Assessing Comprehension in Kindergarten through Third Grade

John P. Sabatini, Laura K. Halderman, Tenaha O'Reilly, Jonathan P. Weeks

Educational Testing Service

Topics in Language Disorders

Published December 2016

Corresponding Author:

John Sabatini

660 Rosedale Road, MS-13E

Princeton, NJ 08541

jsabatini@ets.org

Peer Review Process: *Topics in Language Disorders* (*TLD*) is a double-blind peer-reviewed topical journal.

Acknowledgement

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100005 to the Educational Testing Service as part of the Reading for Understanding Research (RFU) Initiative and Grant (R305A150176). The opinions expressed are those of the authors and do not represent views of Educational Testing Service, the Institute or the U.S. Department of Education. We want to thank Kelsey Dreier, Kim Fryer, Don Powers, Jonathan Steinberg, Zuowei Wang, the journal editors, and anonymous reviewers for helpful comments and editorial assistance.

Abstract

Traditional measures of reading ability designed for younger students typically focus on componential skills (e.g., decoding, vocabulary) and the items are often presented in a discrete and decontextualized format. The current study was designed to explore whether it was feasible to develop a more integrated, scenario-based assessment of comprehension for younger students. A secondary goal was to examine developmental differences in item performance when administration was in listening versus reading modalities. Cross-sectional differences were examined across kindergarten to third grade on a scenario-based assessment comprised of literal comprehension, inference, vocabulary, and background knowledge items. The assessment, originally targeted for third grade, was administered one-on-one to 141 third grade and 485 second grade students. It was adapted for and administered to kindergarten (n = 390) and first grade (n = 419) students by reducing the number of items and switching to a listening comprehension method of administration. Each grade was significantly more accurate than the previous grade on overall performance and background knowledge. A regression analysis showed significant variance associated with background knowledge in predicting comprehension, even after controlling for grade. A deeper analysis of item performance across grades was conducted to examine what elements worked well and where improvements should be made in adapting comprehension assessments for use with young children.

Assessing Comprehension in Kindergarten through Third Grade

Assessing young children's comprehension development can be challenging. Since the passing of the *No Child Left Behind Act of 2001* (2002) school level reading comprehension assessments have been administered beginning in third grade. However, in earlier grades, a componential approach is most common. Following the Simple View of Reading (Gough & Tunmer, 1986; Hoover & Gough, 1990), a traditional starting point is to divide measures between those involving word-level reading recognition versus linguistic comprehension. Although the divided assessment of word reading and language comprehension may provide some understanding regarding young children's reading and language skills, this assessment practice does not necessarily provide insights regarding how children integrate reading and language in order to access deeper meaning. The purpose of the exploratory research presented in this paper was to begin to develop and explore an early comprehension assessment that moves beyond this componential method and yet still readily evaluates students who have limited word reading and language abilities.

One way to accommodate for early or persisting word reading limitations and individual language differences is to assess reading comprehension using a componential approach. There is practicality and efficiency in adopting a componential approach prior to third grade. Component measures may limit the complexity of the task environment, potentially reducing working memory and cognitive load because there are fewer task demands, in comparison to a more integrated assessment that measures multiple skills simultaneously.

Historically, this componential approach was used and considered a sensible and functional way to assess the changing development of children between kindergarten and 3rd grade. This development includes the progression of children from a) limited alphabet knowledge, to understanding of the alphabetic principle; b) from basic decoding of words in their listening lexicon, to acquiring a sizeable sight-word vocabulary; c) from word by word reading, to fluent oral (and silent) reading of continuous texts.

Assessments that provide an indication of a child's ability to assemble the aforementioned component skills into an integrated whole have the potential to provide valuable insights into the skills necessary in literacy activities beyond third grade. One of the aims of the larger project in which this study is embedded was to develop innovative assessments of reading for understanding, and create a new type of computer-based assessment, termed Scenario-Based Assessment (SBA). The use of SBA techniques allowed us to deliver a set of thematically related source materials in a digital environment and potentially enable us to assess reading comprehension and language processes in a more integrated way (Bennett, 2010; 2011).

In order for the SBA results to be interpretable, we examined task difficulty relative to child development. While most kindergarten and first grade children would not yet have the word recognition skills to read, we sought to determine whether younger children would have the language comprehension abilities that were targeted by the SBA form. In this study we explored the performance levels associated with the texts and questions that were read to the students, as well as the impact of changing modality (listening vs. reading). Additionally, we examined the developmental differences in children's background knowledge, memory, and reasoning skills. In short, our goal was to understand—at least in part—how to design SBA comprehension tests that target early developmental reading comprehension abilities in children, and to better understand individual differences as children learn to integrate their language with their reading skills.

Theoretical Background

An evolving construct of reading in the 21st century

While a simplified construct of reading (or listening) comprehension may be justifiable as one type of measure for young children, we agree with the position that the construct of reading comprehension has been changing significantly in the past several decades and that comprehension assessment designs have not kept pace with the changes in how people read in the 21st century (e.g., digital literacy, Coiro, 2009; Leu et al., 2013), nor advances in cognitive science and instruction (Gordon Commission, 2013). This position is aligned with various assessment reforms such as the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), the Partnership for 21st Century Skills (2008), and other seminal works (Bennett, 2010; 2011; Bransford, Brown, & Cocking, 2000; Pellegrino, Chudowsky, & Glaser, 2001).

These sources support the argument that the typical approach to measuring comprehension, one that focuses on students' understanding of a single text in isolation, under-represents the complexity of a modern construct of reading comprehension that emphasizes purpose-driven, multiple document processing (Britt & Rouet, 2012). This is not to say that this form of comprehension test is not valid or should not be used, but rather an acknowledgment that there is more to comprehension than what is covered in traditional, print-based tests of reading. If the construct of comprehension is evolving, ideally, these changes should be reflected in developmentally appropriate content and tasks administered to young, as well as older, children.

From a review and synthesis of these and other literatures, we have been developing a framework for assessing reading for understanding across prekindergarten through twelfth grade

(O’Reilly & Sabatini, 2013; Sabatini, O’Reilly, 2013; Sabatini, O’Reilly & Deane, 2013). In these publications, we provide a definition of reading, outline the constructs underlying the measures and how they might change across the school years, explain the use of scenario-based assessment, and the role of performance moderators.

While the details of the framework are beyond the scope of this paper, we briefly summarize some key points. From our survey of the literature, we hold that reading is a purposeful activity (van den Broek, Linderholm, & Gustafson, 2001), that purposes are used to set standards of coherence (Linderholm, Virtue, Tzeng, & van den Broek, 2004) for determining what is relevant when reading text sources (McCrudden, Magliano & Schraw, 2011). In practical and everyday reading contexts, students must be able to integrate and evaluate multiple sources (Britt & Rouet, 2012) to satisfy their purpose for reading. This process draws upon students’ background knowledge (Shapiro, 2004), as they may be required to interpret texts from different points of view or through the lenses of different disciplines (Goldman, 2012; LaRusso et al., 2016). Skilled readers may also use reading strategies (McNamara, 2012), metacognition, and self-regulation (Hacker, Dunlosky, & Graesser, 2009) to help process text deeply. Encouraging the use of reading strategies during an assessment is one way of modeling good comprehension practices and cognitive habits (e.g., Griffin, Malone, & Kammenui, 1995; Ozuru, Best,O’Reilly, & McNamara, 2007; Meyer & Ray, 2011), as well as an effective means of collecting evidence of reading proficiency. While this portrayal of purposeful, integrative comprehension is often reserved for describing what it means to be college and career ready, one must establish the precursors of these skills in younger children to ensure a trajectory of learning that leads to proficiency by the end of secondary schooling (Goldman, 2004).

To address coverage of this expanded construct, we identified five knowledge and skill targets that span all developmental levels: print, verbal, discourse, conceptual, and social. ***Print*** targets address the skills needed to "get the printed words off the page" including decoding, word recognition, and all other typographical conventions of written language. ***Verbal*** targets address broader language resources such as vocabulary, morphology, syntax, and grammar, with a focus on word to sentence level processes. Moving beyond the sentences and word level, ***discourse*** targets cover elements related to the global understanding of continuous text and discourse sources, including text structure and coherent mental model building. ***Conceptual*** targets include skills of reasoning, critical evaluation of information (e.g., critical thinking), and applying what one understands to solving problems or achieving goals. ***Social*** targets comprise the skills used in mental modeling and reasoning concerning human intent and action, including author intent, agent/character goals and motives, perspective taking, and ethical and social reasoning (e.g., debate, argument, and persuasion). As noted, these targets are present at every developmental level; what changes over time is the complexity of the text and task demands.

Scenario-based Assessment. We do not conceptualize comprehension proficiency as the deployment of isolated, componential skills, but rather as integrated operations that are required when performing literacy tasks such as reading for understanding, learning, or solving a problem using text sources. Students must coordinate multiple task demands simultaneously in real world, 21st century literacy experiences; hence, the discrete, decontextualized texts and items in traditional reading tests under-represent this complexity. How then can tests be used to measure the complex literacy tasks demanded of modern students, while at the same time, providing useful, valid information for students and teachers?

To address the changing nature of reading, we developed scenario-based assessments (SBAs). The general framework for SBAs involves setting an overarching goal for the test taker and providing a broad sequence of steps to get there; gathering a set of thematic sources that are relevant to the goal; measuring or building up background knowledge to gauge its impact on task performance and learning; using sequences, scaffolding, and structured tasks to probe partial and complex skill development; and using simulated peers and agents to engage and model social interactions typical of literacy activities in and outside of school learning.

Although results with the SBA approach for middle and high school students have been promising (e.g., O'Reilly et al., 2015; Sabatini et al. 2014a, b), significant adaptations in task design and administration are needed to apply SBAs with young children. In terms of the construct, we would need to select texts and precursor skills, applied in a literacy context likely familiar to young children, such as shared book reading with a teacher or other adult. In terms of the assessment design, SBAs are intended to be less about discrete skill tasks, and more about natural and integrated activities, hopefully making the experience more welcoming.[1] In this study, the assessment administrator read the electronic book (assessment story) aloud to the younger children, similar to what might happen in a warm and friendly school or home environment. The electronic storybook included pictures for both stimulating interest and for elaboration of the key concepts and themes.

To make the tasks more manageable, the text was broken apart into pages of text like a child would encounter in a storybook at these grade levels. To help reduce working memory

[1] One of the challenges of traditional assessments, is that students may become disengaged, because of the discrete and decontextualized nature of the test and items. On aim of the SBA approach is to make the experience more personal and authentic, with hopes of improving student engagement, reducing stress levels, and thereby giving the students the opportunity to display their best performance.

load for some of the items, the questions were presented one at a time after the associated page was read, with the text still present.[2] Other questions were presented after the story was read. Background knowledge was measured before the student read or heard the passage content. For some of the background knowledge items, the questions were answerable after reading the passages. Some of these same background knowledge questions were asked again after the passages were heard or read, in order to help us identify whether the students learned the information after reading or hearing the text.

The simplified SBA used in this study covered a range of comprehension skills—precursors to those that might appear later in the middle or upper grades versions of SBAs—and some of the scenario design elements characteristic of SBAs. However, adapting the SBA forms for younger children also required that we rethink the relationship of children's developing background knowledge and language skills, in relation to reading comprehension. We discuss these two issues in the next two sections.

Background Knowledge. As noted, the framework for comprehension discusses performance moderators such as background knowledge (BK) and self-regulation skills. These are constructs that are important to proficiency, but we do not consider them to be directly part of the comprehension construct.

One unique target of this study is a focus on the relationship between background knowledge (BK) and comprehension. The literature on older readers shows that BK is associated with higher levels of comprehension (Ozuru et al., 2007), and that this relationship generally becomes stronger with domain expertise and age (Alexander et al., 2004; Murphy &

[2] For some items near the end of the assessment, the text was not available and thus the items require the students to remember some of the information that was read.

Alexander, 2002). Further, the literature suggests that BK can be separated from motivation (Taboada, Tonks, Wigfield, & Guthrie, 2009), yet interacts with general reading ability (O'Reilly & McNamara, 2007a), online reading ability (Coiro, 2011), and text cohesion (McNamara & Kintsch, 1996). BK is malleable to intervention (Vitale & Romance, 2007), and may also compensate for aptitude differences (Schneider, Körkel, & Weinert, 1989). However, there is scant attention to the impact of BK on comprehension tests, especially when using expository content (see Shapiro, 2004 for general discussion of the influence of BK on the interpretation of assessment results), nor the relationship between BK and comprehension for younger children (see Evans et al., 2001 on younger children and adolescence). Given that younger students presumably have less overall knowledge about the world than older students, one might hypothesize that there is little to no relationship between comprehension and BK in young learners, although the inferences about the relationship may depend upon how BK is measured and assessed in text comprehension. We explore this issue in some detail in this study.

The relationship between language/linguistic resources and reading comprehension in developing readers

The theoretical rationale for the modification of an assessment from a reading to a listening comprehension format warrants some explanation. Researchers interested in children's early reading development, often use the Simple View as theoretical grounding (Gough et al., 1996; Hoover & Tunmer, 1993). Hoover & Tunmer (1993) described the Simple View as making "two claims: first, that reading consists of word recognition and linguistic comprehension; and second, that each of these components is necessary for reading, neither being sufficient in itself" (p. 3). The Simple View is fundamentally a language based framework that takes into account how one’s writing system maps onto one’s language system (Perfetti, 2003). That is, how printed letters form words and how the other typographic elements (e.g., punctuation, spaces) form a code that can be mapped onto linguistic structures, i.e., the phonetics, phonology, morphology, syntax, semantics (vocabulary), and pragmatics of language. Vellutino, Tunmer, Jaccard, & Chen (2007) are explicit in this, stating that learning to read in English “entails visual recoding of language in the form of alphabetic characters representing speech segments,” (p. 7). In other words, one can conceptualize word recognition primarily as a language skill; this is a position we support.

Seminal works investigating the relationship between language development and reading acquisition disorders, such as those conducted by Catts and colleagues (Catts, Adlof, Weismer, & Ellis, 2006; Catts, Fey, Zhang, & Tomblin, 1999; Catts, Hogan, & Fey, 2003), appear to support this interpretation of the Simple View as well. The results of these studies suggest that word reading difficulties and disabilities are (or are caused by) language processing deficits or disorders, with phonological processing or awareness deficits at the core (Stanovich, Siegel, & Gottardo, 1997).

The Convergent Skills Model of reading development (Vellutino et al., 2007) presents a more precise and complex representation of the Simple View (see also Just and Carpenter, 1988; Kintsch, 1998; Perfett, Landi, & Oakhill, 2005 for more robust processing models of comprehension). The Convergent Skills Model is premised on multiple language antecedents feeding into to the Simple View components of Context Free Word Recognition and Language Comprehension. Further, developmental evidence for this model shows that the strength of the relationship between language and reading comprehension increases over time while the impact of direct measures of word recognition decreases. The language antecedents of word recognition (e.g., semantic knowledge) continue to be significant, again suggesting that word recognition is a linguistic skill.

The increasing strength of the correlation between language and reading comprehension across early school years may be explained by word recognition skills becoming a more automatized and efficient functional system of language processing itself. After all, the visual, perceptual processing of letters and words map to phonology, morphology (or morphosyntactics), or directly to semantic meaning networks— all of which are linguistic or language processing functions of the brain (Sandak, Mencl, Frost, & Pugh, 2004). Moreover, comprehension processing models suggest that interactions among various linguistic networks occur in parallel, or in very rapid sequences, and are bidirectional (Gerrig & O'Brien, 2005; Kintsch, 1998; Rayner et al., 2001); hence, learning to read (or to write) is essentially the acquisition of automaticity and fluency in the receptive and expressive processing of language in a visual modality/medium. This conception of reading as part of language is how researchers who study language acquisition in non-native speakers think of language learning; that is, as four modalities – speaking, listening, reading, and writing.

One can think of the four modalities of language (reading, writing, speaking, listening) as the ways in which one accesses common language or linguistic resources (e.g., phonological word form, word meaning, syntactic/grammatical knowledge, semantic knowledge network, pragmatics and episodic knowledge, such as the structure of a narrative story). That is not to say that there are no individual differences in knowledge and skills within and across modalities in children (or adults), but rather that underlying language resources are potentially accessible and shareable when an individual processes information in different modalities. If a word, phrase, or sentence is understood when one hears it, we can assume it will be understood when one reads it (further assuming adequate decoding skill) and that it potentially could be used in productive speech or in a written composition.

The language resources typically accessed when a young child is speaking and listening are linked to reading and writing processing through instruction from kindergarten through 3rd grade in the United States (teaching in phonics/decoding and handwriting/typing). An emphasis on foundational reading skill development (mostly decoding, word recognition and their fluent application in text reading) is likely to lead to a successful linking of language resources with the visual modality of reading in print. However, the identification of differences in language skills development will only be evident in reading (or listening) tasks on an assessment when assessment tasks have an appropriate difficulty range to reveal them. That is, if one reads a text and questions aloud to a group of individuals, and all of the group can answer all items accurately, there would be no variance to explain. One could only say that all of the students have listening (language) comprehension above the threshold of difficulty of the spoken test administered. If one then gave those same individuals the written text and questions (no oral reading support), and the performance varies, one can infer that the source of the variance has to do with the print format. However, if one reads a text and questions aloud to a group of individuals, and they vary in their ability to answer them, then one can infer that they vary in their listening comprehension. If that same group, answered all items accurately when reading the text and questions (which one might find in a non-native speaker of English who learns how to read in English), then the variability is in listening comprehension (not reading comprehension). In each instance above, the ability to correctly respond in any modality (reading or listening) is evidence of an underlying language capability that is potentially accessible across modalities, once each can be accessed by one's language processing system.

Put another way, with the exception of some modality differences in processing and information cues (e.g., memory for listening to a story versus the possibility of looking back when reading; intonation, prosody and expression cues when listening versus spacing and punctuation cues when reading), one would expect that one's performance across the modalities would be comparable or at least have certain logical dependencies. That is, if one can answer a question correctly when reading only, then one could also answer that question if one listened to the same text (especially if the written text was simultaneously available, thus, providing the processing and information cues of both modalities).

In addition to examining the use of SBA with children below third grade, the study also explores how the processing of linguistic resources are shared across the modalities of listening and reading. The question of individual differences then can be reformulated to ask: how developed are a child's language resources when that child begins mapping print (decoding/word recognition) onto the phonology and semantics of language processing?

Given the above rationale, one would expect that one can learn about the progress toward successful integration of modalities of listening and reading by examining similarities or differences among tasks presented in different modalities. The design of the study explicitly allows us to explore this question, using a reasonable set of assumptions (described next). A more robust design could provide parallel reading and listening tasks at each grade (or at least the upper grades), but we feel that such a study is a follow up to the current study design, not necessarily the first study one should attempt.

We made some assumptions in designing the modalities of the tasks at each grade level. Kindergarteners, on average, have not yet learned to read continuous text, so one can reasonably expect that there would be near floor performance if we did not read the text to them. In contrast, we can also reasonably expect that if students cannot answer a question in listening to the text/story, then they also would not be able to answer it when reading the text in print.

We assumed that first graders are learning to read and some of the more advanced students may have print reading skills. Hence, by giving them a simultaneous listening and reading version, we maximize the processing and cue information available (across both reading and listening modalities). In the current design, students have access to the visual form of the text, so if their attention or listening fails, they can use their nascent reading skills as back-up. If they still cannot do a task, then we might infer that the source of difficulty could be in their core language resources, their developmental level (too hard for that age/grade), lack of key background knowledge, or a possible item flaw.

The same logic works for second graders, with the exception that we ask them to read aloud, given that this has been found to be a support for reading comprehension, not an adaptation that degrades it (Jenkins & Jewell, 1993) and that second graders, mostly, are not yet silent readers anyway. At some point in development, requiring children to read aloud may interfere with performance (Laitusis et al., 2008), but not at second grade. This approach provides the students with two sources of information (written and listening to their own speaking) to maximize the potential that they can demonstrate their level of comprehension.

Finally, at third grade, we expected that children should be achieving sufficient silent (or subvocal) reading skills to apply their language skills without any oral or read aloud support. Again, if they do not have the underlying language skills, then they would still get the items wrong. Inadequate print skills are now also a potential explanation of low performance (as in 2nd grade), but given the relatively high performance on items we will report in this study, we have evidence that for most of the children in this sample, the print skills (and underlying language resources) were adequate.

Therefore, we can reasonably compare results across modalities and grades, while acknowledging the differences in processing across modalities. If a kindergarten child cannot do an item when listening, we assume that they cannot do it when reading. If a third grader can do an item when reading, then we may also assume that they could answer correctly when listening (or that is, they have the language resources needed to do the item). But for this to work, the items need to align validly with what they are intended to measure. This is evaluated in cross grade comparisons and detailed item analyses.

Research Questions

In this study, we address the following research questions:

1. What are the reliability and basic item properties of the SBA form at each grade level?

2. What is the contribution of BK to comprehension scores when other factors are controlled statistically using linear regression modeling?

3. How frequently do students in the early grades, K-3, respond correctly to background questions posed before and after listening to (or reading) a passage designed for Scenario Based Assessment? Do other patterns ever occur?

4. What did we learn about how items performed across grade levels, and how did this interact with development and modality (e.g., listening versus reading)?

Methods

Participants

A total of 400 Kindergarten (194 female, average age was 6.2 years), 442 first grade (227 female, average age was 7.1 years), 485 second grade (226 female, average age was 7.9 years) and 141 third grade (77 female, average age was 9.2 years) students participated in the study. Ten percent of the birth date information and 2.8% of the gender information was missing for total sample; thus, total gender counts and ages are approximate. One hundred ten kindergarten (50 female), 47 first-grade (24 female), 80 second-grade (41 female) and zero third-grade students were excluded from the analyses, because a) they did not complete all items on the assessment or b) their session did not follow the standard procedures described below. The data were collected in the Spring of 2014 between March and June. The participating students represented a convenience sample drawn from a mix of rural, suburban, and urban schools in seven states in the Midwest, Southwest, Northeast and Southeast of the United States. All students had parental consent to participate. The data were kept confidential and shared only with study staff.

Participants were tested as part of a larger, ongoing intervention study investigating new reading component measures, although some students were recruited solely for the purpose of piloting the current assessment. Schools that were recruited solely for piloting the current assessment were given $15 per student session. To our knowledge, none of the students in the study were in special education, had any diagnosis of speech-language impairment, or were considered to have a pervasive developmental disorder. However, schools were not obligated to report student disability status to us when providing the list of consented, general population students we had requested. As such, we cannot be certain that no students from these subpopulations participated in the study.

Materials

Two assessment forms were created. The original form, designed for third-grade students, consisted of 15 items designed to measure background knowledge (4 general items about animals and 11 specific items about chickens), a narrative story of 557 words (67 sentences), and 35 items designed to measure literal comprehension, inferential comprehension (ability to integrate information across multiple sentences and elaborate beyond the text), vocabulary paraphrasing, fact vs. opinion decision making, and sequencing (i.e., text structure) abilities. Fifteen items were embedded within the story on the page in which the information appeared, and 20 items appeared after the story ended. In the current study, this form was used for children in both second and third grade.

The story received a Text Complexity Score of 310 using the Text Evaluator Tool (Sheehan, 2015; Sheehan, Flor, Napolitano, & Ramineni, 2015) and was classified as a Literary text. A score between 100 – 525 corresponds to the Common Core Grade Level of 2 (Sheehan, 2015). The Flesch-Kinkaid grade level estimate was 2.1. The story was about a girl who visits her cousin on a farm where they have chickens. In the story, the girl learns several facts about chickens, such as what differentiates a male from a female, where and how they sleep, what and how they eat, and how they bathe. While the story has a narrative structure with dialogue, humorous elements, and simple pictures, there are many factual elements that could be learned through an expository format. Therefore, the genre for this text could be considered an informational narrative.

The kindergarten and first grade form was adapted from the original form and modified by omitting three of the specific background knowledge items and 10 of the story-related items. There were 12 background knowledge items and 25 items based on the story. Specifically, the items assessing paraphrase, fact vs. opinion, and sequencing did not appear in the kindergarten and first grade form, as these concepts were judged to be in advance of this age group's typical curriculum.

Procedure

Testing sessions were administered one-on-one in a quiet location within the school with a trained administrator. The assessment was administered on either an iPad running Filemaker Go or a laptop running Filemaker. The assessor sat one-on-one with the student to make sure the student was able to respond to the items and continue to the next item. The administrator would provide support if a student appeared to have trouble operating the electronic device while answering questions.

Students were informally screened for basic hearing and vision requirements to determine their eligibility for inclusion in the study. The kindergarten and first-grade sessions lasted approximately 20 minutes; the second- and third-grade sessions lasted approximately 30 minutes. The session was structured like a shared storybook reading experience. For the kindergarten and first-grade students, the story and all questions were read aloud to the students by the assessor. Assessors only read the introduction of the assessment and background knowledge questions aloud to the second- and third-grade students. Once the story was presented, the second-grade students read the story aloud to the assessor and the third-grade students read the story silently. For kindergarten and first-grade students, the assessor made responses on the iPad or laptop for the students, but second- and third-grade students were allowed to respond to each item with oversight from the assessor. All items were scored automatically by the program.

If the assessors felt the students were struggling or experiencing stress because their native language was not English, these students were allowed to respond in their native language. However, students who were given these accommodations were dropped from the analyses for the current paper, because their administration deviated from the standard protocol.

Results

1. What are the reliability and basic item properties of the form at each grade level?

Table 1 shows the mean and standard deviation of the raw scores for each form at each grade level as well as the percent correct and score reliability (Cronbach's alpha). The first row in the table presents the descriptive statistics for the third grade form based on the data from a large scale, vertical scaling study conducted independently of this study. These statistics are included as a reference for the purpose of comparison. The item difficulties (percent correct) across grades for the shortened form ranged from .44 to .94, with a mean difficulty of .71.

A summary of the item-total correlations (ITCs) are presented in Table 2. These are the point-biserial correlations between the item responses to a given item and the total score. Conceptually, ITCs are an indicator of item discrimination with higher ITCs corresponding to more discriminating items. In operational testing, items with ITCs below .1 or .2 (poorly discriminating items) are generally evaluated to see if they can be revised or should be omitted from the test. Items with negative ITCs are typically excluded. Very easy or difficult items may have low ITCs due to a lack of variability in the responses; however, the decision to exclude an item should not be made solely on the magnitude of the ITC. Some items with low ITCs on our form were deemed valuable and were retained as is. For example, one item tested a necessary piece of information that almost all participants typically know. Although the ceiling performance on the item resulted in a low ITC, the content of the item was deemed important and worth retaining as a screen to identify the few individuals who may lack this knowledge.

Figures 1 and 2 show histograms of raw score distributions for the total sample and for each grade level on the reduced set of 25 items from the kindergarten and first-grade form.

Table 1

Performance on the Full and Shortened Form by Grade.

Chickens form	N	Items	Mean	SD	% Correct	Reliability
Full Form						
Large Scale study	1024	35	24.74	5.66	70.7	0.82
Third Grade	141	35	29.44	3.41	84.1	0.68
Second Grade	405	35	25.56	4.33	73.0	0.71
Shortened Form						
Third Grade	141	25	21.62	2.67	86.5	0.66
Second Grade	405	25	18.93	3.31	75.7	0.67
First Grade	395	25	16.29	3.46	65.2	0.63
Kindergarten	290	25	13.80	3.03	55.2	0.44

Note. All the third and second grade students took the full form. In the analyses that follow, we use the shortened form totals (25 items) for all grades, so that means are comparable across grades.

Table 2

Distribution of Item-Total Correlations by Size for Each Grade

	Item-Total Correlations		
Grade	<.10	.10 - .20	> .20
Full Sample	1	6	18
Third Grade[a]	3	7	14
Second Grade	0	11	14
First Grade	4	7	14
Kindergarten	9	7	9

[a] One ITC for third grade could not be computed due to zero variance (all students got the item correct)

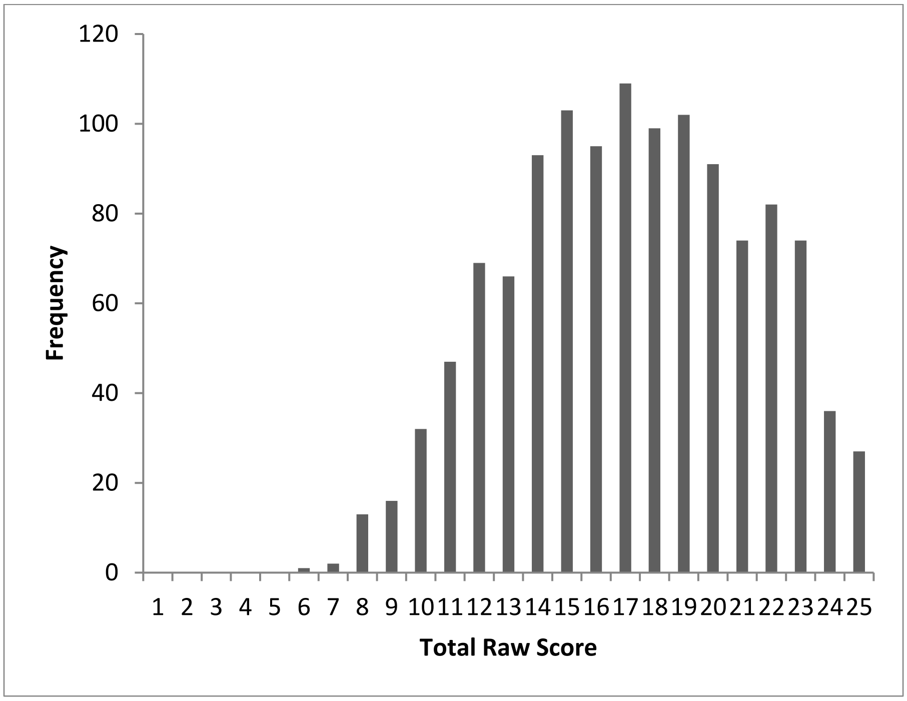
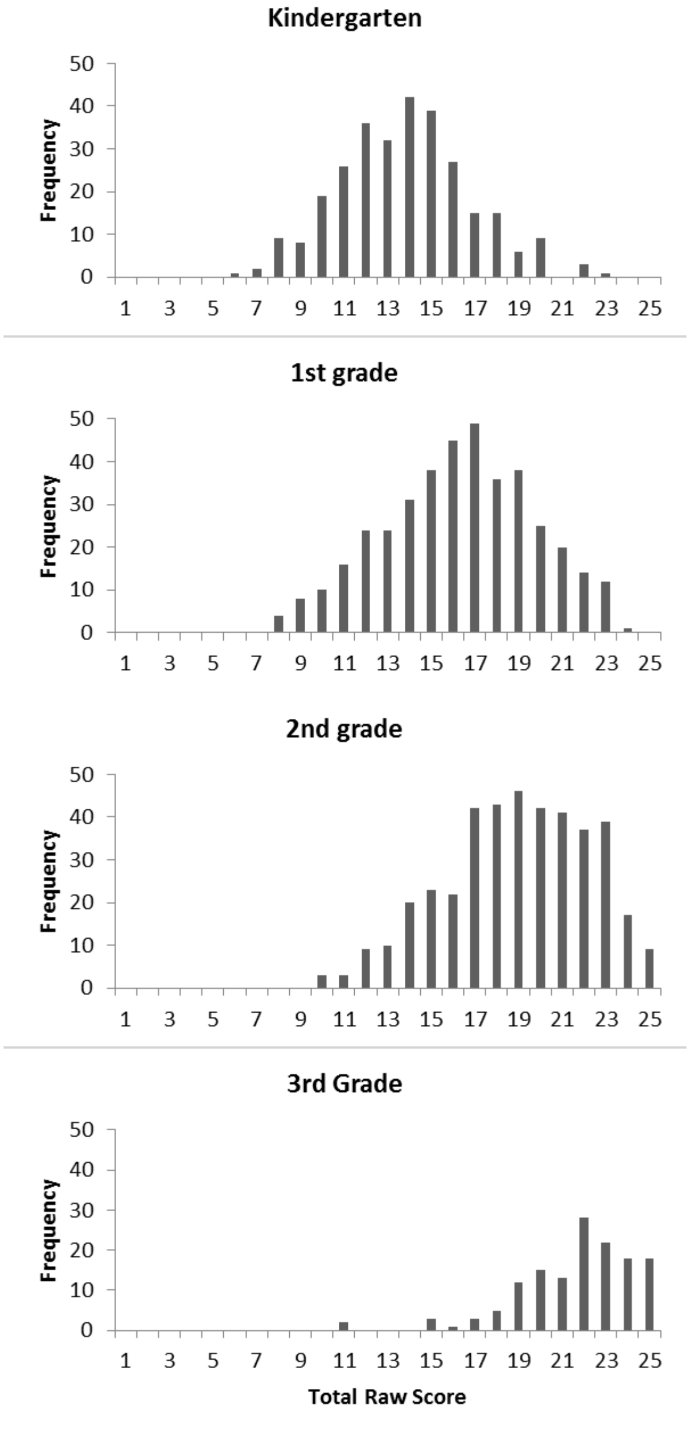


Figure 1. The distribution of Total Raw Scores for the 25 items on the kindergarten and first grade form for the entire sample of kindergarten – third-grade students.



Kindergarten

1st grade

2nd grade

3rd Grade

Figure 2. Distributions of Total Raw Scores for the 25 items on the kindergarten and first-grade form for each grade.

Young children's knowledge about chickens across grades. Pearson correlations were computed to examine the relationship of overall performance (raw total score on the comprehension section) and specific background knowledge about chickens (raw total score on the BK items). The correlation across all grades for the entire sample was moderate and positive (r(1229) = .38, *p* < .001). Correlations were also computed separately for each grade to determine whether the relationship between comprehension and specific chicken background knowledge differed across grades. Third- and first-grade exhibited the strongest relationship between comprehension and background knowledge (r(139) = .36, *p* < .001 and r(393) = .32, *p* < .001, respectively). The correlation was positive for second-grade students, albeit weaker (r(403) = .21, p < .001). For kindergarten students, the correlation was also positive, but even weaker than the other grades (r(197) = .16, *p* < .01).

2. **What is the contribution of BK to comprehension scores when other factors are controlled statistically using linear regression modeling?**

Two fixed-effects regression models were run to evaluate the effect of background knowledge on comprehension. For both models, the raw total score for comprehension was treated as the dependent variable. For Model 1, only grade was included as an independent variable. Grade was a significant predictor with a standardized effect of .61 (see Table 3). That is, for each increase in grade, mean performance is expected to increase by .61 standard deviation units. This model explained 37% of variability in comprehension scores. For Model 2, grade and BK raw total score for the chicken-specific items were included as predictors. Both of these variables were significant predictors. An additional 4% of the variability in comprehension was explained by this model. To illustrate the effect of BK, consider two first-graders, one with low BK (total score = 3) and one with high BK (total score = 9). The first student would be expected to receive a comprehension score of around 16 ≈ 12.15 + 2.31 + (3 × .57), whereas the second student would be expected to receive a comprehension score of around 20 ≈ 12.15 + 2.31 + (9 × .57). This is around a 16% increase in comprehension performance.

Table 3

Regression Results

Variable	B	SE(B)	β	t	R^2
Model 1					
Intercept	13.74	.16			
Grade	2.60	.10	.61	27.01*	.37*
Model 2					
Intercept	12.15	.23			
Grade	2.31	.10	.54	23.52*	
Specific Chickens BK	0.57	.06	.21	8.97*	.41*

*p < .001

3. How frequently do students in the early grades, K-3, respond correctly to background questions posed before and after listening to (or reading) a passage designed for Scenario Based Assessment? Do other patterns ever occur?

Building on the background knowledge analyses, Table 4 shows three questions that we asked all children both before and after the comprehension assessment. Of these, the question about whether chickens use dirt to take a bath has special status. The story includes the surprising fact that chickens do indeed take ‘dirt’ baths, as a way of helping to rid themselves of insects. If children knew this fact beforehand, it would ruin the suspense and surprise element of the story, and we would expect that when we ask the question again after reading the story, they would likely get it correct (again). The other questions are also facts about chickens revealed in the story. We had no *a priori* prediction of which of these facts children of different ages might know, though we would expect an increase of knowledge of some of these facts with grade level.

Table 4

Percent Correct on Background Knowledge Items at pre-test.

	Grade			
Background Knowledge Item	K	1	2	3
What do chickens use to take a bath? Dirt? [**Yes**, No]	6	6	20	8
A female chicken is called a ____. [chick, rooster, **hen**, robin]	40	52	53	82
Chickens sleep in a ___. [**coop**, bed, stable, cave]	43	49	63	90

Note: **Bold** indicates answer used to calculate percent correct in table.

The actual pattern of prior knowledge across grades is shown in columns 2-5 and a developmental pattern is evident. Less than 10% of 3rd, 1st, and kindergarten children knew that chickens take dirt baths. That 20% of 2nd graders knew this fact is an anomaly; some segment of the second grade sample seemed to have learned about chickens prior to the assessment. That a female chicken is called a hen was known by about 82% of the tested 3rd graders, about half of the 1st and 2nd graders (53 and 52 percent respectively), and 40% of kindergarten children in this sample. Higher percentages of 3rd graders (90%) knew that a chicken sleeps in a coop, about 2/3rds of 2nd graders, half of 1st graders, and less in kindergarten.

We also predicted that many students who did not know the answer to the questions beforehand, would learn the answers by reading/listening to the chicken story, and therefore would be able to answer the question at post-test. Table 5 shows the percentages of students who initially got a given item wrong (first question of pair), but answered the item correctly after reading or hearing the story (second question of pair). Most of the children across grades did learn that chickens take dust baths; they answered the first item of the pair incorrectly during pretest, but then answered the item correctly at post-test. They were asked this question in two forms at post-test, 1) the same wording and 2) a reworded version in the form of a true/false question (i.e., Chickens roll in the dust to take a bath.). This change in wording did result in a small decrease in correct performance at post-test across all grades, in comparison to using the exact same wording of the question at pre- and post-test.

Table 5

Percent of Students Who Were Correct at Posttest Among Those Being Incorrect at Pretest

	K	1	2	3
Question Pairs	-/+	-/+	-/+	-/+
What do chickens use to take a bath? Dust? [Yes, No] What do chickens use to take a bath? [water, soap, dust]	84	90	96	97
What do chickens use to take a bath? Dust? [Yes, No] Chickens roll in the dust to take a bath. [True, False]	82	85	82	91
A female chicken is called a ____. [chick, rooster, hen, robin] A hen is a ____. [female chicken, rooster, bright red comb]	39	68	82	73
Chickens sleep in a ___. [coop, bed, stable, cave] Chickens sleep in a coop. [True, False]	80	79	89	93

4. **What did we learn about how items performed across grade levels, and how this interacted with development and modality (e.g., listening versus reading)?**

One-way ANOVAs were conducted to examine grade differences on overall performance and specific chicken background knowledge. The dependent variable was the total raw score for the items that tested information from the story and the specific chicken background knowledge. The results for each one-way ANOVA are presented in Table 6. Bonferroni pair-wise comparisons were made due to the differences in sample sizes between each grade. Each one-way ANOVA revealed a significant effect of grade for overall performance and background knowledge. Pairwise comparisons for overall performance and specific chicken background knowledge revealed a pattern such that each grade performed significantly better than the younger grade.

Table 6

One-way ANOVA Results for Overall Performance and Specific Chicken Background Knowledge Questions

	K	1st	2nd	3rd			
Variable	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	F(3, 1227)	η^2	Bonferroni
Overall Performance	13.8 (3.03)	16.3 (3.46)	18.9 (3.31)	21.6 (2.67)	244.34***	.37	K<1<2<3
Specific Chicken BK	2.9 (1.43)	3.2 (1.41)	3.7 (1.37)	4.5 (1.36)	50.39***	.11	K<1<2<3

As noted previously, we adapted the chickens SBA for use with earlier grades by adapting the administration. Specifically, we asked 2nd grade students to read the text and questions aloud to the administrator. For kindergarten and 1st grade children, the administrator read the text and questions aloud. These children had the same visible access to the text on the screen while it was being read and the text remained available for the first 15 items about the story.

To explore the effect of grade and test modality, we used a variation of differential item functioning (DIF), using the third-grade item parameters estimated in the previous large-scale vertical scaling study. In that study, the psychometric properties of all of the SBA forms across grades were examined and evidence was obtained to support the creation of a unidimensional (developmental), vertical scale. As part of the scaling process, item parameters were estimated via an item response theory (IRT) based approach. Based on the assumptions associated with IRT, the expected performance on an item for students in different groups should be the same, conditional on underlying ability. For example, if we take all of the second- and third-grade students who have roughly the same scale score and compare the expected probability of a correct response for the two groups, we should expect these probabilities to be the same or very similar. The extent to which they differ, particularly across the score range, suggests that the item is functioning differently between the groups. This is sometimes referred to as differential item functioning. By using the item parameters for the third-grade test from the large-scale study as the reference, we were able to examine the performance of students in the lower grades, again conditional on ability, to identify items that may not be functioning in the same way as they were intended on the third grade test.

To identify misfitting items, student scores across the score range were first binned to create several score groups (for each grade). The observed percent correct on each item in each of the score bins was then computed. Two statistics were computed to identify misfit: the root mean squared difference (RMSD) and mean absolute difference (MAD).[i] The RMSD provides an indication of how badly an item misfits overall, whereas the MEAND provides some indication of the direction of the misfit. Items with RMSD values greater than 0.15 and/or items with absolute MAD values greater than 0.10 were flagged as misfitting.

We found 16 items at the kindergarten level that had RMSD values less than a threshold value of 0.15, that is, they behaved as one would predict of a developmentally appropriate item given the third-grade reference group. Ten items were above threshold at kindergarten, and of these, only two were also above threshold at 1st grade, and one other at 2nd grade. On this metric, a little more than half the items performed adequately in the kindergarten sample, and the vast majority at first and second grades.

However, if we also examined items that were unexpectedly more or less difficult than predicted, using a heuristic of mean deviation greater than +/-0.10, then we find an additional six items at kindergarten, five at first grade, and one at 2nd grade. Negative values, suggest that the item is more difficult than expected given individual ability estimates; positive values the opposite.

Discussion

What are the reliability and basic item properties of the form at each grade level?

In comparison to the reference sample collected in the prior large-scale vertical scaling study, there were increased means and decreased variance in third-grade scores, as well as lower alpha reliabilities at all grade levels. We note that we observed higher means in the current third-grade sample collected in this study, in comparison to the reference sample. One explanation of this difference could be due to the timing of each data collection. The data in the current study were collected in the Spring, while the data in the large-scale vertical scaling study were collected in the Fall. Thus, maturation is one possible explanation for the higher scores in this sample, and should be considered in interpreting results.

We also note alpha reliabilities for this sample were lower than that in the large-scale study, both with the third-grade sample, as well as the other grade levels. Reliability in the kindergarten and first-grade samples is likely a result of the reduced number of items in the form for these grades. The low score reliability for the kindergarten form in particular suggests that the measure is not adequate for children in the current, shortened format.

When viewing the entire, cross-grade sample, we observe a relatively normal distribution of scores, with slight evidence of a ceiling effect in the third-grade and perhaps the second-grade subsamples. There seems to be a reasonable spread of scores within each grade (excepting the ceiling effect), and one can see that the mean increases at each grade. Recall that we omitted the items that we considered too challenging for the kindergarten and first grade students in these analyses. We also found more items with ITC values less than .20 in this sample. Examination of cross grade trends provided additional information about ITC values. As shown in Table 2, more items had low ITCs as grade level decreased. In the kindergarten sample, nine items had item-total correlations below .10, suggesting that they were not contributing substantively to the total score. Another seven items had item-total correlations between .1 and .2, suggesting a weak contribution. Later, we discuss sets of these items to explore why this may be, what we can learn about development across the early grades, and implications for SBA design.

The basic properties and reliability of the form as adapted for cross grade use can be seen as mixed. The alpha reliability for each grade in this sample is somewhat lower than what was computed from a larger third-grade sample, with part of this decrease explained by the decrease in the number of items used. The distributions of total score on the shortened form show good variability within and across grades, with a hint of a ceiling effect for the upper grades (that was not as evident in the reference 3rd grade sample, although the full 3rd grade form contained more challenging items). Differences observed on ITCs in comparison to the baseline sample also hint at some issues with whether the items are behaving similarly when adapted for use with the lower grade groups. Before we look at these issues more closely, we discuss results of how knowledge about chickens changed across grades and the impact of this on SBA performance.

What did we learn about young children's knowledge about chickens across grades?

We hypothesized that if a child knows a fact beforehand (as evidenced by answering the question correctly prior to taking the SBA), then they would still know it after reading the story. This was largely true with a couple of exceptions. Fewer than 5% of students across all grades who knew that chickens take dust baths, or that chickens sleep in a coop at pre-test, answered the corresponding item incorrectly after reading and learning about chickens in the comprehension test. However, for kindergarten and first-grade children (who listened versus read the questions), we found larger error rates (19 and 11% respectively) when asked about hens after hearing the story. We hypothesize that the higher errors may have been a result of the wording of the item in combination with the complexity of listening to the set of choices (i.e., A hen is a ____. [female chicken, rooster, bright red comb]). In retrospect, simpler wording might have helped the item to perform better in a listening modality.

To summarize, we found several interesting results from asking young children what they know about chickens before providing a story that gave them the factual answers and asking them the questions again. First, some questions did interact with age/grade level. We found that most third graders knew that female chickens are called hens and sleep in coops, while only about 40-45% of kindergarteners knew these facts. However, the regression analyses suggest that it is not simply safe to assume what children know at any grade level, as prior knowledge still helps predict the total raw score on the assessment above and beyond grade level.

As to whether chickens take dust baths, only a small percentage of the children at any grade knew the fact beforehand, but most children were able to learn this fact by reading or listening to the story. Finally, we also learned that changing the wording or response type from pre to post-test did have some impact on the performance of the item (such as in the rewording between pre and post-test for the female chicken item), although the magnitude of the impact varied across items. This has implications for designing items and how best to evaluate them, particularly in the context of measuring learning within an assessment. We conclude that measuring the background knowledge of children that is subsequently to be presented in expository texts may be an important and useful technique for developing and validating assessment scores for young children.

What did we learn about how items performed across grade levels, and how this interacted with development and modality (e.g., listening versus reading)?

The results of the DIF analysis showed that the item characteristic were retained for some items across grades, while others showed significant changes in relative difficulty; more items were found to have a modality effect in the kindergarten subsample. This may be one reason for the lower internal consistency values for the kindergarten sample. It is valuable to examine some of the items with unexpected difficulty levels. Two of the most difficult items for kindergarten children asked about character feelings. The children were asked to discern whether a character in the story was sad, scared, or angry (p+ = 0.54, 0.78, 0.84, 0.95 respectively in K to 3rd grades), or in another instance, mad, interested, or bored (p+ = 0.52, 0.86, 0.95, 1.00 respectively in K to 3rd grades). In the context of this story, discerning emotions presented an additional challenge for Kindergarten children, although it was relatively easy from 1st grade on. This adds to the findings of individual differences in the development of children's emotional understanding (e.g., Pons, Lawson, Harris, & De Rosnay, 2003).

Another question referred to the fact that chickens make a lot of different sounds. The children were then asked to infer the reason chickens have so many different clucking sounds (i.e., because each sound means something different). It is expected that children as young as 2nd grade can make text based inference (Casteel, & Simpson, 1991). We found this to be true, but that the inference showed a very steep developmental improvement curve (across the listening vs. reading conditions). In short, the evidence suggests it is too difficult for kindergarten children (p+= 0.23, 0.47, 0.74, 0.91 respectively in K to 3rd grades).

One of the relatively easier items for kindergarten children was a vocabulary question that asked what the word prefer meant in the sentence context 'The chickens sleep in here at night because they prefer small spaces.' The difficulty values were p+ = 0.66, 0.48, 0.63, 0.74 for K to 3rd. This item appears to rely heavily on reading vocabulary. That is, reading the entire sentence in print, and seeing the perhaps strange visual word form of 'prefer', made this a more difficult item for 2nd and 3rd graders. Interestingly, it was more difficult for 1st graders than kindergarten children. We hypothesize that 1st graders may have been trying to read along, or simply were confused by the unknown word *prefer* when they heard it in the sentence. Kindergarteners may have simply chosen to ignore the print entirely, and focus on the simple understanding of the context -- that 'chickens like small spaces' (the correct answer). A similar logic may apply to the vocabulary question asking what perch means in the sentence, 'When chickens sleep, they perch on the roosts (p+= 0.69, 0.71, 0.66, 0.82). In both these cases, reading vocabulary in context is more challenging than listening to vocabulary in context, and we observe very little developmental difference between kindergarten and first grade samples.

In sum, if we only interpreted the ANOVA results alone, the SBA form would appear to demonstrate clear ability differences at each grade, as one would predict. This developmental difference occurs despite the fact that 3rd graders are reading silently, 2nd graders are reading aloud, and Kindergarten and 1st graders are listening. To some extent, this demonstrates the viability of SBAs across the range, and provides some evidence that the format is feasible for even young, emergent reading children. Upon closer examination at the item level, however, the story becomes more complex. At the youngest grades, the item difficulties sometimes diverged strongly from expectations. Changing an item from reading to listening sometimes increased and sometimes decreased (or mediated) difficulty. In other cases, it would appear that language, knowledge, or reasoning skills of young children impacted item difficulty, though we will need more research and analyses to tease apart the specific differences at the root of these discrepant results. Whether items perform as expected has implications for whether the item is measuring the skill or construct intended, and therefore, whether we consider it a valid indicator of ability, above and beyond its generic psychometric properties.

General Discussion

Reading comprehension is a complex construct that involves the coordination of many processes (McNamara & Magliano, 2009; Perfetti & Adlof, 2012). Although there is no unified theory of reading comprehension (Cain & Parrila, 2014; Perfetti & Stafura, 2014), most theories agree that comprehension involves both adequate component skills (most of which are predominantly language-based abilities like phonemic awareness, decoding, vocabulary, morphology, grammar/syntax) and integration and coordination of those skills when reading (or listening) for understanding. The orchestration of skills characteristic of comprehension proficiency develops incrementally over time, as students mature and gain experience in building understanding of increasingly complex texts (or of spoken language). We agree with an interpretation of this growing proficiency as intimately intertwined with the growth and development of underlying language resources and ability, that is, the convergence of reading with language or linguistic ability, as reading of the written word becomes integrated with other language functions (Vellutino et al., 2007).

The level of complexity and variety of processes used as one learns to read has resulted in a diverse set of measures to track reading achievement over time. For students in the younger grades, these measures typically focus on foundational skills including word recognition, phonological decoding, fluency, vocabulary, syntax and other language-specific measures. This historical tradition has resulted in the production of reliable and widely used component reading and language measures for use in school and clinical settings.

In recent years, however, educators, researchers, policy makers and the proponents of large scale assessment reforms have documented how the world of print literacy has been changing, mostly as a consequence of the widespread use of digital technologies and devices. They have articulated or provided recommendations for an updated construct of reading comprehension that takes into account changes in our expectations of what it means to be a proficient reader (Goldman, 2012; NGA, CCSSO, 2010, Partnership for 21st Century Skills, 2008). This elaborated construct reflects new demands on students, as they are asked to read and learn from a more diverse set of texts and perform a wider range of purpose-driven tasks, often in a digital environment. The precursors of this new construct, should also be extended downward developmentally, to a new generation of beginning readers growing up in this digital era.

In this paper, we described a new type of assessment, called scenario-based assessment, that was designed to address some of the challenges associated with a changing, expanding construct. SBAs employ techniques to deliver a set of goal oriented, thematically related texts and tasks in a structured, sequenced fashion. While prior work has indicated that this technique is viable for late elementary, middle and high school students, the current study was designed to determine if the approach could be modified appropriately for younger children.

In particular, we wanted to determine what elements of a modified version of an SBA form worked similarly or different with children across kindergarten to 3rd grade. We attempted to maintain as many of the construct and design elements of the SBA form as feasible. With this goal in mind, we measured a range of basic reading skills (literal comprehension, paraphrasing and vocabulary), as well as more demanding skills that involve text structure (sequencing), or drawing inferences across multiple parts of a text.

In general, the assessment for students in third, second, and first grades showed promise in terms of item difficulty, discrimination, though internal consistency reliability was somewhat low for this study sample. While the properties in Kindergarten were less robust, the item analyses indicate several aspects of the comprehension tasks that were viable even with this young, mostly non-reader group, while other aspects clearly require more thought and effort before they are ready for operational use.

One novel aim in this study was to examine how background knowledge and comprehension might interact in an expository assessment context with young children. We found that comprehension scores were associated with the level of students' background knowledge. In particular, there was a moderate correlation between student knowledge of chickens and their comprehension scores. While knowledge seemed to increase over grade level, a linear regression revealed that the effect of knowledge was significant even after student's grade level was accounted for.

The value of measuring background knowledge in a reading test was further bolstered by tracking how students learned over the course of the assessment. When select background knowledge items were presented after students read or listened to the texts, a significant proportion of students learned new material. This design feature could potentially augment the test by allowing the users to measure reading comprehension and learning, as well as the level of student knowledge. These results are generally consistent with the rich literature collected with older students (e.g., Evans et al., 2001). We recommend more research aimed at further understanding the interactions of prior knowledge in young children, especially in the context of expository reading and learning.

Finally, we reviewed literature that discusses the language basis of reading comprehension, suggesting that the skills of reading converge with other language skills as adequacy or proficiency of word recognition develops (Perfetti, 2003; Vellutino et al, 2007). This provided a rationale for exploring adapting listening versions of the same print based reading SBA forms used with 3rd grader readers for younger children. Using IRT and a novel application of DIF analyses, we observed items and tasks that appeared to show stability of parameters and the expected developmental changes in difficulty across grades, while other items seem to function differentially in listening versus reading modalities. We recommend further cross-modality research at the test and item level to better understand how young children integrate and apply their language skills as they learn to read and comprehend printed text developmentally.

Limitations and Conclusions

We note several limitations in the current study. First, the sample was one of convenience, and thus caution should be exercised in generalizing results. Most importantly, we cannot be assured that the cross grade ability distributions are comparable. The cross-grade ability distributions appeared reasonable, however, further research is needed to control for possible contributing factors other than grade. For example, we note that the third-grade cohort in this study may have been shifted towards the higher ability range than a reference sample (though we also noted that this could have also been explained by the late year, Spring administration in this study versus earlier in the year for the reference sample).

Second, we excluded students who needed English language supports during sessions, and were not certain that schools excluded students in special education, with language-based impairments, or other special population differences. In the future, we need to be more stringent in the information we collect from schools. Third, the reliabilities of the assessments were not at levels we would deem as acceptable for operational use and were well below levels we have observed in other SBAs for elementary, middle, and high school samples. Reliability was especially low in the kindergarten sample. A persistent challenge for this age group is testing time. These young students require briefer sessions which naturally limits the amount of testing material and items that can be covered and this may impact reliability. The reliability of an assessment such as the one used in this study for young children may tend to be lower if administered in a single session. For SBA to be a viable assessment option for young children, adequate reliability of measurement must be achieved.

Fourth, while it was surprising that some children were able to handle more complex items that involved inferring and integrating across multiple sentences, some of these items showed evidence of simply being too difficult for the kindergarten and first-grade samples. This may require additional revisions to ensure the items are more appropriate for the grade levels intended. Finally, the design decisions in adapting across listening versus reading modalities need to be evaluated further. Issues of working memory load, attention, and differences in the informational cues associated with each modality need to be carefully considered both in the design and administration of tasks, and in the capture of responses. Along these lines, it seems necessary for the older grades, that both reading and listening versions be prepared, to directly compare differences across modality.

Despite these shortcomings, we find reason to be encouraged in this first trial of SBAs with young children. We learned a great deal about what worked, what did not, and why. Anecdotal reports from study staff and assessors (not always our own staff) suggest many of the children at all grade levels enjoyed participating in the sessions. They were engaged with the story and items. These reports suggest that we were able to create an approachable and naturalistic scenario for an early literacy assessment and experience. Despite the limitations discussed above, the students' receptiveness to the experience and the results of this preliminary study are encouraging and suggest some potential for using more authentic and integrated tasks for measuring reading ability with younger children.

References

Alexander, P., Sperl, C., Buehl, M., & Chiu, S. (2004). Modeling domain learning: profiles from the field of special education. *Journal of Educational Psychology, 96*, 545-557.

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research & Perspectives, 8,* 70–91.

Bennett, R. E. (2011). CBAL: Results from piloting innovative K–12 assessments (Research Report No. RR-11-23). Princeton, NJ: Educational Testing Service.

Bransford, J., Brown, A., & Cocking, R. (2000). How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press.

Britt, M. A., & Rouet, J. F. (2012). Learning with multiple documents: Component skills and their acquisition. In J. R. Kirby & M. J. Lawson, (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 276-314). New York, NY: Cambridge University Press.

Cahalan-Laitusis, C., Cook, L., Cline, F., King, T., & Sabatini, J. (2008). Examining the impact of audio presentation on tests of reading comprehension. Princeton, NJ: Educational Testing Service.

Cain, K., & Parrila, R. (2014). Introduction to the special issue. Theories of reading: what we have learned from two decades of scientific research. *Scientific Studies of Reading, 18*, 1-4.

Casteel, M. A., & Simpson, G. B. (1991). Textual coherence and the development of inferential generation skills. *Journal of Research in Reading*,*14*(2), 116-129.

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language & Hearing Research, 49*(2), 278-293.

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*(4), 331-361.

Catts, H. W., Hogan, T. P., & Fey, M. E. (2003). Subgrouping poor readers on the basis of individual differences in reading-related abilities. *Journal of Learning Disabilities, 36*(2), 151-164.

Coiro, J. (2009). Rethinking reading assessment in a digital age: How is reading comprehension different and where do we turn now? *Educational Leadership, 66*, 59–63.

Coiro, J. (2011). Predicting reading comprehension on the internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of Literacy Research, 43*, 352-392

Evans, J. E., Floyd, R. G., McGrew, K. S., & Leforgee, M. H. (2001). The relations between measures of Cattell–Horn–Carroll (CHC) cognitive abilities and reading achievement during childhood and adolescence. *School Psychology Review, 31*, 246–262.

Gerrig, R. J., & O'Brien, E. J. (2005). The scope of memory-based processing. *Discourse Processes, 39*(2-3), 225-242.

Goldman, S. R. (2004). Cognitive aspects of constructing meaning through and across multiple texts. In N. Shuart-Ferris & D. M. Bloome (Eds.), *Uses of intertextuality in classroom and educational research* (pp. 317–351). Greenwich, CT: Information Age Publishing.

Goldman, S. R. (2012). Adolescent literacy: Learning and understanding content. *The Future of Children, 22*, 89-116.

Gordon Commission (2013). To assess, to teach, to learn: a vision for the future of assessment. Princeton, NJ: Author. Retrieved from: http://www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

Gough, P.B., Tunmer W.E. (1986). Decoding, reading, and reading disability. *Remedial & Special Education, 7,* 6–10.

Gough, P. B., Hoover, W. A., Peterson, C. L., Cornoldi, C., & Oakhill, J. (1996). Some observations on a simple view of reading. *Reading comprehension difficulties: Processes and intervention*, 1-13.

Griffin, C. C., Malone, L. D., & Kameenui, E. J. (1995). Effects of graphic organizer instruction on fifth-grade students. *The Journal of Educational Research, 89*(2), 98-107.

Hacker, D. J, Dunlosky, J., & Graesser, A. C. (2009). Handbook of metacognition in education. Mahwah, NJ: Erlbaum.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2*, 127–160.

Hoover, W. A., & Tunmer, W. E. (1993). 1 The Components of Reading. *Reading acquisition processes*, *4*, 1.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children*, 59(5), 421-432.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, *99*(1), 122.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.

LaRusso, M., Kim, H. Y., Selman, R., Uccelli, P., Dawson, T., Jones, S., .., & Snow, C. (2016). Contributions of Academic Language, Perspective Taking, and Complex Reasoning to Deep Reading Comprehension. *Journal of Research on Educational Effectiveness, 9*(2), 201-222.

Leu, D. J., Kinzer, C. K., Coiro, J., Castek, J., & Henry, L. A. (2013). New literacies: A dual level theory of the changing nature of literacy, instruction, and assessment. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (6th ed., pp. 1150-1181). Newark, DE: International Reading Association.

Linderholm, T., Virtue, S., Tzeng, Y., & van den Broek, P. (2004). Fluctuations in the availability of information during reading: Capturing cognitive processes using the landscape model. *Discourse Processes, 37*, 165-186.

McCrudden, M. T., Magliano, J., & Schraw, G. (Eds). (2011). Text relevance and learning from text. Greenwich, CT: Information Age Publishing.

McNamara, D. S. (Ed.). (2012). Reading comprehension strategies: Theories, interventions, and technologies. Mahwah, NJ: Erlbaum.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*, 247-288.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of learning and motivation, 51*, 297-384.

Meyer, B. J., & Ray, M. N. (2011). Structure strategy interventions: Increasing reading comprehension of expository text. International Electronic Journal of Elementary Education, 4(1), 127.

Murphy, K., & Alexander, P. (2002). What counts? The predictive powers of subject-matter knowledge, strategic processing, and interest in domain-specific performance. *The Journal of Experimental Education, 70*, 197-214.

National Governors Association Center for Best Practices & Council of Chief State School Officers (NGA & CCSSO) (2010). Common Core State Standards for English Language Arts. Washington, DC: Author.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes 43*, 121-152.

O'Reilly, T., & Sabatini, J. (2013). Reading for Understanding: How Performance Moderators and Scenarios Impact Assessment Design (Research Report No. RR-13-31). Princeton, NJ: Educational Testing Service.

O'Reilly T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing Reading Comprehension Assessments for Reading Interventions: How a Theoretically Motivated Assessment Can Serve as an Outcome Measure. *Educational Psychology Review, 26*, 403-424.

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction, 25,* 399-438.

Partnership for 21st Century Skills. (2008). 21st century skills map. Washington, DC: Author. Retrieved from http://www.p21.org/storage/documents/21st_century_skills_english_map.pdf

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

Perfetti, C. A. (2003). The universal grammar of reading. *Scientific studies of reading*, *7*(1), 3-24.

Perfetti, C. A., & Adlof, S. M. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. In J. P. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 3-20). Lanham, MD: Rowman & Littlefield Education.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*, 22-37.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The Acquisition of Reading Comprehension Skill. *The Science of Reading: A Handbook*, 227-247.

Pons, F., Lawson, J., Harris, P. L., & De Rosnay, M. (2003). Individual differences in children's emotion understanding: Effects of age and language. *Scandinavian journal of psychology*, *44*(4), 347-353.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*(2), 31-74.

Sabatini, J., & O'Reilly, T. (2013). Rationale for a New Generation of Reading Comprehension Assessments. In B. Miller, L. Cutting, & P. McCardle (Eds.), *Unraveling Reading Comprehension: Behavioral, Neurobiological, and Genetic Components*, (pp. 100-111). Baltimore, MD: Brookes Publishing.

Sabatini, J., O'Reilly, T., & Deane, P. (2013). Preliminary Reading Literacy Assessment Framework: Foundation and Rationale for Assessment and System Design. (Research Report No. RR-13-30). Princeton, NJ: Educational Testing Service.

Sabatini, J., O’Reilly, T., Halderman, L. & Bruce, K. (2014a). Broadening the Scope of Reading Comprehension using Scenario-based Assessments: Preliminary Findings and Challenges. *International Journal Topics in Cognitive Psychology, 114*, 693-723.

Sabatini, J., O’Reilly, T., Halderman, L. & Bruce, K. (2014b). Integrating Scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disabilities Research & Practice, 29*, 36-43.

Sandak, R., Mencl, W. E., Frost, S. J., & Pugh, K. R. (2004). The neurobiological basis of skilled and impaired reading: Recent findings and new directions. *Scientific Studies of Reading, 8*(3), 273-292.

Schneider, W., Körkel, J., & Weinert, F. E. (1989). Domain-specific knowledge and memory performance: a comparison of high- and low aptitude children. *Journal of Educational Psychology, 81*, 306–12.

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal, 41*, 159–189.

Sheehan, K. M. (2015). Aligning TextEvaluator® Scores With the Accelerated Text Complexity Guidelines Specified in the Common Core State Standards (Research Report No. RR-15-21). Princeton, NJ: Educational Testing Service.

Sheehan, K. M., Flor, M., Napolitano, D., Ramineni, C. (2015). Using TextEvaluator® to quantify sources of linguistic complexity in textbooks targeted at first-grade readers over the past half century. (Research Report No. RR-15-38). Princeton, NJ: Educational Testing Service.

Stanovich, K. E., Siegel, L. S., & Gottardo, A. (1997). Converging Evidence for Phonological and Surface Subtypes of Reading Disability. *Journal of Educational Psychology, 89*(1), 114-127.

Taboada, A., Tonks, S., Wigfield, A., & Guthrie, J. (2009). Effects of motivational and cognitive variables on reading comprehension. *Reading and Writing, 22*, 85–106.

Van den Broek, P., Lorch, R. F., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29*, 1081-1087.

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific studies of reading*, *11*(1), 3-32.

Vitale, M., & Romance, N. (2007). A knowledge-based framework for unifying content-area reading comprehension and reading comprehension strategies. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theory, interventions, and technologies* (pp. 73–104). Mahwah, NJ: Lawrence Erlbaum Associates.

Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology,22*(3), 407-423.

[i] The RMSD characterizes the average squared difference between the observed percent correct and the expected probability based on the item parameters across all of the score bins for a particular item in a given grade; the MEAND is simply the average difference between the observed and expected percent correct across the score bins.