Florida VPK Assessment Measures

Technical Manual

October 2011

Prepared for the Office of Early Learning by
Christopher J. Lonigan, Ph.D.
Florida Center for Reading Research
Florida State University

Acknowledgements

The VPK Assessment Measures were developed for the Office of Early Learning, Florida Department of Education by the Preschool Research Group (PRG) of the Florida Center for Reading Research.

PRG Director

Christopher J. Lonigan, Ph.D.

PRG Co-Director

Beth M. Phillips, Ph.D.

VPK Assessment Development Team

Christopher J. Lonigan, Ph.D.
Beth M. Phillips, Ph.D.
Jennifer Levine Phelan
Jenny Kohn
Kylie S. Flynn
Jeanine Clancy-Menchetti, Ph.D.

Office of Early Learning Facilitators for Project

Shan Goff
Denise Bishop
Melinda Webster

Special thanks to the children and teachers throughout Florida who participated in the various phases of the project to develop the VPK Assessment Measures, to Jenny^2 for their coordination of development studies, and to Jenny Phelan for her tireless efforts coordinating all aspects of the project. Without their participation and help, this work would not have been possible.

Content

1. OVERVIEW & DEVELOPMENT

The Florida VPK Assessment Measures were developed by the Preschool Research Group at the Florida Center for Reading Research for the Florida Department of Education's Office of Early Learning specifically for use by teachers in Florida's Voluntary Preschool Program (VPK). The Florida VPK Assessment Measures include three versions of assessments in four domains including oral language, print knowledge, phonological awareness, and mathematics. The VPK Assessment Measures were designed to be individually administered measures that could be used by VPK teachers to screen children's skills and monitor children's progress in the development of these skills across the preschool year. The three versions of the measures correspond to three assessment periods (fall, winter, and spring) that are designated as AP1, AP2, and AP3. Each version of the measures in each domain includes specific normative information that can be used to identify children who have skill levels suggesting that they are or are not on a developmental trajectory to achieve kindergarten readiness status. These measures were designed and validated for screening and monitoring purposes. They are not intended to be used as diagnostic measures, and no child should be diagnosed on the basis of scores on these measures. These measures are tools for teachers to help them plan and monitor instructional activities and to identify children who may need additional instructional activities to achieve kindergarten readiness as a result of their learning and development experiences in VPK.

This technical manual describes the background for the measures, the development process for the measures, the psychometric characteristics of the measures, and the samples of children and teachers who participated in the development work. Questions about use and administration are described in the assessment materials available from the Office of Early Learning of the Florida Department of Education.

Research Background

Early Literacy Skills

A significant amount of research evidence indicates that skills children develop before school entry (i.e., kindergarten) serve as the foundation for later success in developing reading skills. The collective findings across a growing body of empirical evidence indicate that oral language, phonological processing skills, and print knowledge represent the foundational skills on which later conventional reading and writing skills are built (e.g., Lonigan, 2006; Whitehurst

& Lonigan, 1998). Oral language skills refer to the words in a child's vocabulary as well as his or her ability to use those words to understand and convey meaning (i.e., syntactic and narrative skills). Phonological processing skills refer to children's developing sensitivity to the sound structure of his or her language (e.g., that words are made up of smaller sounds like syllables or phonemes) and the ability to use that information in cognitive processes like memory. Print knowledge refers to a developing understanding about the nature and purpose of books and print (e.g., letters, the sounds letters represent, directionality of print).

Confirmation of the importance of these skills to later reading and writing was provided by the Report of the National Early Literacy Panel (NELP; see Lonigan, Schatschneider, & Westberg, 2008a). The NELP conducted a meta-analysis (a quantitative summary of existing research) of approximately 300 published studies that included data about the predictive relation between a skill measured in preschool or kindergarten and reading outcomes (i.e., word decoding, reading comprehension, spelling) for children learning to read in an alphabetic language like English. The results of this meta-analysis indicated that children's skills related to print knowledge (e.g., alphabet knowledge, print concepts), phonological processing skills (i.e., phonological awareness, phonological access to lexical store, phonological memory), and aspects of oral language (e.g., vocabulary, syntax/grammar, word knowledge) were substantive and independent predictors of children's later reading outcomes—including both word-decoding and reading comprehension outcomes.

In general, results from longitudinal studies indicate that there is a moderate degree of modularity between these early literacy skills and later conventional literacy skills. Some early literacy skills are code-related and other emergent literacy skills are meaning-related (e.g., Sénéchal & LeFevre, 2002; Storch & Whitehurst, 2002). Code-related skills are those skills that facilitate children's abilities to acquire the alphabetic principle successfully and become accurate and fluent decoders of text. Meaning-related skills are those skills, primarily associated with language, that allow children to comprehend text once it is decoded. Whereas skills in these two domains are correlated during development, they are differentially predictive of different aspects of later conventional literacy skills, and they appear to be responsive to different types of instructional activities (Lonigan, Schatschneider, & Westberg, 2008b; Lonigan, Shanahan, & Cunningham, 2008). Consequently, understanding children's patterns of strengths and

weaknesses in early literacy skills can facilitate effective instructional activities (e.g., providing focused instruction in areas identified as a specific weakness).

Early Math Skills

The current knowledge base of early mathematics skills is rooted in the concepts of formal and informal mathematics skills (Greenes, Ginsburg, & Balfanz, 2004; Starkey, Klein, & Wakeley, 2004). Formal mathematics skills are those skills taught in school that require the use of abstract numerical notation such as writing numerals, place-value tasks, knowledge of the base-ten mathematics system, and decimal knowledge (Baroody, Gannon, Berent, and Ginsburg, 1984). Informal mathematics skills are the developmental precursors to formal mathematics skills and do not require specific instruction in abstract mathematical notation. These skills develop as children explore their natural environment (Ginsburg, 1975). Number knowledge and arithmetic operations are the two most studied aspects of informal mathematics skills. Within the domain of number knowledge, two skills, counting and numerical relations appear to be those most necessary for the development of basic formal skills such as addition and subtraction (Jordan, Kaplan, Ramineni, & Locuniak, 2009). Other skills (measurement, shapes, patterns/logical inferences, and spatial concepts) often identified as informal geometric skills are believed to be important for the development of formal geometry skills.

Counting skills include counting objects, subitizing, estimation, counting forward from a number other than one, counting backward, and recognizing counting errors. Numerical relations or “number sense” is typically indicated by skills such as magnitude discrimination, relative size of numbers and quantities, set and number comparison, knowledge of number order, number sequencing, and number reproduction. A child with number sense is able to utilize the concept of a number line to develop basic mathematical understanding and thus develop later skills such as addition and subtraction. For example, recognition that the number ‘four’ is greater than the number ‘two’ is dependent on the knowledge that numbers always follow a specific sequence and the knowledge that numbers that follow other numbers are always larger.

Florida’s Voluntary Pre-Kindergarten Learning Standards

In 2005, the Florida Department of Education (DOE) collaborated with faculty from the Florida Center for Reading Research (FCRR) and an invited group of national experts in early

literacy development to create the *Emergent Literacy* and *Language and Communication* sections of the Florida VPK Education Standards. The best available developmental research data were utilized to support the specific standards. The Emergent Literacy standards included focus on motivation for reading, motivation for writing, phonological awareness, alphabetic knowledge, and print concepts. The *Language and Communication* standards target listening, speaking, vocabulary, sentence structure, and conversation. One innovative aspect of these standards was the inclusion of benchmarks within each standard area that represented stages of mastery within the target area. For example, a benchmark within the phonological awareness standard of "Shows age-appropriate phonological awareness" is that a child can delete a syllable from word.

The math standards to which the VPK Math subtest assessments were aligned are the standards developed in 2008 as a collaborative effort between the Florida Department of Education, FCRR, and an invited team of nationally known experts in early childhood mathematics development. These standards address the six areas of number sense, numerical operations, geometry, patterns and series, spatial relations, and measurement. Within each of these six areas, standards and benchmarks were developed to indicate what children should know and be able to do at the end of their VPK experience. Given the limited time available to teachers for this assessment, the decision was made to focus only on number sense and number and operations within the VPK assessment, as these are the two areas of the standards with the strongest developmental evidence.

Description of the VPK Assessment Measures

The Florida VPK Assessments consist of three versions of measures designed to assess four distinct skills, including oral language, phonological awareness, print knowledge, and mathematics skills. Each of these measures was developed to be able to be used as both a screening measure and a progress monitoring measure. That is, each of the three measures in a skill domain was designed to assess the range of abilities within that domain that is appropriate to the levels of a skill likely to be exhibited by children during their 4-year-old preschool year. The goal of the VPK Assessment Measures is to provide VPK teachers and other professionals with a reliable and valid means of identifying children who are not on a trajectory of success to be "kindergarten ready" in terms of their developing reading-related and math skills during the

children's VPK experience and to monitor children's development of the reading-related and math skills that are strongly related to later academic success across children's participation in VPK programs. It is expected that once children are identified as at risk for meeting kindergarten-readiness standards, teachers will be able to provide enriched experiences and focused instructional activities to help children acquire the skills that will put them on the path to kindergarten readiness, and the VPK Assessment Measures will allow teachers to measure the extent to which these enriched experiences and instructional activities are achieving their goals.

Development of the VPK Assessment Measures

The Florida VPK Assessments were created through an iterative process of item development, testing, and refinement. Initial item development involved creating sets of items that mapped onto the domains of early literacy and early math skills that are included in Florida's VPK Standards. Specifically, items that addressed a range of abilities in oral language, phonological awareness, print knowledge, and mathematics skills were constructed because they mapped onto the specific sets of skills and levels of skills detailed in the Florida VPK standards. Within a skill domain, items using different formats were constructed to identify different ways of assessing children's skills and to provide a range of formats (i.e., question stems, response formats) that would be effective with individuals with limited formal training in assessment administration (e.g., VPK teachers).

Initial Item Development

For the oral language domain, 120 items were generated to assess four different facets/response formats of children's vocabulary and grammatical knowledge. Each of these four item sets initially included 30 items. One item set included items designed to assess children's abilities to understand simple definitions of words. One item set included items designed to assess children's abilities to understand vocabulary that described relations between things (e.g., above, next to, inside). One item set included items designed to assess children's abilities to respond correctly to simple questions indicating an understanding of the meaning of words (e.g., "This bear is lost. Does he know how to get home?"). One item set included items designed to assess children's abilities to respond correctly to different grammatical forms (e.g., regular past tense) in a receptive format.

For the phonological awareness domain, 127 items in 17 variations of linguistic complexity (e.g., word, syllable, subsyllable, phoneme), type of operation (i.e., identification, blending, elision), and response format (i.e., multiple-choice items requiring children to point to a picture of the correct answer, free-response items requiring children to say the correct answer) were developed. Both blending (i.e., putting sounds together to form a new word) and elision (i.e., removing parts of a word to form a new word) were created in multiple-choice and free-response formats to assess children's phonological awareness skills at the word, syllable, onset-rime, and phoneme levels (4 items of each type). Twelve initial sound matching items were created to assess children's ability to identify phonemes in isolation.

For the print knowledge domain, 80 items were developed to assess children's knowledge about print concepts, letter-name knowledge, and letter-sound knowledge. All 16 items developed to assess children's print concepts were multiple-choice items that required children to select a picture that represented the print concept spoken by the examiner. The 32 letter-name knowledge and 32 letter-sound knowledge items were equally divided between items requiring children to say the name or sound of a letter and items requiring children to point to the letter (from a set of four) representing the letter name or letter sound spoken by the examiner.

For the mathematics domain, items were developed to assess children's early mathematics skills in the areas of counting, numerical relations, and arithmetic reasoning. For counting, items included rote counting, counting forward, counting backward, counting error identification, resultative counting, cardinality, estimation, and subitizing. For numerical relations, items included ordinality, relative size, number comparison, sequencing, set reproduction, number identification, and numerals. For arithmetic reasoning, items included addition, subtraction, equivalence, number composition, and number decomposition. For each item type in which it was appropriate, multiple methods of representing quantities (e.g., objects, dots, or pictures) were used.

Initial Item Selection

Following development of items, sets of items were administered to groups of approximately 100 children (e.g., groups of children got subsets of items within the oral language, phonological awareness, print knowledge, and mathematics domains; see sample description below). Based on this initial administration of items, those items that were difficult to

administer or appeared to be confusing to children based on reports of the examiners were eliminated. Additionally, items that were answered correctly by too few children or answered correctly by too many children were eliminated as too difficult or too easy, respectively. Finally, these reduced sets of items were subjected to analyses to determine those items within a set that yielded a relatively high degree of cohesion as indexed by their item-total correlations within the set.

Once the item sets were reduced by eliminating items that were difficult to administer, proved too difficult or too easy for children, or had lower than average cohesion with the other items within a set, the sets were combined into their respective domains (e.g., phonological awareness, print knowledge) and administered to a new sample of approximately 300 children (see sample description below). The data from this round of administration was subjected to Item Response Theory (IRT) analyses to identify the degree to which each item demarcated a specific level of difficulty (a parameters) and the degree of difficulty indexed by specific items (i.e., b parameters). Differential Item Functioning (DIF) analyses were also conducted to identify and eliminate items that did not have similar parameters across major demographic groups (e.g., items with statistically different a or b parameters in different groups for the same underlying ability). Based on these IRT results, items from a domain were grouped into sets of items with similar difficulty levels (i.e., items with similar b parameters were grouped). Items at low, middle, and high levels of difficulty within domain based on their b parameters and that had good a parameters were selected as potential linking items. To create three versions of the assessments for each of the four domains, the remaining items from the same difficulty grouping were randomly distributed into three groups with the restriction that items distributed into the three groups had similar a parameters. These items were combined with the linking items to form the three versions of assessments for oral language, phonological awareness, print knowledge, and math domains.

Final Item Selection

Once the three versions of each measure were created, these versions were administered to a new group of approximately 220 children (see sample description below) to assess the degree to which the cohesion and difficulty of the items sets replicated. Combined data from this round of data collection and the prior round of data collection were used in additional IRT

analyses to determine the final length of each assessment (i.e., required number of items that the minimized standard error of the assessment across a range of ability). The final versions of the assessments included 22-23 items for the three measures of Oral Language Domain, 14 items for the three measures of the Phonological Awareness Domain, 12 items for the three measures of the Print Knowledge Domain, and 13 distinct items (yielding 18 scored responses) for the three measures of the Mathematics Domain. Separate measures from within each domain were designated as the measure for First Pre-K Assessment Period (AP1), the Second Pre-K Assessment Period (AP2), and the Third Pre-K Assessment Period (AP3).

Psychometric Studies of the VPK Assessment Measures

Once the final versions of the three measures of each domain were created, two additional studies were conducted to examine further the psychometric characteristics of the VPK Assessment Measures. Results of these studies of the VPK Assessment Measures are described in Chapter 2 ("Reliability") and Chapter 3 ("Validity") of this manual. In the first study, a group of approximately 300 children completed all three versions of the VPK Assessment Measures as well as two standardized assessment measures. One standardized measure was a nationally normed and validated measure of children's early literacy skills, the *Test of Preschool Early Literacy* (Lonigan, Wagner, Torgesen, & Rashotte, 2007), and the other standardized measure was a nationally normed and validated measure of children's early mathematics skills, the *Test of Early Mathematics Ability* (Ginsburg & Baroody, 2003). Data from this study was used both to establish the concurrently validity of the VPK Assessment Measures (see results in Validity Chapter) and to examine the alternative-forms reliability of the VPK Assessment Measures (see results in Reliability Chapter).

In the second study, a representative sample of VPK teachers in Florida administered the VPK Assessment Measures to over 1,000 students in their VPK classrooms following the expected timing of assessment for a VPK program following a traditional school-year schedule (i.e., September to May). Teachers in this field trial administered the AP1 versions of all of the VPK Assessments except for Oral Language (by design, the Oral Language measures were not completed prior to the start of the field trial, which was pushed forward in time by the Florida Office of Early Learning) during the fall, the AP2 versions of the VPK Assessments in late winter, and the AP3 versions of the VPK Assessments in spring to a selected sample of children

in their classroom. To the extend possible, data from these children's Kindergarten Readiness assessments collected in the fall of their kindergarten year was obtained from the Florida Department of Education (the parents of all children who were part of the field study provided informed consent for identified preschool assessments and access to the Kindergarten Readiness screening scores). Additionally, a subset of children in the field trial also were administered the appropriate VPK Assessment measures either two to three weeks before or after they completed the same measures administered by their classroom teachers. Data from this study was used both to establish the predictive validity/accuracy of the VPK Assessment measures (see results in Validity Chapter) and to examine both the test-retest and examiner-based reliability of the VPK Assessment Measures (see results in Reliability Chapter).

Finally, data across the different development and psychometric studies samples were combined to conduct an IRT analysis on each of the three versions of the VPK Assessment Measures within domain to examine the properties of each test and test items using the broadest and most representative sample of children available (see below/Appendix 1 for item parameters and Chapter 2/Appendix 2 for overall functioning for each measure).

Descriptions of Samples Used in Studies of VPK Assessment Measures

As noted above, several different samples of 4-year-old children attending preschool programs participated in the various phases of development of the VPK Assessment Measures. In this section, the characteristics of these samples are described.

Development Studies Samples

The initial development work on the VPK Assessment Measures involved administering groups of items to children to determine basic item functioning. Following this initial try out of items, data were collected from two groups of children to conduct IRT analyses to select the final item set, to identify and eliminate items that functioned differently between demographic groups, and to test the reliability of the item selection. Descriptive statistics for these two groups of children is shown in Table 1.1 (next page). All of the children in this development work were from the North Florida region. Data collection for these studies took place in the fall and winter of children's preschool year. Although the sample was representative of the area in terms of

race/ethnicity, white children were slightly over-represented and Latino/Hispanic children were under-represented compared the State of Florida as a whole.

Table 1.1. ***Descriptive Statistics for Development Studies***

Variable	Mean or *N*	(*SD*) or %
N	512	
Chronological Age (months)	53.14	(6.05)
Sex		
Male	266	52
Female	246	48
Race/Ethnicity		
White	332	65
African American/Black	135	26
Latino/Hispanic	14	3
Asian	17	3
Multiracial	14	3
Other	0	0

Concurrent Validity Study Sample

All children who participated in the concurrent validity study were from the north Florida region. Data collection for the study took place during the spring of the children's preschool year. There were 302 children who were recruited for the study and who completed the VPK Assessment Measures. Of this 302, 288 also completed the two standardized measures. Data from all children were used in IRT analyses of the VPK Assessment Measures, and only the data from the 288 children who completed the standardized measures were used in the concurrent validity study. Descriptive information on the sample for the concurrent validity study is shown in Table 1.2 (next page). As can be seen in the table, there were not differences in demographics between these two groups. Although the sample was representative of the area in terms of race/ethnicity, white children were slightly over-represented compared the State of Florida. As a group, children in the validity sample had average scores for their ages as indicated by standard scores of approximately 100 on the nationally normed standardized assessments of early literacy and early math skills.

Table 1.2. ***Descriptive Statistics for Children in Concurrent Validity Study***

	Full Sample		Sample with All Measures	
Variable	Mean or *N*	(*SD*) or %	Mean or *N*	(*SD*) or %
N	302		288	
Chronological Age (months)	59.47	(5.27)	59.38	(5.35)
Sex				
Male	154	51	151	52
Female	148	49	137	48
Race/Ethnicity				
White	224	74	213	74
African American/Black	45	15	43	15
Latino/Hispanic	14	5	14	5
Asian	12	4	11	4
Multiracial	6	2	6	2
Other	1	<1	1	<1
TOPEL Standard Scores				
Definitional Vocabulary	---	---	102.02	(11.27)
Phonological Awareness	---	---	100.50	(15.37)
Print Knowledge	---	---	106.51	(12.28)
TEMA Standard Score	---	---	98.18	(14.66)

Note. TOPEL = Test of Preschool Early Literacy; TEMA = Test of Early Mathematics Ability.

Field Test/Predictive Validity Study Sample

Children for the field test of the measure were those children who had parental consent to participate in the study and were selected by their VPK teachers as "representative" of children in the participating teachers classrooms. An initial group of VPK providers was selected from the Florida Office of Early Learning list of VPK providers to be representative of VPK providers in the state. This group was representative of public and private VPK providers and of the density of VPK providers in various regions throughout the state. However, only VPK providers with "school year" programs were included. Not all of the providers identified for participation agreed to participate, and providers were selected to replace those who declined participation to

Table 1.3. ***Descriptive Statistics for Children in Field Trial/Predictive Validity Study***

	Full Sample		Retest Sample	
Variable	Mean or *N*	(*SD*) or %	Mean or *N*	(*SD*) or %
N	1,229		146	
Chronological Age (months) at AP1	55.54	(3.61)	55.88	(3.76)
Sex				
Male	581	47	67	52
Female	648	53	79	48
Race/Ethnicity				
White	595	48	114	78
African American/Black	312	25	18	12
Latino/Hispanic	209	17	7	5
Asian	25	2	0	0
Multiracial	76	6	7	5
Other/Not Reported	12	1	0	0
Region of Florida				
Northwest	188	15	146	100
Northeast	310	25	---	---
West Coast	367	30	---	---
Southeast	216	18	---	---
South	148	12	---	---

maintain the a priori selection criteria. Each participating teacher was asked to provide assessment data on four children in her or his classroom. Although some teachers provided assessment data on the four children requested, some teachers assessed additional children. All data were used in subsequent analyses. Consequently, although the teachers roughly matched the distribution of VPK providers in the state, the children may over-represent one region or another due to the number of children for which each teacher provided assessment data. A subset of children from the field trial (~12%) was administered the VPK Assessment Measures by research staff. Because research staff was located in north Florida, all of these children came from VPK providers in this region. A summary of demographic information on the children included in the field trial and the retest sample is shown in Table 1.3. As can be seen in the table,

the full sample was approximately representative of the population of the state and included representation from all of the major regions of the state. The retest sample was representative of the VPK providers where they attended preschool but were over-representative of white children.

Item Content and Item Functioning for VPK Assessment Measures

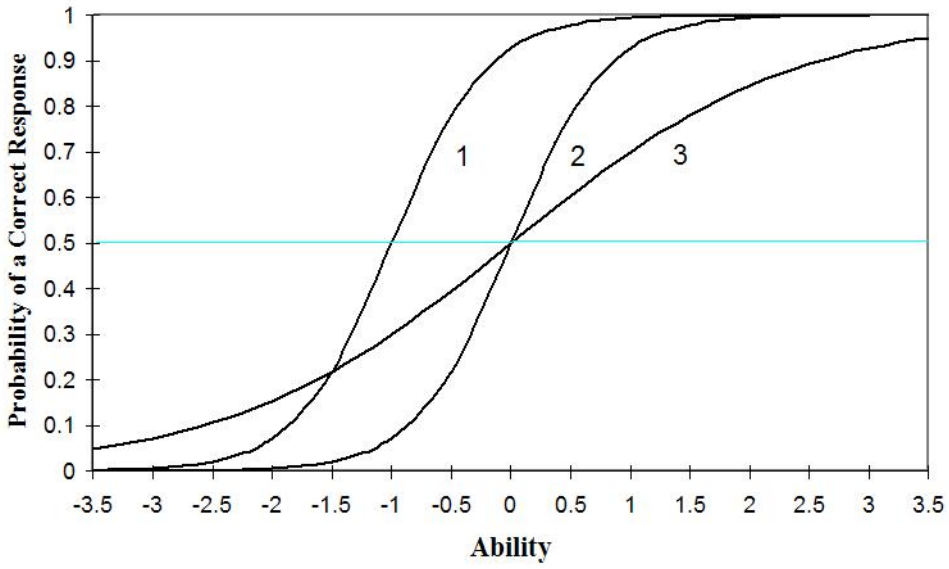
Tables A1.1 to A1.12 in Appendix 1 provide information on the general item content and item-level psychometric information on the final item set selected for the VPK Assessment Measures. All psychometrics reported in these tables were computed on the largest sample available and included children who had been a part of the later-stage-development studies, the concurrent validity study, and the field trial of the measures.

Item-Response Theory Analysis

Item response theory (IRT; Hambleton, Swaminathan, & Rogers, 1991) is a theory of measurement that assumes that the performance on a given item can be explained by the latent ability or abilities of a given examinee. An advantage of IRT is that it allows estimates of item discrimination, difficulty, and guessing parameters, which provide useful information regarding test difficulty and utility across a given range of ability. Within an IRT framework, the relationship between the probability of answering an item correctly and the latent ability of a group of examinees can be described by an ogive shaped function called an item characteristic curve (ICC). Three ICCs for three hypothetical items are shown in Figure 1 (next page). In the figure, the x-axis describes ability levels that have a mean of zero and a standard deviation of one, and the y-axis is the probability of obtaining a correct response on a given item. The monotonically increasing ICCs demonstrate that examinees with higher latent abilities have a higher probability of correctly answering a given item correctly.

The IRT models described by these ICCs contain parameters for both item characteristics and examinee ability estimates. One parameter estimated for each ICC is the difficulty parameter, which is the point on the latent ability scale where the probability of obtaining a correct response on an item is 50%. Based on the information shown in Figure 1, Items 2 and 3 have difficulty parameters of zero (i.e., indexing the average level of ability), whereas the difficulty parameter for Item 1 is -1. These parameters imply that if a group of examinees, all with average ability, were administered Items 2 and 3, 50% of the group would be expected to

Figure 1. Hypothetical Item Characteristic Curves



pass each item, but over 90% would be expected to pass Item 1. Similarly, only 50% of a group of examinees with ability levels one standard deviation below average would be expected to pass Item 1.

The IRT models that characterize the ICCs of Figure 1 also include a discrimination parameter, which is the slope of the ICC curve at that point on the ability distribution where ability equals the difficulty parameter. A steeper slope (i.e., a larger discrimination parameter) implies that a particular item is better at discriminating between levels of ability than an item with a shallower slope. Based on the ICCs shown in Figure 1, Items 2 and 3 have the same difficulty parameter, but Item 2 has a discrimination parameter equal to 1.5 and Item 3 has a discrimination parameter equal to .50. These parameters imply that Item 3 is less able than Item 2 to discriminate between different ability levels.

Item Content of VPK Assessments

Tables A1.1 to A1.12 in Appendix 1 include descriptions of the general item content and response format of the items included in each of the VPK Assessment Measures. It is important to note that items for each of the measures were developed and selected to provide a sample of the specific skill domain and where not developed and selected to represent specific standards or benchmarks in the VPK Learning Standards. Therefore, some items may be easier than the

specific benchmarks detailed in the Learning Standards and some items may be more difficult than the specific benchmarks detailed in the Learning Standards. Whereas the measures in each domain map onto the Learning Standards, they are not intended to represent the Learning Standards as would a criterion-based measure (e.g., a checklist of accomplishments). Measures developed based on psychometric criteria are likely to have stronger reliability and validity characteristics because they allow representation of children along a broader continuum of abilities, including children who have developed a higher level of skills than described by the benchmarks listed in the VPK Learning Standards. Consequently, items in the VPK Assessment Measures should not be used to select targets for instruction. Scores on each measure should be used to identify domain areas (e.g., oral language, print knowledge) that represent strengths and weaknesses for a particular child. These domain scores can be used to determine whether or not a child is on a developmental trajectory to achieve a designation of "ready" on the Kindergarten Readiness Screening administered in the fall of children's kindergarten year.

Item Functioning in VPK Assessments

Tables A1.1 to A1.12 in Appendix 1 include the discrimination and difficulty parameters from IRT analyses as well as the p-level (i.e., percentage of children getting item correct) and the item-total correlation (i.e., strength of relation of item with total score on the measure) for each item. During development of the measures, the goal was to select items that covered a broad range of difficulty with adequate discrimination. In the tables, discrimination parameters with lower values--including negative values--represent easier items and discrimination parameters with higher values represent harder items. In most cases, multiple-choice format items were developed to be easier tasks than items requiring a response without support, and the parameter values in Tables A1.1 to A1.12 tend to support this expectation.

Given the administration requirements of some tasks (i.e., phonological awareness, print knowledge), items with the same response format were grouped together. For other tasks, items with similar content were grouped together. Final item ordering of tasks was based on IRT analyses conducted during the development phases. For the Print Knowledge, Phonological Awareness, and Math measures, this item grouping resulted in measure in which items generally proceed from easiest to most difficult. Consistent with the developmental expectations embedded in the VPK Learning Standards, both print knowledge and phonological awareness followed a

developmental sequence. That is, items assessing letter-name knowledge were easier (came earlier in development) than items assessing letter-sound knowledge. For phonological awareness, items requiring manipulation of larger units of sound (i.e., words, syllables) were easier than items requiring manipulation of smaller units of sound (i.e., onset-rime), and items requiring synthesis (i.e., blending) were easier than items requiring analysis (i.e., elision). Items on the oral language measures were grouped by domain type to avoid confusion in administration and responses; consequently, easier and harder items are intermingled in the administration order.

Although discrimination parameters of 1.0 or higher are preferred, some items in the measures have lower discrimination parameters. During the development of the VPK Assessment Measures, only items with discrimination parameters greater than .60 were retained. In the larger sample of children used in the final IRT analyses of the measures, however, some discrimination parameters were below this level. With lower discrimination parameters, more items are needed to provide a reliable estimate of ability around the difficulty levels of those items. In most case, items with lower discrimination parameters are those using a multiple-choice response format. These items were selected to provide an estimate of children's abilities around lower levels of skills (i.e., these items provide support for children's responses by reducing memory demands and allowing recognition of correct responses). Because some children get these items correct by chance (e.g., a child who does not have the knowledge to answer a question correctly has a one-in-four chance of picking the correct answer by pointing at one of the choices), the degree of precision in the items is less than it is in items in which chance responding plays no role in scoring. Despite their lower discrimination parameters, these items when combined provide reliable estimates of children's abilities--particularly for children who are on the lower end of the skill distribution being measured (see Chapter 2).

2. RELIABILITY

As described previously, the VPK Assessment Measures were developed using an iterative process of item development, field testing of items, and item selection that culminated in a field trial of the measure by VPK Classroom teachers. Final item selection was conducted using an Item-Response Theory (IRT) approach, with item selection governed by both the goal of retaining items that spanned a wide range of difficulty levels and that maximized the reliability of the specific measure. In contrast to Classical Test Theory (CTT), which treats error variance (lack of reliability) as the same for all scores, IRT provides an estimate of error across the range of estimated abilities. Rather than yielding a single number reflecting the reliability of the measure as in CTT, IRT provides an estimate of reliability that is dependant on the level of ability measured (the underlying ability distribution of a measure is referred to as "Theta"). The graphs in Appendix 2 provide the estimates of reliability for each VPK Assessment Measure across the range of ability in terms of standard errors. A standard error of approximately .38 is equivalent to a reliability of .85 in CTT terms. A standard error of approximately .32 is equivalent to a reliability of .90 in CTT terms, and a standard error of approximately .44 is equivalent to a reliability of .80 in CTT terms (Nunnally & Bernstein, 1994). Reliability estimates of .70 or above are considered adequate generally; however, reliability estimates of .80 or higher are typically desired for non-clinical decision making.

IRT Estimates of Measurement Precision

Figures A2.1 to A2.4 in Appendix 2 include the standard error plots for each of the four VPK Assessment Measures for each of the three versions of the measures (i.e., AP1, AP2, AP3). Standard errors are plotted across levels of Theta, a standardized metric of the underlying ability being assessed by a measure (e.g., a Theta value of 0 represents an average level of performance on the measure by all children included in the IRT analysis).

Precision of measurement is influenced both by the characteristics of items and by the number of items included in a specific measure (i.e., in general, more items lead to increased precision of measurement). Therefore, measurement construction involves a trade-off between precision and number of items included--and the corresponding increase in time to administer the measure. Because the intent of the VPK Assessment Measures is, in part, to identify children who are at risk of failing to meet criteria for Kindergarten Readiness, items were selected to

increase precision around scores representing the most likely region of risk. That is, the measures were constructed to be more precise in their abilities to identify children who were likely to fail to meet the Kindergarten Readiness criteria. Consequently, it was expected that standard errors would be lower (more precision) for Theta values from average to below average than for Theta values in the above average range.

Print Knowledge Measures

Standard errors across levels of Theta for the AP1, AP2, and AP3 versions of the Print Knowledge measures are shown in Figure A2.1. For AP1, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -1.50 to 0.75 (range of 2.25 Theta). For AP2, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -1.80 to 0.40 (range of 2.20 Theta), and for AP3, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -2.25 to 0.25 (range of 2.50 Theta). Overall, these results demonstrate that the three versions of the Print Knowledge measure provide adequate measurement precision across a wide range of abilities, with the highest precision obtained for children with print knowledge ability levels below average.

Phonological Awareness Measures

Standard errors across levels of Theta for the AP1, AP2, and AP3 versions of the Phonological Awareness measures are shown in Figure A2.2. For AP1, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -0.25 to 1.25 (range of 1.50 Theta). For AP2, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -0.50 to 0.75 (range of 1.25 Theta), and for AP3, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -1.50 to 0.75 (range of 2.25 Theta). Overall, these results demonstrate that the three versions of the Phonological Awareness measure provide adequate measurement precision across a range of abilities, with the highest precision initially for children with ability levels slightly above average and, by AP3, for children with phonological awareness ability levels below average. A likely explanation for this shift is that the average performance of children at preschool entry on phonological awareness tasks is low; therefore, the absolute score represented by a Theta of 0 increases from AP1 to AP3.

Oral Language Measures

Standard errors across levels of Theta for the AP1, AP2, and AP3 versions of the Oral Language measures are shown in Figure A2.3. For AP1, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -2.80 to -0.25 (range of 2.55 Theta). For AP2, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -3.00 to -0.50 (range of 2.50 Theta), and for AP3, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -3.00 to -0.25 (range of 2.25 Theta). Overall, these results demonstrate that the three versions of the Oral Language measure provide adequate measurement precision across a wide range of abilities, with the highest precision obtained for children with oral language ability levels below average.

Math Measures

Standard errors across levels of Theta for the AP1, AP2, and AP3 versions of the Math measures are shown in Figure A2.4. For AP1, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -1.50 to 1.00 (range of 2.50 Theta). For AP2, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -2.50 to 0.60 (range of 3.10 Theta), and for AP3, the figure indicates that measurement precision greater than or equal to a reliability estimate of .80 was obtained for Theta values of -2.75 to 0.30 (range of 3.05 Theta). Overall, these results demonstrate that the three versions of the Math measure provide adequate measurement precision across a wide range of abilities, with the highest precision obtained for children with math ability levels below average.

Internal Consistency Reliabilities

Although the standard errors obtained from IRT analyses provide estimates of precision for particular ability levels on a measure, it may also be useful to know the values of reliability estimates obtained from CTT analyses. Table 2.1 (next page) provides the CTT estimates of internal consistencies (i.e., Coefficient Alpha) for each version of the four VPK Assessment measures for children in each phase of the development and validation of the VPK Assessment Measures and for all children across phases. As can be seen in the table, each version of each measure both for each sample used in development work and for all samples combined achieved

Table 2.1. ***Internal Consistency Reliabilities for Each Version of VPK Assessment Measures for Samples Used in Different Phases of VPK Measure Development***

Measure	Phase of VPK Measure Development				
	Development	Validity	Teacher Field Trial	Field Trial Retest	All Samples Combined
Approximate *N* in Sample[1]	512	302	1,075 to 1,227	124 to 146	2,013 to 2,189
Print Knowledge AP1	.87	.86	.84	.80	.85
Print Knowledge AP2	.89	.89	.86	.81	.89
Print Knowledge AP3	.89	.89	.86	.84	.90
Phonological Awareness AP1	.82	.84	.82	.82	.83
Phonological Awareness AP2	.75	.84	.84	.81	.84
Phonological Awareness AP3	.82	.84	.85	.82	.88
Oral Language AP1	.85	.77	---[2]	---[2]	.82
Oral Language AP2	.85	.79	.75	.73	.79
Oral Language AP3	.86	.81	.75	.70	.81
Math AP1	.91	.84	.84	.83	.85
Math AP2	.93	.88	.85	.83	.86
Math AP3	.94	.87	.83	.79	.86

Notes. [1]Number of children in sample varies by time of assessment (AP1 typically higher than AP3). [2]Oral Language measure was not administered during AP1 assessment of Field Trial.

adequate to high levels of internal consistency. Generally, there were no substantial differences in the internal consistency of the measures when administered by development staff (i.e., Development, Validity, and Retest phases) or when administered by classroom teachers. Consequently, the nature of the measures was not substantially influenced dependant on the "assessment expertise" of the person administering the measures.

Alternate-Forms Reliability

The VPK Assessments were created to provide assessments of children's print knowledge, phonological awareness, oral language, and math skills across the preschool year. Three versions of the measures were created to allow the VPK Assessments to be used both to screen children who may be at risk of failing to meet the Kindergarten Readiness Criteria and to allow teachers to monitor children's progress in each of these four areas. Consequently, the goal was to create three parallel versions of each measure. To evaluate the degree to which each version (i.e., AP1, AP2, and AP3) of the measures evaluated the same underlying set of abilities, scores from each version of the four VPK Assessment Measures for children who participated in the concurrent validity study were compared. As a part of the concurrent validity study, children completed all three versions of the VPK measures within a relatively narrow window of time; therefore, their scores allow an estimate of how well the measures function as alternative forms of the same assessment.

Correlations between the different versions of the VPK Assessment Measures for the 288 children from the concurrent validity study who completed all three versions of all four measures are shown in Table 2.2 (next page). As can be seen in the table, all correlations between versions of the measures were moderate to high. There was a very high degree of consistency for the different versions of the print knowledge measure and the different versions of the math measure (i.e., between 74 and 85 percent common variance between forms). The oral language measure had moderately high consistency between forms (i.e., between 55 and 59 percent common variance between forms). For phonological awareness, consistency ranged from moderate to high, with the AP2 version of the measure sharing less common variance with the AP1 (38 percent common variance) and the AP3 (48 percent common variance) versions than the AP1 and AP3 versions did (64 percent common variance). In general, these cross-version correlations were similar to the levels of internal consistency estimated for each of the versions of each measure. Consequently, these data support the expectation that the AP1, AP2, and AP3 versions of the measures would be parallel forms, assessing the same underlying abilities.

Table 2.2. ***Correlations Between Scores on Different AP Versions of VPK Assessment Measures for Children in Concurrent Validity Study***

	AP1 with AP2	AP1 with AP3	AP2 with AP3
Print Knowledge	.88	.86	.91
Phonological Awareness	.62	.80	.69
Oral Language	.76	.77	.74
Math	.88	.88	.92

Notes. N = 288; sample only includes children who completed all four VPK Assessment Measures.

Test-Retest Reliability

A final type of reliability of measurement reflects consistency of measurement across time (i.e., test-retest reliability). Because the skills measured by the VPK Assessment Measures are individual difference variables that are assumed to represent accumulated knowledge and skills of young children, it is expected that there will be some degree of cross-time consistency of children's scores. That is, children who have high levels of skills at the beginning or middle of the preschool year are expect to continue to have high levels of skills through the end of the preschool year. Such stability would be reflected in moderate to high correlations between scores on the measures of the different domains assessed by the VPK Assessment Measures at different assessment times. Of course, it is expected that children in VPK programs will be gaining skills in each of the domains assessed by the VPK Assessment Measures; therefore, it is expected that the longer the interval between assessments, the lower the correlations between children's scores from one assessment period to the other will be. For example, children's scores from the AP1 assessment should be more highly correlated with their scores from the AP2 assessment than they are with scores from the AP3 assessment.

Two sets of data were used to estimate the test-retest reliabilities of the VPK Assessment Measures. The first set of data included the 957 children whose teachers assessed them at AP1, AP2, and AP3 as a part of the field trial of the VPK Assessment Measures. Consequently, the test-retest period for the assessments in this set of data was relatively long (i.e., three to seven

months). The second set of data included approximately 150 children who were assessed by both teachers and the measure development staff as a part of the field trial of the VPK Assessment Measures. For these children, the development project staff completed assessments within a two- to three-week period before or after the children were assessed by their teachers. Half of this sample received teacher assessments first and half of the sample received the project staff assessment first. Consequently, the test-retest period for these assessments was relatively brief (i.e., two to three weeks). Neither set of data provides a pure estimate of the test-retest reliabilities of the measures, however. In the first set of data, the teachers used the different forms of the measures (i.e., AP1, AP2, AP3) at each assessment; therefore, the estimated cross-time stability of scores also includes variability of scores due to alternative forms of the four measures. However, in this set of data, the assessments were conducted by the same individuals (i.e., children's classroom teachers). In the second set of data, the same version of the assessment was used; however, different individuals conducted the assessments. Therefore, in this set of data, the estimated cross-time stability of scores also includes variability of scores due to the type of assessor (e.g., project staff had more training and experience conducting assessments; children were more familiar with their classroom teachers than with the project staff conducting the assessments).

Table 2.3 (next page) includes correlations between assessment periods for the children assessed by their teachers at all three assessment periods in the field trial. As can be seen in the table, the correlations for scores in each domain assessed between adjacent assessment periods were moderate to high, particularly for the retest interval from AP2 to AP3. As expected, the correlations between scores from AP1 to AP3 were lower than the correlations across the other intervals; however, the test-retest interval for these scores was six months in most cases, compared to the two- to three-month retest intervals for the correlations between scores from AP1 to AP2 and from AP2 to AP3. The magnitude of these correlations--particularly those for the shorter retest intervals approaches the magnitude of the correlations obtained from the alternate-forms reliability analysis reported above. Consequently, these results indicate the expected level of cross-time consistency for the measures.

Table 2.4 (next page) includes correlations between scores obtained from teacher-administered assessments and scores obtained from project staff administered assessments. As can be noted in the table, correlations between scores from teachers and project staff were

Table 2.3. ***Correlations Between Scores at Different Assessment Periods for Children Administered VPK Assessment Measures by Teachers at AP1, AP2, and AP3***

	Test-Retest Interval		
Measure	AP1 to AP2	AP1 to AP3	AP2 to AP3
Print Knowledge	.70	.59	.81
Phonological Awareness	.64	.56	.73
Oral Language	---[1]	---[1]	.73
Math	.76	.69	.82

Notes. N = 957; Sample only includes children assessed by teachers at all three periods during field trial. [1]Oral Language measure was not administered in field trial at AP1.

Table 2.4. ***Correlations Between Scores from Measures Administered by teachers and Measures Administered by Project Staff at Different Assessment Periods for Children in VPK Assessment Measure Field Trial***

	Assessment Period During Field Trial		
Measure	AP1	AP2	AP3
N in Sample	146	148	125
Print Knowledge	.80	.80	.90
Phonological Awareness	.67	.61	.75
Oral Language	---[1]	.64	.71
Math	.75	.83	.84

Notes. [1]Oral Language measure was not administered in field trial at AP1.

moderate to high, and, in most cases, approached or equaled the estimates of internal-consistency reliabilities for the measures. Correlations between scores on the teacher versus project-staff administered measures of phonological awareness measures were somewhat lower than the

correlations for the measures of the other skills, suggesting a degree of variability in how the phonological awareness measures were administered or scored by teachers versus project staff. However, even for these measures, the correlations were well within the acceptable range for test-retest correlations.

Overall Summary

Results from both IRT and CTT analyses reveal that the VPK Assessment Measures provide a reliable assessment of children's skills in the domains of print knowledge, phonological awareness, oral language, and math skills. IRT analyses revealed that each version of the measures of the four domains provides a high degree of precision of measurement in the region of the ability distribution most relevant for identifying children who have weak early language, literacy, or math skills and who are, therefore, at high risk of failing to meet the State of Florida's Kindergarten Readiness Criteria. For all measures, but particularly for the print knowledge, oral language, and math measures, precise measurement was obtained over a wide range of abilities that spanned from around average to well below average levels. The range of ability for which the phonological awareness measure provided precise measurement increased from AP1 to AP3, most likely as a result of children's development of phonological awareness skills.

CTT analyses concerning internal-consistency reliability confirmed the results of the IRT analyses and showed that all versions of the measures of the four skill areas had moderate to high levels of internal consistency in several independent samples. Analyses of alternate-forms reliability demonstrated that the three versions of each measure were assessing the same underlying ability. These findings support the expectation that the three forms of each measure (i.e., AP1, AP2, AP3) represent parallel forms of the same measure, and they provide strong support for using the three forms of each measure as a way to monitor children's development of skills in the four skill domains. Finally, analyses of test-retest reliability indicated that each measure had moderate to high levels of cross-time stability at a level expected given the measures' internal-consistency reliabilities and the length of time between assessments in the test-retest analyses.

3. VALIDITY

The validity of a test refers to the extent to which it measures what it is intended to measure. Validity is a concept that is difficult to separate completely from the theoretical underpinnings of the construct being measured because the theory specifies the expected pattern of relations involving the construct. Establishing the validity of an assessment typically involves demonstrating that the test (a) includes items that adequately sample the domain of the construct to be measures (content validity), (b) correlates with other measures of the same or a highly related construct or that it identifies known groups (convergent or concurrent validity), (c) predicts construct-related outcomes over time (predictive validity), and (d) is unrelated to different constructs (discriminant validity). For example, the validity of an assessment of print knowledge would be supported if it correlated with a measure of letter knowledge (concurrent validity), correlated with a concurrent or later measure of decoding (predictive validity), and correlated less with a measure of receptive vocabulary than with a measure of letter knowledge (discriminant validity).

Defining the level of evidence required to determine that an assessment is valid is difficult because of the multitude of possible theoretical relations (most of which are not evaluated) and the strength of relations specified by the theory. Assessments designed to measure similar constructs can be more or less valid than each other depending on the sizes of their validity coefficients with theoretically specified constructs. In general, however, assessments with low concurrent or predictive coefficients, or high relative discriminative coefficients would not be considered valid (i.e., a print knowledge assessment that correlated as highly with a measure of general intelligence as it did with an assessment of letter knowledge would have poor discriminant validity). An assessment with low reliability also would have validity problems because an assessment cannot be expected to correlate with related constructs at a higher level than it can correlate with itself (i.e., the more measurement error in an assessment, the less able it will be to predict important outcomes accurately).

For the VPK Assessment Measures, content validity for the measure was derived by aligning the measures with content from the Florida VPK Standards. Concurrent and predictive validity of the measures were specifically evaluated in two studies. In the first study, the relations between children's scores on the VPK Assessment Measures were compared to children's scores on diagnostic measures of the same constructs. In the second study, the

predictive validity of the VPK Assessment Measures were evaluated by examining how well children's scores on the VPK Assessment Measures predicted scores on other measures administered to children when they entered kindergarten, included the measures used to establish kindergarten readiness (i.e., FAIR-K, ECHOS).

Concurrent Validity

To establish the concurrent validity of the VPK Assessment Measures, a group of 288 children attending preschools were administered the three versions (i.e., AP1, AP2, AP3) of the Print Knowledge, Phonological Awareness, Oral Language, and Mathematics subtests of the VPK Assessment Measures. Additionally, each of these children completed two nationally standardized diagnostic tests, the Test of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen, & Rashotte, 2007) and the Test of Early Mathematics Abilities, 3rd Edition (TEMA-3; Ginsberg & Baroody, 2003).

The TOPEL includes three subsets, Print Knowledge (PK), Phonological Awareness (PA), and Definitional Vocabulary (DV), and each subtest has high reliability for this age group (i.e., αs of .86 to .96 for 3- to 5-year-old children). The 36-item PK subtest measures print concepts, letter discrimination, word discrimination, and letter-name and sound discrimination. The 27-item PA subtest uses both blending and elision tasks to measure the developmental continuum of phonological awareness skills. The 36-item DV subtest measures children's single-word spoken vocabulary and their abilities to define words. The TEMA-3 yields a single score reflecting children's abilities to complete counting and basic arithmetic tasks, and it has high levels of reliability for this age group. Both the TOPEL and the TEMA-3 have strong evidence for validity (e.g., substantial concurrent and predictive correlations with other measures of similar constructs). Both the standardized tests and the VPK Assessment Measures were administered in the children's schools by a trained assessment team. Children completed all of these measures over a 2- to 4-week period to avoid fatigue and to minimize practice effects, and the order of test administration was varied across children to prevent order effects.

Zero-order correlations between scores on the AP1, AP2, and AP3 versions of the four VPK Assessment Measures and scores on the TOPEL and TEMA-3 are shown in Table 3.1 (next page). All correlations between the VPK measures and the similar subtest score on the TOPEL and TEMA-3 were high and statistically significant (all *p*s < .001). Across the three versions of

Table 3.1. ***Correlations between Scores on AP1, AP2, and AP3 versions of VPK Assessment Measures and Scores on Diagnostic Tests of Language, Early Literacy, and Early Mathematics Skills***

	TOPEL Subtests			
	Print Knowledge	Phonological Awareness	Definitional Vocabulary	TEMA-3
		VPK AP1 Measures		
Print Knowledge	**.86**	.44	.43	.56
Phonological Awareness	.40	**.68**	.45	.55
Oral Language	.51	.55	**.68**	.52
Mathematics	.73	.55	.48	**.73**
		VPK AP2 Measures		
Print Knowledge	**.87**	.44	.41	.54
Phonological Awareness	.37	**.57**	.49	.46
Oral Language	.44	.55	**.66**	.45
Mathematics	.70	.54	.48	**.76**
		VPK AP3 Measures		
Print Knowledge	**.86**	.41	.40	.57
Phonological Awareness	.44	**.72**	.52	.55
Oral Language	.45	.51	**.63**	.48
Mathematics	.72	.53	.48	**.76**

Notes. TOPEL = Test of Preschool Early Literacy; TEMA-3 = Test of Early Mathematics Ability, 3rd Edition; correlations in bold represent convergent validity coefficients for VPK Assessment Measures; $N = 288$; all correlations are significant at $p < .001$.

VPK Assessment Measures, the average convergent validity coefficient was .86 for Print Knowledge, .66 for Phonological Awareness, .66 for Oral Language, and .75 for Math. The average discriminant validity coefficients for Print Knowledge (.47), Phonological Awareness

(.47), Oral Language (.49), and Math (.58) were lower than these convergent validity coefficients. Statistical tests of the differences between convergent and discriminant coefficients within AP1, AP2, and AP3 revealed that in all but three instances, the convergent coefficients for the VPK Assessments Measures were significantly higher than their discriminant coefficients. For the AP1 and the AP3 versions, the VPK Math Assessment was not more highly correlated with scores on the TEMA-3 (*r*s = .73 and .76, for AP1 and AP3, respectively) than with scores on the PK subtest of the TOPEL (*r*s = .73 and .72, for AP1 and AP3, respectively), and for the AP2 version, the VPK Oral Language measure was not more highly correlated with scores on the DV subtest of the TOPEL ($r = .57$) than with scores on the PA subtest of the TOPEL ($r = 49$; $.10 > p > .05$).

Overall, these results indicate that the four VPK Assessment Measures have good concurrent validity. Correlations between the VPK Measure and its corresponding measure on one of the standardized diagnostic measures were high. Moreover, 92% of the 36 discriminant coefficients were significantly lower than their companion convergent coefficients, indicating that the VPK Assessment Measures specifically measure the constructs that they are intended to measure.

Predictive Validity

To establish the predictive validity of the VPK Assessment Measures, data from the Field Trial of the measures were matched to children's scores on measures collected as a part of the initial kindergarten assessment using the Florida Assessment of Reading for Instruction (FAIR-K). Of the 1,307 children who were assessed using the VPK Assessment Measures during at least one of the assessment periods (i.e., AP1 - AP3), approximately 1,095 could be matched in the Florida Department of Education's database and at least some relevant data from the kindergarten screening retrieved[1]. These children attended VPK in 191 VPK classrooms distributed across 186 VPK Providers. When they were screened in kindergarten, these children attended 516 schools in 48 different school districts. At the time of initial preschool assessment (i.e., AP1), the children with kindergarten screening data were 56.1 months of age ($SD = 3.60$), and boys made up 48 percent of the sample. The majority of this group of children was White

[1]Parental consent/permission was obtained for all children who participated in the Field Study and for whom identified data was obtained from the Florida Department of Education's database.

(48.6%), and the remainder of the sample was African American/Black (25.5%), Latino/Hispanic (16.2%), Asian (2.0%), multiracial (6.2%), or other (1.5%).

Prediction of FAIR-K Scores

As part of the FAIR-K assessment at Kindergarten AP1, children complete assessments of letter-name knowledge, phonological awareness, vocabulary, and listening comprehension (both implicit and explicit questions). Scores on the letter-name knowledge measure and the phonological awareness measure (i.e., FAIR-K AP1 Broad Screen) are used to establish an index of Probability of Reading Success (PRS). The PRS for the FAIR was established by predictive analysis to scores on end-of-year standardized assessment of reading (for FAIR-K this was the Stanford Early School Achievement Test [SESAT]). The PRS was benchmarked for at-risk status in reading, defined as scoring below the 40th percentile on the SESAT at the end of the kindergarten year. According to the FAIR K-2 Technical Manual, scores on the Broad Screen at AP1 in kindergarten accounted for 17% of the variance in SESAT scores. Classification analyses indicated that the Broad Screen at AP1 correctly classified 63.3% of children into "at-risk" and "not at-risk" categories.

Zero-order correlations between scores on the AP1, AP2, and AP3 versions of the Print Knowledge, Phonological Awareness, and Oral Language VPK Assessment Measures and scores on the FAIR-K measures collected at kindergarten AP1 are shown in Table 3.2. All correlations between the VPK Assessment Measures and FAIR-K scores were statistically significant (ps < .001), and, in general, the correlations between the VPK Assessment Measures and the FAIR-K scores increased in size over time. Correlations between VPK Print Knowledge and FAIR Letter-Name Knowledge scores increased significantly from AP1 to AP2, AP2 to AP3, and AP1 to AP3 (all ps < .001). Correlations between VPK Phonological Awareness and FAIR Phonological Awareness scores increased significantly from AP1 to AP2, AP2 to AP3, and AP1 to AP3 (all ps < .05). Correlations between VPK Oral Language and FAIR Vocabulary scores did not increase from AP2 to AP3. Correlations between scores on the VPK Assessment Measures at AP1, AP2, and AP3 also were significantly correlated with the Probability of Reading Success index of the FAIR-K. Correlations of scores on the VPK Print Knowledge and the VPK Phonological Awareness with the FAIR-K Probability of Reading Success index increased significantly from AP1 to AP2, AP2 to AP3, and AP1 to AP3 (all ps < .05); however, correlations between scores

Table 3.2. ***Correlations between VPK Assessment Measures Administered by Teachers at AP1, AP2, and AP3 and Scores from Kindergarten Readiness Administration of FAIR***

	Outcome Indices from FAIR-K				
VPK Measures	Letter Names	Phonological Awareness	Listening Comprehension	Vocabulary	Probability of Reading Success
VPK Assessments at AP1					
Print Knowledge	**.39**	.41	.22	.31	.43
Phonological Awareness	.19	**.37**	.29	.40	.31
VPK Assessments at AP2					
Print Knowledge	**.56**	.39	.15	.23	.53
Phonological Awareness	.24	**.44**	.30	.37	.39
Oral Language	.23	.36	.37	**.50**	.34
VPK Assessments at AP3					
Print Knowledge	**.63**	.35	.13	.20	.55
Phonological Awareness	.28	**.48**	.31	.38	.45
Oral Language	.21	.37	.34	**.49**	.34

Notes. All correlations are significant at $p < .001$; correlations in bold represent convergent validity coefficients for VPK Assessment Measures; $N = 1{,}008$ at AP1; $N = 914$ at AP2; $N = 898$ at AP3.

on the VPK Oral Language and the FAIR-K Probability of Reading Success index did not increase from AP2 to AP3.

Statistical tests of the differences between convergent and discriminant coefficients within AP1, AP2, and AP3 revealed that in all but two instances, the convergent coefficients for the VPK Assessments Measures were significantly higher than their discriminant coefficients. Scores on the VPK Print Knowledge measure at AP1 were not more highly correlated with scores on the FAIR-K Letter-Name measure (r = .39) than with scores on the FAIR-K Phonological Awareness measure (r = .41), and scores on the VPK Phonological Awareness measure at AP1 were not more highly correlated with scores on the FAIR-K Phonological Awareness measure (r = .37) than with scores on the FAIR-K Vocabulary measure (r = .40).

Multiple regressions were used to examine the joint and unique predictive relations between scores on the VPK Assessment Measures at AP1, AP2, and AP3 and the FAIR-K PRS Index. Results from these analyses are shown in Table 3.3. At AP1 in pre-K, only the Print Knowledge and Phonological Awareness measures were administered. Together, these two measures accounted for 21 percent of the variance in the FAIR-K PRS Index, and both measures contributed significant unique variance to the prediction of FAIR-K PRS. At AP2 and AP3 in pre-K, Print Knowledge, Phonological Awareness, and Oral Language measures were

Table 3.3. ***Summary of Multiple Regressions for VPK Assessment Measures Administered by Teachers at AP1, AP2, and AP3 Predicting Probability of Reading Success from FAIR-K***

	Variance Accounted for in Model				
Predictor Variables	Overall R^2	Print Knowledge	Phonological Awareness	Oral Language	Shared
VPK Assessments at AP1	.21	.12	.02	---	.07
VPK Assessments at AP2	.33	.16	.02	.01	.14
VPK Assessments at AP3	.38	.17	.04	.001[a]	.17

Notes. [a]p = .33; unless otherwise marked, all total and unique variance components were significant at $p < .001$; N = 1,008 at AP1; N = 914 at AP2; N = 898 at AP3.

administered. At AP2, these three measures accounted for 33 percent of the variance in the FAIR-K PRS Index, and at AP3, these three measures accounted for 38 percent of the variance in the FAIR-K PRS Index. All three measures at AP2 contributed significant unique variance to the prediction of FAIR-K PRS; however, at AP3, only the VPK Print Knowledge and the VPK Phonological Awareness measures contributed significant unique variance to the prediction of FAIR-K PRS.

Logistic regression models were used to determine the joint and unique predictive relations between scores on the VPK Assessment Measures at AP1, AP2, and AP3 and the Kindergarten Readiness Classification that is based on the FAIR-K PRS Index. For these analyses, children with PRS scores less than 67 were classified as "at-risk" (coded as 1), and children with PRS scores of 67 or higher were classified as "not at-risk" (coded as 0). The results of these logistic regressions are summarized in Table 3.4 (next page). Both VPK Assessment Measures administered at AP1 uniquely contributed to the prediction of risk status based on the FAIR-K PRS. The VPK Print Knowledge measure and the VPK Phonological Awareness measure administered at AP2 and AP3 uniquely contributed to the prediction of risk status based on the FAIR-K PRS. For both AP2 and AP3, the VPK Oral Language measures were not significant unique predictors of risk status based on the FAIR-K PRS; however, the AP2 VPK Oral Language measure just failed to achieve conventional levels of significance ($p = .06$) as a unique predictor.

Classification Accuracy for Kindergarten Readiness

The results described above provide evidence that the VPK Assessment measures are valid measures of their intended constructs. In the majority of cases, convergent validity coefficients were significantly higher than discriminant validity coefficients, indicating specificity of measurement for the VPK Print Knowledge, Phonological Awareness, and Oral Language Measures. In addition, the VPK Assessment Measures were both singularly and jointly related to one of the outcome metrics used to determine VPK children's kindergarten readiness (i.e., Probability of Reading Success). Multivariate and logistic regressions demonstrated that these relations held both for the continuous FAIR-K PRS outcome at the start of kindergarten and for the risk-status classification derived from it. The final step in evaluating the relations between the VPK Assessment Measures and the FAIR-K was to determine the accuracy of

Table 3.4. ***Summary of Logistic Regressions for VPK Assessment Measures Administered by Teachers at AP1, AP2, and AP3 Predicting Kindergarten Readiness (Probability of Reading Success < 67) on FAIR***

							95% Confidence Interval for Exp(B)	
Assessment Period Measure	Parameter Estimate	Standard Error	Wald	df	p-value	Exp(B)	Lower	Upper
VPK Assessments at AP1								
Intercept	.83	.26	10.30	1	.001	2.89		
Print Knowledge	-.29	.04	67.64	1	<.001	.75	.70	.80
Phonological Awareness	-.12	.03	13.21	1	<.001	.88	.83	.94
VPK Assessments at AP2								
Intercept	3.29	.69	22.94	1	<.001	26.91		
Print Knowledge	-.30	.04	70.83	1	<.001	.74	.69	.80
Phonological Awareness	-.14	.04	10.45	1	.001	.87	.80	.95
Oral Language	-.08	.04	3.54	1	.06	.92	.85	1.00
VPK Assessments at AP3								
Intercept	4.34	.76	32.63	1	<.001	76.77		
Print Knowledge	-.36	.04	82.85	1	<.001	.70	.65	.75
Phonological Awareness	-.23	.05	26.15	1	<.001	.79	.73	.87
Oral Language	-.04	.05	0.54	1	.46	.97	.88	1.06

Notes. N = 1,008 at AP1; N = 912 at AP2; N = 898 at AP3.

predictive classifications of risk status for failing to achieve kindergarten readiness using the VPK Assessment Measures at the different assessment periods.

Receiver operator characteristic (ROC) curves were used to identify appropriate cut points on the VPK Assessment Measures at each assessment period (i.e., AP1 - AP3) with respect to risk status on the FAIR-K PRS (PRS < 67). Dichotomization of continuous measures introduces substantial error into prediction because even highly reliable measures contain measurement error (i.e., variation in observed scores around true scores). Use of a single cut score to dichotomize a measure does not take into account this measurement error, and, therefore, some children will be misclassified. This misclassification creates a choice between over- and under-identification. Given that the purpose of screening using the VPK Assessment Measures is to identify children at risk of failing to achieve kindergarten readiness so that instructional adaptations can be implemented, it was determined that under-identification of children (i.e., false negatives--children classified as not at risk who end up classified as not kindergarten ready) was more problematic than over-identification of children (i.e., false positives--children classified as at risk who end up classified as kindergarten ready). Therefore, cut scores were selected to reduce the incidence of false negatives.

Using ROC curve analyses, the score on each of the VPK Assessment Measures at AP1, AP2, and AP3 at which specificity of approximately .90 was achieved was determined. A summary of the classification results is shown in Table 3.5 (next page). Analyses determined the predictive accuracy of using each VPK Assessment Measure and combinations (i.e., scoring below the cut score on any of the measures in the combination) of measures for kindergarten readiness classification. As shown in Table 3.5 (next page), each measure and measure combination resulted in statistically significant prediction across the range of scores (i.e., Area Under the Curve). Overall classification accuracy (i.e., Overall Correct Classification) for individual measures ranged from .80 for using either VPK Print Knowledge or VPK Phonological Awareness at AP1 to predict kindergarten readiness to .88 for using VPK Print Knowledge at AP2 to predict kindergarten readiness. Combinations of measures (i.e., children classified as at risk if they scored below the cut score on any measure included in the combination) resulted in a slightly lower overall correct classification accuracy (range: .75 - .84) because of the reduction in false negatives and a corresponding increase in false positives.

Table 3.5. ***Predictive Accuracy Between VPK Assessment Measures Administered by Teachers at AP1, AP2, and AP3 and Kindergarten Readiness Based on FAIR-K Probability of Reading Success***

VPK Assessment Screen Used	Cut Score	AUC	OCC	NPP	PPP	Specificity	Sensitivity
***VPK Assessment Measures Administered at AP1* (*N* = 1,008)**							
Print Knowledge	4	.77	.80	.89	.36	.87	.36
Phonological Awareness	4	.68	.80	.88	.28	.89	.28
Print Knowledge + Phonological Awareness	---	.77	.75	.91	.53	.79	.53
***VPK Assessment Measures Administered at AP2* (*N* = 922)**							
Print Knowledge	7	.79	.88	.91	.53	.95	.39
Phonological Awareness	5	.73	.83	.90	.28	.92	.23
Oral Language	16	.70	.83	.91	.31	.90	.33
Print Knowledge + Phonological Awareness	---	.81	.83	.93	.38	.87	.54
Print Knowledge + Phonological Awareness + Oral Language	---	.80	.78	.94	.32	.80	.63
***VPK Assessment Measures Administered at AP3* (*N* = 898)**							
Print Knowledge	8	.79	.87	.93	.49	.92	.51
Phonological Awareness	7	.78	.86	.91	.42	.93	.35
Oral Language	16	.72	.83	.90	.30	.91	.26
Print Knowledge + Phonological Awareness	---	.85	.84	.95	.42	.87	.67
Print Knowledge + Phonological Awareness + Oral Language	---	.84	.80	.95	.36	.82	.71

Notes. AUC = Area under curve in Receiver Operator Characteristic Curve Analysis; OCC = Overall classification accuracy; NPP = negative predictive power; PPP = positive predictive power.

Across the three assessment periods and VPK Assessment Measures, between 68 and 87 percent of children could be classified accurately as kindergarten ready or not kindergarten ready using these cut scores.

As seen in Table 3.5, the rates of false negatives (i.e., 1 - negative predictive power) ranged from 5 percent to 12 percent--between 5 and 12 percent of children who were actually not kindergarten ready based on FAIR-K PRS were classified as not at-risk using these cut scores on the VPK Assessment Measures. Consistent with a trade off between false negatives and false positives, the rate of false positives ranged from 53 percent to 73 percent--between 53 and 73 percent of children who were actually kindergarten ready based on FAIR-K PRS were classified as at-risk using these cut scores on the VPK Assessment Measures. In general, use of the combination of the VPK Print Knowledge and VPK Phonological Awareness measures to classify children resulted in the best balance of minimizing false negatives and minimizing false positives. These classification results are consistent with the results of the multivariate and logistic regression in which the VPK Oral Language measure contributed the least unique predictive influence to FAIR-K PRS. Such a result is not surprising given that the FAIR-K PRS is derived from the FAIR-K Letter Name measure and the FAIR-K Phonological Awareness measure. However, the other predictive analyses demonstrate that the VPK Oral Language measure is specifically predictive of FAIR-K Oral Language, which is not used in deriving the kindergarten readiness classification.

Prediction of Score on Early Childhood Observation System

In addition to completing the FAIR-K as a part of kindergarten readiness screening, kindergarten teachers of children who attended VPK complete the Early Childhood Observation System (ECHOS). The ECHOS contains 19 items that require teachers to indicate the stage a child is at within different developmental domains. Three levels of classification are used for each domain--“Not Yet Demonstrating,” “Emerging/Progressing,” and “Demonstrating.” Ratings of children in the VPK Field Trial sample showed little variability on the majority of ECHOS items (see Table 3.6, next page). Few children were rated in the lowest category across all items, and for many items, most children were rated in the highest category. The majority of children (i.e., 60.7%) received an average rating across all ECHOS items that placed them in the

Table 3.6. ***Distribution of Kindergarten Children's Ratings in Developmental Categories on Early Childhood Observation System (ECHOS) Items and Total Score in Validity Sample***

Item	Status		
	Not Yet Demonstrating	Emerging/ Progressing	Demonstrating
1. Concepts of Print	0.8	29.1	70.1
2. Oral Language & Vocabulary	3.2	46.3	50.5
3. Comprehension 1	1.7	53.6	44.7
4. Comprehension 2	6.9	16.5	76.6
5. Writing	6.7	16.6	76.7
6. Number Sense & Operations	1.7	16.3	82.1
7. Geometry	5.6	27.6	66.8
8. Algebraic Thinking	5.8	58.6	35.6
9. Data Analysis	5.1	47.4	47.5
10. Responsive Decision Making	2.3	26.7	71.0
11. Social Problem Solving	0.5	37.3	62.3
12. Approaches to Learning	7.2	39.3	53.5
13. Scientific Inquiry	8.1	58.2	33.7
14. Production, Distribution, Consumption	3.5	48.9	47.5
15. Civil Ideas and Participation	0.7	32.4	66.9
16. Fitness	0.3	23.7	76.0
17. Fine Motor Skills	1.1	5.4	93.5
18. Dance	1.4	33.8	64.8
19. Visual Arts	5.6	38.7	55.7
ECHOS Total Score Average	0.5	38.8	60.7

Demonstrating category. Analyses of the 19 ECHOS items indicated that they were best represented as a single construct. Factor analysis of the 19 ECHOS items resulted in a 1-factor solution, and the internal consistency of the 19 ECHOS items was high (i.e., $\alpha = .92$). Regardless, three ECHOS composite scales were created for analyses. A 5-item Language/Literacy Scale included ECHOS items 1 - 5 ($\alpha = .80$). A 4-item Math Scale included

ECHOS items 6 - 9 (α = 76), and a 3-item Socio-Emotional Scale included ECHOS items 10 - 12 (α = .70).

Zero-order correlations between scores on the VPK Assessment Measures at AP1, AP2, and AP3 and scores on the ECHOS total score and Language/Literacy, Math, and Socio-Emotional scales are shown in Table 3.7 (next page). All correlations were in the low to moderate range, and all correlations were statistically significant at the $p < .001$ level. For comparison, Table 3.7 also includes the correlations between the four measures from the FAIR-K and ECHOS scores. In general, the correlations of the Print Knowledge, Phonological Awareness, and Oral Language VPK measures at AP1, AP2 and AP3 with the ECHOS Language/Literacy scale were similar to the correlations of the Letter Name, Phonological Awareness, Vocabulary, and Listening Comprehension FAIR-K measures with the ECHOS Language/Literacy scale. Statistical comparisons of the strength of these correlations revealed that the correlation between FAIR-K Phonological Awareness and ECHOS Language/Literacy was significantly higher than the correlation between VPK Phonological Awareness at AP1 and ECHOS Language/Literacy ($p < .01$). Additionally, the correlation between VPK Oral Language at AP2 and ECHOS Language Literacy was higher than the correlations between ECHOS Language/Literacy and both FAIR-K Vocabulary and FAIR-K Listening Comprehension ($ps < .01$).

Statistical tests of the differences between convergent and discriminant coefficients within AP1, AP2, and AP3 provided mixed evidence for specificity of prediction. Although 11 of 12 convergent coefficients for the VPK Assessment Measures (i.e., Print Knowledge, Phonological Awareness, Oral Language, Math) were significantly larger than the discriminant coefficients with the ECHOS Socio-Emotional scale, three of the three convergent coefficients for the VPK Phonological Awareness Measure were significantly larger than discriminant coefficients with the ECHOS Math scale, and both convergent coefficients for the VPK Oral Language Measure were significantly larger than the discriminant coefficients with the ECHOS Math scale, none of the three convergent coefficients for the VPK Print Knowledge Measure were significantly larger than the discriminant coefficients with the ECHOS Math scale and only one of the three convergent coefficients for the VPK Math Measure were significantly larger than the discriminant coefficients with the ECHOS Language/Literacy scale. Significantly, a similar pattern of results was obtained when examining convergent and discriminant coefficients

Table 3.7. ***Correlations between Scores from Early Childhood Observation System (ECHOS) and VPK Assessment Measures Administered by Teachers at AP1, AP2, and AP3 and Scores from Kindergarten Readiness Administration of FAIR***

	ECHOS Scales			
Assessment Period Measures	Total	Language / Literacy	Math	Socio-emotional
VPK Assessments at AP1				
Print Knowledge	.29	**.30**	.32	.22
Phonological Awareness	.20	**.26**	.21	.15
Math	.32	.34	**.32**	.25
VPK Assessments at AP1				
Print Knowledge	.26	**.30**	.29	.18
Phonological Awareness	.22	**.28**	.22	.16
Oral Language	.28	**.35**	.22	.22
Math	.31	.35	**.30**	.23
VPK Assessments at AP1				
Print Knowledge	.27	**.28**	.28	.21
Phonological Awareness	.25	**.29**	.21	.18
Oral Language	.24	**.29**	.18	.19
Math	.32	.35	**.32**	.26
FAIR-K Assessments				
Letter Names	.32	**.34**	.36	.23
Phonological Awareness	.29	**.35**	.29	.23
Listening Comprehension	.23	**.26**	.15	.20
Vocabulary	.22	**.26**	.20	.18

Notes. N =1,004 at AP1; N = 910 at AP2; N = 894 at AP3; N = 1,064 for FAIR-K; all correlations are significant at $p < .001$; correlations in bold represent convergent validity coefficients for VPK Assessment Measures and FAIR-K measures.

between the FAIR-K measures and the ECHOS. All FAIR-K measures were more highly correlated with the ECHOS Language/Literacy Scale than with the ECHOS Socio-emotional scale. The Phonological Awareness, Listening Comprehension, and Vocabulary measures from the FAIR-K were more highly correlated with the ECHOS Language/Literacy scale than with the ECHOS Math scale; however, the Letter Name measure from the FAIR-K was not more highly correlated with the ECHOS Language/Literacy Scale than with the ECHOS Math Scale.

Summary. Validity coefficients between scores on the VPK Assessment Measures and scores on the ECHOS were only moderate. However, with only a few exceptions, these correlations were similar to the correlations between scores on the ECHOS and scores on the FAIR-K measures that were administered concurrently, rather than 12-, 7- or 5-months before the ECHOS as was the case with the VPK Assessment Measures. In part, these results may be the result of the limited variability of children's scores on the ECHOS. Despite this limited variability and the absence of a clear pattern of item covariance that would indicate that the items on the ECHOS assess different constructs, there was some evidence of the expected pattern of convergent and discriminant relations. All but one of the convergent correlations for the VPK Assessment Measures were significantly higher than the correlations between the VPK measures and the socio-emotional scale constructed from the ECHOS, although this ECHOS scale had the fewest items and lowest reliability of the three constructed scales and this may have contributed to the lower correlations. Regardless, the overall pattern of results for the VPK Assessment Measures was similar to the pattern of results for the FAIR-K measures, which suggests that the VPK measures are at least as valid with respect to the ECHOS as the FAIR-K measures. Future investigations should compare performance on the VPK Math measure and a specific and valid indicator to children's math performance in kindergarten to determine more clearly the validity of the VPK Math measure.

References

Ginsburg, H. P. (1975). Young children's informal knowledge of mathematics. *Journal of Children's Mathematical Behavior*, 1, 63-156.

Ginsburg, H. P. & Baroody, A. J. (2003). *Test of Early Mathematics Ability (3rd ed.)*, Pro-ed, Austin, TX.

Greenes, C., Ginsburg, H. P., & Balfanz, R. (2004). Big math for little kids. *Early Childhood Research Quarterly*, 19, 159 - 166.

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

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, 45, 850-867.

Lonigan, C. J. (2006). Development, assessment, and promotion of pre-literacy skills. *Early Education and Development, 17,* 91-114.

Lonigan, C. J., Schatschneider, C., & Westberg, L. (2008a). Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling. In *Developing Early Literacy: Report of the National Early Literacy Panel* (pp. 55-106). Washington, DC: National Institute for Literacy.

Lonigan, C. J., Schatschnieder, C., Westberg, L., & Smith, L. S. (2008b). Impact of code-focused interventions on young children's early literacy skills. In *Developing Early Literacy: Report of the National Early Literacy Panel* (pp. 107-151). Washington, DC: National Institute for Literacy.

Lonigan, C. J., Shanahan, T., & Cunningham, A. (2008). Impact of shared-reading interventions on young children's early literacy skills. In *Developing Early Literacy: Report of the National Early Literacy Panel* (pp. 153-171). Washington, DC: National Institute for Literacy.

Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. (2007). *Test of Preschool Early Literacy*. Austin, TX: ProEd.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd Ed.)*. New York: McGraw-Hill.

Sénéchal, M., & LeFevre, J. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development, 73,* 445–460.

Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly,* 19, 99 - 120.

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

APPENDIX 1

Table A1.1

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP1 Print Knowledge Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
PK1. Print Concepts (MC)	0.59	-0.99	0.69	0.34
PK2. Print Concepts (MC)	0.58	-0.88	0.67	0.37
PK3. Letter-name (MC)	1.15	-1.10	0.79	0.47
PK4. Letter-names (MC)	1.52	-1.12	0.82	0.49
PK5. Letter-names (MC)	1.37	-0.91	0.76	0.53
PK6. Letter-names (MC)	1.06	-0.63	0.67	0.52
PK7. Letter-sounds (MC)	0.81	-0.32	0.58	0.47
PK8. Letter-sound (MC)	0.95	-0.41	0.60	0.50
PK9. Letter-names (FR)	1.72	-0.26	0.57	0.66
PK10. Letter-sounds (MC)	1.59	-0.21	0.55	0.65
PK11. Letter-sounds (MC)	1.57	0.03	0.48	0.65
PK12. Letter-sounds (MC)	1.66	0.39	0.36	0.62

Note. MC = Multiple-choice item requiring only pointing response; FR = Item requiring spoken response.

Table A1.2

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP2 Print Knowledge Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
PK1. Print Concepts (MC)	0.59	-2.47	0.89	0.38
PK2. Print Concepts (MC)	0.80	-0.86	0.70	0.48
PK3. Letter-names (MC)	1.52	-1.39	0.87	0.55
PK4. Letter-names (MC)	1.55	-1.34	0.86	0.56
PK5. Letter-names (MC)	1.44	-1.23	0.84	0.58
PK6. Letter-sounds (MC)	1.02	-0.66	0.68	0.54
PK7. Letter-names (FR)	2.13	-0.86	0.77	0.71
PK8. Letter-names (FR)	1.86	-0.62	0.70	0.70
PK9. Letter-names (FR)	2.18	-0.66	0.72	0.73
PK10. Letter-names (FR)	2.81	-0.36	0.63	0.74
PK11. Letter-sounds (FR)	1.47	-0.70	0.71	0.66
PK12. Letter-sounds (FR)	2.30	-0.13	0.54	0.68

Note. MC = Multiple-choice item requiring only pointing response; FR = Item requiring spoken response.

Table A1.3

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP3 Print Knowledge Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
PK1. Letter-names (MC)	0.59	-2.55	0.90	0.57
PK2. Letter-names (MC)	1.41	-1.69	0.91	0.51
PK3. Letter-names (MC)	1.69	-1.37	0.87	0.60
PK4. Letter-names (MC)	1.24	-1.24	0.82	0.56
PK5. Letter-sounds (MC)	1.41	-1.58	0.89	0.53
PK6. Letter-sounds (MC)	1.63	-1.09	0.81	0.65
PK7. Letter-sounds (MC)	1.05	-0.80	0.71	0.54
PK8. Letter-names (FR)	1.53	-0.77	0.73	0.66
PK9. Letter-names (FR)	1.61	-0.90	0.77	0.67
PK10. Letter-names (FR)	1.93	-0.58	0.69	0.71
PK11. Letter-sounds (FR)	1.72	-0.69	0.71	0.69
PK12. Letter-sounds (FR)	1.92	-0.40	0.63	0.69

Note. MC = Multiple-choice item requiring only pointing response; FR = Item requiring spoken response.

Table A1.4

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP1 Phonological Awareness Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
PA1. Blending Syllable (MC)	0.59	-2.58	0.90	0.25
PA2. Blending Syllable (MC)	0.64	-2.38	0.90	0.27
PA3. Blending Onset-Rime (MC)	0.53	-1.48	0.76	0.32
PA4. Blending Onset-Rime (MC)	0.48	-1.90	0.80	0.29
PA5. Blending Syllable (FR)	0.75	-0.46	0.61	0.49
PA6. Blending Word (FR)	0.81	-0.29	0.58	0.52
PA7. Blending Syllable (FR)	0.77	-0.02	0.51	0.52
PA8. Blending Onset-Rime (FR)	0.82	0.07	0.49	0.55
PA9. Blending Onset-Rime (FR)	0.81	0.56	0.36	0.52
PA10. Elision Word (FR)	2.24	0.52	0.32	0.57
PA11. Elision Word (FR)	2.41	0.33	0.39	0.58
PA12. Elision Syllable (FR)	2.60	0.45	0.34	0.59
PA13. Elision Syllable (FR)	1.26	0.93	0.23	0.46
PA14. Elision Onset-Rime (FR)	1.29	1.66	0.09	0.36

Note. MC = Multiple-choice item requiring only pointing response; FR = Item requiring spoken response.

Table A1.5

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP2 Phonological Awareness Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
PA1. Blending Onset-Rime (MC)	0.59	-3.38	0.95	0.26
PA2. Blending Word (MC)	0.47	-3.74	0.94	0.20
PA3. Blending Onset-Rime (MC)	0.63	-2.22	0.88	0.31
PA4. Blending Phoneme (MC)	0.67	-2.03	0.87	0.33
PA5. Blending Syllable (FR)	0.64	-0.70	0.65	0.47
PA6. Blending Onset-Rime (FR)	0.87	-0.38	0.60	0.57
PA7. Blending Syllable (FR)	0.90	-0.29	0.58	0.60
PA8. Blending Word (FR)	0.81	-0.71	0.68	0.55
PA9. Blending Phoneme (FR)	0.74	0.29	0.43	0.51
PA10. Elision Word (FR)	3.59	0.14	0.45	0.66
PA11. Elision Word (FR)	3.97	0.03	0.50	0.66
PA12. Elision Syllable (FR)	2.59	0.21	0.43	0.62
PA13. Elision Syllable (FR)	1.31	0.42	0.37	0.51
PA14. Elision Onset-Rime (FR)	1.31	1.34	0.14	0.41

Note. MC = Multiple-choice item requiring only pointing response; FR = Item requiring spoken response.

Table A1.6

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP3 Phonological Awareness Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
PA1. Blending Syllable (MC)	0.59	-3.90	0.97	0.24
PA2. Blending Word (MC)	0.79	-2.26	0.91	0.32
PA3. Blending Onset-Rime (MC)	0.81	-1.97	0.89	0.36
PA4. Blending Syllable (MC)	0.73	-1.96	0.87	0.36
PA5. Blending Word (FR)	1.23	-0.85	0.74	0.60
PA6. Blending Syllable (FR)	1.37	-0.64	0.70	0.63
PA7. Blending Syllable (FR)	1.19	-0.64	0.69	0.61
PA8. Blending Word (FR)	1.32	-0.79	0.74	0.62
PA9. Blending Onset-Rime (FR)	1.00	-0.06	0.52	0.56
PA10. Elision Word (FR)	2.59	-0.18	0.57	0.70
PA11. Elision Word (FR)	2.42	-0.22	0.58	0.69
PA12. Elision Syllable (FR)	1.77	0.02	0.50	0.62
PA13. Elision Syllable (FR)	1.74	0.03	0.49	0.63
PA14. Elision Onset-Rime (FR)	1.44	0.84	0.25	0.49

Note. MC = Multiple-choice item requiring only pointing response; FR = Item requiring spoken response.

Table A1.7

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP1 Oral Language Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
OL1. Definitions (MC)	0.59	-3.17	0.94	0.35
OL2. Expressive Vocabulary	1.04	-2.13	0.93	0.38
OL3. Definitions (MC)	1.22	-1.62	0.89	0.47
OL4. Expressive Vocabulary	1.31	-1.02	0.79	0.55
OL5. Definitions (MC)	0.98	-1.77	0.89	0.43
OL6. Expressive Vocabulary	0.90	0.68	0.33	0.39
OL7. Definitions (MC)	0.58	-1.21	0.73	0.36
OL8. Expressive Vocabulary	0.59	2.49	0.11	0.20
OL9. Grammar (MC)	0.65	-2.78	0.93	0.29
OL10. Grammar (MC)	0.77	-1.29	0.78	0.42
OL11. Grammar (MC)	0.70	-0.85	0.69	0.41
OL12. Receptive Vocabulary (MC)	1.07	-2.69	0.97	0.28
OL13. Receptive Vocabulary (MC)	0.62	-1.86	0.83	0.33
OL14. Receptive Vocabulary (MC)	1.04	-1.67	0.88	0.45
OL15. Receptive Vocabulary (MC)	0.96	-1.56	0.85	0.45
OL16. Receptive Vocabulary (MC)	1.17	-1.01	0.77	0.56
OL17. Word Knowledge	0.99	-2.17	0.93	0.38
OL18. Word Knowledge	0.23	-5.20	0.88	0.12
OL19. Word Knowledge	0.69	-1.55	0.81	0.39
OL20. Word Knowledge	0.84	-1.09	0.76	0.45
OL21. Word Knowledge	0.69	-0.17	0.54	0.41
OL22. Word Knowledge	0.64	0.95	0.30	0.33

Note. MC = Multiple-choice item requiring only pointing response.

Table A1.8

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP2 Oral Language Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
OL1. Definitions (MC)	0.59	-3.25	0.94	0.36
OL2. Expressive Vocabulary	1.09	-1.80	0.90	0.44
OL3. Definitions (MC)	0.94	-2.07	0.92	0.34
OL4. Expressive Vocabulary	1.51	-1.47	0.88	0.52
OL5. Definitions (MC)	1.31	-1.80	0.92	0.44
OL6. Expressive Vocabulary	1.06	-1.14	0.79	0.48
OL7. Definitions (MC)	0.94	-2.30	0.94	0.36
OL8. Expressive Vocabulary	0.65	0.51	0.39	0.34
OL9. Definitions (MC)	0.65	-1.75	0.83	0.35
OL10. Expressive Vocabulary	0.65	1.67	0.18	0.26
OL11. Grammar (MC)	0.63	-1.87	0.84	0.34
OL12. Grammar (MC)	0.73	-1.90	0.86	0.36
OL13. Grammar (MC)	0.60	-1.05	0.71	0.35
OL14. Receptive Vocabulary (MC)	0.78	-2.98	0.96	0.26
OL15. Receptive Vocabulary (MC)	0.84	-2.67	0.95	0.29
OL16. Receptive Vocabulary (MC)	0.78	-2.61	0.94	0.30
OL17. Receptive Vocabulary (MC)	0.95	-2.16	0.93	0.37
OL18. Receptive Vocabulary (MC)	0.67	0.20	0.46	0.36
OL19. Word Knowledge	0.23	-3.82	0.81	0.13
OL20. Word Knowledge	0.79	-1.86	0.87	0.38
OL21. Word Knowledge	0.84	-1.81	0.87	0.40
OL22. Word Knowledge	0.76	-1.77	0.86	0.38
OL23. Word Knowledge	0.51	1.39	0.26	0.26

Note. MC = Multiple-choice item requiring only pointing response.

Table A1.9

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP3 Oral Language Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
OL1. Definitions (MC)	0.59	-3.93	0.97	0.29
OL2. Expressive Vocabulary	0.94	-1.72	0.88	0.43
OL3. Definition (MC)	1.32	-1.91	0.93	0.43
OL4. Expressive Vocabulary	1.23	-1.30	0.84	0.51
OL5. Definitions (MC)	1.58	-1.15	0.83	0.50
OL6. Expressive Vocabulary	1.73	-1.13	0.83	0.53
OL7. Definitions (MC)	0.96	-1.96	0.91	0.39
OL8. Expressive Vocabulary	1.01	-0.11	0.53	0.48
OL9. Definitions (MC)	0.75	-1.97	0.88	0.36
OL10. Expressive Vocabulary	0.66	1.52	0.20	0.28
OL11. Grammar (MC)	0.74	-2.65	0.94	0.30
OL12. Grammar (MC)	0.65	-2.82	0.93	0.27
OL13. Grammar (MC)	0.68	-1.82	0.85	0.36
OL14. Grammar (MC)	0.66	-1.42	0.78	0.36
OL15. Receptive Vocabulary (MC)	1.12	-2.72	0.98	0.29
OL16. Receptive Vocabulary (MC)	0.82	-2.13	0.91	0.37
OL17. Receptive Vocabulary (MC)	0.87	-2.14	0.92	0.35
OL18. Receptive Vocabulary (MC)	0.76	-1.95	0.88	0.35
OL19. Word Knowledge	0.21	-4.20	0.81	0.12
OL20. Word Knowledge	0.79	-1.95	0.88	0.38
OL21. Word Knowledge	0.91	-1.81	0.88	0.41
OL22. Word Knowledge	0.40	0.94	0.36	0.27
OL23. Word Knowledge	0.61	0.79	0.34	0.33

Note. MC = Multiple-choice item requiring only pointing response.

Table A1.10

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP1 Math Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
MA1a. Counting	0.59	-3.23	0.94	0.37
MA1b. Counting	1.73	-1.36	0.87	0.47
MA1c. Counting	1.82	0.04	0.49	0.62
MA1d. Counting	3.23	0.25	0.41	0.66
MA1e. Counting	14.43	0.30	0.39	0.70
MA1f. Counting	3.45	0.73	0.24	0.58
MA2. Counting Sets	0.62	-2.22	0.88	0.31
MA3. Number Identification	1.05	-1.56	0.87	0.43
MA4. Number Identification	1.12	-1.07	0.79	0.50
MA5. Counting Up from Number	0.64	-0.07	0.52	0.44
MA6. Counting	0.81	-0.78	0.69	0.49
MA7. Ordinality	0.49	-0.91	0.66	0.37
MA8. Counting	1.19	-0.08	0.53	0.57
MA9. Number Sense	0.15	1.06	0.44	0.12
MA10. Number Sense	0.63	0.79	0.34	0.41
MA11. Subtraction	0.44	1.31	0.29	0.31
MA12. Addition	0.63	0.68	0.36	0.46
MA13. Addition	0.52	1.07	0.31	0.38

Table A1.11

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP2 Math Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
MA1a. Counting	0.59	-4.26	0.98	0.28
MA1b. Counting	1.59	-2.06	0.95	0.39
MA1c. Counting	1.76	-0.55	0.69	0.63
MA1d. Counting	3.07	-0.30	0.62	0.68
MA1e. Counting	7.04	-0.25	0.61	0.71
MA1f. Counting	2.58	0.22	0.42	0.61
MA2. Number Identification	0.75	-3.54	0.98	0.21
MA3. Number Identification	1.24	-2.09	0.94	0.38
MA4. Number Identification	1.58	-1.65	0.91	0.48
MA5. Number Identification	1.46	-1.62	0.90	0.49
MA6. Counting	0.80	-0.78	0.69	0.53
MA7. Counting	0.70	-1.14	0.75	0.45
MA8. Counting	0.96	-0.85	0.72	0.52
MA9. Counting Up from Number	1.18	-0.04	0.51	0.56
MA10. Addition	0.55	0.15	0.47	0.41
MA11. Number Sense	0.66	0.31	0.43	0.43
MA12. Addition	0.70	0.90	0.30	0.41
MA13. Addition	0.49	0.77	0.36	0.34

Table A1.12

Item-Response Theory Parameters and Selected Classical Test Theory Indices for AP3 Math Measure

	IRT Parameters			
Item/Item Content	discrimination	difficulty	p-level	Item-total *r*
MA1a. Counting	0.59	-4.62	0.98	0.31
MA1b. Counting	2.54	-2.06	0.97	0.43
MA1c. Counting	1.66	-0.87	0.77	0.60
MA1d. Counting	3.14	-0.58	0.72	0.69
MA1e. Counting	7.57	-0.50	0.70	0.72
MA1f. Counting	2.41	-0.09	0.54	0.61
MA2. Counting	1.16	-2.64	0.97	0.33
MA3. Number Identification	1.32	-2.36	0.96	0.39
MA4. Number Identification	1.21	-1.95	0.93	0.46
MA5. Counting	1.02	-1.99	0.92	0.42
MA6. Number Identification	1.05	-1.45	0.85	0.51
MA7. Count Down from Number	0.63	-0.78	0.66	0.41
MA8. Counting	1.24	-0.94	0.77	0.58
MA9. Number Identification	0.94	-0.40	0.61	0.54
MA10. Number Sense	0.60	-0.45	0.59	0.41
MA11. Addition	0.66	0.08	0.48	0.43
MA12. Ordinality	0.58	0.04	0.49	0.38
MA13. Addition	0.70	0.20	0.45	0.42

Appendix 2

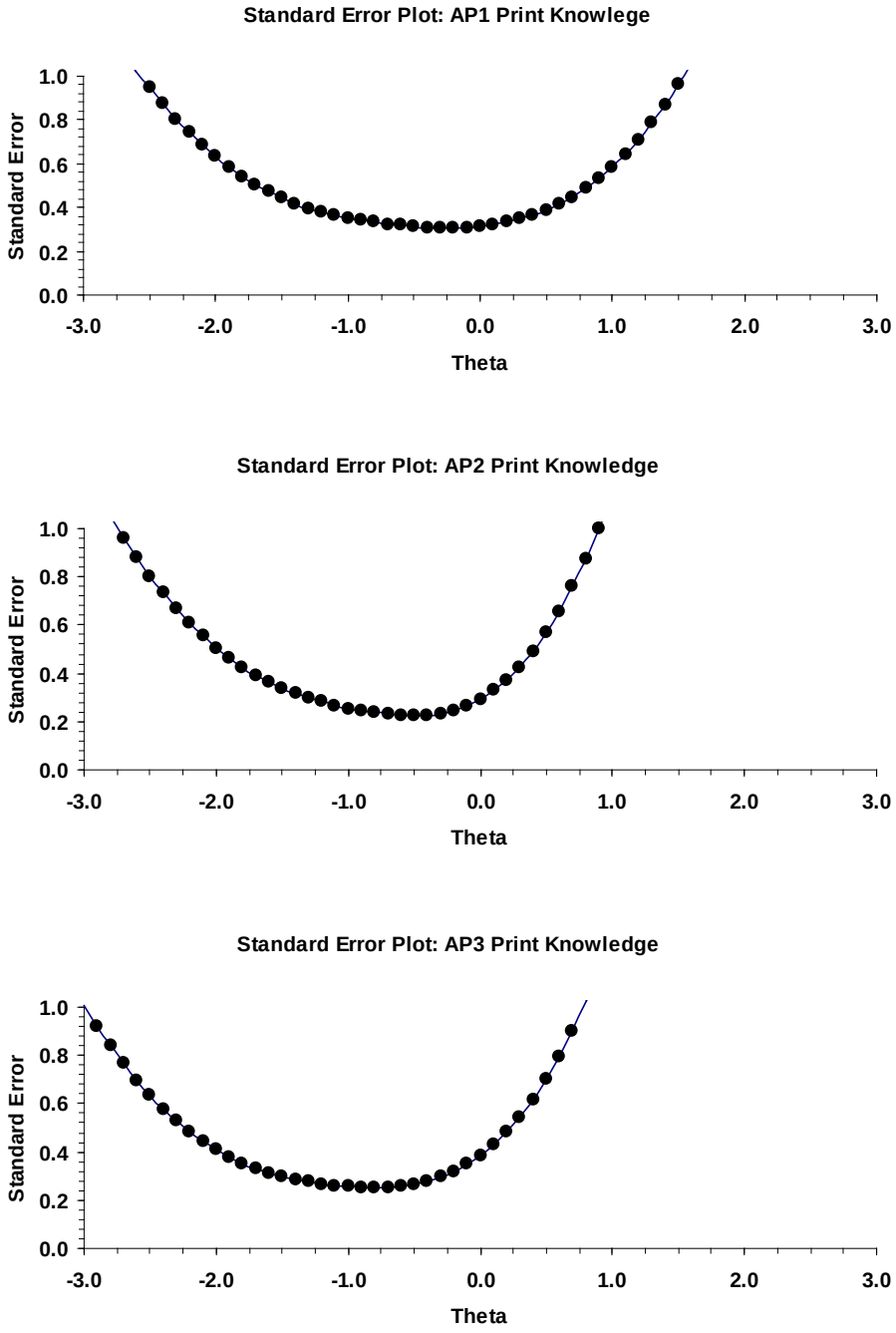


Figure A2.1. Standard error plots from 2PL IRT Models for Print Knowledge measures at AP1 (*N* = 2,042), AP2 (*N* = 2,029), and AP3 (*N* = 2,013).

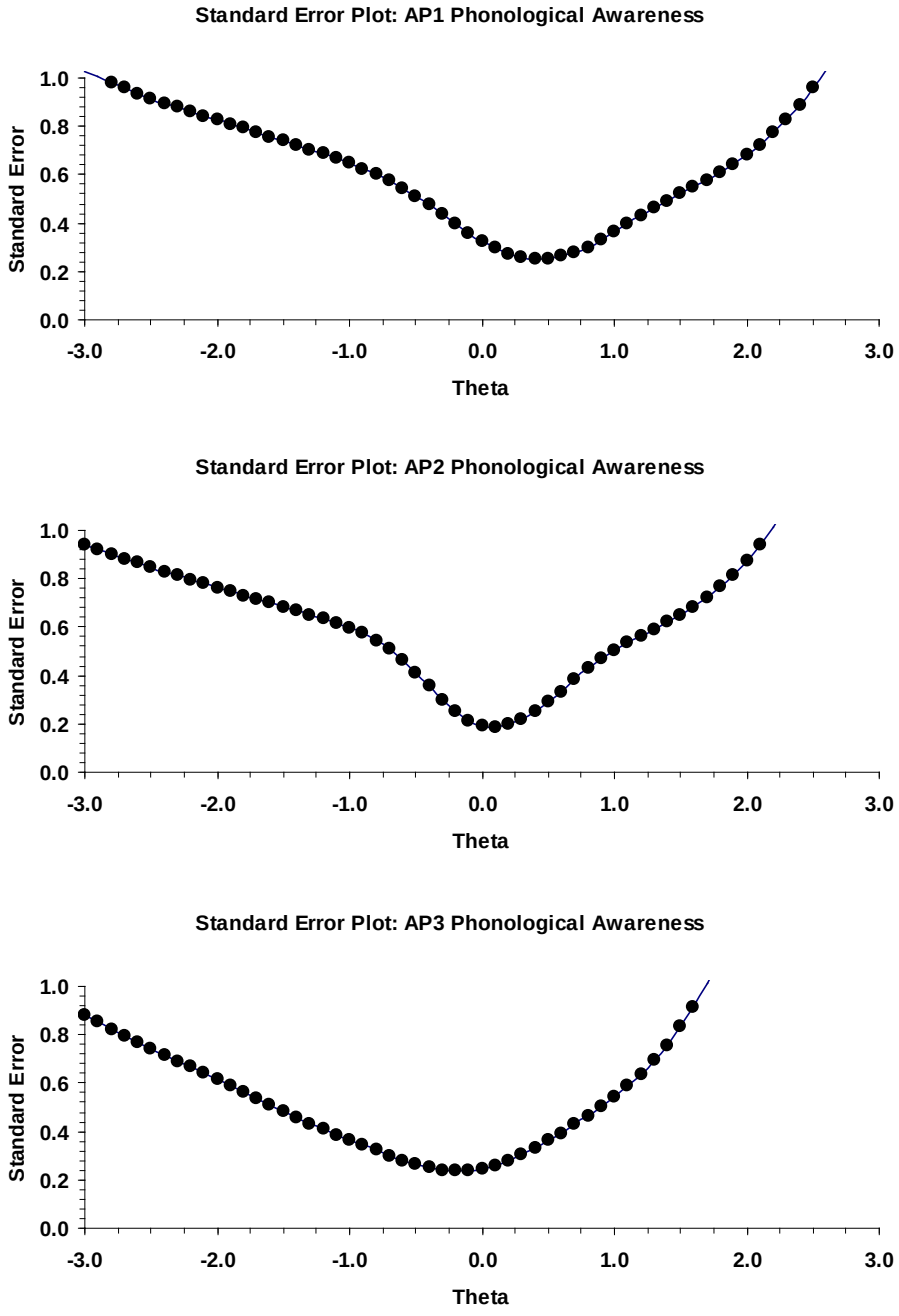


Figure A2.2. Standard error plots from 2PL IRT Models for Phonological Awareness measures at AP1 (*N* = 1,939), AP2 (*N* = 1,949), and AP3 (*N* = 1,786).

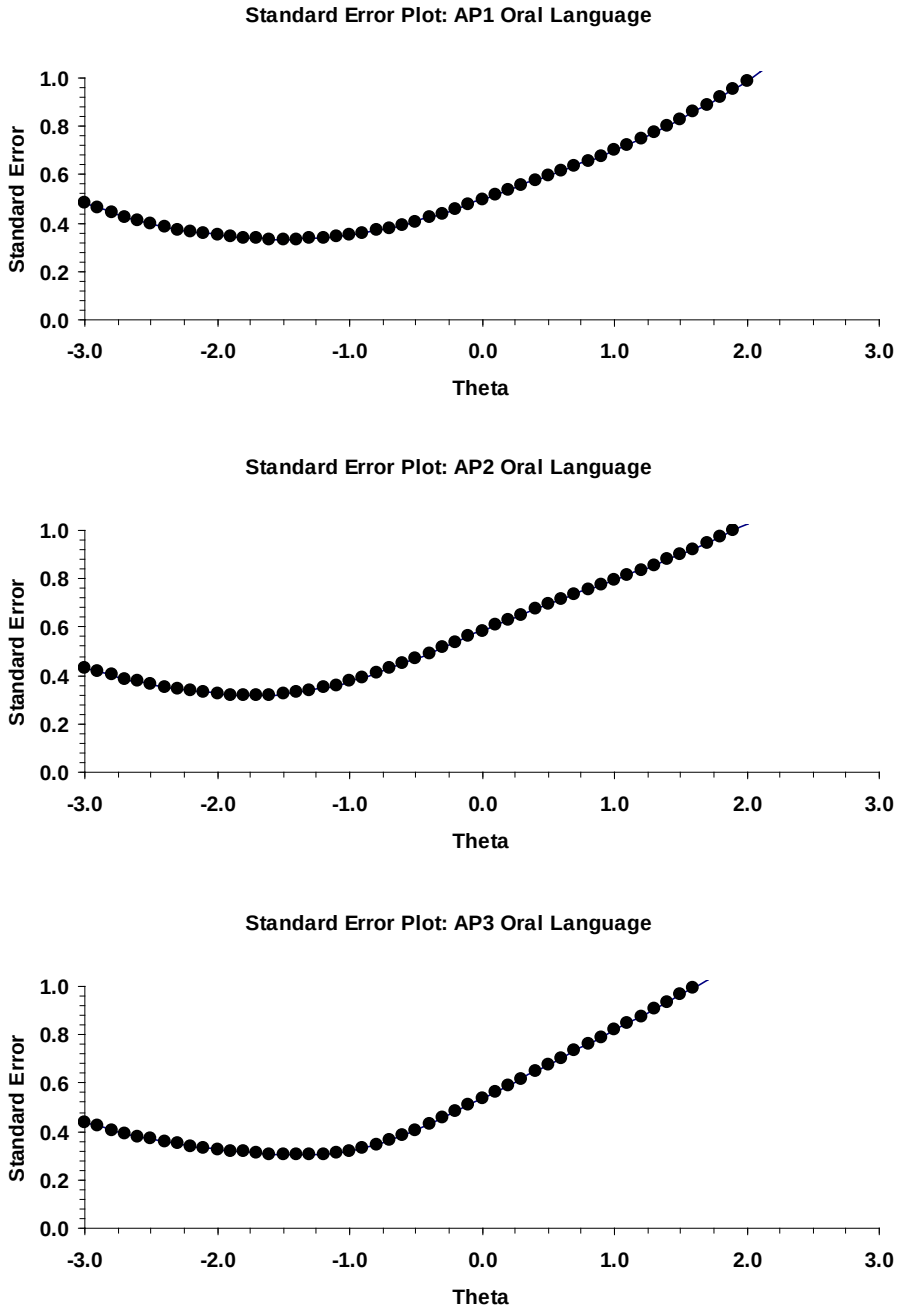


Figure A2.3. Standard error plots from 2PL IRT Models for Oral Language measures at AP1 (*N* = 596), AP2 (*N* = 1,714), and AP3 (*N* = 1,674).

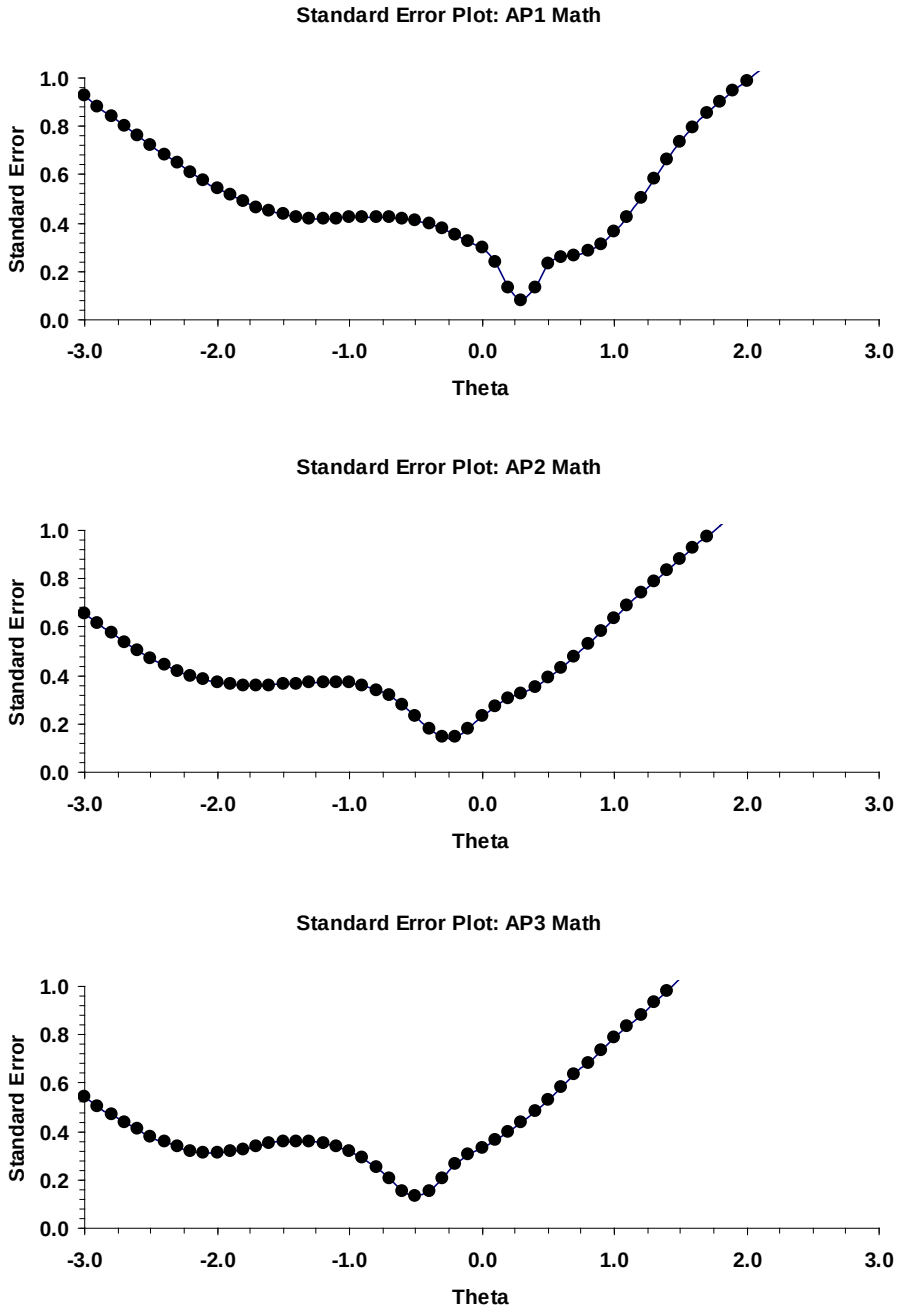


Figure A2.4. Standard error plots from 2PL IRT Models for Math measures at AP1 (*N* = 1,590), AP2 (*N* = 1,468), and AP3 (*N* = 1,436).