Hammill Institute on Disabilities

Assessment for Effective Intervention
2019, Vol. 44(4) 267–280
© Hammill Institute on Disabilities 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1534508418761232
aei.sagepub.com
SAGE

Internal Structure and Item Characteristics of the Phonological Awareness Literacy Screening in Spanish for Preschool

J. Patrick Meyer, PhD[1], Marcia A. Invernizzi, PhD[1], and Karen L. Ford, PhD[1]

Abstract

We used multidimensional item response theory to test the internal structure of the Phonological Awareness Literacy Screening in Spanish for Preschool and test for item parameter drift. The measure is aligned with the simple view of reading, which defines reading as consisting of two equally important dimensions: decoding and language comprehension. It involves 134 items grouped into nine different tasks. We administered a pilot version of the measure to 677 students in 2014 and a final version to 968 students in 2015. We did not find evidence that code-related skills and oral language are distinct but correlated dimensions among preschoolers. Rather, a single general dimension of early literacy with task-related specific dimensions fit best. A benefit of our modeling approach is that we can examine the internal structure while also evaluating the difficulty and discrimination of individual items. Results suggest a stable ordering of items within tasks and that some Spanish letters are learned more easily than others.

Keywords

early literacy, measurement theory, English as second language

An important component of effective preschool language and literacy programs is age-appropriate, research-based assessment. The National Association for the Education of Young Children (NAEYC) underscored the importance of providing effective assessment for students who are dual language learners (DLLs), stating that “young English-language learners have the right to be assessed for the same reasons and benefits as all children. Moreover, they have the right to be assessed with high-quality assessments and under assessment conditions responsive to their needs” (NAEYC, 2005, p. 2). To create assessment conditions that are responsive to the needs of DLL students, we must develop psychometrically sound assessments in children’s home languages that can be used along with assessments in English to provide a complete picture of children’s early language and literacy development (Peña & Halle, 2011). The current study examines the internal structure of the Phonological Awareness Literacy Screening in Spanish for Preschool (PALS español PreK). PALS español PreK is a measure of foundational early language and literacy skills that yields a composite score and individual task scores that can guide teachers in planning instruction to meet individual children’s literacy needs.

In this study, we used multidimensional item response theory (MIRT) to explore two research questions. The primary question was, “What is the internal structure of PALS español PreK?” Answering this question will provide validity evidence and help ensure accurate interpretation of test scores. Based on theories of early literacy development, we anticipated that either a bifactor or a two-tier model would fit the data best and represent a single general dimension of early literacy or the dominant constructs of oral language and code-based skills. Our secondary research question was, “What are the characteristics of individual items?” We wanted to know the ordering of items by difficulty and the extent to which items related to a dimension. Answering this question will help teachers plan instruction methods that optimize the use of time by spending more time on difficult skills and less time on those skills more easily achieved or by scaffolding these skills by difficulty.

Theoretical Framework

The theoretical framework for early literacy instruction and assessment has been dominated by the simple view of

[1] University of Virginia, Charlottesville, USA

Corresponding Author:
J. Patrick Meyer, PhD, Associate Professor, Curry School of Education, University of Virginia, P.O. Box 400265, 405 Emmet Street South, Charlottesville, VA 22904, USA.
Email: jpm4qs@virginia.edu

reading, which defines reading as consisting of two equally important dimensions: decoding and language comprehension. In terms of early literacy, decoding has been defined as all of the code-related criteria necessary to learn to read in an alphabetic language. These include concepts about print (i.e., directionality, parts of a book, conventions of writing, etc.), alphabet knowledge (i.e., letter names and letter sounds), phonological awareness (i.e., awareness of the sound structures of spoken language that can be mapped to print), and the alphabetic principle (i.e., the insight that speech can be segmented into smaller units of sound and represented by letters in a systematic way; Catts, Fey, Zhang, & Tomblin, 1999; Evans, Bell, Shaw, Moretti, & Page, 2006; Foulin, 2005; Foy & Mann, 2006; Lonigan, Burgess, & Anthony, 2000; National Early Literacy Panel, 2008; Storch & Whitehurst, 2002). Language comprehension has been variously defined as just those aspects of oral language necessary for the comprehension of concepts and vocabulary or, more broadly, as ideas expressed aurally through words, sentences, and larger discourse-level structures (Dickinson, McCabe, Anastasopoulos, Peisner-Feinberg, & Poe, 2003; Kendeou, van den Broek, White, & Lynch, 2009; Lonigan et al., 2000; NICHD Early Child Care Research Network, 2005; Storch & Whitehurst, 2002; van Kleeck, 1998; Whitehurst & Lonigan, 1998). The latter entails language competencies related to syntax, semantics, pragmatics, and narrativity.

For more than 20 years, researchers have conceptualized early literacy as progressing along these two dimensions, often referred to as oral language and code-based competencies, based on longitudinal studies of early literacy development (Kendeou et al., 2009; Storch & Whitehurst, 2002; Whitehurst & Lonigan, 1998). Constructs of oral language and code-related skills have been shown to be highly intercorrelated and reciprocal such that development along one dimension influences development of the other (Dickinson et al., 2003; Kendeou et al., 2009; Lonigan et al., 2000; NICHD Early Child Care Research Network, 2005). Nevertheless, both constructs have been shown to make their own contributions to literacy development over time, and both contribute in prediction studies independently as well as together (Kendeou et al., 2009; Lonigan et al., 2000).

By defining oral language broadly to include not only vocabulary knowledge but also discourse characteristics such as narrativity and language production, researchers have come to a more nuanced understanding of the relationship of these two constructs of early literacy. For example, studies using longitudinal structural equation modeling have shown that the relationship of oral language, more broadly defined, to the code-related construct changes over the course of development, such that oral language exerts a powerful and direct effect on the development of code-related skills in the preschool years but may have a more indirect effect as children are actually learning to decode text. Once the reading code is cracked, however, oral language once again has a direct effect as the focus shifts to the comprehension of text (Dickinson, Golinkoff, Hirsh-Pasek, 2010; Kendeou et al., 2009; NICHD Early Child Care Research Network, 2005; Storch & Whitehurst, 2002). While the simple view of reading still holds up in light of these developmental shifts, we now have a more refined understanding of the ebb and flow of oral language and code-related dimensions across time and development. More specifically, it appears that in the preschool years oral language provides a direct influence on code-related skills such as phonological awareness, which, with print concepts, contributes to developing an understanding of the alphabetic principle. Understanding the alphabetic principle is an important milestone in early literacy development that is often thought to signal the transition to conventional reading (Liberman, Shankweiler, & Liberman, 1989).

As we developed the tasks and items for PALS español PreK, we adopted a theoretical framework, based on the simple view of reading that posits two dimensions to early literacy development: oral language and code-related skills and concepts. We expanded our definition of oral language to include narrativity and language production, both of which have been shown to be particularly useful in predicting potentially later-emerging comprehension-based reading problems with discourse-level aspects of text (Kendeou et al., 2009). We expanded our definition of code-related skills and concepts to include name writing because the letters children are most likely to learn first are the letters in their first name, particularly the initial letter, and they use their knowledge of those letter names to learn and remember letter sounds (Huang, Tortorelli, & Invernizzi, 2014; Treiman & Broderick, 1998; Treiman, Tincoff, Rodriguez, Mouzaki, & Francis, 1998). Below we describe the tasks included in PALS español PreK.

PALS español PreK Tasks

PALS español PreK includes nine tasks that measure aspects of children's oral language and code-related skills. The measure as a whole has a reliability of 0.92, but there are differences in the reliability of each individual task (see Meyer, Ford, & Invernizzi, 2017).

Oral Language

The PALS español PreK *Language and Listening Comprehension* (LL) task measures listening comprehension, narrative skill, and language production. Children listen to a story read orally and then retell the story while placing picture cards representing story events in the correct order. Scores are based on how many story details children include in their retelling and the extent to which they relate the events of the story in the right order. The teacher also rates children's oral language production during the retelling on a scale ranging from no response to complete sentences with story embellishments.

Code-Related Skills

PALS español PreK measures code-related skills in the domains of phonological awareness, alphabet knowledge, writing, and concepts about print. Although phonological awareness refers to a child's ability to attend to sound units in spoken words, it is considered a code-related skill because of the metalinguistic nature of reflecting on sound structures within spoken language that can be matched to corresponding structures in print.

Phonological awareness. Children develop phonological sensitivity over time, first attending to larger sound units such as syllables and rhyme, and later to smaller sound units such as beginning phonemes. This progression appears to be universal, varying only in the rate with which children move through the sequence (Anthony & Lonigan, 2004). Spanish-speaking children, for example, develop syllable awareness much sooner than children in other languages where syllables are less salient (Anthony & Francis, 2005). Clapping out syllables in words and identifying words that rhyme or begin with the same sound are activities that teachers typically use to introduce children to the sound elements in words. In the PALS español PreK *Syllable Clapping* (SC) task, children clap to syllables as they repeat words aloud (e.g., mo-ne-da, cho-co-la-te). In the *Rhyme Awareness* (RA) task, they match pictures of rhyming words (e.g., gato/pato). In the *Beginning Sound Awareness* (BS) task, children repeat words aloud, then isolate the beginning sound (e.g., mesa, /m/).

Alphabet knowledge. The alphabet knowledge tasks test children's knowledge of the 29 letters and digraphs in the Spanish alphabet. There are three subtasks: *Uppercase Alphabet and Digraph Recognition* (UC), *Lowercase Alphabet and Digraph Recognition* (LC), and *Letter Sounds* (LS), and each subtask represents an increasing level of difficulty for preschoolers. Children begin with the easiest subtask, UC. They are administered each subsequent task (LC then LS) only if they demonstrate at least minimal success on the previous one.

Name writing. Name writing is often a child's first experience with printing letters, and research has demonstrated that the development of name writing correlates with the development of other early literacy skills such as alphabet knowledge and beginning sound awareness (Bloodgood, 1999; National Early Literacy Panel, 2008; Welsch, Sullivan, & Justice, 2003). Treiman and Broderick (1998) argued that children's interest in analyzing and reproducing the printed form of their own name provides the key to learning letter names, letter sounds, and ultimately to analyzing letters and letter sounds in other words. Clay (1979) and Chomsky (1971) in English and Ferreiro and Teberosky (1982) in Spanish have demonstrated that children's writing develops along a predictable continuum, starting with scribbles and concluding with traditional spelling using decipherable letters. On the PALS español PreK *Name Writing* (NW) task, teachers use just such a continuum to evaluate and describe children's name writing attempts. The child is asked to draw a self-portrait and write his or her name. The task is scored based on an eight-level rubric that allows teachers to analyze the sophistication of the name writing attempt, ranging from a scribble that represents both the name and the picture to a name that is written correctly and is separate from the picture.

Concepts about print. Part of children's preparation for learning to read is developing an understanding of what are referred to as concepts about print (e.g., knowing that it is print that conveys the message in text, understanding that text is organized from left to right and top to bottom, knowing the difference between numbers, letters, and words, etc.; National Early Literacy Panel, 2008; Snow, Burns, & Griffin, 1998). The PALS español PreK *Print and Word Awareness* (PW) task assesses concepts about print through a shared book reading activity. The teacher reads a book individually with each child and asks him or her to point to text components such as the title, individual letters, and words. The teacher also explores the child's understanding of book orientation and directionality of print. The score for the task is based on the child's ability to respond to a series of print-related directives (e.g., identify the letter T, point to the words in the title, demonstrate left to right directionality, etc.).

While these tasks represent the foundational components of early literacy development, it is important to acknowledge that these components may be "necessary but insufficient" by themselves for children to progress in literacy development. Developmental perspectives describe how these understandings merge across time and become integrated as they are put to use toward the authentic purposes of reading and writing (Clay, 1977; Ehri, 2005; Morris, Bloodgood, Lomax, & Perney, 2003). For example, oral language (e.g., vocabulary, sentence complexity, and narrativity) is associated with the development of phonological awareness in the early stages of learning to read, becomes less important in the early stages of learning to decode text, but becomes a powerful predictor of reading comprehension once children become independent readers (Kendeou et al., 2009; Storch & Whitehurst, 2002). There is also a reciprocal relationship between phonological awareness and decoding in that phonological awareness is essential to the ability to decode text, and learning to decode an alphabetic writing system increases phonological awareness (Wagner, Torgesen, & Rashotte, 1994). Once children begin to receive formal literacy instruction, however, print-related skills gradually become better predictors of later reading achievement than phonological awareness (Ford, Cabell, Konold, Invernizzi, & Gartland, 2013; Hammill, 2004; Morris, Bloodgood, & Perney, 2003; Warley, Landrum, Invernizzi, & Justice, 2005). Because these early literacy skills tend to be highly intercorrelated, and because the relationships between skills tend to shift across time, a central issue in the assessment of early literacy development in the preschool years pertains to a more nuanced understanding of the general dimension of early literacy development in relation to its component skills. In other words, this raises questions as to whether there are distinct, but highly correlated components or whether these components are parts of a single dimension. Furthermore, the relationship among these parts may be different at different points in time. This study explores these issues by examining the internal structure of PALS español PreK administered in the fall and in the spring.

Factor Structure of PALS Assessments

PALS español PreK is part of a suite of assessments (PALS PreK, K, and 1-3 in English; PALS español K and 1-3 in Spanish) that measure children's progress toward developing essential early literacy skills necessary for becoming successful readers. They are all designed according to the same theoretical model presented above. Townsend and Konold (2010) explored the factor structure of PALS PreK in English and found that a model with two correlated factors fit the data better than a one-factor model. The correlated factors consisted of Print/Phonological Awareness and Alphabet Knowledge. However, Townsend and Konold's final model did not involve simple structure. The LS task loaded on both factors. Yaden, Marx, Cimetta, Alkhadim, and Cutshaw (2017) replicated the work of Townsend and Konold using a short, draft version of PALS español PreK. They too found that the same two correlated factors fit better than a single-factor model. In two studies of PALS for kindergarteners, researchers found that a higher-order model fit the data better than a one-factor or two-correlated-factor model both in English (Huang & Konold, 2014) and in Spanish (Huang, Ford, Invernizzi, & Fan, 2013). Finally, Huang (2014) championed a bifactor model over a correlated factor model; he conceptualized the overall dimension as including both code-related and phonological components.

Prior factor analytic work with PALS employed item parcels of assessment items, a practice of computing the sum or mean of multiple items and using this parcel score in factor analysis instead of individual item scores. Item parceling is done to create indicators that are normally distributed and safe to analyze by factor analysis. However, Bandalos (2008) notes that item parceling may not even be necessary with recent advances in estimation techniques that do not require the assumption of multivariate normality. Limited information methods such as the WLSMV estimator in Mplus (Muthén & Muthén, 2010), and full-information item factor analysis methods such as marginal maximum likelihood (Bock, Gibbons, & Muraki, 1988) and the Metropolis-Hastings Robbins-Monro (MHRM) algorithm (Cai, 2010) do not require the assumption that manifest indicators be multivariate normal. These estimation methods are ideal for items that are binary (e.g., right/wrong) or ordinal in nature and are routinely part of item response theory (IRT).

PALS assessments involve multiple tasks, and each task may be considered to be a testlet—a group of items developed as a single unit that is intended to be administered together (Wainer, Bradlow, & Wang, 2007, p. 53). Testlets are a defensible way to create item parcels (Bandalos & Finney, 2001), and it was the approach adopted in prior research on the factor structure of PALS assessments. Parceling has limitations and disadvantages even when it is justified. One limitation is that the influence of a testlet cannot be tested or quantified because items that form the testlet are combined into a single parcel score, and the model does not explicitly account for relationships among items within a testlet. Another limitation is that the factor structure or dimensionality of the measure cannot be evaluated at the item level. Finally, parceling only results in estimates for two or more items combined. It does not allow for the estimation of individual item parameters, which is necessary for understanding item difficulty and the relationship between each item and the latent trait. To overcome these limitations, we chose to examine the internal structure with MIRT methods.

MIRT

In a model with bifactor latent structure, there is a single general dimension, two or more orthogonal specific dimensions, and each item loads on the general dimension but only one specific dimension (Gibbons & Hedeker, 1992). The general dimension is the main dimension of interest, and the specific dimensions represent groups of items that have something in common after accounting for the primary dimension (i.e., testlets). Confirmatory factor analysis provides a way to fit a model with bifactor structure to the data when item responses are continuous, but MIRT is a more suitable choice when item responses are binary or ordered categories. Most of the items on PALS español are binary, but there are four items that are polytomously scored. For binary items, we used the multidimensional extension of the two-parameter logistic item response model as given by,

$$P\left(y_{ij}=1\mid\theta_i\right)=\frac{1}{1+\exp\left[-\left(\sum_{v=1}^{m}a_{jv}\theta_{jv}+d_j\right)\right]}.$$

In this equation, each of the $i=1,\ldots,n$ examinees have $v=1,\ldots,m$ proficiency parameters, θ_v, where m is the number of dimensions. The parameters a_{jv} and d_j are slope and intercept parameters, respectively. The slope parameter is analogous to a factor loading in factor analysis. There is a slope parameter for every dimension, but only one intercept per item j. The intercept parameter is not the same as the difficulty parameter in unidimensional item response models. However, Reckase (2009) defines the multidimensional difficulty as

$$\text{MDIFF}_j=\frac{-d_j}{\sqrt{\sum_{v=1}^{m}a_{jv}^2}},$$

which has the same interpretation as the difficulty parameter in unidimensional IRT. For polytomous items, we used the multidimensional generalized partial credit model.

A model with bifactor structure has multiple dimensions, but only two dimensions contribute to an item response. That is, the summation in Equation 1 reduces to $\sum_{v=1}^{m}a_{jv}\theta_{jv}=a_{jGeneral}\theta_{iGeneral}+a_{jSpecific}\theta_{iSpecific}$, where one slope and one proficiency parameter pertains to the general dimension and one slope and one proficiency parameter applies to a specific dimension. All item slopes are freely estimated in the bifactor MIRT model, but the variance of each dimension is constrained to one. The top panel in Figure 1 illustrates a model with bifactor structure. The circles in the diagram indicate latent variables, squares indicate observed variables, and arrows indicate item slopes. Notice that only two arrows point to an item, one from the general dimension and one from a specific dimension. This feature is a defining characteristic of bifactor structure.

The testlet model is a special case of the bifactor model in which an item's slope for the specific dimension is constrained to equal the item's slope for the general dimension (Cai, 2010). In terms of Equation 1, the summation becomes $\sum_{v=1}^{m}a_{jv}\theta_{iv}=a_j(\theta_{iGeneral}+\theta_{iSpecific})$ in the testlet model. This constraint allows the specific dimension variances to be freely estimated, although the general dimension variance remains fixed to unity. Specific dimension variance indicates the magnitude of the testlet effect for its group of items. Larger values indicate a stronger testlet effect. The middle panel in Figure 1 illustrates the path diagram of the testlet model; two arrows point to an item but they have the same slope, and the circles pointing to the specific dimensions that indicate free estimation of the variance.

Cai (2010) generalized the bifactor model to a two-tier model, where there may be multiple correlated general dimensions and multiple specific dimensions for each general dimension. An advantage of a two-tier model over multiple independent bifactor models is improved estimation because the general dimensions "borrow strength" from each other. Another advantage is that additional dimensions can be added to the model without compromising computational efficiency. The bottom panel in Figure 1 illustrates a model with a two-tier latent structure that has two correlated general dimensions. The present study involved unidimensional, bifactor, testlet, and two-tier models.

Method

Participants

Data were collected in the fall and spring of 2014 and 2015. In 2014, data collection involved an assessment with an expanded number of test items (about 50% more than were needed for the final test form). We selected the best of these items for the final test form that we used in data collection during 2015. Item selection considered item fit statistics and test content (details not shown herein). The general structure of the assessment remained the same in both years.

A total of 344 students from 16 different schools located in seven school districts completed the assessment in Spring 2014. They had a mean age of 56.67 months with a standard deviation of 8.04 months. Female students comprised half of the sample. Forty-seven students did not report an ethnicity. A large majority were Hispanic (93%), but White (6.4%) and Asian students (0.3%) were also included. Spanish was the primary language spoken in the home for 91% of the 308 students whose parents reported a language. Some parents reported that English (6.8%), another language (1%), or two or more languages (1%) were the primary language in the home.

In Fall 2014, 333 students from 19 schools and 9 school districts completed the assessment. They were 52.18 months of age on average with a standard deviation of 5.86 months. Male students comprised a slightly large portion of the sample (54%). Of the 292 students reporting an ethnicity, 90% were Hispanic, 8.6% were White, 0.7% were Black, and 0.3% were Asian. Spanish was the primary language spoken in the home for 90% of the sample. English (9.3%), another language (0.3%), or bilingual (0.3%) was also reported as the primary language in the home.

The Spring 2015 sample included 519 examinees. Students were from 22 schools located in 13 different school districts. The mean age was 58.9 months with a standard deviation of 8 months. A slight majority were male (52.5%). Hispanic students represented the largest ethnicity (97.6%), while White/non-Hispanic (1.6%) and Black/non-Hispanic (0.8%) students were also included. Spanish was

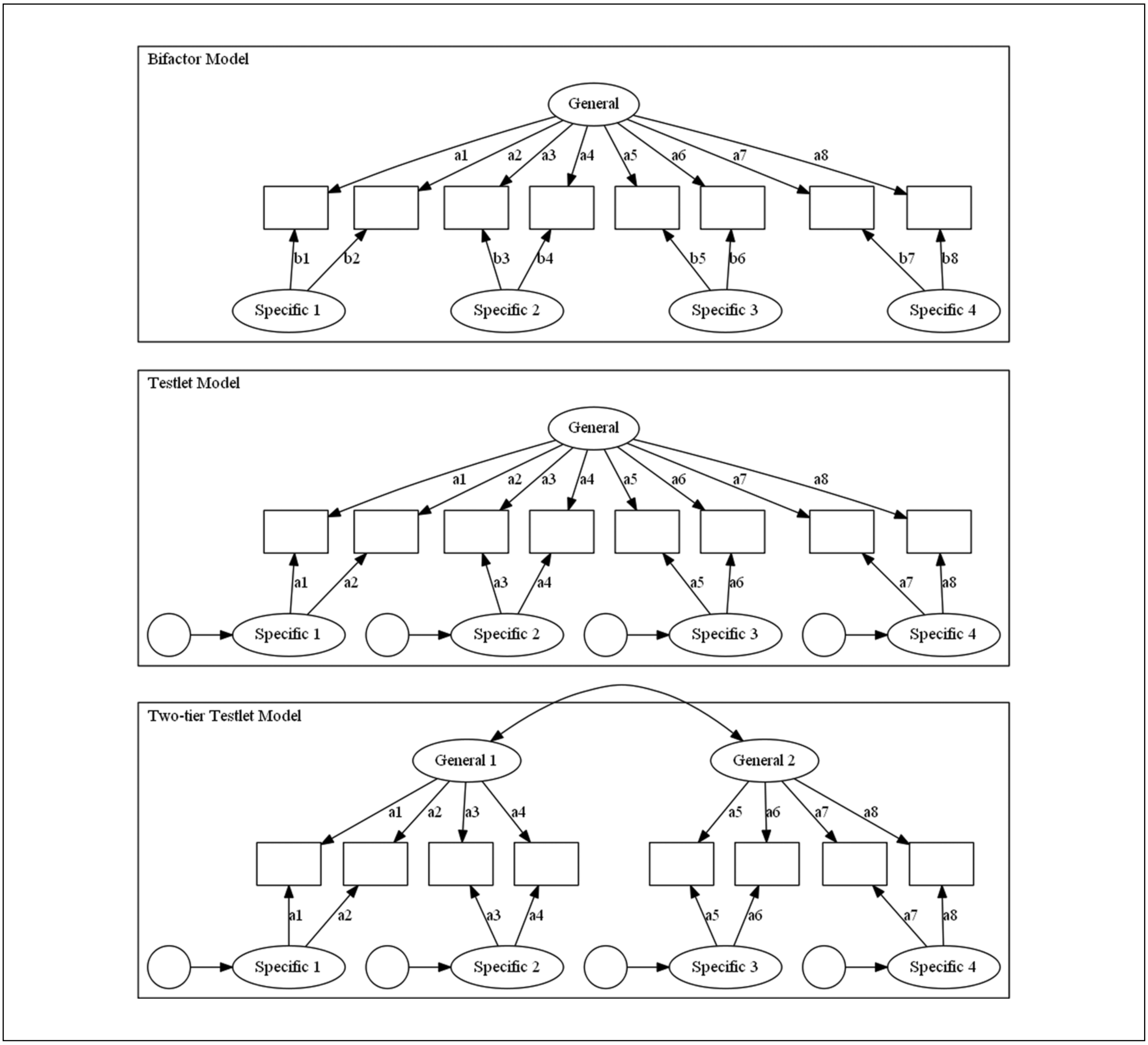

Figure 1. Path diagrams for bifactor, testlet, and two-tier models.

the primary language in the home for 96% of the sample, but some students spoke English (3%) or English and Spanish (0.7%) in the home.

Finally, the Fall 2015 sample consisted of 449 examinees from 23 different schools located in 11 school districts. Examinees had an average age of 52.4 months with a standard deviation of 7 months. Fifty-one percent were female. Of the 295 students who reported their ethnicity, 98% were Hispanic, 1% were Black/non-Hispanic, less than 1% were White /non-Hispanic, and only one student was Asian.

The number of participating schools and school districts varied with each administration. In Fall 2014, participants were from 13 different schools (4 private, 9 public) located in four separate districts across two states. Participants were from 14 schools (5 private and 9 public) located in three districts and two states in Spring 2014. The numbers increased in 2015. In the fall, examinees were from 19 different schools (4 private, 15 public) within seven districts and four states. Finally, Spring 2016 participants were from 17 schools (3 private, 14 public) located in nine districts from five different states.

Procedure

In 2014, trained researchers visited each school and administered the exam to students. A portion of students were tested by two raters to evaluate interrater reliability (see Meyer et al., 2017). In 2015, trained teachers administered most of the assessments but some were given by trained researchers. Data from each administration were manually entered into a spreadsheet and double-checked for accuracy. The 2015 data collection involved operational procedures where the LC and LS tasks are only administered to high-scoring students. As a result, 88% to 95% of examinees did not complete these two tasks. Given this large amount of missing data, we omitted the LC and LS tasks from the 2015 analysis.

Our sample size was too small to randomly split the data in half and conduct an exploratory analysis on one half and a confirmatory analysis on the other. Therefore, we considered theory, the design of PALS español PreK, and research on the factor structure of other PALS measures to conceive of various alternatives for the internal structure of the test. The most basic model (Model 1) was a unidimensional model where all items contributed to a single dimension.

A second type of model (Model 2) was a task-based model, where each item loaded on the general dimension (i.e., early literacy) and only one task-specific dimension. The only exception was the Name Writing task, which is only a single item. It loaded on the general dimension only; there was no specific dimension for Name Writing. We fit both a task-based bifactor model (Model 2a; see top panel of Figure 1) and a task-based testlet model (Model 2b; see middle panel of Figure 1).

The third type of model (Model 3; see bottom panel of Figure 1) was a two-tier model that represented the simple view of early literacy. It had a Code-related general dimension that was allowed to be correlated with an Oral Language general dimension. Each task represented a specific dimension that loaded on only one general dimension. The exception was Name Writing, which was a single item that loaded on the Code-related general dimension only. We ran Model 3 as both a bifactor model (Model 3a) and a testlet model (Model 3b).

Finally, the most complex model (Model 4) was a two-tier model with three correlated general dimensions: Phonological Awareness, Alphabet Knowledge, and Oral Language. Tasks also constituted the specific dimensions and items for each task loaded on only one specific dimension and only one general dimension. Because of the complexity of Model 4, we only ran it as a testlet model.

For each type of model, we used multigroup models, which allowed us to combine data from the same year into a single analysis. In these models, groups are nonequivalent and the group means are estimated as part of the model. The 2014 data were analyzed together because they involved a preliminary version of the assessment that included try-out items. Similarly, the 2015 data were analyzed together and separate from the 2014 data because they used a refined, operational version of the assessment. Constraints added to the multigroup models forced item parameters estimates from fall to be equal to item parameter estimates for spring. The only parameter allowed to differ from fall to spring was the latent mean of the general dimension(s).

These constraints assume there is no fall-to-spring item parameter drift (i.e., differential item functioning [DIF] across time points). To test this assumption, we conducted a DIF sweep procedure based on the Wald test using a significance level of .01 because this procedure is know to have an inflated type I error rate (Woods, Cai, & Wang, 2013). To compute a raw estimate of the magnitude of drift, we re-estimated the testlet model but freed the equality constraints for items with significant level of drift and freely estimated item parameters for fall and spring separately. We computed the fall to spring difference in estimates to quantify the amount of drift.

We fit each model to the data and estimated parameters with the MHRM Algorithm in flexMIRT 3.0 (see Houts & Cai, 2015). We identified the best-fitting model as the one with the lowest Bayesian information criterion (BIC) statistic. We focused on relative model fit because absolute model fit is certain to fail given the large number of parameters in a MIRT model (Maydeu-Olivares, 2015). In addition, there is currently no research to support the use of fit statistics from the factor analysis literature when using MHRM algorithm in MIRT.

Results

The task-based testlet model (Model 2b) fit better than other models in both the 2014 and 2015 data (see Table 1). The two-tier testlet model (Model 3b) was the second best-fitting model. All statistics and follow-up analyses reported below are based on the testlet model as it had the best fit both years. Latent means from the best-fitting model show that student scores on the general dimension are 1.02 to 1.76 logits higher in the spring than they are in the fall. These differences indicate that the odds of correctly answering an item are 2.8 to 5.8 greater in the spring than in the fall.

According to standard deviation estimates for the specific dimensions shown in the last column of Table 2, LL consistently had the largest testlet effects. In 2014, the phonological awareness tasks (e.g., SC, RA, and BS) also had strong testlet effects, but they were less prominent in 2015 when the alphabet knowledge subtask (UC) had the strongest testlet effects. By comparison, PW had the smallest testlet effects in both years. Taken together, there is some variability in the testlet effect by task, and the size of the effects suggests that tasks contribute a notable portion of variance.

Item Parameter Drift

According to the DIF sweep procedure, 20 out of 154 items (13%) in 2014 had statistically significant levels of item

Table 1. Model Fit Statistics.

Year	Name	Description	BIC (95% CI)
2014	Model 1	Unidimensional	[80,110.19, 80,113.33]
	Model 2a	Bifactor, one general dimension	[73,278.91, 73,306.05]
	Model 2b	**Testlet,** one **general dimension**	**[66,112.06, 66,127.62]**
	Model 3a	Two-tier bifactor, two general dimensions	[73,476.39, 73,505.62]
	Model 3b	Two-tier testlet, two general dimensions	[68,661.37, 68,680.92]
	Model 4	Two-tier testlet, three general dimensions[a]	[69,677.50, 69,702.98]
2015	Model 1	Unidimensional	[67,666.20, 67,669.98]
	Model 2a	Bifactor, one general dimension	[62,195.51, 62,212.85]
	Model 2b	**Testlet,** one **general dimension**	**[57,380.95, 57,392.93]**
	Model 3a	Two-tier bifactor, two general dimensions	[62,344.89, 62,362.82]
	Model 3b	Two-tier testlet, two general dimensions[a]	[60,877.90, 60,894.93]
	Model 4	Two-tier testlet, three general dimensions[a]	[61,177.14, 61,194.04]

Note. Bold font indicates the best fitting model. BIC = Bayesian information criterion; CI = confidence interval.

[a] The convergence criterion was satisfied, but the model did not converge on a maximum.

Table 2. Descriptive Statistics for Item Parameter Estimates and Testlet Effects.

			Slope			MDIFF			Testlet
Year	Task	No. of Items	Minimum	*M*	Maximum	Minimum	*M*	Maximum	*SD*
2014	NW	1	0.52	0.52	0.52	—	—	—	—
	UC	29	0.73	1.76	2.62	−0.37	1.41	4.30	1.07
	LC	29	1.08	1.90	2.76	−0.14	1.42	3.31	0.97
	LS	25	0.90	1.73	2.86	−0.40	1.09	3.53	1.05
	BS	15	0.82	1.04	1.19	−0.77	−0.30	0.17	1.90
	LL	10	0.08	0.73	1.10	−2.09	−0.37	1.83	2.69
	PW	15	0.37	0.80	1.32	−2.52	−0.33	2.55	0.81
	RA	15	0.28	0.52	0.88	−1.85	−0.62	0.82	1.57
	SC	15	0.50	0.79	1.15	−2.85	−1.40	0.25	1.96
2015	NW	1	1.00	1.00	1.00	—	—	—	—
	UC	29	1.10	1.94	2.51	−0.06	1.54	3.05	1.47
	BS	10	1.32	1.62	1.75	−0.12	0.04	0.37	1.21
	LL	8	0.13	1.32	2.03	−1.94	−0.70	0.00	1.95
	PW	12	0.42	1.01	1.34	−1.46	−0.01	1.56	0.62
	RA	10	0.50	0.72	0.88	−1.09	−0.25	0.53	1.40
	SC	10	0.80	1.18	1.58	−2.08	−0.99	0.20	1.15

Note. NW = name writing; UC = uppercase alphabet and digraph recognition; LC = lowercase alphabet and digraph recognition; LS = letter sounds; BS = beginning sound awareness; LL = language and listening comprehension; PW = print and word awareness; RA = rhyme awareness; SC = syllable clapping.

parameter drift in 2014. The drift was in the intercept parameter for 18 items and in the slope parameter for two items. Six of the drifting items belonged to the BS task, five belonged to the LL task, and another five belonged to the LS. Three other tasks (NW, UC, LC, and RA) each had one drifting item. The largest amount of drift in the MDIFF statistic occurred for a LL item (−2.11). For the remaining 17 items with drift in the intercept, the amount of drift ranged from −0.53 to 0.75 with an average of 0.004. The two items with drift in the slope that ranged from −0.91 to 0.60 with an average of −0.16.

In 2015, 11 out of 80 (14%) items showed drift. Only two items, NW and a LL item, showed drift in both administrations. The remaining items with drift were unique to each year. Two items showed significant amount of drift in the slope, while the remaining nine items showed significant drift in the intercept. A PW item showed the largest amount of drift in MDIFF (1.25). For the remaining items, MDIFF drift ranged from −0.68 to 0.49 with an average of 0.00. For the two items with drift in the slope, the difference in slopes ranged from 0.32 to 0.78 with an average of 0.54.

Table 3. Ordering of Uppercase Letters According to Spring Multidimensional Difficulty.

2014 Uppercase Letters			2015 Uppercase Letters		
Letter	MDIFF	Slope	Letter	MDIFF	Slope
O	−0.37	1.31	O	−0.06	1.10
M	0.64	1.82	A	0.86	1.97
F	0.73	1.97	M	1.04	1.85
A	0.77	1.50	B	1.12	2.12
B	0.77	2.13	E	1.16	1.55
N	0.78	2.29	N	1.21	2.51
L	0.78	2.62	S	1.21	2.20
S	0.87	1.93	F	1.24	2.29
P	0.94	2.02	D	1.27	2.48
T	0.94	2.33	L	1.28	2.51
E	0.95	1.52	X	1.34	1.75
D	1.04	2.47	C	1.36	2.13
C	1.04	2.23	T	1.47	2.06
H	1.20	1.60	U	1.48	1.87
R	1.21	1.86	P	1.49	1.81
V	1.26	2.45	H	1.49	2.11
X	1.39	1.28	R	1.54	2.11
U	1.40	1.58	V	1.60	2.19
K	1.57	1.78	I	1.70	1.47
Z	1.61	1.81	J	1.77	1.74
I	1.78	1.24	K	1.77	2.00
Q	1.80	1.44	Z	1.78	2.06
W	1.88	1.69	Y	1.83	2.07
Y	1.90	1.89	W	1.84	1.72
Ñ	1.95	1.56	Q	1.91	1.88
J	2.08	1.29	G	2.04	2.15
G	2.29	1.55	Ñ	2.21	1.64
LL	3.27	1.03	Ch	2.74	1.48
Ch	4.30	0.73	LL	3.05	1.34

Note. Ordering by MDIFF is the same as ordering by a response probability of 0.5.

Although about 14% of the items showed statistically significant levels of drift in both years, the magnitude of drift may not be practically significant. The average amount of drift in MDIFF is near zero, which may lead to DIF cancelation (see Wyse, 2013). Moreover, Wells, Subkoviak, and Serlin (2002) showed that a lack of invariance (i.e., DIF or drift) had little effect on estimation of examinee proficiency. To determine whether drift had any practical effect, we estimated person ability using the testlet model with equality constraints and again using the testlet model that allowed for time-variant item parameters. The correlation of general dimension ability estimates was 1.0 and the correlations among specific dimension estimates were all above 0.97 in 2014. We observed similar results for 2015. The correlation among estimated abilities for the general dimension was 1.0 and the specific dimension correlations we all above 0.98. Given the negligible impact of item parameter drift on estimated abilities, we selected the testlet model with equality constraints (Model 2b) as our best and most parsimonious model.

Item Discrimination and Difficulty

Item discrimination provides information about the relationship between an item and the general dimension. NW and RA had low slope parameters on average in 2014 and 2015 (see Table 2). Tasks with the largest average slope in 2014 were LC, UC, and LS. UC items also had large slopes, on average, in 2015, but the other two tasks with large slopes in 2015 were BS and LL.

The range of item slope estimates was largest for LS, UC, and LC in 2014, but in 2015 the largest range of slope estimates occurred for items belonging to LL, UC, and LS (see Table 2). In both years, RA has the lowest range of slope parameter estimates. The range of discrimination values within each task was larger than the range of mean discrimination values between tasks, which suggests that tasks alone do not account for the range of discrimination values.

Item difficulty is a useful way to arrange items from easiest to most difficult, relative to a particular response probability. Table 2 shows summary statistics for MDIFF statistics for the testlet model in 2014 and 2015. Average MDIFF statistics show that Syllable Clapping (SC) was the easiest task[1] in both years. At the other end of the spectrum, UC and LC were the most difficult, on average, in 2014. UC and BS tasks were the two most difficult tasks, on average, in 2015. All tasks have item difficulties that span a wide range of the scale. In 2014, PW and UC had the largest range of MDIFF values. They were followed closely by LC, LS, and LL items that had a range of MDIFF values close to four. BS had the smallest range of MDIFF values in both years. Items belonging to LC and LS had the largest range of MDIFF values in 2015.

Table 3 shows item difficulty (MDIFF) values for individual items on the UC letters task for the 2014 and 2015 administrations. Letters in Table 3 are listed in ascending order of MDIFF values. The letter O was the easiest to recognize. Other letters consistently located at the easy end of the spectrum were M, F, A, and B. Conversely, the most difficult letters or digraphs to name were LL, Ch, J, Ñ, and G. We observed a similar pattern for LC and LS tasks in 2014. The lowercase letter o was the easiest to name and the second easiest letter sound to make. Other letters that were consistently easy to name or sound were s, f, and t. Conversely, the most difficult letters or digraphs to sound or identify in lower case were ll, ch, ñ, and j.

Discussion

We evaluated the internal structure of PALS español PreK using multidimensional IRT. One advantage of our approach

Table 4. Ordering of Lowercase Letters and Letter Sounds According to Spring Multidimensional Difficulty.

2014 Lowercase Letters			2014 Letter Sounds		
Letter	MDIFF	Slope	Letter	MDIFF	Slope
o	−0.14	1.27	S	−0.40	2.04
s	0.53	2.51	O	−0.11	1.07
f	0.74	2.44	F	0.19	2.01
m	0.79	2.76	P	0.22	2.43
n	0.89	1.81	T	0.26	2.77
a	0.92	2.02	B	0.33	1.71
c	0.94	2.32	K	0.51	2.14
t	1.07	2.52	D	0.61	2.41
e	1.08	1.97	L	0.64	2.26
v	1.15	2.45	A	0.65	1.34
x	1.16	1.46	N	0.75	2.86
p	1.16	2.29	V	0.90	1.92
r	1.20	2.06	R	0.93	1.50
u	1.27	2.11	E	0.96	1.82
h	1.35	1.73	Z	0.96	1.25
l	1.37	1.64	C	0.99	1.67
i	1.39	1.70	W	1.20	1.77
k	1.47	2.02	G	1.25	1.92
b	1.49	1.66	Y	1.61	1.81
d	1.55	1.67	U	1.69	1.01
z	1.59	1.89	I	1.94	1.00
w	1.66	2.04	Ch	2.18	1.22
y	1.78	2.27	Ñ	2.49	1.47
ñ	1.81	1.52	J	2.90	0.99
j	1.89	1.61	Ll	3.53	0.90
g	2.34	1.46			
q	2.54	1.46			
ll	2.93	1.36			
ch	3.31	1.08			

Note. Ordering by MDIFF is the same as ordering by a response probability of 0.5.

over the use of item parcels with confirmatory factor analysis is that we can test the internal structure of the measure while also studying the contribution of each individual item to the latent trait and evaluating the difficulty of each individual item. Results indicated that a task-based testlet model (Model 2b) fit the data better than other models we applied to the data. About 14% of the items had statistically significant levels of drift, but the practical significance was small and had little to no effect on estimated abilities.

In the testlet model, all of the items have a direct impact on the general dimension (early literacy), while at the same time maintaining a unique relationship to each other within each task. Although we did not find that the two parts of the simple view (Oral Language and Code-Related Skill) constituted separate factors, we did find large testlet effects and evidence that our oral language task contributed substantial variance after accounting for early literacy. Thus, oral language appears to have a strong contribution to test score variance, but not to the point of being a separate but correlated general dimension. It may be that oral language does not develop into a separate factor until a later age. Additional research is needed to explore this possibility.

Another reason for not finding more support for the two-tier model that represented the simple view of early literacy may be attributed to the analytic procedures and the way the relationship between the two dimensions is represented. Prior factor analytic work with PALS assessment involved item parcels at the task level, which prevented the use of either a bifactor or testlet model (see Townsend & Konold, 2010; Yaden et al., 2017). Huang (2014) used item parcels formed from pairs of items, and this method enabled him to fit correlated factor models and a bifactor model to the data. He did not fit a testlet model to the data. Nevertheless, his work championed the bifactor model, and it is consistent with our results. Considering that Yaden et al. (2017) found support for a model with two correlated factors when using task-level item parcels but we did not when conducting an item-level analysis, the analytic procedure may be affecting the results. It is difficult to know for sure as Yaden et al. used a shorter, preliminary version of the measure.

It is important to note that the correlated factor model, higher-order models, bifactor model, and testlet model are conceptually related. Although we found that a task-specific testlet model with one general dimension fit best, the result does not contradict past research or prior factor analytic work with PALS assessments. Reise, Moore, and Haviland (2010) explain the similarity of a model with correlated factors and one with bifactor structure, which includes the testlet model. Both types of models account for variance among traits, but they do it in different ways. In a model with correlated factors, the correlation accounts for what is shared among traits. By comparison, a model with bifactor structure uses the general dimension to account for variance shared among all items, and the specific dimensions allow for task- or domain-specific shifts in the construct. The simple view of reading entails two traits (Oral Language and Code-Related Skills), which have been shown to be highly correlated (Dickinson et al., 2003; Kendeou et al., 2009; Lonigan et al., 2000; NICHD Early Child Care Research Network, 2005). This relationship among traits seems to be adequately captured as either a model with correlated factors (Townsend & Konold, 2010; Yaden et al., 2017) or a model with a single general dimension and multiple specific dimension that reflect either domain-specific (Huang, 2014) or task-specific shifts in the construct. Thus, the main difference in these types of models may be more a matter of measurement utility than theoretical import. Advantages of the testlet model are that we know how each item contributes to the general dimension, we know the impact of each specific dimension (i.e., testlet effect), we can order all items by difficulty, and we can consider the relative difficulty of each task.

Item Discrimination and Item Difficulty

Item discrimination (i.e., slopes) varied by task. Those related to alphabet knowledge (i.e., UC, LC, and LS) tended to have larger slopes than those related to phonological awareness (e.g., SC, RA, BS). These results suggest that alphabet knowledge items make a stronger contribution to the general dimension of early literacy than do phonological awareness items. This is in keeping with previous research that has consistently shown alphabetic knowledge to be an important early literacy skill (Foulin, 2005; Hammill, 2004; National Early Literacy Panel, 2008; Whitehurst & Lonigan, 1998). Letter-name knowledge remains one of the strongest predictors of later reading success and later reading difficulty (Catts et al., 1999; Hammill, 2004; Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004; Share, Jorm, Maclean, & Matthews, 1984; Snow et al., 1998). Knowledge of letter names and letter sounds also helps children understand the alphabetic principle, the insight that language is made up of individual speech sounds and that letters represent those sounds in a systematic way. Achieving insight into the alphabetic principle is a turning point in early literacy development and is often associated with the transition to conventional reading (Liberman et al., 1989).

Phonological awareness may contribute to letter-name and letter-sound knowledge by making the sounds embedded within letter names and in associated letter sounds more noticeable and thus easier to remember (Evans et al., 2006; Foy & Mann, 2006; Piasta & Wagner, 2010). Previous research has suggested a reciprocal relationship between phonological awareness and the learning of letter sounds, in which competence in either skill facilitates competence in the other (Burgess & Lonigan, 1998). The fact that item slopes related to alphabet knowledge were larger than those related to phonological awareness makes sense given the reciprocal nature of these two constructs.

Our use of MIRT also provides information about the relative difficulty of each item, which offers insight into the order in which student are likely to learn to name and pronounce letters as well as other skills. Phillips, Piasta, Anthony, Lonigan, and Francis (2012) demonstrated this utility of IRT in their analysis of letter-name knowledge. Results of their unidimensional two-parameter logistic model allowed them to order items by difficulty and study the developmental sequence of letter naming skill. Their results were consistent with findings by Alonzo, Liu, and Tindal (2007), who found that English letters A and B were among the easiest, while English letters U and V were some of the most difficult. Taken together, their findings suggest that children are likely to learn the English letters A and B before they learn the English letters U and V. In a similar fashion, our analysis of PALS español PreK indicates a developmental path in Spanish. The uppercase letters O, A, B, and M are easy for children to name, and thus likely to be among the first learned. In contrast, the letters LL, Ch, and Ñ are the most difficult for preschoolers to name. We found some evidence that the difficulty of letter naming coincides with the difficulty of letter sound knowledge. For example, the easiest letter sounds were S, O, and F, and these were among the easiest letters to name. Two of the most difficult letter sounds (LL and Ñ) were also among the most difficult letters to name.

In summary, we used MIRT to explore the internal structure of PALS español PreK. Based on theories of early literacy development, we anticipated that the best-fitting model would be a two-tier model with two correlated dimensions representing the dominant early literacy constructs of oral language and code-based skills. We found that the best-fitting model was actually a testlet model with one general dimension, early literacy, and a specific dimension for each task. Tasks for code-based skills and the oral language task contributed to the same general dimension; our result did not support the notion that the oral language and code-based skills were distinguishable correlated constructs. However, the nature of the testlet model suggests that we also did not have just a single dimension. Each task contributed substantial variance even after accounting for the early literacy dimension. Thus, tasks represent skills that should be given attention.

Implications for Instruction

PALS español PreK yields both a composite score and individual task scores. The composite score can be seen as a measure of the general dimension of early literacy. Teachers can use this score to determine which children are on track for developing the early literacy skills that are the foundation for reading and which children need additional support if they are to become successful readers. The individual task scores, which represent the specific dimensions examined in this study, provide diagnostic information that aids instructional decisions. For example, a student with low performance on alphabet knowledge tasks has different instructional needs than a student with low performance in language and listening comprehension. Thus, the task-based testlet model provides information about overall early literacy development and, after taking overall development into account, it also provides information necessary for targeted instruction in specific areas.

Knowing the comparative difficulty of each PALS español task and item also provides useful information that can guide instruction. Our research has shown that some tasks measure skills that are easier to achieve than other tasks. For example, the phonological awareness tasks (e.g., SC and RA) were easier than the alphabet tasks (UC, LC, and LS), underscoring the importance of explicit, systematic instruction in alphabet skills early on (National Early

Literacy Panel, 2008). Our results also indicate an order of difficulty for specific letter names and sounds. In Spanish, the vowel sounds are taught first in combination with what are considered to be the more salient consonant sounds. Our research suggests a possible progression of consonant sounds from easiest to most difficult (see Table 3).

Finally, the fact that our testlet model is a compensatory model suggests that the most efficient way to improve overall literacy development might be to teach specific skills in an integrated fashion to make the most of strengths in all specific dimensions and to harness the power of the general underlying trait. In that way, instructional activities that target early literacy skills in general and/or those that target skills required by each task will improve overall literacy development, which will be reflected in children's performance on PALS español PreK. For example, a student with poorly developed alphabet knowledge may benefit not only from direct instruction in alphabet skills, but also from more contextual engagement, such as print referencing techniques that focus on further developing concepts about print (including alphabet knowledge) in the context of shared reading (Justice & Ezell, 2004). Having a clearer understanding of this relationship between individual literacy skills and the underlying trait of early literacy is particularly important now, given the recent discourse on the potentially negative side effects of assessment, such as narrowing of the early literacy curriculum and the resulting fragmentation of early literacy skills (Meisels, 2007; Paris, 2005).

Limitations

The main limitation of this work was the relatively small sample size. We were unable to randomly split the data into one part of an exploratory analysis and another part of a confirmatory analysis. We fit models consistent with theoretical expectation and prior work with PALS assessments to overcome this limitation. We continue to collect data and that will allow us an opportunity to replicate our work on independent samples.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130469 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Note

[1] Our discussion of item difficulty ordering is based on the location of an item at a response probability of .5 (i.e., a 50% chance of correctly answering the item).

ORCID iD

J. Patrick Meyer https://orcid.org/0000-0002-0044-7959

References

Alonzo, J., Liu, K., & Tindal, G. (2007). *Estimating the technical adequacy of reading comprehension measures in a progress monitoring assessment system* (Technical Report No. 41). Eugene: Behavioral Research & Teaching, University of Oregon.

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science*, *14*, 255–259.

Anthony, J. L., & Lonigan, C. J. (2004). The nature of phonological awareness: Converging evidence from four studies of preschool and early grade school children. *Journal of Educational Psychology*, *96*, 43–55.

Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling*, *15*, 211–240. doi: 10.1080/10705510801922340

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Mahwah, NJ: Lawrence Erlbaum.

Bloodgood, J. W. (1999). What's in a name?: Children's name writing and literacy acquisition. *Reading Research Quarterly*, *34*, 342–367.

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

Burgess, S. R., & Lonigan, C. J. (1998). Bidirectional relations of phonological sensitivity and prereading abilities: Evidence from a preschool sample. *Journal of Experimental Child Psychology*, *70*, 117–141.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612. doi:10.1007/S11336-010-9178-0

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading*, *3*, 331–361.

Chomsky, C. (1971). Write first, read later. *Childhood Education*, *47*, 296–299.

Clay, M. M. (1977). *Reading: The patterning of complex behavior*. Exeter, NH: Heinemann.

Clay, M. M. (1979). *What did I write? Beginning writing behavior*. Portsmouth, NH: Heinemann.

Dickinson, D. K., Golinkoff, R. M., & Hirsh-Pasek, K. (2010). Speaking out for language: Why language is central to reading development. *Educational Researcher*, *39*, 305–310. doi: 10.3102/0013189X10370204

Dickinson, D. K., McCabe, A., Anastasopoulos, L., Peisner-Feinberg, E. S., & Poe, M. D. (2003). The comprehensive language approach to early literacy: The interrelationships among vocabulary, phonological sensitivity, and print knowledge among preschool-aged children. *Journal of Educational Psychology*, *95*, 465–481. doi:10.1037/0022-0663.95.3.465

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, *9*, 167–188.

Evans, M. A., Bell, M., Shaw, D., Moretti, S., & Page, J. (2006). Letter names, letter sounds and phonological awareness: An examination of kindergarten children across letters and of letters across children. *Reading and Writing*, *19*, 959–989.

Ferreiro, E., & Teberosky, A. (1982). *Literacy before schooling*. Exeter, NH: Heinemann.

Ford, K. L., Cabell, S. Q., Konold, T. R., Invernizzi, M., & Gartland, L. B. (2013). Diversity among Spanish-speaking English language learners: Profiles of early literacy skills in kindergarten. *Reading and Writing*, *26*, 889–912. doi:10.1007/s11145-012-9397-0

Foulin, J. N. (2005). Why is letter-name knowledge such a good predictor of learning to read? *Reading and Writing*, *18*, 129–155. doi:10.1007/s11145-004-5892-2

Foy, J. G., & Mann, V. (2006). Changes in letter sound knowledge are associated with development of phonological awareness in pre-school children. *Journal of Research in Reading*, *29*, 143–161. doi:10.1111/j.1467-9817.2006.00279.x

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika*, *57*, 423–436.

Hammill, D. D. (2004). What we know about correlates of reading. *Exceptional Children*, *70*, 453–468.

Houts, C. R., & Cai, L. (2015). *flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring user's manual version 3.0RC*. Retrieved from https://www.vpgcentral.com/software/irt-software/

Huang, F. L. (2014). Using a bifactor model to assess the factor structure of the Phonological Awareness Literacy Screening for grades 1 through 3. *Journal of Psychoeducational Assessment*, *32*, 638–650. doi:10.1177/0734282914525026

Huang, F. L., Ford, K. L., Invernizzi, M., & Fan, X. (2013, April). *Measuring early Spanish literacy: Factor structure and measurement invariance of the "Phonological Awareness Literacy Screening for Kindergartners" in Spanish ("PALS español K")*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA) San Fancisco, CA. Retrieved from ERIC database. (ED 546651)

Huang, F. L., & Konold, T. R. (2014). A latent variable investigation of the Phonological Awareness Literacy Screening-Kindergarten assessment: Construct identification and multigroup comparisons between Spanish-speaking English-language learners (ELLs) and non-ELL students. *Language Testing*, *31*, 205–221.

Huang, F. L., Tortorelli, L. S., & Invernizzi, M. A. (2014). An investigation of factors associated with letter-sound knowledge at kindergarten entry. *Early Childhood Research Quarterly*, *29*, 182–192.

Justice, L. M., & Ezell, H. K. (2004). Print referencing: An emergent literacy enhancement strategy and its clinical applications. *Language, Speech, and Hearing Services in Schools*, *35*, 185–193.

Kendeou, P., van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology*, *101*, 765–778. doi:10.1037/a0015956

Liberman, I. Y., Shankweiler, D., & Liberman, A. M. (1989). The alphabetic principle and learning to read. In D. Shankweiler & I. Y. Liberman (Eds.), *Phonology and reading disability: Solving the reading puzzle* (pp. 1–34). Ann Arbor: University of Michigan Press.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: From a latent-variable longitudinal study. *Developmental Psychology*, *36*, 596–613. doi:10/1037/0012-1649.36.5.596

Maydeu-Olivares, A. (2015). Evaluating the fit of IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling* (pp. 111–127). New York, NY: Routledge.

Meisels, S. J. (2007). Accountability in early childhood: No easy answers. In R. C. Pianta, M. J. Cox & K. L. Snow (Eds.), *School readiness and the transition to kindergarten in the era of accountability* (pp. 31–47). Baltimore, MD: Paul H. Brookes.

Meyer, J. P., Ford, K. L., & Invernizzi, M. A. (2017). *The phonological awareness literacy screening in Spanish for preschool (PALS español PreK): Estimating reliability using multivariate generalizability theory*. Manuscript submitted for publication.

Morris, D., Bloodgood, J., & Perney, J. (2003). Kindergarten predictors of first- and second-grade reading achievement. *The Elementary School Journal*, *104*, 93–109.

Morris, D., Bloodgood, J. W., Lomax, R. G., & Perney, J. (2003). Developmental steps in learning to read: A longitudinal study in kindergarten and first grade. *Reading Research Quarterly*, *38*, 302–328.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

National Association for the Education of Young Children. (2005). *Screening and assessment of young English-language learners: Supplement to the NAEYC and NAECS/SDE joint position statement on early childhood curriculum, assessment, and program evaluation*. Washington, DC: Author.

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy.

NICHD Early Child Care Research Network. (2005). Pathways to reading: The role of oral language in the transition to reading. *Developmental Psychology*, *41*, 428–442. doi:10.1037/0012-1649.41.2.428

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, *40*, 184–202.

Peña, E. D., & Halle, T. G. (2011). Assessing preschool dual language learners: Traveling a multiforked road. *Child Development Perspectives*, *5*, 28–32. doi:10.1111/j.1750-8606.2010.00143.x

Phillips, B. M., Piasta, S. B., Anthony, J. L., Lonigan, C. J., & Francis, D. J. (2012). IRTs of the ABCs: Children's letter name acquisition. *Journal of School Psychology*, *50*, 461–481.

Piasta, S. B., & Wagner, R. K. (2010). Learning letter names and sounds: Effects of instruction, letter type, and phonological processing skill. *Journal of Experimental Child Psychology*, *105*, 324–344.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544–559. doi:10.1080/00223891.2010.496477

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, *96*, 265–282. doi:10.1037/0022-0663.96.2.265

Share, D. L., Jorm, A. F., Maclean, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Educational Psychology*, *76*, 1309–1324.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology*, *38*, 934–947. doi:10.1037/0012-1649.38.6.934

Townsend, M., & Konold, T. R. (2010). Measuring early literacy skills: A latent variable investigation of the Phonological Awareness Literacy Screening for Preschool. *Journal of Psychoeducational Assessment*, *28*, 115–128. doi:10.1177/0734282909336277

Treiman, R., & Broderick, V. (1998). What's in a name: Children's knowledge about the letters in their own names. *Journal of Experimental Child Psychology*, *70*, 97–116.

Treiman, R., Tincoff, R., Rodriguez, K., Mouzaki, A., & Francis, D. J. (1998). The foundations of literacy: Learning the sounds of letters. *Child Development*, *69*, 1524–1540.

van Kleeck, A. (1998). Preliteracy domains and stages: Laying the foundations for beginning reading. *Journal of Children's Communication Development*, *20*, 33–51.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental Psychology*, *30*(1), 73–87.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge.

Warley, H., Landrum, T., Invernizzi, M., & Justice, L. (2005). Prediction of first grade reading achievement: A comparison of kindergarten predictors. In B. Maloch, J. V. Hoffman, D. L. Schaller, C. M. Fairbanks & J. Worthy (Eds.), *National reading conference yearbook* (pp. 428–442). Oak Creek, WI: National Reading Conference.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*, 77–87.

Welsch, J., Sullivan, A., & Justice, L. (2003). That's my letter: What preschoolers' name writing representations can tell us about emergent literacy knowledge. *Journal of Literacy Research*, *35*, 757–776.

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

Woods, C., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*, 532–547.

Wyse, A. E. (2013). DIF cancellation in the Rasch model. *Journal of Applied Measurement*, *14*, 118–128.

Yaden, D. B., Marx, R. W., Cimetta, A. D., Alkhadim, G. S., & Cutshaw, C. (2017). Assessing early literacy with Hispanic preschoolers: The factor structure of the Phonological Awareness Literacy Screening—Español. *Hispanic Journal of Behavioral Sciences*, *39*, 1–18. doi:10.1177/0739986316688877