Hammill Institute on Disabilities

Assessment for Effective Intervention
2017, Vol. 42(4) 209–223

DOI: 10.1177/1534508417694121
aei.sagepub.com
SAGE

A Curriculum-Based Measure of Language Comprehension for Preschoolers: Reliability and Validity of the Assessment of Story Comprehension

Trina D. Spencer, PhD[1], Howard Goldstein, PhD[2], Elizabeth Spencer Kelley, PhD[3], Amber Sherman, MS[4], and Luke McCune, MA[5]

Abstract

Despite research demonstrating the importance of language comprehension to later reading abilities, curriculum-based measures to assess language comprehension abilities in preschoolers remain lacking. The Assessment of Story Comprehension (ASC) features brief, child-relevant stories and a series of literal and inferential questions with a focus on causal and predictive inference skills surrounding the main story grammar components and a novel vocabulary word. Following an overview of the iterative development process and pilot studies, this article presents preliminary evidence of the fidelity of administration, reliability of scoring, alternate form reliability, and validity of the ASC. In all, 237 preschoolers, ages 3 to 5 years old, participated in this study. Fidelity of administration and scoring reliability averaged over 90%. Concurrent validity with two established language measures revealed correlations ranging from .67 to .81. Test–retest reliability and internal consistency also indicated high levels of reliability for this new tool; however, alternate form reliability results suggest further work is needed. Preliminary results indicate that the ASC holds promise as a viable curriculum-based measure that early childhood educators can use for monitoring preschoolers' progress in language comprehension.

Keywords

language, early childhood, curriculum-based measurement, comprehension

Functional literacy is arguably the most important skill for academic success. Unfortunately, developing literacy skills is difficult for many students. More than 60% of fourth graders read below grade level, and for culturally and linguistically diverse subgroups, the proportion of students performing poorly on national reading tests is closer to 80% (National Center for Education Statistics, 2015). Reading comprehension relies on both decoding and language comprehension skills (Gough & Tunmer, 1986; Hoover & Gough, 1986; Storch & Whitehurst, 2002; Tunmer & Hoover, 1992). There is considerable research on early interventions to promote the development of decoding (e.g., Bailet, Repper, Murphy, Piasta, & Zettler-Greeley, 2013; Haager, Klingner, & Vaughn, 2007; Hurry & Sylva, 2007; Koutsoftas, Harmon, & Gray, 2009; VanDerHeyden, Snyder, Broussard, & Ramsdell, 2008), but there is limited research on early interventions designed to promote language comprehension (e.g., Spencer et al., 2012; Spencer, Petersen, Slocum, & Allen, 2015; Zucker, Solari, Landry, & Swank, 2013). Effective intervention to promote oral language development, including language comprehension, is as important as decoding interventions.

Early language experiences are important for later reading comprehension (Dooley & Matthews, 2009), making a focus on oral language an important component of early childhood education (Dickinson, Golinkoff, & Hirsh-Pasek, 2010; Whitehurst & Lonigan, 1998). Because preschool children are not yet readers, language comprehension, rather than reading comprehension, is the focus of instruction, intervention, and assessment. Skilled language comprehension as it relates to stories is linked to school achievement (Bishop & Edmundson, 1987; Feagans & Appelbaum, 1986) and specifically predicts later reading comprehension (Catts, Fey, Tomblin, & Zhang, 2002;

[1]Northern Arizona University, Flagstaff, AZ, USA
[2]University of South Florida, Tampa, FL, USA
[3]University of Missouri, Columbia, MO, USA
[4]The Ohio State University, Columbus, OH, USA
[5]Commerce Bancshares, Inc., Kansas, MO, USA

Corresponding Author:
Trina D. Spencer, Institute for Human Development, Northern Arizona University, P.O. Box 5630, Flagstaff, AZ 86011-5630, USA.
Email: Trina.Spencer@nau.edu

Dickinson & McCabe, 2001; Griffin, Hemphill, Camp, & Wolf, 2004).

Within the broader domain of language comprehension, inferential language is one subskill that contributes to later reading comprehension skills (Cain & Oakhill, 1999; Cain, Oakhill, & Lemmon, 2004; Kendeou, Bohn-Gettler, White, & Van Den Broek, 2008; Lepola, Lynch, Laakkonen, Silvén, & Niemi, 2012). Children's ability to make inferences is predictive of later reading comprehension abilities in school-age children, contributing unique information beyond what is explained by working memory (Cain et al., 2004), and children who have poor reading comprehension also have poor inference making skills (Cain & Oakhill, 1999). This is especially true for children with language difficulties. Although children with language impairment have difficulty answering literal and inferential questions about a story (Bishop & Adams, 1992), inferential language is particularly challenging for them (Blank, Rose, & Berlin, 2003; Ford & Milsoky, 2003). Van Kleeck (2008) argued that inference making contributes to later text comprehension by encouraging children to make connections between information in the text and their own knowledge. Indeed, text comprehension strategies taught to older children frequently include strategies for generating inferences and making connections with background knowledge. For younger children, strategies to promote inference skills are appropriately embedded in storybook reading activities.

Understanding and answering questions about stories is a common experience for preschool children. When reading storybooks with young children, adults use a range of literal and inferential language in a discussion about the book (Hargrave & Sénéchal, 2000; van Kleeck, 2008; van Kleeck, Gillam, Hamilton, & Cassandra, 1997). Examples of literal questions might include "What is this?" and "What is he doing?" while inferential questions might involve connecting events in the story to children's background knowledge such as "How do you think she feels?" "Why did he do that?" and "What do you think he will do next?" These types of questions also have been included in intervention strategies to improve language abilities in young children with strong effects (Cain & Oakhill, 1999; Cain, Oakhill, Barnes, & Bryant, 2001; Tompkins, Guo, & Justice, 2013; van Kleeck, 2008; van Kleeck, Vander Woude, & Hammett, 2006; Whitehurst et al., 1994). In studies of parent–child interactions during storybook sharing activities, children who were exposed to more inferential language used more inferential language and improved in inferential language abilities (van Kleeck et al., 1997). Importantly, children who participated in storybook sharing with inferential language had higher scores on measures of reading comprehension in the third grade than peers who were less engaged (Serpell, Baker, & Sonnenschein, 2005).

Given the evidence that inferential language is important for reading comprehension and that children with language difficulties may struggle with inferential language, it is important to assess both literal and inferential language abilities to identify children who may need additional support. Unfortunately, there are only a few existing preschool assessment instruments that address language comprehension; fewer that provide information about both literal and inferential language; and none designed to fulfill the purposes of universal screening and progress monitoring. To help educators address language comprehension of young children, new assessment tools are needed.

Curriculum-Based Measurement (CBM) Applied to Language

To effectively implement oral language intervention in educational settings, there is a need for assessment tools that can identify children who would benefit from additional instruction and to monitor their progress with respect to instruction. CBM refers to a specific set of standard assessment procedures that reflect curricular content addressed in classrooms (Deno, 2003; Deno, Mirkin, & Chiang, 1982; Missall & McConnell, 2004). CBM is often used to assist educators in making data-based decisions in differentiated instructional systems such as Response to Intervention (RtI) or Multi-Tiered Systems of Support (MTSS) educational models, which are extending into early childhood settings (Greenwood et al., 2011; Greenwood et al., 2013). CBM was designed to help educators identify which students may benefit from additional intervention, monitor students' progress once intervention has commenced, and determine when curricular objectives have been achieved (Deno, 2003; Fuchs, Fuchs, Hamlett, & Phillips, 1994). CBM has been an indispensable tool for promoting achievement in decoding (Christ, Zopluoglu, Long, & Monaghen, 2012), math (Foegen, Jiban, & Deno, 2007), and writing (McMaster & Espin, 2007), but few researchers have applied CBM to oral language (see Bradfield et al., 2014; Petersen & Spencer, 2012; Wackerle-Hollman, Rodriguez, Bradfield, Rodriguez, & McConnell, 2014).

Instructionally relevant assessment of language comprehension skills is needed to promote reading outcomes for a wide range of children (i.e., at risk to high achieving) and to identify children with particular deficits in those skills (Bagnato & Neisworth, 2005). In addition, the increasing prominence of MTSS models of differentiated intervention heightens the importance of CBM tools that (a) have strong reliability and validity, (b) are time efficient and easy to administer and score, (c) include alternate forms of standardized tasks, (d) measure socially important outcomes, and (e) are sensitive to growth due to intervention and change over time (Deno, 2003; Deno et al., 1982; Missall & McConnell, 2004). Teachers who employ CBM are more likely to identify students in need of intervention, adjust instruction to meet students' needs, and produce better student achievement (Fuchs, Deno, & Mirkin, 1984; Fuchs & Fuchs, 2007).

Existing Measures of Language Comprehension for Preschool Children

Only a few existing assessments that qualify as CBM tools address constructs subsumed under the umbrella of oral language. Set within an early childhood RtI context (Bradfield et al., 2014; Greenwood et al., 2011; Greenwood et al., 2013), Which One Doesn't Belong (WODB) is a new comprehension Individual Growth and Development Indicator (IGDI) designed to help early childhood educators identify children in need of comprehension intervention (Wackerle-Hollman et al., 2014). To administer WODB, an examiner shows an individual child a series of cards with three colored pictures and asks the child to identify the one that does not belong. It has 68 items and is scored using a total number of items answered correctly. It is a brief assessment that is easy to administer and engaging for the children. Validity correlations with the Clinical Evaluation of Language Fundaments–Preschool (CELF-P; Wiig, Secord, & Semel, 2004) range from .57 to .74 (Wackerle-Hollman et al., 2014). WODB is still in development and the authors consider the core construct to be inference. However, the nature of WODB is less authentic than story-based comprehension tasks and does not take into account different ranges of comprehension skills afforded by literal and inferential questions. Furthermore, the current version of WODB does not have multiple forms for progress monitoring.

A second related measure is the Test of Narrative Retell (TNR) for preschoolers (Petersen & Spencer, 2012). The TNR was specifically designed to be a curriculum-based measure of oral language skills for use in preschools using RTI/MTSS frameworks. It includes 25 short stories comprising multiple forms for repeated sampling of oral language. It takes approximately 2 min to administer; it can be scored in real time and it is not necessary that the scorer has advanced knowledge of language. It is brief and easy to use, and measures important oral language skills known to predict later reading comprehension (Catts et al., 2002; Fazio, Naremore, & Connell, 1996; Griffin et al., 2004). In a factor analysis, the TNR and its companion assessments, the Test of Story Comprehension (TSC) and the Test of Personal Generation (TPG), were investigated. Two factors were found—comprehension and production. The TPG loads exclusively on narrative production, the TSC loads exclusively on comprehension, and the TNR loads on both. As the TNR was intended to be a CBM tool (but the TSC and TPG were not), its psychometric properties have been examined more closely than the TSC and the TPG. The TNR correlates with other narrative retell measures (the Renfrew Bus Story, r = .88, and the Index of Narrative Complexity, r = .93; Petersen & Spencer, 2012) and the CELF-P (Wiig et al., 2004; r = .70; Spencer & Petersen, 2016). Exact scoring agreement for the TNR yielded a mean of 94% and the mean alternate form correlations was .77 (Petersen & Spencer, 2012). While the TNR is a promising measure for identifying preschoolers with oral language needs and monitoring the progress of these students, inference is not one of its core constructs. The TNR stories are very simple and brief without the necessary contextual engineering to assess inference skills related to comprehension.

Introduction to the Assessment of Story Comprehension (ASC)

To effectively identify children who need instruction and intervention in language comprehension, and to monitor the effects of these programs, there is a need for high quality assessment tools that can be administered with fidelity and scored reliably in authentic educational settings by busy preschool teachers. In this section, we describe a new measure to assess language comprehension for use in tiered models of instruction in early childhood education, called the Assessment of Story Comprehension (ASC; pronounced "ask").

Development of the ASC followed an iterative process (Diamond & Powell, 2011; Kern, Evans, & Lewis, 2011) that began with a review of the literature on language comprehension and initial creation of stories and questions following frameworks put forth by van Kleeck (2008) and Paris and Paris (2003) and principles of comprehension assessment in young children (Van den Broek et al., 2005). For the purpose of creating parallel forms for repeated administration, nine stories with consistent story grammar (i.e., character, setting, initiating event, emotion, attempt, resolution) and linguistic elements (i.e., coordinating conjunctions, temporal markers, causal and temporal subordination) were written. Each story has 158 to 160 words. One less common word (e.g., *injure*, *tidy*, *annoy*) was embedded in each story along with supportive clues (e.g., *After he cleaned his room*, *he looked around. His room was very tidy.*) so children can figure out the meaning of the word from context. Stories feature personally relevant experiences such as being bothered by a sibling, getting hurt, or breaking a toy. The purpose of using realistic topics was to help neutralize the effects of background knowledge across stories. All of the stories were written to support answering four inferential questions that included prestory prediction based on the title, causality between problem and feeling, explaining a character's motivation based on background information, and prediction about subsequent events (see Table 1 for framework for ASC questions). The last of the ASC's eight items requires children to provide a definition of the less common word or to choose between two definitions if they are unable to provide a definition. This item assesses a child's ability to infer the meaning of an unfamiliar word from the context clues embedded in the story.

Table 1. Framework for ASC Questions.

Questions		Descriptions	Examples
1	Inferential	Make a prediction based on the title	Let's think about the title, *Jenny and the Mud Puddle.* What do you think will happen?
2	Literal	Setting information (where or what)	Where was Jenny playing in this story?
3	Inferential	Infer causal relation between problem and feeling	In this story, Jenny was sad. Why was Jenny sad?
4	Literal	Attempt	Jenny's teddy bear fell in the mud. What happened next?
5	Inferential	Explain character's motivation using background knowledge	Why do you think Jim wanted to help Jenny?
6	Literal	Consequence/resolution	What happened at the end of the story?
7	Inferential	Make a prediction about subsequent events	The next time Jenny plays outside, do you think she will take her teddy bear? Why/why not?
8	Vocabulary	Define a word	Tell me, what does *filthy* mean?
8a	Vocabulary	Choose between two definitions	Does *filthy* mean very tall or very dirty?

Note. ASC = Assessment of Story Comprehension.

Within the primary construct of language comprehension, we focused on the ability to answer literal and inferential questions about stories. In the case of answering questions about stories, literal questions can be answered using the information explicitly stated in the text. For example, the literal question, "What was he doing?" could be answered using information from the story text (e.g., *Danny was riding his bike*). To answer inferential questions, information beyond what is presented explicitly in the text is necessary. In many cases, to respond appropriately, children will need to make a connection between story events and their own knowledge. For example, to respond to the inferential question, "Why was he sad?" children could make a connection between the events of the story (e.g., *Danny fell off his bike*) and their knowledge of emotions related to a similar event.

There were a number of reasons for including both literal and inferential questions in the ASC, even though the primary need is in the area of inference skills. First, this balance reflects the balance of literal and inferential language typical in adult–child book reading (Hammett, Van Kleeck, & Huberty, 2003; van Kleeck et al., 1997). Second, the literal questions help to broaden the scale so that it would be sensitive to developmental changes and to create a more normal distribution of scores among diverse children who may not be able to use inference. We wanted the ASC to be useful for younger children whose inference skills are only emerging and the inclusion of literal questions ensures a lower floor. Third, the variety in test items was intended to maintain children's attention and motivation during the administration without using pictures.

The ASC includes a single item related to vocabulary knowledge. Although vocabulary learning was not the primary construct of interest, this item was included because vocabulary knowledge is highly correlated with other oral language skills and reading comprehension (Cunningham & Stanovich, 1997; Scarborough, 2001; Storch & Whitehurst, 2002; Tunmer & Chapman, 2012). The ability to learn new words from stories is related to general language ability, and there is evidence that children with strong language skills are better at this type of incidental word learning than children with poor language skills (Cain, Oakhill, & Elbro, 2003; Cain et al., 2004; Daneman & Green, 1986; Nippold, 2002). The manner in which the ASC stories were written also ensures that to define the word correctly, the child has to use context information to generate a definition, which depends on inference skills (Cain et al., 2004).

To administer the ASC, examiners follow a simple standardized script. Scoring is standardized and, for Questions 1 to 7, involves rating each answer on a 3-point scale, where 2 points are given for clear and complete responses; 1 point is given for correct, but incomplete or unclear responses; and 0 points are given for incorrect answers. For the eighth item (i.e., "What does _____ mean?"), 3 points are possible if a complete and accurate definition is provided and 2 points are given for answers that are correct but incomplete, unclear, or an example of the word without a definition. In situations in which an incorrect definition is given or no response is provided, the examiner asks a follow-up question giving a choice between two definitions (i.e., "Does ___ mean ___ or ____?"). Correct responses are given 1 point and incorrect responses are given 0 points. A total of 17 points are possible for the ASC; 8 points for inferential questions, 6 points for literal questions, and 3 points for defining the less common word.

The intended purposes of the ASC are to identify children who could benefit from supplemental language intervention and to monitor language growth of children who participate in language intervention. The nine ASC stories and the consistent pattern of literal and inferential questions across stories represent nine parallel forms, the development of which was described above and the examination of which occurs in the current study. As a CBM tool, the ASC

was designed to be administered in two ways. First, a single ASC form can be administered on a schedule relevant to intervention (e.g., weekly or monthly) for monitoring language comprehension growth. Parallel forms are critical to fulfill this purpose of monitoring progress over time. If forms were not interchangeable, then growth (or lack of growth) could be attributed to variations between forms and not learning. Second, when important decisions are dependent on ASC results such as for benchmarking (fall, winter, spring) or the identification of children who would benefit from language intervention, three ASC forms are administered in a single session. The best or the median score of the three ASC forms is used for decision making. The reason for this is to reduce possible confounds (e.g., distraction, content unfamiliarity) that is common when assessing young children. The delivery of three ASC forms and selecting the median or best score are adaptations designed to maximize the validity of the ASC results while maintaining its authenticity. Regardless of how well a test performs, young children are less consistent test performers than school-age children. Their performance can be easily influenced by emotional, environmental, and motivational factors (Spencer & Slocum, 2010).

Current Study

Once initial stories and questions were developed, we conducted a series of small, pilot studies (Spencer & Goldstein, 2011). In the first study, 36 preschoolers received a random selection of three ASC forms and criterion measures. Trained undergraduate students served as testers and scorers. As a result of preliminary scoring reliability, fidelity, and validity analyses, we rewrote three stories, eliminated a question (there were initially five inferential questions), and developed story-specific scoring guides with examples drawn from children's responses. In a second pilot study with the revised ASC forms, undergraduate research assistants administered all nine ASC forms to 20 preschoolers in a random sequence. Parallel form reliability and validity results suggested that the ASC was a promising tool for measuring language comprehension. Minor revisions to the scoring guides were made and administration and scoring manuals were created.

In a third study, the ASC's sensitivity to intervention effects was examined. Kelley, Goldstein, Spencer, and Sherman (2015) implemented a 9-week intervention that targeted vocabulary and answering inferential questions about stories. The ASC was administered across four assessment waves; only one ASC form was administered for each wave to reflect progress-monitoring conventions. Because the ASC was not directly aligned with the intervention targets, it served as a distal outcome measure. Using a randomized control group design with repeated measures, researchers found a statistically significant group by time interaction ($F = 4.86$, $p < .01$) with a moderate effect size ($\eta_p^2 = .20$) for the inferential questions, but there were no group by time differences for the literal questions of the ASC. This evidence suggests that children made growth in the exact area of comprehension that was targeted in the intervention (i.e., inference skills).

Through these pilot studies, we established that it was possible to assess language comprehension within an authentic context such as listening to stories. It took 2 to 3 min to administer one ASC and approximately 8 min to administer three in one session. Scoring required less than 1 min per story. The administration fidelity and preliminary scoring reliability results were strong suggesting that the ASC was easy to learn. Results of the intervention study indicated that the ASC was sensitive to interventions targeting inference skills. Overall, the ASC appeared to be an efficient, economical, and easy-to-use tool for sampling key child behaviors such as listening to a story and answering questions about it. Its promising evidence of reliability and validity indicated that a more rigorous examination of its psychometric properties was warranted. Thus, to evaluate the ASC as a viable CBM tool, we examined its technical adequacy using a larger sample of preschool children. Specifically, the current study addressed the following research questions:

Research Question 1: To what extent can the ASC be administered with fidelity?
Research Question 2: To what extent can the ASC be scored reliably?
Research Question 3: To what extent does the ASC positively correlate with other measures of oral language?
Research Question 4: To what extent are the ASC forms and items reliable?

Method

Participants

Children attending preschools in a Western state were recruited to participate in this study. Anticipating that the ASC ultimately would be used in needs-based preschool environments like Head Start or district sponsored pre-kindergarten programs for students with risk factors, we recruited participants from eight Head Start classrooms. However, given the need to establish a more normative sample of children for proper validation, we also recruited children from five community-based, for-profit day care centers that implemented a preschool curriculum and one public school special education classroom for preschoolers with developmental disabilities. All of the classrooms served children between the ages of 3 to 5 years. To obtain parent permission, researchers spoke to the parents about

Table 2. Description of Participants.

Demographic	*n* = 230
Age in months, *M* (range)	45 (37–67)
Ethnicity, *n* (%)	
Caucasian	78 (34)
Latino/Hispanic	71 (31)
Native American	42 (18)
African American	7 (3)
Asian American	2 (1)
Multi-ethnic	23 (10)
Other	7 (3)
Dominant language, *n* (%)	
English	193 (84)
Other	25 (11)
Bilingual	12 (5)
Disability, *n* (%)	
Parent concern	16 (7)
Individualized education plan	21 (9)
Mother's education, *n* (%)	
Graduate degree	41 (18)
Bachelor's degree	39 (17)
Associate's degree	23 (10)
High school diploma or equivalent	92 (40)
Less than high school	28 (12)
Did not report	7 (3)

the study during drop off and pick up times rather than relying on the teacher to send home informed consent packets. Initially, 237 children attending the 14 different classrooms received parent permission to participate in the study, representing a 98% response rate. Only four families did not provide permission for their child to participate. For some questions, multiple data sets from a portion of the children were included because they participated in multiple ASC test sessions several months apart.

When parents/guardians provided permission, they also completed a brief survey about the child's age, ethnicity, dominant language, and mother's education and concerns about the child's language development. Table 2 shows the results of the demographic survey and includes a description of the participants. Seven parents/guardians did not complete a survey.

General Procedures

Each child received at least one ASC session (with three forms administered in a single session) and two other oral language measures—the Clinical Evaluation of Language Fundamentals Preschool–Second Edition (CELF-P; Wiig et al., 2004) and the TNR (Spencer & Petersen, 2011). A subset of participants ($n = 56$) received all nine ASC forms within 2 weeks (in three sessions) as well as the CELF-P and the TNR. Ten undergraduate research assistants served as examiners for this study and all testing took place in preschool classrooms or in hallways near the classrooms. Half of the research assistants were taught to administer the CELF-P and half were taught to administer the TNR; all were taught to administer the ASC. Given the logistical challenges of testing in preschool classrooms and that testers were trained to only administer the CELF-P or the TNR, not both, the order of ASC, CELF-P, and TNR administration was not preplanned (i.e., counterbalanced). The order of tests depended on the availability of research assistants and the cooperation of children. Thus, the order of test administrations varied. In addition, further analyses of order effects were not conducted because it was not possible to track the order of assessments for each participant given the available resources.

Prior to administering tests, the first author provided approximately 2 hr of training for all research assistants. The research assistants also read the testing manuals and practiced administering the tests. After several practice administrations, research assistants were required to demonstrate 100% fidelity of administration during check out sessions with an experienced school psychology doctoral student. All research assistants achieved this criterion on their first try for all tests. Each research assistant practiced scoring the ASC and the TNR and received feedback from the first author on a minimum of three forms for each test.

Criterion Measures

The CELF-P is norm-referenced measure of oral language ability that is administered individually. The core language subtests (Sentence Structure, Word Structure, and Expressive Vocabulary) reflect general oral language ability. The CELF-P has satisfactory correlations with other measures of oral language, internal consistency (.61–.96), and adequate test–rest reliability (.77–.92; Wiig et al., 2004). The CELF-P took 10 to 15 min to administer to each child. Scoring was completed in real time using the CELF-P protocols and raw scores were calculated later. The CELF-P was selected as a concurrent criterion oral language measure because it is a well-established instrument that is widely used in preschool settings. Importantly, it is most often used to identify children with language impairment and is commonly used to examine the validity of newer oral language measures (e.g., WODB, TNR).

The TNR (Spencer & Petersen, 2011) is a criterion-referenced, CBM tool (Deno, 2003) that utilizes a story retell format to assess oral language comprehension and production. The TSC is a companion to the TNR, which on the surface appears to be more closely related to the ASC. However, it was not selected as a criterion measure for a number of reasons. The TSC contains only very basic literal questions about story structure based on dramatically simpler stories than used in the ASC. The TSC does not assess

higher level comprehension that can be captured using story retells or the ASC. Furthermore, it has not been validated through empirical examinations of its technical properties like the TNR has. The TNR has 25 parallel forms and involves standardized administration and scoring procedures. In previous research, the TNR was shown to have adequate alternate form reliability (r = .77, p < .001) and strong evidence of concurrent validity with other story retell instruments (r = .88–.93, p < .001; Petersen & Spencer, 2012) and the CELF-P (r = .70, p < .001; Spencer & Petersen, 2016). In the current study, a randomly selected set of three TNR forms was administered to each participant in a single session. Administration and scoring took 5 to 7 min for the TNR. Children's story retells were recorded using digital voice recorders and scored by examiners who listened to the audio files. Scoring took place after the administration was complete. In the current sample, the mean alternate form reliability of the TNRs was .49 and all comparisons were statistically significant.

ASC

Using Microsoft Excel, a random sequence of ASC forms was generated for each child. Three ASC forms were administered in each session. This was important for evaluating the reliability and validity of using either the median score or the best score of the set of three. ASC testing sessions lasted 8 to 10 min. Examiners recorded every administration session using digital voice recorders, but did not rely on the audio files for scoring. During administration, the examiners wrote children's responses to the questions on the ASC protocols. Scoring was completed in real time or within 3 days of the ASC administration if scoring guides were needed as a reference. The examiner listened to the audio file only if the child's speech was unintelligible and the examiner was unable to write the child's answer during administration.

Data Analysis and Results

A total of 237 children participated in this study. Subsamples of these children were used to answer some of the research questions. For example, two Head Start sites and one community-based preschool received all nine forms within 2 weeks to help answer Question 4 (alternate form reliability). All of the other participants received their ASCs seasonally using the method of administering three ASCs and using the best or median for analysis, which allowed for multiple sets of qualifying data per child (fall and spring CELF-P, TNR, and ASCs) to answer Question 3 (e.g., construct validity). All 237 children were assessed using the ASC, CELF-P, and TNR in the fall (September or October), but not all of the children were available for retesting in the spring (April or May). Altogether, the ASC

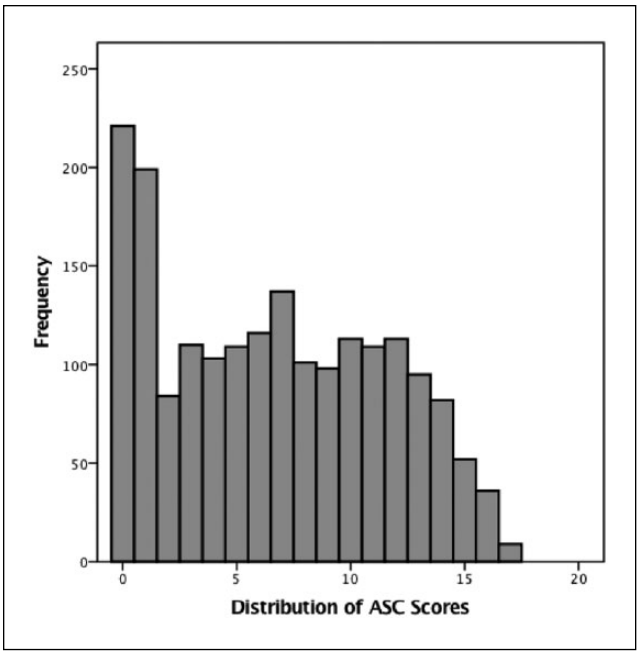

Figure 1. Frequency distribution of ASC scores.
Note. ASC = Assessment of Story Comprehension.

was administered 1,701 times and each of these administrations was available for the scoring reliability and fidelity analyses (Questions 1 and 2). Sample sizes for each analysis are indicated below. Figure 1 shows a frequency distribution of ASC scores. The mean score was 6.66 with a standard deviation of 4.8 and a range of 0 to 17.

ASC Administration Fidelity

To document the examiners' adherence to the ASC's standardized administration protocol, we examined a randomly selected sample of the ASC administrations (29.4% = 501 of 1,701). The same 10 research assistants who administered the ASC served as independent observers of fidelity of ASC administration; however, the research assistants did not observe fidelity for their own ASC administrations. To observe fidelity, research assistants listened to the audio files of ASC administrations and completed a fidelity checklist. A checklist was created for each of the nine ASC forms to include the exact script/prompt that was to be used during administration, which varied slightly depending on the content of the stimulus story. Using the checklists, research assistants rated 12 to 14 items (depending on the need to use prompts) for correct delivery of questions and the appropriate use of standardized prompts. For each item, the research assistant answered the question, "Did the examiner say (scripted line/question)?" Possible checklist answers were *yes, exactly; paraphrased with only minor changes; paraphrased with major changes*; and *no*. Ratings of *yes, exactly* and *paraphrased with only minor changes* counted as correct

Table 3. Kappa and Intraclass Correlation Coefficients.

Item	Kappa coefficients	Intraclass correlation coefficients
1	0.69	.82
2	0.77	.87
3	0.79	.91
4	0.74	.88
5	0.68	.75
6	0.7	.87
7	0.6	.79
8	0.81	.91
8a	0.94	.94

and *paraphrased with major changes* and *no* were incorrect. An example of a minor change is "I will ask you some questions" instead of "I'm going to ask you some questions." An example of a major change is "What did he do then?" instead of "What happened next?" Percent fidelity was calculated by dividing the number of items completed correctly out of the total number of items administered, multiplied by 100. Across all 501 administrations evaluated for fidelity, the mean fidelity was 99.6% (range = 78.6%–100%).

ASC Scoring Reliability

Research assistants listened to a randomly selected sample of ASC audio files for the purpose of providing an independent score for children's responses. Research assistants rescored 24.5% (421 of 1,701) of the ASCs. There were nine opportunities for scoring agreement for each administration (counting 8a as a separate item). Percent agreement was calculated by dividing the number of agreements by the total number of items scored, multiplied by 100. Mean scoring agreement was 92% with a range of 52% to 100%. Kappa coefficients were calculated for each of the items specifically to examine whether some items were more difficult to score reliably than others. Overall, the range of coefficients was from .60 to .94 suggesting moderate to high scoring reliability. With the exception of Item 3 (causal relationship between problem and feeling), the inferential questions had the lowest reliability coefficients while the definitional vocabulary items had the highest coefficient. Intraclass correlation coefficients (ICCs) show that, on average, 86% of the variation between item scores was due to differences between children, as opposed to differences between raters. Ranging from 75% for Item 5 to 94% for Item 8a, these high ICCs indicate the ASC is reliably measuring child characteristics. These coefficients are displayed in Table 3.

ASC Construct Validity

Pearson correlations were used to examine the extent to which the ASC measures a construct similar to known measures of oral language production and comprehension. ASC scores, which are not converted into scaled or standard scores, were compared with the raw score total of Core Language Composite of the CELF-P. Pearson correlations were completed comparing the CELF-P raw scores with the median and the best score of the three ASCs that were administered closest in time with the CELF-P (concurrent-related validity). Means were not used due to the skewed nature of the scoring. Testing environments are often chaotic affecting the validity of a single performance and thus use of a *median* score from a set is common practice. In addition, young children may have reduced attention and motivation, suggesting the *best* performance from a set might be the most valid. We included analyses for both the median and best scores to inform recommendations about best practice. The same analyses were completed comparing the median ASC score to the TNR median score and the best ASC score to the best TNR score, taken from the set of ASCs that were administered closest in time to the TNR. All 237 children were administered the ASC and the CELF-P in the fall and 117 of them were administered the CELF-P and the ASC again in the spring yielding 354 total number of cases available for this analysis. Correlations for the CELF-P raw score total and the median score and the best ASC score were large, $r = .79$, $p < .001$, and $r = .81$, $p < .001$, respectively. All 237 children were administered the ASC and the TNR in the fall and 181 of them were administered the TNR and the ASC again in the spring yielding 418 total number of cases available for this analysis. Correlations for the TNR and ASC were moderately large, $r = .69$, $p < .001$ (median scores), and $r = .67$, $p < .001$ (best scores).

Reliability of ASC Forms and Items

Parallel form reliability was examined using Pearson correlations among all nine forms of the ASC. Only 56 children received all nine ASC forms within 2 weeks and qualified for inclusion in this analysis. A correlation matrix (see Table 4) revealed moderate to large correlations among ASC forms ($r = .65$–$.83$, $p < .01$). Another way to assess reliability is by examining the correlation between the median or best scores from three different ASC sessions for a single child close in time ($n = 56$). This yielded a variation of test–retest reliability, which resulted in a mean correlation of .82, $p < .01$ (using median scores), and .78, $p < .01$ (using best scores). When parallel form reliability was examined for forms that were administered within the same session, albeit a random selection of forms, the mean correlation was .78, $p < .01$.

Cronbach's alphas were calculated to examine the internal consistency among all items within each form, individual items across forms, and all items in all forms. The alpha coefficients according to ASC forms and ASC items,

Table 4. Alternate Form Reliability of Nine ASC Forms.

ASC Forms	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8	Form 9
1	—	—	—	—	—	—	—	—	—
2	.69	—	—	—	—	—	—	—	—
3	.70	.71	—	—	—	—	—	—	—
4	.65	.73	.69	—	—	—	—	—	—
5	.72	.66	.73	.70	—	—	—	—	—
6	.67	.79	.83	.68	.80	—	—	—	—
7	.76	.75	.75	.73	.78	.78	—	—	—
8	.67	.81	.70	.71	.66	.82	.77	—	—
9	.71	.73	.68	.60	.68	.73	.73	.74	—

Note. ASC = Assessment of Story Comprehension.
All correlations are significant at the .01 level.

presented in Table 5, indicate high internal consistency. The mean coefficient across the nine forms was .83 (range = .79–.86). The mean coefficient for the eight items was .81 (range = .71–.89). The overall, alpha coefficient when considering all items in all forms was .83, but when totals from each of the forms were used as items, the alpha coefficient was .96. Altogether, these Cronbach's alpha statistics indicate high consistency within and across forms.

While relative scores were consistent across forms according to Cronbach's alpha, paired-samples *t* tests indicated that the scores themselves were significantly different between forms after accounting for the repeated sampling of children across forms. In particular, ASC Forms 1, 2, and 5 tended to have significantly higher scores than other forms, while Forms 6 and 7 tended to have significantly lower scores than other forms (*p* value ranges from <.001 to .047 for 19 out of 36 possible tests). These data are presented in Table 6. When testing mean differences across the form set scores (i.e., both the maximum/best score and the median score across all forms in the set), the set containing Forms 1 through 3 (Set 1) showed significantly higher scores than the set containing Forms 7 through 9 (Set 3; paired-samples $t = 3.73$, $df = 55$, $p < .001$, for best score; paired-samples $t = 3.76$, $df = 55$, $p < .001$, for median score). When using specifically the best score across all forms in the set, the average Set 1 score was significantly higher than Set 2 (paired-samples $t = 2.10$, $df = 55$, $p = .040$), and when using the median score across all forms in the set, the average Set 2 score was significantly higher than Set 3 (paired-samples $t = 2.47$, $df = 55$, $p = .017$). These results are presented in Table 7. In general, these tests reveal significant differences in the absolute (vs. relative) score between forms/sets.

Table 5. Internal Consistency Results for Items Within Forms and Items Across Forms.

Cronbach's α for ASC forms		Cronbach's α for ASC items	
Form 1	.79	Item 1	.82
Form 2	.81	Item 2	.80
Form 3	.84	Item 3	.89
Form 4	.82	Item 4	.71
Form 5	.84	Item 5	.82
Form 6	.82	Item 6	.80
Form 7	.86	Item 7	.84
Form 8	.85	Item 8	.82
Form 9	.82		

Note. ASC = Assessment of Story Comprehension.

Discussion

The purpose of this study was to examine the technical adequacy of the ASC as a curriculum-based measure of language comprehension for preschoolers. Initial measurement development and preliminary evidence of feasibility were established during iterative pilot studies (Spencer & Goldstein, 2011). To evaluate the utility of this tool for early childhood educators interested in fostering the development of their students' language comprehension skills, the psychometric properties of the ASC were examined with 237 preschoolers from a variety of early childhood education settings.

The first research question addressed the fidelity with which the ASC can be administered. Undergraduate research assistants administered the ASCs and obtained an extremely high fidelity of administration ($M = 99.6\%$). Although the research assistants received training to administer the ASCs, it was minimal in comparison with what is typically required to administer most norm-referenced standardized tests. It is important to note that early childhood educators are the intended end users of the ASC and may differ significantly from undergraduate research assistants with respect to age, experience, and education. Nonetheless, the unusually high fidelity of administration suggests the ASC is an easy instrument to deliver. It is likely that similarly high levels of fidelity would be achieved when administered by early childhood educators. By design, the ASC

Table 6. Paired-Sample *t* Test Results for ASC Form Mean Comparison.

Form	1	2	3	4	5	6	7	8	9
1									
Mean of differences	—	—	—	—	—	—	—	—	—
t	—	—	—	—	—	—	—	—	—
p	—	—	—	—	—	—	—	—	—
Cohen's*d*	—	—	—	—	—	—	—	—	—
2									
Mean of differences	0.39	—	—	—	—	—	—	—	—
t	0.90	—	—	—	—	—	—	—	—
p	.373	—	—	—	—	—	—	—	—
Cohen's *d*	0.12	—	—	—	—	—	—	—	—
3									
Mean of differences	**−0.91**	**−1.30**	—	—	—	—	—	—	—
t	**−2.03**	**−2.88**	—	—	—	—	—	—	—
p	**.047**	**.006**	—	—	—	—	—	—	—
Cohen's *d*	**−0.27**	**−0.38**	—	—	—	—	—	—	—
4									
Mean of differences	−0.57	**−0.96**	0.34	—	—	—	—	—	—
t	−1.25	**−2.30**	0.73	—	—	—	—	—	—
p	.216	**.025**	.470	—	—	—	—	—	—
Cohen's*d*	−0.17	**−0.31**	0.10	—	—	—	—	—	—
5									
Mean of differences	−0.05	−0.45	0.86	0.52	—	—	—	—	—
t	−0.13	−0.96	1.97	1.19	—	—	—	—	—
p	.897	.341	.054	.239	—	—	—	—	—
Cohen's*d*	−0.02	−0.13	0.26	0.16	—	—	—	—	—
6									
Mean of differences	**−1.72**	**−2.13**	**−0.82**	**−1.16**	**−1.68**	—	—	—	—
t	**−3.85**	**−5.71**	**−2.34**	**−2.57**	**−4.75**	—	—	—	—
p	**<.001**	**<.001**	**.023**	**.013**	**<.001**	—	—	—	—
Cohen's*d*	**−0.51**	**−0.76**	**−0.31**	**−0.34**	**−0.64**	—	—	—	—
7									
Mean of differences	**−1.79**	**−2.18**	**−0.88**	**−1.21**	**−1.73**	−0.05	—	—	—
t	**−4.42**	**−5.20**	**−2.03**	**−2.79**	**−4.46**	−0.14	—	—	—
p	**<.001**	**<.001**	**.047**	**.007**	**<.001**	.892	—	—	—
Cohen's*d*	**−0.59**	**−0.70**	**−0.27**	**−0.37**	**−0.60**	−0.02	—	—	—
8									
Mean of differences	**−1.23**	**−1.63**	−0.32	−0.66	**−1.18**	0.50	0.55	—	—
t	**−2.54**	**−4.26**	−0.66	−1.42	**−2.34**	1.37	1.31	—	—
p	**.014**	**<.001**	.510	.160	**.023**	.176	.197	—	—
Cohen's*d*	**−0.34**	**−0.57**	−0.09	−0.19	**−0.31**	0.18	0.17	—	—
9									
Mean of differences	−0.71	**−1.11**	0.20	−0.14	−0.66	**1.02**	**1.07**	0.52	—
t	−1.77	**−2.71**	0.43	−0.29	−1.51	**2.50**	**2.52**	1.19	—
p	.083	**.009**	.671	.770	.138	**.015**	**.015**	.240	—
Cohen's*d*	−0.24	**−0.36**	0.06	−0.04	−0.20	**0.33**	**0.34**	0.16	—

Note. Degrees of freedom for all ASC Form paired-samples *t* tests are 55. ASC = Assessment of Story Comprehension.
Bold-faced values indicate significant differences ($p < .05$).

has simple scripted instructions for the examiner to say, including how and when to prompt. This feature seems to be an appealing feature of the ASC because easy to administer tests are more likely to be used by educators.

It is important that examiners (and end users) can score the ASC reliably. Other researchers have sought to assess oral language and language comprehension skills via proxy items such as picture naming, receptive vocabulary, and

Table 7. Paired Samples *t* Test Results for ASC Form Sets Mean Comparison.

Form set		Best score sets			Median score sets		
		1	2	3	1	2	3
1	Mean of differences	—	—	—	—	—	—
	t	—	—	—	—	—	—
	p	—	—	—	—	—	—
2	Mean of differences	**−0.71**	—	—	−0.43	—	—
	t	**−2.10**	—	—	−1.42	—	—
	p	**.040**	—	—	.161	—	—
3	Mean of differences	**−1.09**	−0.38	—	**−1.11**	**−0.68**	—
	t	**−3.73**	−1.10	—	**−3.76**	**−2.47**	—
	p	**<.001**	.274	—	**<.001**	**.017**	—

Note. Degrees of freedom for all ASC Form Set paired-samples *t* tests are 55. ASC = Assessment of Story Comprehension.
Bold-faced values indicate significant differences ($p < .05$).

pointing to a picture that does not belong (Bradfield et al., 2014; Wackerle-Hollman et al., 2014). It is more common to use discrete responses in curriculum-based measures because they are easier to score reliably than responses that require judgments about inferences, definitions, or event recall. Scoring open-ended responses of young children whose language coherence is emerging is challenging. However, with scoring guidelines using the story-specific examples and formulaic scoring rules, sufficient scoring reliability was achieved ($M = 92\%$). Children's answers were sometimes difficult to hear if they spoke very softly (live or on the audio recorders). Given their age, many produce unintelligible speech. This accounts for a few lower than desired scoring agreements (e.g., 52%) even though the mean is sufficiently high.

Concurrent evidence of construct validity was examined by correlating the ASC with two other types of instruments that measure language and specifically aspects of language comprehension. We found strong correlations with the CELF-P, a norm-referenced standardized test of oral language abilities, and the TNR, a criterion-referenced test of narrative language using a retell format. The correlations between the ASC and the CELF-P were slightly larger than between the ASC and the TNR. We expected stronger correlations between the ASC and the TNR because they both use a brief personally relevant story as the basis for test administration. It may be important that the TNR, which requires a large sample of connected speech to retell the story, loads on two factors: language production and language comprehension (Petersen & Spencer, 2012), whereas the ASC and the CELF-P have minimal language production requirements and may represent language comprehension more directly without the confound of expressive language. In the CELF-P, children point to pictures or respond using only a few words, and in the ASC, children answer questions verbally but responses of two to five words are sufficient to earn the maximum points possible. It may also be important to note that correlations between the ASC and the CELF-P were similar to or slightly better than correlations between the ASC forms. Because the ASC forms use different stories in their administration, it is not surprising that correlations between them are not higher. However, this finding indicates that regardless of how well the ASC forms relate to each other, they all tend to measure the same thing as the CELF-P. Another aspect of validity that we examined was whether the median or best score of three ASCs within a session was a more valid measure of children's language comprehension. There were negligible differences when the median or the best scores were used, suggesting that either approach would produce valid scores for interpretation.

The final research question addressed the reliability of the ASC with respect to parallel forms and internal consistency. To be useful as a progress-monitoring tool, equivalent forms are needed for repeated sampling over time. If forms differ in difficulty, for example, then changes in children's scores over time cannot confidently be attributed to learning. When each ASC form was correlated with the other forms, reliability coefficients were medium to high (range = .65–.83). The *t*-test analyses reveal significant differences between some of the ASC forms, suggesting they are not equivalent. Given the small sample size for these analyses ($n = 56$), it is difficult to determine whether the variability is related to the ASC, the testing environment, or the population. Logic suggests that highly stable responding is unlikely for such young children. Most of these children qualify for enrollment in Head Start preschools because they come from low income households and many have additional risk factors such as cultural and language differences. Motivation and attention are influenced by children's lack of experiences with formal testing activities and limited background experiences. We compensated for this limitation by writing story themes that are generally applicable to young children, but there are unavoidable differences in their experiences.

In addition, testing conditions in preschools can be challenging. If preschool teachers are going to be the end users, they will likely be testing during class time while many other activities are happening in the classroom, which is how the ASCs were administered in the current study. Despite efforts to reduce the variability between ASC forms, the results of this early study indicate that revisions to the forms are needed to increase their equivalence. A careful examination of what makes Forms 1, 2, and 5 easier and Forms 6 and 7 harder is needed. Future research should re-examine alternate form reliability following revisions to these stories and/or questions.

Because we expect that the ASC will be used for screening purposes, we examined reliability of ASCs within the same session and across sessions, independent of forms. When three ASCs that were administered within the same session were correlated, this indicator of test–retest reliability revealed a correlation of .78. This suggests that from one story to the next even within a few minutes of each other, there are slight differences in performance. This supports our observations that children's variable experiences with story content and/or motivational and attention factors reduce the stability of responding. Although this is an interesting finding, we do not imagine the ASC being used in this manner. More realistically, one session of three ASCs would be administered seasonally, aligning with benchmark screening common in differentiated intervention models such as RTI or MTSS. Therefore, we examined the correlation if the median scores from three sessions were used. This analysis produced the strongest reliability coefficient of .82, which suggests this strategy may help reduce variable responding among young children. However, significant differences were noted when paired-samples *t* tests were conducted between Sets 1, 2, and 3 using the best or median scores. These results indicate that further work is needed to increase the equivalence of the forms and sets.

The examination of internal consistency reliability yielded strong Cronbach's alphas (.71–.89) no matter how the items and forms were analyzed. The results indicate that all of the items appear to contribute to the construct of language comprehension. However, if individual children's scores will be compared with criterion scores, some standardization will need to be conducted due to the mean differences across forms/sets. In other words, a child's relative standing to his or her peers or even to his or her past performance can be assessed with some moderate reliability, though making judgments based on the student's score versus some absolute benchmark will require more investigation and standardization.

Limitations and Future Directions

This study represents an intermediate phase of the validation of the ASC. Future research is needed to account for some of the limitations in the current study. For example, the analysis of parallel form reliability had the smallest sample size, making it difficult to distinguish between differences among ASC forms versus variability in responding among young children. These results also suggest some additional development may be needed to be able to use the ASC as a criterion-referenced assessment. In future research, alternate forms should be examined with a larger sample of children. It also will be important to examine the ASC's feasibility and acceptability when end users such as early childhood educators serve as examiners. Currently, we do not know how useful and doable preschool teachers will find the ASC. Finally, research is needed to determine whether the ASC can accurately identify children who would benefit from language comprehension intervention. Longitudinal data collection would be needed to assess the predictive validity of the ASC.

Conclusion

The ASC is a promising assessment tool for measuring and monitoring preschoolers' language comprehension. One of its greatest strengths is that the testing environment is extremely authentic. The use of stories and answering questions about the stories reflects common preschool activities, which may enhance preschool teachers' adoption of the ASC. Based on the results of this study, the ASC has not only preliminary evidence of validity and reliability but also room for improvement. The results provide a direction for appropriate refinement so that it can be included in early childhood education MTSS models and so that educators can effectively attend to language comprehension in addition to decoding-related skills.

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: This work was supported by Cooperative Agreement R324C080011, the Center for Response to Intervention in Early Childhood, from the Institute of Education Sciences, U.S. Department of Education.

References

Bagnato, S. J., & Neisworth, J. T. (2005). Recommended practices in assessment. In S. Sandall, M. E. McLean, & B. J. Smith (Eds.), *DEC recommended practices in early intervention/early childhood special education* (pp. 17–28). Longmont, CO: Sopris West.

Bailet, L. L., Repper, K., Murphy, S., Piasta, S., & Zettler-Greeley, C. (2013). Emergent literacy intervention for prekindergarteners at risk for reading failure: Years 2 and 3

of a multiyear study. *Journal of Learning Disabilities*, *46*, 133–153. doi:10.1177/0022219411407925

Bishop, D. V. M., & Adams, C. (1992). Comprehension problems in children with specific language impairment: Literal and inferential meaning. *Journal of Speech and Hearing Research*, *35*, 119–129. doi:10.1044/jshr.3501.119

Bishop, D. V. M., & Edmundson, A. (1987). Language impaired 4-year-olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorders*, *52*, 156–173. doi:10.1044/jshd.5202.156

Blank, M., Rose, S. A., & Berlin, L. J. (2003). Preschool language assessment instrument (2nd ed.). Austin, TX: Pro-Ed.

Bradfield, T. A., Besner, A. C., Wackerle-Hollman, A. K., Albano, A. D., Rodriguez, M. C., & McConnell, S. R. (2014). Redefining individual growth and development indicators: Oral language. *Assessment for Effective Intervention*, *39*, 233–244. doi:10.1177/1534508413496837

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing*, *11*, 489–503. doi:10.1023/A:1008084120205

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition*, *29*, 850–859. doi:10.3758/BF03196414

Cain, K., Oakhill, J. V., & Elbro, C. (2003). The ability to learn new word meanings from context by school-age children with and without language comprehension difficulties. *Journal of Child Language*, *30*, 681–694.

Cain, K., Oakhill, J. V., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology*, *96*, 671–681. doi:10.1037/0022-0663.96.4.671

Catts, H. W., Fey, M. E., Tomblin, J. B., & Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research*, *45*, 1142–1157. doi:10.1044/1092-4388(2002/093)

Christ, T. J., Zopluoglu, C., Long, J. D., & Monaghen, B. D. (2012). Curriculum-based measurement of reading: Quality of progress monitoring outcomes. *Exceptional Children*, *78*, 356–373. doi:10.1177/001440291207800306

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*, 934–945.

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language*, *25*, 1–18.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*, 184–192. doi:10.1177/00224669030370030801

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, *49*, 36–45.

Diamond, K. E., & Powell, D. R. (2011). An iterative approach to the development of a professional development intervention for Head Start teachers. *Journal of Early Intervention*, *33*, 75–93. doi:10.1177/1053815111400416

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. (2001). Bringing it all together: The multiple origins, skills, and environmental supports of early literacy. *Learning Disabilities Research & Practice*, *16*, 186–202.

Dooley, C. M., & Matthews, M. W. (2009). Emergent comprehension: Understanding comprehension development among young literacy learners. *Journal of Early Childhood Literacy*, *9*, 269–294. doi:10.1177/1468798409345110

Fazio, B. B., Naremore, R. C., & Connell, P. J. (1996). Tracking children from poverty at risk for specific language impairment: A 3-year longitudinal study. *Journal of Speech, Language and Hearing Research*, *39*, 611–624. doi:10.1044/jshr.3903.611

Feagans, L., & Appelbaum, M. I. (1986). Validation of language subtypes in learning disabled children. *Journal of Experimental Psychology*, *78*, 358–364. doi:10.1037/0022-0663.78.5.358

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, *41*, 121–139. doi:10.1177/00224669070410020101

Ford, J. A., & Milsoky, L. M. (2003). Inferring emotional reactions in social situations: Differences in children with language impairment. *Journal of Speech, Language, and Hearing Research*, *46*, 21–30. doi:10.1044/1092-4388(2003/002)

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, *21*, 449–460. doi:10.3102/00028312021002449

Fuchs, L. S., & Fuchs, D. (2007). A model for implementing responsiveness to intervention. *Teaching Exceptional Children*, *39*(5), 14–20. doi:10.1177/004005990703900503

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Phillips, N. B. (1994). Classwide curriculum-based instruction: Helping general educators meet the challenge of student diversity. *Exceptional Children*, *60*, 518–537. doi:10.1177/001440299406000605

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*, 6–10. doi:10.1177/074193258600700104

Greenwood, C. R., Bradfield, T., Kaminski, R., Linas, M., Carta, J. J., & Nylander, D. (2011). The Response to Intervention (RTI) approach in early childhood. *Focus on Exceptional Children*, *43*(9), 1–22.

Greenwood, C. R., Carta, J. J., Atwater, J., Goldstein, H., Kaminski, R., & McConnell, S. (2013). Is a Response to Intervention (RTI) approach to preschool language and early literacy instruction needed? *Topics in Early Childhood Special Education*, *33*, 48–64. doi:10.1177/0271121412455438

Griffin, T. M., Hemphill, L., Camp, L., & Wolf, D. P. (2004). Oral discourse in the preschool years and later literacy skills. *First Language*, *24*, 123–147. doi:10.1177/0142723704042369

Haager, D. E., Klingner, J. E., & Vaughn, S. E. (Eds.). (2007). *Evidence-based reading practices for response to intervention*. Baltimore, MD: Paul H. Brooks.

Hammett, L. A., Van Kleeck, A., & Huberty, C. J. (2003). Patterns of parents' extratextual interactions during book sharing

with preschool children: A cluster analysis study. *Reading Research Quarterly*, *38*, 442–468. doi:10.1598/RRQ.38.4.2

Hargrave, A. C., & Sénéchal, M. (2000). A book reading intervention with preschool children who have limited vocabularies: The benefits of regular reading and dialogic reading. *Early Childhood Research Quarterly*, *15*, 75–90. doi:10.1016/S0885-2006(99)00038-1

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

Hurry, J., & Sylva, K. (2007). Long-term outcomes of early reading intervention. *Journal of Research in Reading*, *30*, 227–248. doi:10.1111/j.1467-9817.2007.00338.x

Kelley, E. S., Goldstein, H., Spencer, T. D., & Sherman, A. (2015). Effects of automated Tier 2 storybook intervention on vocabulary and comprehension intervention for preschool children with limited oral language skills. *Early Childhood Research Quarterly*, *31*, 47–61.

Kendeou, P., Bohn-Gettler, C., White, M. J., & Van Den Broek, P. (2008). Children's inference generation across different media. *Journal of Research in Reading*, *31*, 259–272.

Kern, L., Evans, S., & Lewis, T. J. (2011). Description of an iterative process for intervention development. *Education and Treatment of Children*, *34*, 593–617. doi:10.1353/etc.2011.0037

Koutsoftas, A. D., Harmon, M. T., & Gray, S. (2009). The effect of Tier 2 intervention for phonemic awareness in a response-to-intervention model in low-income preschool classrooms. *Language, Speech, and Hearing Services in Schools*, *40*, 116–130. doi:10.1044/0161-1461(2008/07-0101)

Lepola, J., Lynch, J., Laakkonen, E., Silvén, M., & Niemi, P. (2012). The role of inference making and other language skills in the development of narrative listening comprehension in 4-6-year-old children. *Reading Research Quarterly*, *47*, 259–282.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing a literature review. *The Journal of Special Education*, *41*, 68–84. doi:10.1177/00224669070410020301

Missall, K. N., & McConnell, S. R. (2004). *Psychometric characteristics of individual growth and development indicators: Picture naming, rhyming & alliteration* (Technical report). Minneapolis, MN: Center for Early Education and Development.

National Center for Education Statistics. (2015). *Reading assessment*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://nces.ed.gov/nationsreportcard/reading

Nippold, M. A. (2002). Lexical learning in school-age children, adolescents, and adults: A process where language and literacy converge. *Journal of Child Language*, *29*, 474–478. doi:10.1017/S0305000902275340

Paris, A. H., & Paris, S. G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly*, *38*, 36–76. doi:10.1598/RRQ.38.1.3

Petersen, D. B., & Spencer, T. D. (2012). The narrative language measures: Tools for language screening, progress monitoring, and intervention planning. *Perspectives on Language Learning and Education*, *19*(4), 119–129. doi:10.1044/lle19.4.119

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. Neuman & D. Dickinson (Eds.), *Handbook for research in early literacy* (pp. 97–110). New York, NY: Guilford Press.

Serpell, R., Baker, L., & Sonnenschein, S. (2005). *Becoming literate in the city: The Baltimore early childhood project*. Oxford, UK: Cambridge University Press.

Spencer, E. J., Goldstein, H., Sherman, A., Noe, S., Tabbah, R., Ziolkowski, R., & Schneider, N. (2012). Effects of an automated vocabulary and comprehension intervention: An early efficacy study. *Journal of Early Intervention*, *34*, 195–221.

Spencer, T. D., & Goldstein, H. (2011, July). The Assessment of Story Comprehension (ASC): A preliminary investigation of reliability and validity. In S. Piasta (Chair), *New measures for investigating emergent literacy environments and skill development*. Symposium presented at the annual meeting of the Society for the Scientific Study of Reading, Fort Lauderdale, FL.

Spencer, T. D., & Petersen, D. B. (2011). *The test of narrative retell*. Available from www.LanguageDynamicsGroup.com

Spencer, T. D., & Petersen, D. B. (2016). *Reliability and validity of the Narrative Language Measures*. Unpublished manuscript.

Spencer, T. D., Petersen, D. B., Slocum, T. A., & Allen, M. M. (2015). Large group narrative intervention in Head Start preschools: Implications for response to intervention. *Journal of Early Childhood Research*, *13*, 196–217.

Spencer, T. D., & Slocum, T. A. (2010). The effect of a narrative intervention on story retelling and personal story generation skills of preschoolers with risk factors and narrative language delays. *Journal of Early Intervention*, *32*, 178–199. doi:10.1177/1053815110379124

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

Tompkins, V., Guo, Y., & Justice, L. M. (2013). Inference generation, story comprehension, and language skills in the preschool years. *Reading and Writing*, *26*, 403–429. doi:10.1007/s11145-012-9374-7

Tunmer, W. E., & Chapman, J. W. (2012). The simple view of reading redux: Vocabulary knowledge and the independent components hypothesis. *Journal of Learning Disabilities*, *45*, 453–466.

Tunmer, W. E., & Hoover, W. A. (1992). Cognitive and linguistic factors in learning to read. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 175–214). Hillsdale, NJ: Lawrence Erlbaum.

Van den Broek, P. D., Kendeou, P., Kremer, K., Lynch, J., Butler, J., White, M. J., & Lorch, E. P. (2005). Assessment of comprehension abilities in young children. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 107–130). Mahwah, NJ: Lawrence Erlbaum.

VanDerHeyden, A. M., Snyder, P. A., Broussard, C., & Ramsdell, K. (2008). Measuring response to early literacy intervention with preschoolers at risk. *Topics in Early Childhood Special Education*, *27*, 232–249. doi:10.1177/0271121407311240

van Kleeck, A. (2008). Providing preschool foundations for later reading comprehension: The importance of and ideas for

targeting inferencing in storybook-sharing interventions. *Psychology in the Schools*, *45*, 627–643. doi:10.1002/pits.20314

van Kleeck, A., Gillam, R. B., Hamilton, L., & Cassandra, M. (1997). The relationship between middle-class parents' book-sharing discussion and their preschoolers' abstract language development. *Journal of Speech, Language, and Hearing Research*, *40*, 1261–1271. doi:10.1044/jslhr.4006.1261

van Kleeck, A., Vander Woude, J., & Hammett, L. (2006). Fostering literal and inferential language skills in Head Start preschoolers with language impairment using scripted book-sharing discussions. *American Journal of Speech-Language Pathology*, *15*, 85–95. doi:10.1044/jslhr.4006.1261

Wackerle-Hollman, A. K., Rodriguez, M. I., Bradfield, T. A., Rodriguez, M. C., & McConnell, S. R. (2014). Development of early measures of comprehension innovation in individual growth and development indicators. *Assessment for Effective Intervention*, *40*, 81–95. doi:10.1177/1534508414551404

Whitehurst, G. J., Arnold, D. S., Epstein, J. N., Angell, A. L., Smith, M., & Fischel, J. E. (1994). A picture book reading intervention in day care and home for children from low-income families. *Developmental Psychology*, *30*, 679–689.

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development*, *69*, 848–872. doi:10.1111/j.1467-8624.1998.tb06247.x

Wiig, E. H., Secord, W., & Semel, E. M. (2004). *Clinical evaluation of language fundamentals–preschool 2*. San Antonio, TX: Psychological Corporation.

Zucker, T. A., Solari, E. J., Landry, S. H., & Swank, P. R. (2013). Effects of a brief tiered language intervention for prekindergartners at risk. *Early Education and Development*, *24*, 366–392. doi:10.1080/10409289.2012.664763