.33	.36	.18	.07	.09	.35	.33	.15	.09	.11	.38	.30	.12

Ratings for ELS tasks: emergent (1), novice (2), apprentice (3), developing (4), independent (5)

Results of the IRT procedure show that there were distinct differences among the groups across latent ability levels. An inspection of the tables shows that the expected proportion of students with the highest underlying ability, that is, those students rated ‘consistently’ in degree of use of all eight kindergarten Language Arts behaviors and at the ‘independent’ literacy level for each of ELS tasks, appeared greater for the groups in Treatment than Comparison schools. In addition, the expected proportion of students rated at the highest level for both the kindergarten and ELS rating scales generally appeared greater for students with Head Start experience compared to those students with non-Head Start experience.

Chi-square tests for the expected proportion (converted to percent) of students rated at the highest levels across experimental groups were calculated according to the following formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O = expected percent for each experimental group

E = mean percent across groups

Results of the chi-square analysis for the kindergarten data showed that for all items but ‘Listens with understanding,’ there were significant differences between the expected percent of students at the highest ability level across experimental groups [ $\chi^2(3) \geq 8.441$ ,  $p \leq 0.038$ ]. On the other hand, results of a chi-square analysis showed no significant differences in expected percents for the highest ability students across groups for the ELS data [ $\chi^2(3) \leq 4.667$ ,  $p \geq 0.198$ ]. For ‘running records,’ however, there was a significant difference in the expected percent of students at the highest ability level between the Treatment/Head Start (27%) and Comparison/non-Head Start (14%) groups [ $\chi^2(1) = 4.122$ ,  $p = 0.042$ ].

Results of the IRT procedure suggest that there may be a Treatment by ability factor. In other words, distinctions across experimental groups were noted for students with latent abilities at the higher end of the continuum underlying the kindergarten assessment. In addition, there was a tendency for students with higher underlying abilities on the first grade assessment to perform differently across experimental groups, although not statistically significant.

#### *Reading Comprehension Subtest of the Metropolitan Achievement Tests*

Scores for the standardized MAT6 Reading Comprehension subtest were submitted to two different procedures to identify the instrument’s ability to distinguish among the four experimental groups in the study. First, an analysis of covariance (ANCOVA) of MAT6 total raw scores was performed using SPSS for Windows, Release 6.1 (1994). The groups were: Treatment/Head Start, Treatment/non-Head Start, Comparison/Head Start, and Comparison/non-Head Start. Second, the MAT6 data was fitted to a three-parameter IRT model using MULTILOG, Release 6.30 (Thissen, 1991). The procedure was executed to determine whether differences in expected proportions of correct responses among the four experimental groups could be identified using item response theory techniques.

Of the 170 Cohort II MAT6 records available for this study, 115 were retained for the analyses. Fifty-five (55) records were eliminated because either kindergarten and ELS data were not available for the students or the students had moved from a Treatment to a Comparison school, or vice versa, between taking the ELS in first grade and the MAT6 in second grade. Table 4.23 shows the results of an ANCOVA by experimental status using enrollment in ESL or special education programs as covariates and the MAT6 total raw score as the dependent variable. Total raw scores were used instead of factor scores because they are the traditional method for reporting test results.

**Table 4.23.** ANCOVA on MAT6 Reading Comprehension total raw scores by Head Start and school status with ESL and special education as covariates (n=115).

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Covariates	219.807	2	109.903	1.109	.333
ESL	175.931	1	175.931	1.776	.185
Special Education	43.876	1	43.876	.443	.507
Main Effects	331.827	3	110.609	1.117	.346
Groups	331.827	3	110.609	1.117	.346
Explained	551.634	5	110.327	1.114	.357
Residual	10797.357	109	99.058		
Total	11348.991	114	99.553		

Results of the ANCOVA indicated that there were no significant differences in mean raw MAT6 scores among experimental groups. [F(3)=1.117, p=0.346] after accounting for the effects of enrollment in English as a second language (ESL) and special education programs.

*IRT Procedures for the MAT6 Reading Comprehension Subtest*

The MAT6 data were also fitted to one-, two-, and three-parameter IRT models. The data fit all three IRT models poorly (p<0.01) as shown in the chart below. The poor fit for the three models may be due to the fact that the data violate the assumption inherent in these IRT models, namely, that the measure is unidimensional. A factor analysis on the data revealed sixteen factors with eigenvalues in excess of 1.0, an indication of the multidimensionality of the measure.

Thissen (1991) noted that in fitting data to an IRT model, *twice the negative loglikelihood* is asymptotically distributed as a chi-square with ([n-number of groups] - [parameters fitted]) degrees of freedom. When the p-value associated with the chi-square statistic approaches 1, the data fits the model, that is, observed and expected values are

essentially equivalent and measurement error is minimal. A significant chi-square, on the other hand, indicates a poor model fit.

IRT Model	Twice the negative loglikelihood	# of parameters	p-value
one-parameter	4032.6	54	<0.01
two-parameter	3899.2	106	<0.01
three-parameter	3780.4	159	<0.01

The difference between the *twice the negative loglikelihood* values for different models can be used to indicate the relative fit for these models. When comparing the values for two different IRT models, the model associated with the smaller *twice the negative loglikelihood* value has the better fit. The difference between the two values approximates a chi-square distribution with the number of degrees of freedom equal to the difference in the number of parameters fitted. As a result, the chi square test indicates whether the data significantly fit one model better than another model (Thissen, 1991).

Comparisons were made to determine the better fit for the MAT6 data among the three models. First, the data fit for the three-parameter model was compared to that for the one-parameter model. Using the critical value of chi-square with 105 degrees of freedom at  $\alpha=0.05$  of 130, the difference in the *twice the negative loglikelihood* values for the two models (252.2) was statistically significant. Second, the data fit for the three-parameter and two-parameter models were compared. Using the critical value of chi-square with 53 degrees of freedom at  $\alpha=0.05$  of 71, the difference in the *twice the negative loglikelihood* values for the two models (118.8) was also statistically significant.

Because smaller *twice the negative loglikelihood* values imply a better fit, a significant chi-square in the difference score indicates that the model with the smaller fit statistic provides a significantly better fit than the model with the larger fit statistic. Consequently, the MAT6 data fit the three-parameter model better than the one- or two-parameter models; therefore, this model was selected for the subsequent procedure.

The three-parameter model estimates the difficulty, discrimination, and pseudo-chance-level parameters. The difficulty parameter (*b*) as noted previously, is the point on the underlying ability continuum at which 50% of the respondents are expected to identify the correct response. The item discrimination parameter (*a*), defined earlier, represents the slope of the item characteristic curve (ICC) at point *b*. The pseudo-chance-level parameter (*c*) “is incorporated into the model to take into account performance at the low end of the ability continuum where guessing is a factor in test performance on selected-response (e.g., multiple choice) test items” (Hambleton, et al., 1991, p.17). Table 4.24 shows the results of the IRT procedure using the MAT6 responses for students who also had kindergarten and ELS scores (n=115).

**Table 4.24.** MAT6 (Reading Comprehension) IRT parameter estimates and expected proportion of correct responses by experimental groups (n=115).

Item #	Item Discrimination (a)	Item Difficulty (b)	Pseudo-Chance-Level (c)	Treatment Head Start	Treatment Non-Head Start	Comparison Head Start	Comparison Non-Head Start
1	6.85	-1.86	0.46	1.00	1.00	.99	.98
2	4.42	-1.25	0.00	0.97	.98	.94	.89
3	18.13	-1.78	0.00	0.99	1.00	.98	.96
4	1.14	-0.54	0.55	0.91	.93	.89	.85
5	-0.17	-0.54	0.00	0.97	.97	.97	.98
6	9.02	-0.30	0.82	.97	.98	.95	.93
7	0.64	-1.31	0.00	.85	.87	.82	.76
8	0.60	-3.24	0.00	.97	.98	.96	.95
9	0.88	-1.02	0.00	.86	.89	.82	.75
10	1.73	-0.01	0.36	.81	.85	.76	.68
11	1.34	-0.45	0.21	.84	.88	.80	.72
12	0.69	0.11	0.00	.61	.66	.56	.47
13	1.05	-0.66	0.00	.82	.86	.77	.68
14	6.11	0.88	0.49	.70	.74	.65	.59
15	7.82	0.63	0.35	.66	.71	.59	.51
16	0.84	0.62	0.17	.58	.63	.53	.45
17	0.99	-0.69	0.00	.82	.86	.77	.69
18	0.81	-0.14	0.17	.74	.78	.69	.62
19	1.25	-0.23	0.13	.78	.82	.72	.63
20	1.49	0.12	0.06	.68	.74	.60	.49
21	2.70	0.35	0.41	.76	.81	.71	.63
22	3.10	-0.02	0.26	.80	.85	.74	.63

**Table 4.24 (continued).** MAT6 (Reading Comprehension) IRT parameter estimates and expected proportion of correct responses by experimental groups (n=115).

Item #	Item Discrimination (a)	Item Difficulty (b)	Pseudo-Chance-Level (c)	Treatment Head Start	Treatment Non-Head Start	Comparison Head Start	Comparison Non-Head Start
23	1.59	0.41	0.63	.84	.87	.81	.77
24	1.76	-0.34	0.63	.92	.94	.90	.86
25	0.98	-0.10	0.42	.82	.85	.78	.72
26	3.72	-0.26	0.16	.83	.88	.77	.66
27	8.88	-0.04	0.20	.80	.85	.74	.63
28	0.84	-0.22	0.00	.71	.75	.65	.56
29	2.84	0.11	0.67	.89	.92	.87	.82
30	2.23	-0.02	0.45	.84	.88	.80	.73
31	2.04	0.57	0.40	.71	.76	.66	.58
32	1.13	-0.51	0.00	.80	.84	.74	.65
33	1.04	-0.05	0.00	.69	.74	.62	.52
34	0.86	0.90	0.33	.62	.65	.57	.52
35	0.47	-0.07	0.00	.62	.65	.58	.51
36	0.59	0.22	0.42	.75	.78	.73	.68
37	1.50	0.08	0.39	.80	.84	.75	.68
38	1.33	1.44	0.17	.39	.43	.34	.28
39	0.78	-0.93	0.00	.83	.86	.79	.72
40	1.52	0.41	0.21	.66	.71	.59	.50
41	2.09	0.56	0.32	.68	.73	.62	.53
42	1.29	0.84	0.35	.63	.67	.58	.52
43	1.23	0.68	0.38	.68	.72	.63	.57

**Table 4.24 (continued).** MAT6 (Reading Comprehension) IRT parameter estimates and expected proportion of correct responses by experimental groups (n=115).

Item #	Item Discrimination (a)	Item Difficulty (b)	Pseudo-Chance-Level (c)	Treatment Head Start	Treatment Non-Head Start	Comparison Head Start	Comparison Non-Head Start
44	0.96	0.05	0.23	.73	.77	.68	.60
45	1.97	0.99	0.35	.59	.64	.54	.48
46	1.76	0.92	0.13	.47	.54	.40	.31
47	0.95	0.66	0.22	.60	.65	.55	.47
48	2.58	1.12	0.23	.48	.53	.42	.34
49	1.79	1.42	0.18	.38	.43	.33	.27
50	0.29	1.42	0.00	.41	.43	.38	.34
51	1.53	1.83	0.22	.35	.38	.31	.27
52	0.12	1.68	0.00	.45	.46	.44	.42
53	0.36	1.33	0.00	.40	.43	.37	.32

Table 4.24 shows that the MAT6 Reading Comprehension items varied in difficulty and discrimination power across the latent ability continuum. In addition, the pseudo-chance-level parameter, was a factor in the responses to several items. For example, Item 1 was a relatively easy item. Fifty percent (50%) of respondents at -1.86 on the ability continuum would be expected to answer the item correctly. The item discriminated well, as the slope of the ICC at -1.86 was 6.85. Further, the pseudo-chance-level parameter value of 0.46 shows that ‘guessing’ may have influenced responses of students at the lower end of the ability continuum. On the other hand, Item 52 was a relatively difficult item ( $b=1.68$ ). Unlike Item 1, the slope for this item at point  $b$  was less steep ( $a=0.12$ ). As a result, the item did not discriminate sharply between respondents above and below 1.68 on the latent ability continuum. The pseudo-chance-level factor was 0.00 indicating that students at the lower end of the ability continuum did not identify the correct response by ‘guessing.’ It is most likely that those students did not attempt the item.

To determine whether a significant difference existed across experimental groups for each MAT6 item, a chi-square analysis was performed on the values of the expected proportion correct converted to percent correct. Results of the analysis indicated that there were no significant differences across experimental groups in expected percent correct for each of the 53 items in the MAT6 Reading Comprehension subtest [ $\chi^2(3) \leq 6.744$ ,  $p \geq 0.081$ ].

## **CHAPTER 5**

### **SUMMARY, CONCLUSIONS AND RECOMMENDATIONS**

Using a variety of classical test theory (CTT) and item response theory (IRT) procedures and a data set consisting of responses from Cohort II participants at the local site of the National Transition Project to four literacy measures, each of the purposes of the research were addressed. The literacy measures included a variety of early childhood literacy assessments developed by a local school system as well as a formal reading achievement measure. The purposes of the research were: to confirm the developmental nature of the assessments, to verify the ability of the assessments to locate students on the developmental continuum underlying the assessments, to investigate the feasibility of an early literacy assessment package for program evaluation, and to explore the usefulness of an IRT analysis of a standardized measure for program evaluation.

#### **Summary and Conclusions**

Traditionally, standardized achievement tests have been used to monitor program effectiveness. Recently, however, educators have questioned the appropriateness of standardized tests for this purpose, especially for programs designed for young children. Early childhood advocates, namely the National Association for the Education of Young Children (NAEYC) and the National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE), proposed guidelines for developmentally appropriate assessment in programs serving children between the ages of three and eight. Their suggestions were based on the current literature for programs implementing developmentally appropriate practices. The NAEYC and NAECS/SDE suggest using developmentally appropriate assessments instead of standardized achievement tests for making classroom-level decisions about children. In addition, researchers in the field of early childhood education advocate the limited use of standardized achievement measures for young children because formal tests narrow the definition of emergent literacy, lack consideration for literacy development and the characteristics of young learners, require children to perform tasks unrelated to classroom practices, and provide a limited use for instruction. To date, these proponents have not identified the psychometric properties of the assessments to confirm that they are developmental assessments. Further, they have not investigated the use of the assessments for evaluating the effectiveness of developmentally appropriate programs.

This study found that the Concepts About Print portion of the Early Childhood Assessment Package (ECAP), the Language Arts component of the kindergarten developmental progress reports, and the Early Literacy Scale (ELS) tasks are, in fact, developmental assessments. They were designed to measure emergent literacy behaviors according to the research on child development and literacy acquisition. They adhere to the criteria established by the NAEYC and NAECS/SDE (1991) for developmentally appropriate assessments in that they are based on observations within the context of the

classroom instructional program, tap a variety of literacy behaviors, and are age-appropriate.

Descriptive statistics were generated, and the items within each of the instruments were rank ordered by mean scores. For the dichotomously scored ECAP items, the means represented item difficulty levels. It was shown that the items in the instrument ranged from very easy to difficult.

Because classical item difficulty indices determine to what extent items distinguish between high and low scoring individuals but not how response behavior is related to a developmental sequence of skills underlying a measure, the assessment data were fitted to IRT models. Results of those procedures showed that response behavior for the three locally developed literacy assessments could be equated to underlying, or latent, literacy ability levels. As a result, a developmental sequence of skills underlying each of the assessments was identified. The ECAP items represented a developmental sequence of Concepts About Print. The teacher ratings for degree of use of kindergarten Language Arts behaviors over time and the first grade ELS tasks also represented developmental sequences of literacy behaviors.

Using IRT procedures, latent ability levels for the Cohort II participants from the local site of the National Transition Project were estimated for each of the assessments. Accordingly, the Cohort II students were located on the developmental sequences underlying each of the instruments.

Whether the assessments afford utility for program evaluation is dependent upon the precision with which the assessments are able to measure literacy development and their ability, in fact, to distinguish across experimental groups. Traditionally, standardized tests have been used to evaluate programs because they are highly reliable measures with enough power to detect program differences.

Reliability estimates for the four instruments were generated through classical test theory analyses. Results showed that internal consistency reliability estimates (Cronbach's  $\alpha$ ) were strong suggesting that the items within each measure are homogenous, that is, representative of the same general content domain (Crocker and Algina, 1986, p.135). Corresponding standard errors of measurement indicated that with the exception of the first grade Early Literacy Scale, the instruments were sufficiently reliable.

Using factor scores for teacher ratings on the Language Arts kindergarten developmental progress reports and those for the first grade ELS tasks, analyses of covariance (ANCOVAs) by experimental groups were performed. The groups were: Treatment/Head Start, Treatment/non-Head Start, Comparison/Head Start, and Comparison /non-Head Start. Enrollment in English as a second language (ESL) and special education programs served as covariates. Results of the ANCOVA with kindergarten factor scores as the dependent variable showed both a Treatment and a Head Start effect suggesting that the procedure for randomly assigning Treatment and Comparison status may have resulted in groups that differed in emergent literacy ability. No significant differences among the groups were noted with the ELS factor scores as the dependent variable. Results of an ANCOVA using raw score totals for the second grade

Reading Comprehension subtest of the MAT6 also showed no differences among the mean scores for the four experimental groups.

Because item response theory procedures equate response behavior to the developmental sequence of skills underlying a measure, two data sets were fitted to IRT models. The first data set consisted of teacher ratings of kindergarten and first grade literacy behaviors represented by the kindergarten progress reports and the Early Literacy Scale. The second set of data was the second grade standardized MAT6 (Reading Comprehension subtest) responses.

An early childhood literacy assessment package consisting of the kindergarten and first grade measures differentiated experimental groups by their location on the developmental continuum underlying those assessments. Among the students with the highest latent ability, that is, those rated 'consistently' in degree of use of kindergarten Language Arts behaviors and at the 'independent' literacy level for first grade ELS tasks, differences across experimental groups were noted. These differences were statistically significant for the kindergarten data. Although results from the IRT procedure with the first grade assessment data showed a tendency for students with the highest latent ability to perform differently across experimental groups, the differences were not statistically significant. On the 'running records' task, however, there was a statistically significant difference between expected proportions of high ability students in the Treatment/Head Start and Comparison/non-Head Start groups.

Results of the procedure using second grade MAT6 (Reading Comprehension subtest) responses, however, differed from those obtained with the kindergarten and first grade assessments. Results showed that for the MAT6 (Reading Comprehension subtest) items, differences in the expected proportion correct for students were not statistically significant across experimental groups.

Although the MAT6, like other standardized tests, has high reliability as measured by classical procedures, it lacks unidimensionality. In addition, a score on a standardized measure is a function of raw score position relative to a mean of a norming group. It is not a position on an established sequence relative to a developmental trait. As a result, the data from standardized measures may not fit IRT models as well as assessment data. Consequently, the utility of employing IRT procedures on data from standardized tests is limited for evaluating program effectiveness.

The difference in IRT results between the early childhood assessment package and the standardized MAT6 raises two important issues. One has to do with validity. The other issue concerns the source of the difference between the two measures.

The kindergarten and first grade measures were developed to assess early literacy behavior according to current research in that field. The MAT6, on the other hand, was developed to reflect the research on reading behavior prevalent in the 1970's. As a result, the kindergarten and first grade assessments possess construct and content validity for today's early reading programs. It can be logically inferred that the developmental sequence of skills underlying these assessments is, therefore, more valid for the emergent literacy program implemented by the school system in this study. Consequently, a package of early childhood developmental assessments is a more valid indicator of literacy

behavior than a formal achievement measure and, therefore, should be used for evaluating the program at the local site of the National Transition Project.

Further, the items within an assessment equate to a developmental sequence of skills underlying the measure. Because items are not designed to be of equal difficulty, assessments lack the precision of standardized measures. As a result, classical methods for identifying the properties of assessments and their ability to discern differences for program evaluation are unsuitable. Item response theory procedures, on the other hand, provide a more appropriate method for determining program effectiveness when using assessment data. IRT procedures offer a method for identifying differences across latent abilities, and as such, provide insights into differences between groups of students that classical analyses are unable to discern.

Whether the difference in results between the kindergarten and first grade assessments reported in this research is due to student literacy development and/or variability in teacher rating behavior remains problematic. In addition, whether the source of the discrepancy in results between the early childhood assessments and the second grade standardized measure is due to student literacy development or to the nature of the measures themselves is unclear. Further research may clear up these ambiguities.

### **Recommendations**

Based on the results and conclusions of this study, several recommendations for further research are proposed. Included among those recommendations are replicating the study with a larger sample, investigating whether a second grade literacy assessment maintains the differences among experimental groups identified by the kindergarten and first grade assessment package, and exploring the improvement of the accuracy of the assessments through generalizability theory procedures.

First, the study should be replicated to verify the results and conclusions presented in the present research. In doing so, it is suggested that a larger sample be used. Several statistical analyses were unable to be performed in this study because the data set was too small to reliably draw conclusions. It is suggested that the sample be enlarged to increase the statistical power of the analyses to detect differences in measures with high measurement error and to account for problems associated with attrition from one grade level to another as well as listwise deletions. As a result, the ECAP data could be incorporated into the early childhood assessment package for program evaluation. In addition, with a larger sample the study could be replicated without the effects of ESL and special education enrollment. Further, an increase in the sample size would allow for a longitudinal study rather than the cross-sectional studies presented in this research.

Second, the Cohort II students provided a writing sample in the spring of their second grade year which was graded by a team of independent raters. These data were not available for use in the current study. It is proposed that both classical and IRT procedures be applied to the data and compared with the results of the kindergarten and first grade assessments presented in this research to determine whether any differences among experimental groups for students at the higher end of the underlying ability continuum continue into the second grade.

Third, a great deal of research has been reported in the literature on the use of generalizability theory methods with alternative, or authentic, assessments. It is proposed that the variance of the literacy assessments investigated in this study be partitioned among the various facets using g-theory methods. Results of that research may provide suggestions for improving both the reliability and precision of the assessment instruments. Consequently, their potential as valid and reliable program evaluation measures would increase. Limitations with the data prevented the use of this procedure in the present study.

Further research studies to address the recommendations outlined above may determine whether these literacy assessments provide a viable alternative to standardized measures for program evaluation. Then, not only will the assessments be valuable for making instructional decisions within the classroom, but their ability to evaluate instructional programs will increase their utility. Hence, the overall costs inherent in alternative assessments will decrease, thereby, making them more desirable for widespread use.

The pivotal contribution of this study is the rendering of additional support to the limited research in utilizing developmentally appropriate assessments for program evaluation. Although further investigation is required, this study supports the findings of the research on developmentally appropriate practices and literacy acquisition. The theory underlying developmentally appropriate assessments that can be utilized for instructional decisions in the classroom may also be used for evaluating program effectiveness. This research contributes to the methodology for discerning program effects with developmentally appropriate assessments using item response theory procedures rather than traditional classical analyses. Clearly, findings of this research warrant considerations by administrators and policy makers in determining evaluation indicators for early childhood programs.

## REFERENCES

- Allen, P.D. & Watson, D.J. (Eds.). (1976). Findings of research in miscue analysis: Classroom implications. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills & National Council of Teachers of English.
- Archbald, D.A. & Newmann, F.M. (1988). Beyond standardized testing: Assessing authentic achievement in the secondary school. Reston, VA: National Association of Secondary School Principals.
- Baker, F.B. (1992). Item response theory: Parameter estimation techniques. New York: Marcel Dekker, Inc.
- Bergan, J.R. (1988). *Latent variable techniques for measuring development*. In R.Langeheine & J.Rost (Eds.) Latent trait and latent class models (pp.233-261). New York: Plenum Press.
- Casteen, J.T., III. (1982). *The public stake in proper test use*. In C.W. Daves (Ed.), The uses and misuses of tests: Examining current issues in educational and psychological testing (pp.1-11). San Francisco: Jossey-Bass Publishers.
- Chittenden, E.A. & Courtney, R. (1989). *Assessment of young children s reading: Documentation as alternative to testing*. In D.S. Strickland & L.M. Morrow (Eds.), Emerging Literacy: Young children learn to read and write (pp.107-120). Newark, DE: International Reading Association.
- Clay, M.M. (1970). *An increasing effect of disorientation on the discrimination of print: A developmental study*. Journal of Experimental Child Psychology, 9, 297-306.
- Clay, M.M. (1985). The early detection of reading difficulties: A diagnostic survey with reading recovery procedures (3<sup>rd</sup> ed.). Auckland, NZ: Heinemann.
- Clay, M.M. (1991). Becoming literate: The construction of inner control. Portsmouth, NH: Heinemann.
- Clay, M.M. (1993). An observation survey of early literacy achievement. Portsmouth, NH: Heinemann.
- Cohn, D. (1995, July 9). *Minorities settle down in suburbs*. The Washington Post, pp. B1, B6.

- Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Harcourt Brace Jovanovich College Publishers.
- Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E. (1995). Generalizability analysis for educational assessments. Los Angeles: University of California, Center for the Study of Evaluation & The National Center for Research on Evaluation, Standards, and Student Testing.
- Cunningham, P. (1982). *Diagnosis by observation*. In J.J. Pikulski & T. Shanahan (Eds.), Approaches to the informal evaluation of reading (pp. 12-22). Newark, DE: International Reading Association.
- Day, H.D. & Day, K. (1978). The reliability and validity of the concepts about print and record of oral language. Arlington, VA: Resources in Education. (ERIC Documentation Reproduction Service No. ED 179 932)
- Early childhood assessment package. (1993). Fairfax, VA: Fairfax County Public Schools.
- Early childhood assessment package. (1994). Fairfax, VA: Fairfax County Public Schools.
- Early literacy scale. (1993). Fairfax, VA: Fairfax County Public Schools.
- Ebel, R.L. (1977). The uses of standardized testing. Bloomington, IN: Phi Delta Kappa Educational Foundation.
- Engel, B.S. (1991). *An approach to assessment in early literacy*. In C. Kamii, M.M. Manning, & G.L. Manning, (Eds.), Early literacy: A constructivist foundation for whole language (pp. 129-148). Washington, DC: National Education Association.
- Farr, R. & Carey, R.F. (1986). Reading: What can be measured? (2nd ed.). Newark, DE: International Reading Association.
- Ferreiro, E. (1990). *Literacy development: Psychogenesis*. In Y.M. Goodman (Ed.), How children construct literacy (pp.12-25). Newark, DE: International Reading Association.
- Goodman, Y.M. & Burke, C. (1972). The reading miscue inventory. New York: Macmillan.
- Goodman, Y.M. & Burke, C.L. (1970). Reading miscue inventory manual procedure for diagnosis and evaluation. New York: Macmillan.

- Green, K.E. (Ed.). (1991). Educational testing: Issues and applications. New York: Garland Publishing, Inc.
- Gullo, D.F. (1994). Understanding assessment and evaluation in early childhood education. New York: Teachers College Press.
- Hambleton, R.K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijoff Publishing.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publications, Inc.
- Hiebert, E.H. & Calfee, R.C. (1992). *Assessing literacy: From standardized tests to portfolios and performances*. In S.J. Samuels & A.E. Farstrup (Eds.), What research has to say about reading instruction (2nd ed.). (pp.70-100). Newark, DE: International Reading Association.
- Hills, T.W. (1992). *Reaching potentials through appropriate assessment*. In S. Bredekamp & T. Rosegrant (Eds.), Reaching potentials: Appropriate curriculum and assessment for young children (Volume 1, pp. 43-63). Washington, DC: National Association for the Education of Young Children.
- Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item response theory. Homewood, IL: Dow Jones-Irwin.
- Johns, J.L. (1982). *The dimensions and uses of informal reading assessment*. In J.J. Pikulski & T. Shanahan (Eds.), Approaches to the informal evaluation of reading (pp.1-11). Newark, DE: International Reading Association.
- Johns, J.L. (1980). *First graders concepts about print*. Reading Research Quarterly, 15(4), 529-549.
- Johnston, P.H. (1983). Reading comprehension assessment: A cognitive basis. Newark, DE: International Reading Association.
- Johnston, P.H. (1984). *Assessment in reading*. In P.D. Pearson, R. Barr, M.R. Kamil, & P. Mosenthal (Eds.), Handbook of reading research (Volume 1, pp.147-182). New York: Longman, Inc.
- Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1993). Interim report: The reliability of the Vermont portfolio scores (Technical Report No. 370). Los Angeles: University of California, Center for the Study of Evaluation & The National Center for Research on Evaluation, Standards, and Student Testing.

- Madaus, G.F. & Tan, G.A. (1993). *The growth of assessment*. In G. Cawelti (Ed.), Challenges and achievements of American education (pp. 53-79). Alexandria, VA: Association for Supervision and Curriculum Development.
- Maeroff, G.I. (1991). *Assessing alternative assessment*. Phi Delta Kappan, 73(4), 272-281.
- Maine State Department of Educational and Cultural Services. (1988). Developmentally appropriate practice: A guide to change. Augusta, ME: Author.
- Mehrens, W.A. & Lehmann, I.J. (1969). Standardized tests in education. New York: Holt, Rinehart and Winston.
- Meyer, C.A. (1992). *What s the difference between authentic and performance assessment?* Educational Leadership, 49(8), 39-40.
- Mislevy, R.J. (1994). Test Theory Reconceived (Technical Report No. 376). Los Angeles: University of California, Center for the Study of Evaluation & The National Center for Research on Evaluation, Standards, and Student Testing.
- Moore, R.E. (1992). *Developmentally appropriate assessment: Alternatives to standardized testing*. Journal of Humanistic Education and Development, 30, 122-130.
- Morrow, L.M. (1988). *Retelling stories as a diagnostic tool*. In S.M. Glazer, L.W. Searfoss, & L.M. Gentile (Eds.), Reexamining reading diagnosis: New trends and procedures (pp.128-149). Newark, DE: International Reading Association.
- National Association for the Education of Young Children & The National Association of Early Childhood Specialists in State Departments of Education. (1991). *Guidelines for appropriate curriculum content and assessment in programs serving children ages 3 through 8*. Young Children, 46, 21-38.
- Nurss, J.R. & McGauvran, M.E. (1986). Metropolitan readiness tests, Level 1. (5<sup>th</sup> ed.). Abstract. (on-line) `gopher_root_eric_ae:[tc]e0594.txt;1'`.
- Office of Technology Assessment. (1992). Testing in American schools: Asking the right questions (OTA-Set-520). Washington, DC: Congress of the U.S. Office of Technology Assessment.
- Pearson, P.D. & Stallman, A.C. (1993). Approaches to the future of reading assessment: Resistance, complacency, reform (Technical Report No. 575). Champaign: University of Illinois, Center for the Study of Reading.

- Prescott, G.A., Balow, I.H., Hogan, T.P., & Farr, R.C. (1989). Test Manual. Metropolitan Achievement Tests, Sixth edition. New York: Harcourt, Brace, & Jovanovich, Inc.
- Ravich, D. (1982). *Value of standardized tests in indicating how well students are learning*. In C.W. Daves (Ed.), The uses and misuses of tests: Examining current issues in educational and psychological testing (pp.1-11). San Francisco: Jossey-Bass Publishers.
- Readence, J.E. & Martin, M.A. (1988). *Comprehension assessment: Alternatives to standardized tests*. In S.M. Glazer, L.W. Searfoss, & L.M. Gentile (Eds.), Reexamining reading diagnosis: New trends and procedures (pp.67-80). Newark, DE: International Reading Association.
- Resnick, L., Resnick, D., & DeStefano, L. (1993). Cross-scorer and cross-method comparability and distribution of judgements of student math, reading, and writing performance: Results from the new standards project Big Sky scoring conference (Technical Report No. 368). Los Angeles: University of California, Center for the Study of Evaluation & The National Center for Research on Evaluation, Standards, and Student Testing.
- Sattler, J.M. (1992). Assessment of children (3<sup>rd</sup> ed.). San Diego: Jerome M. Sattler, Publisher, Inc.
- Simmons, J. (1990). *Adapting portfolios for large-scale use*. Educational Leadership, 74(6), 28.
- Southern Association on Children Under Six. (1990). Five position statements of the southern association on children under 6 (SACUS): Developmentally appropriate assessment. Little Rock, AK: Southern Association on Children Under Six. (ERIC Document Reproduction Service No. ED 319 493)
- SPSS, Inc. (1994). SPSS for windows, Release 6.1. Chicago: Author.
- Stallman, A.C. & Pearson, P.D. (1990). Formal measures of early literacy. (Technical Report No. 511). Champaign: University of Illinois, Center for the Study of Reading. (ERIC Documentation Reproduction Service No. ED 324 647)
- Stiggins, R.J. (1995). *Assessment Literacy for the 21<sup>st</sup> century*. Phi Delta Kappan, 77(3), 238-245.
- Strickland, D.S. (1990). *Emergent literacy: How young children learn to read and write*. Educational Leadership, 47(6), pp.18-23.

- SYSTAT, Inc. (1989). SYSTAT: The system for statistics for the PC (2<sup>nd</sup> ed.). Evanston, IL: Author.
- Teale, W.H. (1988). *Developmentally appropriate assessment of reading and writing in the early childhood classroom*. Elementary School Journal, 89(2), 173-183.
- Teale, W.H. & Sulzby, E. (1989). *Emergent literacy: New perspectives*. In D.S. Strickland & L.M. Morrow (Eds.), Emerging Literacy: Young children learn to read and write (pp.107-120). Newark, DE: International Reading Association.
- Thissen, D. (1991). MULTILOG: Item analysis and scoring with multiple category response models. Mooreville, IN: Scientific Software, Inc.
- Tierney, R.J., Carter, M.A., & Desai, L.E. (1991). Portfolio assessment in the reading-writing classroom. Norwood, MA: Christopher-Gordon Publishers, Inc.
- Valencia, S.W., Hiebert, E.H., & Afflerbach, P.P. (Eds.). (1994). Authentic reading assessments: Practices and possibilities. Newark, DE: International Reading Association.
- Valencia, S.W., Pearson, P.D., Peters, C.W., & Wixson, K.K. (1989). *Theory and practice in statewide reading assessment: Closing the gap*. Educational Leadership, 46(7), 57-63.
- VanLeirsburg, P. (1991). Socio-educational influences on standardized reading tests, 1900-1991 Literacy Research Report No.11. Dekalb, IL: Northern Illinois University. (ERIC Document Reproduction Service No. ED 335 622)
- Wigdor, A.K. & Garner, W.R. (Eds.). (1982). Ability testing: Uses, consequences, and controversies (Part 1). Washington, DC: National Academy Press.
- Wiggins, G. (1990). *A case for authentic assessment*. ERIC Digest. Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation.
- Wiggins, G.P. (1993). Assessing student performance. San Francisco: Jossey-Bass Inc., Publishers.
- Williams, E.J. (1991). *Curriculum-based assessment*. In K.E. Green (Ed.), Educational testing: Issues and applications (pp.109-123). New York: Garland Publishing, Inc.

## **APPENDIX**

**Table A.1.** t-tests for independent samples of Head Start and non-Head Start mean factor scores by program enrollment.

Program (#HeadStart-#Non-HeadStart)	Head Start		Non-Head Start		t-value(df)	p
	m	sd	m	sd		
ECAP Factor 1 (n=19)						
Special Education						
none (12-4)	-.13	1.22	-.25	1.27	.17(14)	.864
enrolled (3-0)	.63	.14	na	na	na	na
ESL						
none (9-3)	.42	.83	.26	.94	.29(10)	.774
enrolled (6-1)	-.58	1.32	-1.78	na	na	na
ECAP Factor 2 (n=19)						
Special Education						
none (12-4)	.18	.19	-1.18	2.70	1.01(3.01)	.389
enrolled (3-0)	.09	.05	na	na	na	na
ESL						
none (9-3)	.10	.13	-1.71	3.04	1.03(2)	.410
enrolled (6-1)	.26	.19	.43	na	na	na
3 <sup>rd</sup> quarter Kindergarten Language Arts (n=155)						
Special Education						
none (101-40)	.21	.91	-.05	1.02	1.49(139)	.137
enrolled (11-3)	-.14	1.58	.01	.10	-.30(10.26)	.478
ESL						
none (71-29)	.34	.98	.22	.91	.57(98)	.573
enrolled (41-14)	-.12	.96	-.62	.90	1.72(53)	.092
ELS (n=192)						
Special Education						
none (124-49)	.08	.96	.10	.90	-.10(171)	.919
enrolled (15-4)	-1.0	.93	-.14	1.52	-1.44(17)	.168
ESL						
none (87-34)	.07	1.07	.36	.73	-1.68(88.4)	.097
enrolled (52-19)	-.21	.93	-.41	1.09	.77(69)	.446
MAT6 (n=92)						
Special Education						
none (62-24)	.02	1.02	.01	.98	.03(84)	.978
enrolled (5-1)	-.08	1.08	-.85	na	na	na
ESL						
none (45-20)	.04	.96	.01	.89	.15(63)	.885
enrolled (22-5)	-.06	1.15	-.14	1.36	.15(25)	.885

'na' indicates value not calculated because there is no variability ( $\leq 1$  student in cell)

**Table A.2.** Correlation coefficients for Concepts About Print items in the Early Childhood Assessment Package (n=31).

	Front Back	Top Bottom	Left Right	Return Sweep	Letter	Word	One-to- One	Pictures	Letter Sound	Predicts
Front-to-Back	1.0000									
Top-to-Bottom	.2854	1.0000								
Left-to-Right	1.0000	.2854	1.0000							
Return Sweep	.2646	.9269	.2646	1.0000						
Letter	.3381	.8444	.3381	.7826	1.0000					
Word	.2646	.7748	.2646	.7048	.7826	1.0000				
One-to-One	.2297	.6589	.2297	.5850	.6796	.8683	1.0000			
Pictures	.1886	.5184	.1886	.4365	.5578	.5746	.6883	1.0000		
Letter Sound	.2646	.7748	.2646	.7048	.7826	1.0000	.8683	.5746	1.0000	
Predicts	.2148	.6086	.2148	.5323	.4791	.6722	.8009	.4853	.6722	1.0000

**Table A.3.** Correlation coefficients for the 3<sup>rd</sup> quarter Language Arts scores on the kindergarten progress reports (n=190).

	LA31A	LA31B	LA32	LA33	LA34	LA35	LA36	LA37
LA 31A (Listens with understanding)	1.0000							
LA31B (Communicates ideas verbally)	.6738	1.0000						
LA32 (Explores language through rhyme, poetry, and movement)	.6004	.5429	1.0000					
LA33 (Shares in group reading & writing activities)	.6575	.6178	.6728	1.0000				
LA34 (Understands ideas from literature)	.7085	.6608	.7206	.8022	1.0000			
LA35 (Uses beginning reading strategies)	.6475	.5696	.6266	.6559	.7552	1.0000		
LA36 (Communicates ideas with drawings and words)	.5715	.5070	.4793	.5331	.5239	.5169	1.0000	
LA 37 (Chooses reading and writing as independent activities)	.4858	.3542	.5384	.4692	.5166	.5929	.5056	1.0000

**Table 4.4.** Inter-item correlations for Reading Comprehension items in the MAT6 (n=170).

Item	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16
1	1.0000														
2	-.0158	1.0000													
3	.7032	-.0111	1.0000												
4	.2182	-.0327	.3395	1.0000											
5	.3335	-.0226	-.0226	-.0666	1.0000										
7	-.0666	-.0468	-.0468	-.0333	.0492	1.0000									
8	.5642	-.0280	.3967	.2303	.3750	.1207	1.0000								
9	.1354	.2373	.2373	.2804	-.0953	-.0377	.1207	1.0000							
10	.1074	.2052	-.0541	.1270	.0217	-.0096	.1904	.2090	1.0000						
11	.1014	.1986	.1986	.1164	.0156	.1218	.2867	.4079	.3150	1.0000					
12	.0285	.1273	-.0873	.2170	.1499	.1146	.1407	.0543	.1803	.2090	1.0000				
13	.3246	-.0487	.2282	.0605	.0418	.1060	.2263	.2616	.4020	.3823	.0291	1.0000			
14	.0667	-.0685	.1623	.2231	-.0219	.1653	.1182	.1653	.0215	.2943	.1970	.0792	1.0000		
15	.0543	.1503	.1503	.0299	-.0363	.2555	.0963	.1295	.2149	.1923	.2761	.0445	.2609	1.0000	
16	.0183	.1190	-.0934	.0378	.0261	.2630	.1216	.1436	.0898	.1716	.4839	.0562	.2487	.3209	1.0000
17	.1277	.2282	.2282	.3663	-.0990	.1060	.1099	.3394	.1889	.3126	.3229	.1658	.2056	.4744	.2307
18	.1014	.1986	.1986	.4912	.0156	.0503	.2867	.4079	.1191	.3592	.2090	.1731	.2943	.1359	.0646
19	.2578	.1812	.1812	.2659	.1220	.1505	.3549	.4231	.3856	.4851	.1878	.3292	.2865	.2914	.2491
20	.0543	.1503	-.0739	.2776	.0777	.3185	.0963	.2555	.2149	.3617	.3713	.1674	.3121	.3038	.3209
21	.0853	-.0613	.1812	-.0913	-.0014	.2187	.0493	.2187	.0745	.2409	.2908	.3292	.2311	.2914	.2491
22	.2655	.1867	.1867	.1873	.1293	.1642	.1603	.3717	.3414	.3821	.3686	.3460	.1951	.3684	.2756
23	.2655	.1867	.1867	.0966	.1293	.2334	.1603	.0258	.0888	.1342	.3164	.2786	.1389	.4230	.2239
24	.4525	-.0349	.3182	.0271	.1085	.1505	.2086	.0512	.0109	.1797	.1005	.2338	.2683	-.0757	.1516
25	.2578	.1812	.1812	-.0020	-.0014	.1505	.1512	.2868	.2612	.3020	.1878	.3292	.1204	.2914	.2491
26	.1137	.2123	-.0523	.3329	.1627	.2255	.2016	.2255	.2200	.3360	.4876	.2055	.2814	.1799	.3384
27	.3374	-.0468	.2373	.3850	.0492	.3614	.2400	.2018	.0633	.1933	.2955	.1838	.3598	.1925	.1436
28	-.0872	-.0613	-.0613	.0873	-.1247	.0824	-.0526	.0142	.0745	.0577	.3937	.1964	.0097	.1838	.1472
29	.1438	-.0449	.2472	.1903	-.0914	.2210	.0094	.3852	.1556	.2885	.0191	.2831	.3231	.0249	.0519
30	.3128	-.0505	.2199	.0503	.1723	.1670	.3271	.1670	.2394	.2903	.2920	.2230	.3059	.2636	.1436
31	.0543	-.0739	.1503	.2776	-.0363	.2555	.1905	.0665	.2149	.3052	.1809	.2902	.2609	.5525	.3209
32	.0853	.1812	.1812	.1766	-.0014	.2868	.2530	.2868	.1367	.4241	.2908	.1299	.3419	.3452	.3510
33	.0667	.1623	-.0685	.2231	-.0219	.1653	.0213	.1653	.1990	.0619	.1970	.0792	.2101	.2098	.0549
34	.1941	.1365	.1365	.0009	-.0548	.1467	.1609	-.0370	.2739	.1386	.4243	.0610	-.0044	.2798	.2313
35	.0116	.1139	-.0976	.1018	.0165	.2422	.1093	.1234	.2295	.2006	.2209	.0935	.2190	.1010	.3348

**Table 4.4 (continued).** Inter-item correlations for Reading Comprehension items in the MAT6 (n=170).

Item	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16
36	.2824	-.0559	.1986	.0227	.1451	.1218	.2867	.1218	.1844	.2310	.0469	.2428	.1200	.1923	.0646
37	-.0796	-.0559	-.0559	-.1648	-.1138	.1933	-.1410	.1218	-.2074	.1028	.1550	.1731	.1200	.2488	.3320
38	-.0356	.0833	-.1333	.0061	-.0509	.2295	.0280	.1686	.1838	.2560	.4245	.1278	.2663	.4583	.4726
39	.3516	.2472	.2472	.0827	.0571	.2210	.1322	-.0253	.0058	.0679	.0811	.2031	.0564	.1545	.1747
40	.1896	.1333	.1333	.1534	-.0593	.1967	.1540	.1358	.4831	.1259	.2653	.2876	.1293	.3585	.2556
41	.0543	.1503	.1503	.1950	-.0363	.1295	.1905	.3185	.1574	.3052	.4189	.2902	.2609	.3038	.3209
42	.0466	-.0776	-.0776	.1777	.1789	.2311	.1753	.1070	.2445	.0537	.0463	.2040	.0758	.3158	.0433
43	.2086	.1466	.1466	.1044	.0721	.1181	.0893	.2431	.2012	.2344	.1131	.2771	.1930	.2851	.1109
44	.2138	.1503	.1503	.1950	.0777	.1295	.0963	.1295	.0999	.2488	.1809	.1674	.4144	.1049	.0854
45	.0251	.1245	.1245	.2093	-.1816	.1043	-.0454	.0442	.1681	.1425	.2064	.0772	.2793	.4016	.3276
46	-.0049	.1020	-.1089	.0674	-.0071	.0742	.0798	.1928	.2806	.1949	.2189	.2736	.2915	.2575	.3252
47	.2086	-.0758	.1466	.2681	-.0410	.2431	.0893	.3056	.2583	.0665	.1131	.3380	.2438	.3344	.1576
48	-.0183	.0934	.0934	.2749	-.2421	.1548	-.1216	.2144	.0192	.1492	.2823	.1764	.2844	.3385	.2044
49	.1025	.0721	.0721	.0458	-.0834	.1132	.0867	.0496	.1192	.1921	.1340	.1300	.1861	.2790	.1781
50	.1298	.0913	.0913	.1120	.0774	-.0342	.0511	-.0940	.0624	.1379	.0387	.1083	.1736	-.1485	-.0383
51	.0999	.0703	.0703	.2070	-.0893	.0395	.0812	.2320	.2253	.0663	.2125	.0577	.1724	.2143	.1587
52	-.0016	.1043	.1043	-.0034	-.0024	.0247	.0856	.0840	.0214	-.0058	-.0762	.1681	.0169	.1326	.1231
53	-.0216	.0913	-.1217	-.1232	.0774	.0256	.0511	.0256	.0078	.1379	.3100	-.0083	.0278	.0877	.3196

**Table 4.4 (continued).** Inter-item correlations for Reading Comprehension items in the MAT6 (n=170).

Item	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
17	1.0000														
18	.3126	1.0000													
19	.4621	.5462	1.0000												
20	.4130	.3052	.4528	1.0000											
21	.1964	.1798	.3018	.1838	1.0000										
22	.4134	.3201	.5575	.4230	.4984	1.0000									
23	.2112	.0722	.3213	.3684	.4394	.3406	1.0000								
24	.1371	.1797	.3155	.2375	.2308	.4148	.1568	1.0000							
25	.1964	.1798	.3018	.2914	.4182	.4394	.3213	.2308	1.0000						
26	.3504	.3360	.4095	.5320	.3460	.5572	.2351	.2975	.2825	1.0000					
27	.2616	.3364	.4231	.3815	.2868	.3717	.3717	.3489	.1505	.4486	1.0000				
28	.1964	.0577	.1272	.1300	.4763	.2622	.4394	.1461	.1854	.2825	.3550	1.0000			
29	.2031	.2150	.1725	.3489	.2426	.3283	.1149	.3689	.3127	.2470	.3031	.0323	1.0000		
30	.2971	.2222	.4349	.3836	.3700	.4537	.2562	.4079	.4998	.4698	.3950	.1754	.3426	1.0000	
31	.3516	.1923	.2376	.4033	.2376	.2046	.3684	.0026	.2376	.2386	.1295	.2914	.1545	.2037	1.0000
32	.4621	.3630	.4182	.4528	.4182	.4984	.3213	.3155	.3600	.4095	.2187	.1272	.2426	.4349	.3990
33	.3319	.1781	.2311	.4144	.0651	.1951	.2513	.1877	.1204	.2814	.4246	.1204	.2564	.3059	.0562
34	.2400	.1386	.0735	.0866	.2826	.1475	.1475	.1246	.1258	.1867	.2079	.1258	-.0145	.1543	.0866
35	.2094	.2006	.2731	.2886	.1717	.1465	.1465	.1363	.3238	.3703	.1234	.1210	.2773	.3480	.3355
36	.0337	.1669	.2409	.1359	.2409	.1962	.1962	.0020	.2409	.0695	.1218	-.0644	.1414	.0860	.0794
37	.2428	.1028	.0577	.1923	.5462	.1342	.3201	.1797	.2409	.2694	.1218	.3020	.1414	.2222	.1923
38	.1871	.0923	.2519	.3142	.1999	.1826	.2881	.0349	.1999	.2225	.1686	.1479	.1493	.0892	.2181
39	.2831	.2150	.2426	.1545	.1024	.2572	.3995	.3689	.2426	.0941	.1389	-.0377	.1558	.1082	.1545
40	.2876	.0713	.3718	.3105	.3198	.2922	.3449	.1921	.3198	.3446	.3184	.2679	.2889	.3745	.3105
41	.3516	.4746	.3452	.3038	.3990	.3684	.3684	.1592	.2914	.4147	.3815	.3452	.3489	.4436	.2541
42	.0830	.2761	.2601	.1200	.1013	.2292	.1217	.0644	.2601	.2121	.1690	.1542	.0687	.2375	.1689
43	.3380	.2344	.3823	.2851	.3823	.4606	.2982	.1502	.3823	.3416	.2431	.0621	.1434	.3694	.0878
44	.2902	.2488	.3990	.4033	.1300	.4230	.1500	.3941	.1838	.4147	.2555	.0225	.4137	.3236	.0552
45	.2528	.1425	.1741	.2120	.1741	.0944	.2505	.0180	.3792	.1389	.1643	.1228	.1948	.3374	.3542
46	.1581	.1949	.2085	.2108	.2085	.2382	.1355	.1732	.2085	.3700	.0742	.2085	.1688	.1817	.2575
47	.2771	.2344	.3823	.2851	.1689	.4606	.1357	.3056	.2756	.3416	.4306	.2756	.2076	.3694	.2358
48	.4090	.3096	.2094	.3385	.2603	.1898	.1898	.1451	.1584	.3286	.3338	.2603	.1935	.2541	.3385
49	.0681	.1921	.1807	.0281	.2350	.1659	.2760	.1476	.1264	.0436	.1767	.2350	.0955	.0858	.2790
50	.0500	.1379	-.0584	.0405	-.0584	-.0296	-.0815	.2869	-.0073	.0956	.0855	-.0584	.1231	.0733	.0405
51	.3079	.2388	.2234	.3156	.2234	.2096	.2096	.1411	.1138	.2115	.1678	.1138	.1523	.1363	.1130
52	-.0051	.1534	-.0317	-.0545	-.0823	-.0062	-.0062	-.1877	.0695	-.1158	.0247	-.0823	.0562	-.0899	.0390
53	.1667	.0306	.0438	-.0540	-.0073	.0741	-.0815	.1381	-.0073	.1513	-.0342	.0438	-.1231	.1302	-.0540

**Table 4.4 (continued).** Inter-item correlations for Reading Comprehension items in the MAT6 (n=170).

Item	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
32	1.0000														
33	.0651	1.0000													
34	.2826	.1448	1.0000												
35	.3746	.1708	.1507	1.0000											
36	.1188	.0038	.0838	.0942	1.0000										
37	.4241	.1200	.1935	.2006	.0387	1.0000									
38	.1999	.2663	.1904	.1882	.2560	.3105	1.0000								
39	.3828	.1898	.3003	.2162	.0679	.2885	.1493	1.0000							
40	.3198	.2282	.4166	.3102	.1259	.2350	.3000	.2263	1.0000						
41	.3990	.2609	.3282	.2886	.0230	.3617	.3142	.2841	.3105	1.0000					
42	.1542	.1766	.0972	.0568	.2205	-.0018	.2984	.0049	.2693	.1689	1.0000				
43	.2756	.3453	.1638	.0787	.2904	.2904	.1393	.1434	.3373	.1371	.3933	1.0000			
44	.2376	.3121	.0382	.1479	.0794	.0230	.2181	.1545	.3105	.3038	.2179	.3838	1.0000		
45	.2254	.2305	.0830	.1994	.1425	.1425	.3488	.1330	.3840	.2120	.2631	.2847	.1172	1.0000	
46	.2591	.2433	.2927	.2341	.0887	.3011	.3199	.2297	.4485	.3979	.2981	.2779	.3043	.2396	1.0000
47	.1689	.3453	.0679	.0322	.0105	.1224	.2346	.2076	.4326	.3344	.4418	.3640	.3344	.2377	.4171
48	.3113	.2844	.2721	.1981	.0423	.3631	.3011	.1935	.2450	.3856	.1885	.2628	.2443	.4356	.3837
49	.2892	.1345	.2357	.2074	.2491	.1921	.2834	.2263	.2983	.2288	.1579	.0935	.1285	.3878	.3293
50	.0949	.1250	.1639	.0445	.1915	.0306	-.0913	.1231	.0913	.0405	.0332	-.0335	.0405	.1479	.0952
51	.1686	.2245	.1211	.1874	.2388	.1813	.3539	.0203	.1845	.2143	.1918	.2782	.1636	.2748	.1649
52	.0189	.0169	-.0994	-.0102	.2596	-.0058	.0765	-.0047	-.0765	.0858	.0840	-.0775	-.1480	.2581	.0106
53	.0438	.2222	.1639	.1782	-.1302	.1379	.2282	.0615	-.1369	.1822	-.0598	-.0335	.0405	.1029	.0952

**Table 4.4 (continued).** Inter-item correlations for Reading Comprehension items in the MAT6 (n=170).

Item	47	48	49	50	51	52	53
47	1.0000						
48	.3563	1.0000					
49	.1930	.2971	1.0000				
50	.0134	.1726	.2179	1.0000			
51	.2782	.3690	.1061	.0000	1.0000		
52	.0617	.0097	.1254	.1206	.0549	1.0000	
53	.0602	.1278	.0749	.0577	.0962	.0317	1.0000

**Table A.5.** ECAP (Concepts About Print) Rasch ability estimates (n=13).

Case Number	Total Score	Mean Score	Rasch Ability	SE
85	1	.125	-2.469	1.204
178	1	.125	-2.469	1.204
58	3	.375	-.648	.823
61	4	.500	.000	.796
56	5	.625	.648	.823
57	5	.625	.648	.823
217	5	.625	.648	.823
68	7	.875	2.469	1.204
78	7	.875	2.469	1.204
79	7	.875	2.469	1.204
174	7	.875	2.469	1.204
228	7	.875	2.469	1.204
283	7	.875	2.469	1.204

**Table A.6.** Estimated ability levels for third quarter kindergarten Language Arts behaviors (n=190).

Case Number	KGN	
	THETAHAT	SE
108	-3.558	0.397
112	-3.558	0.397
56	-2.266	0.319
173	-1.779	0.221
289	-1.744	0.220
54	-1.674	0.231
66	-1.669	0.233
106	-1.476	0.281
107	-1.476	0.281
251	-1.392	0.306
247	-1.355	0.303
162	-1.333	0.304
57	-1.301	0.310
70	-1.297	0.307
213	-1.232	0.309
182	-1.208	0.309
258	-1.208	0.309
203	-1.198	0.310
104	-1.139	0.307
265	-1.086	0.310
223	-1.075	0.303
234	-1.075	0.303
292	-1.052	0.301
248	-0.967	0.291
235	-0.958	0.288
171	-0.950	0.294
110	-0.935	0.289
183	-0.929	0.283
210	-0.929	0.283
73	-0.917	0.281
185	-0.882	0.272
293	-0.882	0.272
65	-0.834	0.261
123	-0.800	0.252
44	-0.795	0.255

**Table A.6 (continued).** Estimated ability levels for third quarter kindergarten Language Arts behaviors (n=190).

Case Number	KGN	
	THETAHAT	SE
281	-0.785	0.248
14	-0.744	0.238
34	-0.744	0.238
102	-0.744	0.238
204	-0.744	0.238
280	-0.709	0.233
59	-0.697	0.229
266	-0.668	0.222
120	-0.639	0.217
21	-0.630	0.221
111	-0.600	0.213
114	-0.589	0.212
277	-0.512	0.216
177	-0.510	0.213
169	-0.509	0.211
221	-0.498	0.213
156	-0.469	0.216
257	-0.469	0.216
58	-0.443	0.221
262	-0.402	0.229
222	-0.385	0.234
113	-0.363	0.237
121	-0.345	0.243
10	-0.320	0.251
208	-0.320	0.251
30	-0.282	0.262
64	-0.282	0.262
184	-0.282	0.262
189	-0.274	0.263
51	-0.264	0.275
36	-0.208	0.284
25	-0.202	0.290
38	-0.193	0.291
79	-0.162	0.296
97	-0.162	0.296

**Table A.6 (continued).** Estimated ability levels for third quarter kindergarten Language Arts behaviors (n=190).

Case Number	KGN	
	THETAHAT	SE
137	-0.162	0.296
144	-0.162	0.296
149	-0.162	0.296
268	-0.162	0.296
62	-0.144	0.303
252	-0.144	0.303
60	-0.056	0.326
9	-0.024	0.326
13	-0.024	0.326
18	-0.024	0.326
23	-0.024	0.326
27	-0.024	0.326
99	-0.024	0.326
115	-0.024	0.326
118	-0.024	0.326
199	-0.024	0.326
206	-0.024	0.326
52	0.020	0.343
260	0.104	0.368
245	0.147	0.364
136	0.155	0.374
61	0.160	0.352
134	0.165	0.349
139	0.165	0.349
141	0.165	0.349
142	0.165	0.349
151	0.165	0.349
153	0.165	0.349
211	0.165	0.349
129	0.270	0.383
205	0.314	0.366
4	0.334	0.351
6	0.334	0.351
24	0.334	0.351
31	0.334	0.351

**Table A.6 (continued).** Estimated ability levels for third quarter kindergarten Language Arts behaviors (n=190).

Case Number	KGN	
	THETAHAT	SE
55	0.334	0.351
69	0.334	0.351
74	0.334	0.351
116	0.334	0.351
147	0.334	0.351
170	0.334	0.351
179	0.334	0.351
186	0.334	0.351
187	0.334	0.351
198	0.334	0.351
215	0.334	0.351
220	0.334	0.351
233	0.334	0.351
282	0.334	0.351
284	0.334	0.351
154	0.406	0.360
157	0.406	0.360
270	0.406	0.360
172	0.462	0.356
35	0.510	0.357
103	0.534	0.342
138	0.534	0.342
46	0.535	0.346
17	0.570	0.341
155	0.570	0.341
163	0.600	0.359
45	0.629	0.352
190	0.661	0.333
274	0.680	0.322
127	0.742	0.322
135	0.742	0.322
48	0.756	0.316
63	0.858	0.291
212	0.923	0.273
214	0.923	0.273

**Table A.6 (continued).** Estimated ability levels for third quarter kindergarten Language Arts behaviors (n=190).

Case Number	KGN	
	THETAHAT	SE
224	0.924	0.287
158	0.938	0.273
131	0.986	0.259
283	0.986	0.259
122	0.993	0.256
20	1.045	0.242
238	1.057	0.239
254	1.057	0.239
174	1.115	0.228
193	1.145	0.218
287	1.157	0.217
33	1.171	0.215
98	1.205	0.209
286	1.259	0.202
16	1.308	0.201
12	1.313	0.201
166	1.330	0.212
273	1.330	0.212
100	1.338	0.201
228	1.372	0.204
101	1.394	0.207
88	1.420	0.211
188	1.454	0.218
80	1.497	0.227
197	1.516	0.232
49	1.533	0.237
78	1.536	0.238
167	1.563	0.249
168	1.563	0.249
196	1.563	0.249
83	1.579	0.252
85	1.579	0.252
91	1.579	0.252
272	1.582	0.256
285	1.596	0.257

**Table A.6 (continued).** Estimated ability levels for third quarter kindergarten Language Arts behaviors (n=190).

Case Number	KGN	
	THETAHAT	SE
75	1.762	0.312
164	1.821	0.340
90	1.832	0.335
230	1.832	0.335
240	1.832	0.335
244	1.832	0.335
92	1.909	0.358
71	2.135	0.437
81	2.135	0.437
82	2.135	0.437
87	2.135	0.437
89	2.135	0.437
93	2.135	0.437
191	2.135	0.437
241	2.135	0.437

**Table A.7.** Estimated ability levels for Early Literacy Scale scores (n=192).

Case Number	ELS	
	THETAHAT	SE
197	-2.341	0.544
80	-2.035	0.433
175	-2.035	0.433
34	-1.985	0.450
43	-1.965	0.410
54	-1.873	0.395
245	-1.857	0.405
208	-1.801	0.370
179	-1.657	0.365
91	-1.607	0.376
132	-1.580	0.366
189	-1.522	0.386
281	-1.452	0.386
30	-1.388	0.435
73	-1.325	0.381
251	-1.314	0.394
2	-1.246	0.391
28	-1.202	0.385
127	-1.202	0.385
3	-1.164	0.405
14	-1.157	0.392
15	-1.157	0.392
21	-1.138	0.393
8	-1.049	0.391
280	-1.049	0.391
210	-1.049	0.391
10	-0.999	0.376
284	-0.999	0.376
173	-0.968	0.374
185	-0.950	0.371
277	-0.950	0.371
84	-0.948	0.389
79	-0.935	0.377
31	-0.934	0.387
42	-0.934	0.387
123	-0.920	0.369
12	-0.920	0.369

**Table A.7 (continued).** Estimated ability levels for Early Literacy Scale scores (n=192).

Case Number	ELS	
	THETAHAT	SE
200	-0.920	0.369
260	-0.912	0.391
99	-0.807	0.383
107	-0.807	0.418
118	-0.782	0.373
196	-0.703	0.373
39	-0.703	0.373
6	-0.703	0.373
103	-0.686	0.374
157	-0.686	0.374
259	-0.636	0.390
106	-0.612	0.402
248	-0.612	0.402
4	-0.605	0.377
1	-0.592	0.388
46	-0.546	0.390
148	-0.537	0.381
186	-0.525	0.403
105	-0.485	0.398
289	-0.481	0.403
140	-0.475	0.415
235	-0.473	0.385
70	-0.434	0.384
27	-0.434	0.406
100	-0.434	0.406
266	-0.396	0.393
168	-0.372	0.387
222	-0.305	0.436
169	-0.295	0.394
207	-0.295	0.394
145	-0.279	0.414
203	-0.279	0.414
88	-0.257	0.430
274	-0.257	0.430
68	-0.194	0.431

**Table A.7 (continued).** Estimated ability levels for Early Literacy Scale scores (n=192).

Case Number	ELS	
	THETAHAT	SE
44	-0.188	0.429
182	-0.178	0.411
233	-0.111	0.425
121	-0.100	0.430
183	-0.100	0.430
163	-0.100	0.430
81	-0.100	0.430
156	-0.100	0.430
171	-0.100	0.430
133	-0.100	0.430
262	-0.100	0.430
223	-0.100	0.430
184	-0.100	0.430
192	-0.094	0.421
136	-0.064	0.470
18	-0.055	0.443
177	0.035	0.429
221	0.035	0.429
23	0.035	0.429
24	0.035	0.429
216	0.035	0.429
131	0.096	0.464
35	0.108	0.506
36	0.108	0.506
293	0.115	0.431
238	0.115	0.431
63	0.124	0.421
102	0.124	0.448
87	0.124	0.421
7	0.124	0.421
120	0.124	0.421
52	0.124	0.421
265	0.124	0.421
193	0.124	0.421
271	0.124	0.421

**Table A.7 (continued).** Estimated ability levels for Early Literacy Scale scores (n=192).

Case Number	ELS	
	THETAHAT	SE
261	0.124	0.421
82	0.179	0.432
144	0.180	0.447
33	0.247	0.403
17	0.247	0.403
20	0.247	0.403
244	0.247	0.403
272	0.247	0.403
90	0.262	0.424
66	0.262	0.424
220	0.262	0.424
278	0.265	0.428
77	0.273	0.420
78	0.273	0.420
201	0.273	0.420
25	0.324	0.398
206	0.324	0.398
270	0.345	0.406
167	0.345	0.406
92	0.376	0.430
108	0.387	0.431
142	0.458	0.390
147	0.458	0.390
137	0.458	0.390
170	0.475	0.398
64	0.475	0.398
198	0.475	0.398
115	0.540	0.385
128	0.540	0.385
85	0.548	0.386
114	0.548	0.386
125	0.572	0.402
32	0.572	0.402
110	0.582	0.403
187	0.645	0.387

**Table A.7 (continued).** Estimated ability levels for Early Literacy Scale scores (n=192).

Case Number	ELS	
	THETAHAT	SE
264	0.645	0.387
75	0.653	0.388
83	0.657	0.387
139	0.715	0.386
211	0.715	0.386
241	0.718	0.416
240	0.718	0.416
69	0.741	0.404
212	0.775	0.440
273	0.821	0.398
74	0.821	0.398
199	0.821	0.398
215	0.821	0.398
209	0.821	0.398
292	0.821	0.398
214	0.821	0.398
159	0.908	0.433
149	0.908	0.433
174	0.910	0.442
97	0.910	0.442
16	0.922	0.424
141	0.922	0.424
188	0.922	0.424
172	0.974	0.441
158	0.974	0.441
254	0.974	0.441
155	1.042	0.457
111	1.042	0.457
49	1.111	0.463
164	1.111	0.463
93	1.111	0.463
230	1.111	0.463
224	1.111	0.463
48	1.187	0.493
205	1.187	0.493

**Table A.7 (continued).** Estimated ability levels for Early Literacy Scale scores (n=192).

Case Number	ELS	
	THETAHAT	SE
19	1.187	0.493
89	1.195	0.515
166	1.267	0.511
129	1.267	0.511
134	1.267	0.511
55	1.432	0.505
122	1.432	0.505
101	1.432	0.505
287	1.610	0.493
143	1.610	0.493
283	1.610	0.493
98	1.918	0.425
285	1.959	0.416
191	2.089	0.406
71	2.089	0.406

**Table A.8.** MAT6 (Reading Comprehension) Rasch ability estimates (n=90).

Case Number	Total Score	Mean Score	Rasch Ability	SE
37	16	0.308	-1.079	0.347
30	18	0.346	-0.846	0.336
169	19	0.365	-0.735	0.332
189	20	0.385	-0.626	0.329
258	20	0.385	-0.626	0.329
77	21	0.404	-0.518	0.326
115	21	0.404	-0.518	0.326
3	22	0.423	-0.413	0.324
42	22	0.423	-0.413	0.324
289	23	0.442	-0.309	0.322
135	24	0.462	-0.205	0.321
223	24	0.462	-0.205	0.321
14	25	0.481	-0.102	0.32
257	25	0.481	-0.102	0.32
123	26	0.5	0	0.32
183	27	0.519	0.102	0.32
148	27	0.519	0.102	0.32
198	27	0.519	0.102	0.32
235	27	0.519	0.102	0.32
173	27	0.519	0.102	0.32
57	28	0.538	0.205	0.321
201	28	0.538	0.205	0.321
179	28	0.538	0.205	0.321
162	29	0.558	0.309	0.322
153	29	0.558	0.309	0.322
177	30	0.577	0.413	0.324
4	30	0.577	0.413	0.324
81	30	0.577	0.413	0.324
171	30	0.577	0.413	0.324
185	31	0.596	0.518	0.326
292	31	0.596	0.518	0.326
58	32	0.615	0.626	0.329
70	32	0.615	0.626	0.329
182	32	0.615	0.626	0.329
245	32	0.615	0.626	0.329
248	32	0.615	0.626	0.329
271	32	0.615	0.626	0.329

**Table A.8 (continued).** MAT6 (Reading Comprehension) Rasch ability estimates (n=90).

Case Number	Total Score	Mean Score	Rasch Ability	SE
88	34	0.654	0.846	0.336
168	34	0.654	0.846	0.336
12	34	0.654	0.846	0.336
75	35	0.673	0.961	0.341
186	35	0.673	0.961	0.341
188	36	0.692	1.079	0.347
244	37	0.712	1.202	0.353
207	37	0.712	1.202	0.353
141	38	0.731	1.329	0.361
99	39	0.75	1.462	0.369
79	40	0.769	1.602	0.38
69	40	0.769	1.602	0.38
172	41	0.788	1.751	0.392
64	41	0.788	1.751	0.392
187	41	0.788	1.751	0.392
209	41	0.788	1.751	0.392
130	41	0.788	1.751	0.392
163	42	0.808	1.91	0.406
36	42	0.808	1.91	0.406
97	42	0.808	1.91	0.406
134	42	0.808	1.91	0.406
273	43	0.827	2.082	0.423
33	44	0.846	2.269	0.443
61	44	0.846	2.269	0.443
206	44	0.846	2.269	0.443
78	45	0.865	2.477	0.469
32	45	0.865	2.477	0.469
212	45	0.865	2.477	0.469
184	45	0.865	2.477	0.469
211	45	0.865	2.477	0.469
60	46	0.885	2.711	0.501
74	46	0.885	2.711	0.501
101	46	0.885	2.711	0.501
238	46	0.885	2.711	0.501
125	47	0.904	2.982	0.543

**Table A.8 (continued).** MAT6 (Reading Comprehension) Rasch ability estimates (n=90).

Case Number	Total Score	Mean Score	Rasch Ability	SE
110	47	0.904	2.982	0.543
166	47	0.904	2.982	0.543
87	47	0.904	2.982	0.543
154	47	0.904	2.982	0.543
147	48	0.923	3.307	0.6
129	48	0.923	3.307	0.6
287	48	0.923	3.307	0.6
56	49	0.942	3.718	0.686
93	49	0.942	3.718	0.686
240	49	0.942	3.718	0.686
230	49	0.942	3.718	0.686
98	49	0.942	3.718	0.686
254	49	0.942	3.718	0.686
89	50	0.962	4.284	0.832
65	50	0.962	4.284	0.832
143	50	0.962	4.284	0.832
191	51	0.981	5.233	1.165
146	51	0.981	5.233	1.165

**Table A.9.** Factor loadings for the 3<sup>rd</sup> quarter kindergarten Language Arts ratings (n=148).

	Factor Loadings
Understands ideas from literature	.89043
Uses beginning reading strategies	.84642
Shares in group reading & writing activities	.83318
Listens with understanding	.82505
Explores language through rhyme, poetry, and movement	.80800
Communicates ideas verbally	.76134
Communicates ideas with drawings and words	.71463
Chooses reading and writing as independent activities	.69338
Percent of Variance	63.9

**Table A.10.** Factor loadings for the first grade ELS ratings (n=148).

	Factor Loadings
Writing Sample	.83838
Retell	.77185
Running Record	.75037
Writing Conference	.53727
Percent of Variance	53.8

## VITA

Stephanie Jacobson is a reading specialist with the Fairfax County Public Schools. Over her career, she has worked as a classroom teacher, grades one through four, a Chapter 1 Math Teacher, and a reading specialist in the Chicago Public Schools and school systems across Virginia. In addition, Mrs. Jacobson has been a research assistant at Virginia Polytechnic Institute and State University and has taught both graduate and undergraduate level classes for Old Dominion University, George Mason University, and the Fairfax County Public Schools' Office of Staff Development and Training. On numerous occasions she has been an invited speaker for classes at Old Dominion University, George Mason University, and Virginia Polytechnic Institute and State University. Mrs. Jacobson has also conducted a number of training workshops for classroom teachers and administrators throughout the state of Virginia.

Mrs. Jacobson's areas of expertise include reading and writing processes, emergent literacy development, alternative assessments, educational research, and program evaluation.

### ***Education***

Ph.D. in Educational Research and Evaluation, 1997, VPI & SU

M.S. in Education (Reading), 1982, Old Dominion University

B.A. in Elementary Education, 1971, University of Illinois Chicago Circle

### ***Publications and Paper Presentations:***

Cline, M.G. & Jacobson, S. (October, 1996). *Standardized tests and alternative assessments: Strange bedfellows*. (paper presented at the annual Northeastern Educational Research Association conference in Ellenville, New York).

Cline, M.G. & Jacobson, S. (October, 1995). *Parent involvement: Head start to public school transition*. (paper presented at the annual Northeastern Educational Research Association conference in Ellenville, New York).

Jacobson, S. (November, 1992). *The home-school connection: Parents lending us support*. The Apple. Fairfax, VA: Fairfax County Public Schools.

Jacobson, S. (February, 1993). *Training program is a real PLUS for school volunteers*. What's Working in Parent Involvement? Fairfax Station, VA: The Parent Institute.

Meeks, J., Jacobson, S., & Duffy, T. (1984). *Information processing in the military: An ethnographical investigation of high and low literacy enlisted personnel*. In J.A. Niles & L.A. Harris (Eds.), Changing Perspectives on Research in Reading/Language Processing and Instruction (Thirty-third Yearbook of the National Reading Conference). Rochester, NY: The National Reading Conference, Inc.

DOB: 12/03/48