LSHSS

Research Article

Development of the Narrative Assessment Protocol-2: A Tool for Examining Young Children's Narrative Skill

Ryan P. Bowles,[a] Laura M. Justice,[b] Kiren S. Khan,[c] Shayne B. Piasta,[c] Lori E. Skibbe,[a] and Tricia D. Foster[a]

Purpose: Narrative skill, a child's ability to create a temporally sequenced account of an experience or event, is considered an important domain of children's language development. Narrative skill is strongly predictive of later language and literacy and is emphasized in curricula and educational standards. However, the need to transcribe a child's narrative and the lack of psychometrically justified scoring methods have precluded broad consideration of narrative skill among practitioners. We describe the development and validation of the Narrative Assessment Protocol-2 (NAP-2), an assessment of narrative skill for children ages 3–6 years, which uses event-based frequency scoring directly from a video recording of a child's narrative.
Method: The NAP-2 underwent a rigorous development process involving creation of four wordless picture books and associated scripts and identification of a broad item pool, including aspects of narrative microstructure and macrostructure. We collected two narratives from each of 470 children using the NAP-2 elicitation materials and scored each with the 60 items in the initial item pool.
Results: Cross-validated exploratory factor analyses indicated a single narrative skill factor. Rasch measurement analysis led to selection of 20 items that maintained high reliability while having good fit to the model and no evidence of differential item functioning across books and gender.
Conclusions: The NAP-2 offers a psychometrically sound and easy-to-use assessment of narrative skill for children ages 3–6 years. The NAP-2 is available freely online for use by speech-language pathologists, educational practitioners, and researchers.
Supplemental Material: https://doi.org/10.23641/asha.11800779

Narrative skill, which represents a child's ability to create a temporally sequenced production of a fictional or real account of an experience or event (Engel, 1995), is considered an important domain of children's language development (Boudreau & Hedberg, 1999; Curenton & Justice, 2004) that can be reliably measured and improved through intervention (e.g., Nicolopoulou & Trapp, 2018; Pesco & Kay-Raining Bird, 2016; Petersen, 2011; Price et al., 2006). Narratives can be about personal experiences or fictional events (i.e., fictional narratives) and can be prompted or unprompted. As with other domains of language, children show developmental changes in their narrative skill over time (Curenton & Justice, 2004).

Assessment of a child's narrative can provide information about many aspects of his or her expressive language, including general language productivity (e.g., total number of utterances), vocabulary (e.g., number of different words), syntax (e.g., percentage of utterances containing multiple clauses), and morphology (e.g., accuracy of word inflections). Measures derived from narrative assessment provide generally strong and reliable indices of children's concurrent and future language competence (e.g., Boudreau & Hedberg, 1999; Gardner-Neblett & Iruka, 2015; Pankratz et al., 2007; Tilstra & McMaster, 2007). For instance, one study reported correlations of .77 and

[a]Department of Human Development and Family Studies, Michigan State University, East Lansing
[b]Crane Center for Early Childhood Research and Policy, The Ohio State University, Columbus
[c]Department of Teaching and Learning and Crane Center for Early Childhood Research and Policy, The Ohio State University, Columbus
Correspondence to Ryan P. Bowles: bowlesr@msu.edu
Kiren S. Khan is now at the Department of Psychology, Rhodes College, Memphis, TN.
Tricia D. Foster is now at the School of Health Sciences, Eastern Michigan University, Ypsilanti, MI.

Editor-in-Chief: Holly L. Storkel
Editor: Sherrie Hill
Received April 25, 2019
Revision received May 24, 2019
Accepted October 16, 2019
https://doi.org/10.1044/2019_LSHSS-19-00038

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

.61 between a measure of 5- to 6-year-old children's narrative complexity and measures of receptive vocabulary and reading comprehension, respectively, 3 years later (Pankratz et al., 2007). Studies also show some measures of narrative ability to have good to excellent accuracy for identifying presence of language impairment (Liles et al., 1995; Pankratz et al., 2007; Peña et al., 2006), with sensitivity and specificity values in the 78%–99% range (Pankratz et al., 2007; Peña et al., 2006).

Narrative assessment can be an important tool within clinical and educational practices. For the former, identification of children with language impairment, based on current consensus statements, should rely on information from a variety of sources that includes, for instance, implementation of tasks that "tax both expressive and receptive skills" (Bishop et al., 2016). Narrative tasks represent a key way to "tax" children's language skills, as production of a narrative requires the complex integration of numerous lower level language skills, including grammar, morphology, vocabulary, and pragmatics. Use of a valid, reliable, and easy-to-use narrative task may provide clinicians with information about a child's narrative skills that augments information from other tasks, such as standardized testing and observations (Bishop et al., 2016). In addition, researchers have long emphasized the value of narrative assessments in terms of their ecological validity and cultural sensitivity relative to more traditional forms of language assessment (Gardner-Neblett et al., 2012; Peña et al., 2006; Price et al., 2006). This is due, in part, to the nature of narrative-based assessment tasks, in which a child's language skills are examined in an authentic, contextualized task that exemplifies a typical linguistic task in which many children will have considered experience (Schraeder et al., 1999). To this end, some argue that the use of narrative assessment can show linguistic strengths of children from culturally and linguistically diverse backgrounds (De Villiers & Burns, 2003; Gardner-Neblett et al., 2012; Govindarajan & Paradis, 2019; Laing & Kamhi, 2003; Muñoz et al., 2003). For instance, traditional vocabulary assessments that examine only single-word receptive vocabulary may miss important features of children's vocabulary knowledge; similarly, a child may have a nuanced understanding of how to take a listener's perspective into account when conveying information, yet this skill likely would not be apparent in a more traditional assessment situation.

Narrative assessments can also provide a valuable tool for educators. Increasingly, students' use of complex oral language and narrative-related skills are emphasized for young children in state curriculum standards (Calkins et al., 2012; Neuman & Roskos, 2005; Petersen & Spencer, 2016b) and practice guides (Foorman et al., 2016). For instance, the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) expects children at the kindergarten level to demonstrate the following narrative abilities: retell familiar stories with prompts; identify characters, settings, and key events in familiar stories; and compare and contrast the adventures and experiences of characters in familiar stories. Likewise, the What Works Clearinghouse recently provided four recommendations for educator practices to improve reading comprehension in the primary grades, one of which emphasized the need to "explicitly engage students in developing narrative language skills" (Foorman et al., 2016, p. 9). The Head Start Early Learning Outcomes Framework explicitly expects preschool children to "demonstrate an understanding of narrative structure," including retelling stories, appropriately sequencing events, and identifying story characters and key events (U.S. Department of Health and Human Services, Administration for Children and Families, & Office of Head Start, 2015, p. 47). These recommendations are based on empirical findings showing a robust association between early narrative skill and later academic skills (e.g., Griffin et al., 2004; Reese et al., 2010).

Clinical and educational use of narrative assessments is grounded, at least in part, in the increased emphasis on improving children's narrative skills in clinical therapies and everyday educational practices. To do so, there is a great need for (a) developmental work to improve our understanding of early narrative development and (b) applied research investigating effective approaches to improving narrative skill. Both of these lines of inquiry necessitate the development of psychometrically strong measures of narrative skill.

Challenges With Narrative Assessment

Many researchers contend that narrative assessment should have a much more prominent place in the language assessment practices used in clinical and educational settings (e.g., Justice et al., 2006; Pankratz et al., 2007; Price et al., 2006). However, predominant approaches to narrative assessment are not functionally usable for many educators and allied professionals, as they rely largely on traditional language sampling analysis (LSA) procedures, in which a narrative sample must be transcribed before scoring (e.g., Gagarina et al., 2012; Hux et al., 1993; Washington & Craig, 2004). LSA requires great investments of time, identified by speech-language pathologists (SLPs) as the biggest obstacle to the use of LSA (Pavelko et al., 2016). Transcription requires specialized skills and software, and the products of transcription can be unreliable (Gavin & Giles, 1996). Furthermore, following transcription, one must decide what aspects of language to analyze so that assessment outcomes can be interpreted in some manner. There is no consensus on what the construct of narrative skill encompasses, and the possibilities regarding what to analyze are numerous, including mean length of utterances, total number of words, number of complete episodes, and number of story grammar components, to name only a few possibilities. Although the construct of narrative skill and what it encompasses are not well defined, narrative skill is commonly broken down into two interrelated but conceptually distinct aspects of a child's narrative: *macrostructure* and *microstructure* (see Justice et al., 2006).

Narrative macrostructure involves analysis of the child's use or understanding of causal networks, event

representations (i.e., scripts), and story grammar elements within narratives and is greatly influenced by the seminal work of Labov (1972). Macrostructural analysis is based on the perspective that children's narrative abilities are influenced by their "mental representations of events and the verbalizations of such scripts" (Berman, 1995, p. 287), and this approach to analysis examines global characteristics of the narrative, such as adherence to traditional story grammar rules (i.e., if the story contains a series of episodes, comprising an initiating event, goal, plot, and resolution, called *episodic analysis*; Botvin & Sutton-Smith, 1977; McCabe & Peterson, 1984). Macrostructure analysis may also consider the extent to which the child's narrative (a) leads sequentially up to a high point (called *high point analysis*), (b) is coherent (*dependency analysis*; Deese, 1983; Liles et al., 1995; McCabe & Peterson, 1984), and/or (c) contains evaluative devices and other elements of "artfulness" (Ukrainetz et al., 2005; Zevenbergen et al., 2003). This breadth of conceptualization highlights that macrostructure has not been established as a unified and measurable construct.

Narrative microstructure involves the more granular aspects of a narrative, such as the specific sentences, phrases, clauses, and words. Analysis of narrative microstructure often examines, for instance, the total number of T-units within a narrative (a T-unit is one independent clause and any dependent phrases and clauses) and the percentage of these T-units that contain complex syntax. This can also include examining how the narrator builds cohesion across the narrative, through use of pronominal references, coordinating and subordinating conjunctions, and other morphosyntactic devices. Some work has suggested that microstructure is a multidimensional construct, representing productivity and complexity (e.g., Justice et al., 2006).

Narrative Assessment Protocol-2

In light of the challenges of assessing narrative skill among clinical and educational professionals, the primary goal of this article is to develop and provide validity evidence for the Narrative Assessment Protocol-2 (NAP-2) as a new easy-to-use tool to measure narrative skill for 3- to 6-year-old children. We designed the NAP-2 to have three key design features to increase its usability compared to classical approaches to assessing narrative skill, for SLPs, educators, and allied professionals, as well as researchers. First, the NAP-2 scores children's narratives directly from a video recording, therefore eliminating the need for transcription. Although transcription offers information not readily available directly from video recordings (e.g., total number of words), eliminating transcription greatly reduces the time and effort needed to score the narrative. Second, the NAP-2 requires relatively little time to administer (typically less than 10 min to collect the narrative and less than 10 min to score). Third, the NAP-2 is cost-effective, as all materials are available for free online, including elicitation materials, training materials for users, and scoring tools. The NAP-2 is derived from the NAP (see Justice et al., 2010), improving on the earlier version in several important, defining ways. First, whereas the NAP relied on eliciting narratives via a single commercially available wordless picture book, the NAP-2 features multiple researcher-developed books with well-defined story grammar and scripts, thus providing multiple test forms with evidence of invariance, all freely available. Second, the NAP-2 was designed to substantially broaden the NAP's focus, in that the prototype assessed only microstructural features of children's narratives, whereas the NAP-2 captures both microstructural and macrostructural features.

The first aim of this article is to describe the development process of the NAP-2 in order to provide evidence of *content validity* of the NAP-2 as addressed through a rigorous development process for both the narrative elicitation materials and the pool of items used to score the narratives (i.e., test content validity evidence in the Standards for Educational and Psychological Measurement; American Educational Research Association et al., 2014). A secondary goal of this article is to examine the structure of narrative skill for young children and to identify a set of items that reliably and coherently assess narrative skill. To that end, the second aim of this article is to describe the *construct validity* of the NAP-2 by understanding how narrative skill is expressed through the items used to score the narratives (i.e., internal structure validity evidence; AERA et al., 2014). We used exploratory factor analysis (EFA) and Rasch measurement methods to examine the nature of narrative skill and to select a set of items from a larger item pool that can measure children's narrative skill reliably. Although some analytic activities have been used to understand the distinctiveness of and interrelations among microstructural and macrostructural elements of narratives (Liles et al., 1995), there is generally little empirical understanding of the underlying factor structure of narrative skill and what scoring methods are associated with each underlying factor (Justice et al., 2006). The final version of the NAP-2 items is provided in Appendix A, and the elicitation and scoring materials, along with training modules, are available for free on our website (https://www.narrativeassessment.com).

The NAP-2 joins a small family of recently developed and validated narrative assessments for young children that have taken different approaches to deal with the challenges of narrative assessment. The CUBED Narrative Language Measures (NLM; Petersen & Spencer, 2016a) is freely available and eliminates transcription by having the child retell short, simple, paragraph-length narratives that are scored with a constrained set of items. The CUBED NLM has good evidence of validity through relations with alternative language measures (Language Dynamics Group, 2019; relations to other measures validity evidence: AERA et al., 2014). To our knowledge, however, the CUBED NLM has little evidence of validity that the scored items work together to comprehensively and validly measure a child's narrative skills (internal structure validity evidence; AERA et al., 2014). The Test of Narrative

Language–Second Edition (TNL-2; Gillam & Pearson, 2017) is relatively expensive ($200 for the kit plus $2 per examinee booklet) and involves a single form consisting of one narrative retelling with a comprehension task intervening between the script and the retelling and two narrative generations to go with a wordless five-panel comic strip. It is appropriate only for children as young as 4 years of age. Narratives are scored from audiotape on a variety of items. The TNL-2 has good evidence of diagnostic power for language learning disability and some limited evidence of validity through relations to other measures and internal structure.

Study 1: Development of the NAP-2

The aim of Study 1 was to develop the NAP-2, including the elicitation materials and initial item pool used to score children's narratives. In this, typical of early phases of other measures, our focus was to develop a psychometrically sound and feasible tool for assessing children's narrative skill, with potential use by SLPs, educators, and researchers. To this end, early measurement work did not focus on the diagnostic accuracy of the tool for identifying children with language impairment; we anticipate that this diagnostic measurement work will be pursued in the future by numerous research teams. We engaged in a robust instrument development process involving multiple layers of expert review to ensure that the NAP-2 is capturing narrative skill as intended, in order to yield strong evidence of content validity. Instrument development for the NAP-2 included two components: developing materials for eliciting a narrative from a child and developing an item pool for scoring the narrative.

Development of Elicitation Materials

Four sets of elicitation materials (books and associated story scripts) were developed in order to have multiple forms for administration of the NAP-2, as might be needed in a repeated testing situation (e.g., longitudinal measurement of a child's skills). Note that our protocol relied on story retelling rather than on story generation for eliciting narratives from children. Story generation procedures are often associated with limited linguistic output from young children (McCabe & Rollins, 1994), whereas story retelling supports young children in providing richer and more robust narratives (Merritt & Liles, 1989). We created the books and scripts to be as parallel as possible while ensuring that the stories were coherent and engaging. To that end, all four books followed the same temporal and causal sequence, with a title page and 16 pages with simple black-and-white illustrations, which clearly represented the salient visual elements of the plot. The content of each story was chosen to be representative of an everyday situation with which children from a variety of backgrounds could identify: going on a bike ride, cleaning a bedroom, getting ready for bed, and making lemonade. Each of the four stories followed the same plot sequence, representing identical plot lines according to a typical story grammar sequence (Stein & Glenn, 1979). Specifically, each story began with a page establishing the setting (e.g., playing outside on a beautiful day), followed by a page presenting the overall goal for the main character of the story (e.g., going on a bike ride with friends). Subsequently, there were three sets of a four-page sequence presenting a subgoal (e.g., trying to put on a helmet), a problem (helmet does not fit over hair), its solution (changing hair ties), and resolution (putting the helmet on). Finally, there was a resolution of the overall problem and a closing page including "The End" in the script. All books contained anthropoid animal characters; two stories included a girl as the main character, and two stories included a boy as the main character.

In addition to having each story be parallel in terms of the overall sequence and plot, we also developed scripts for each book that were similar in terms of microstructure elements of language; a sample script from one of the four stories appears in Appendix B. Each story script had one to three sentences per page and approximately the same number of words per story (± 10 words). Within each story, every scorable item (see item development below) occurred at least one time. Items that were used with less frequency (e.g., emotional references or direct quotes) were represented in similar locations in each script and occurred approximately the same number of times. Other items, such as pluralized nouns or irregular past tense verbs, occurred frequently and were represented throughout the stories. Scripts were piloted with five children between the ages of 3 and 6 years to make sure that children were engaged throughout the story. Engagement was determined by independent qualitative judgment of the videotaped pilot samples by three members of the research team; all three members had full agreement that the children were fully engaged throughout the length of all four books and associated scripts. Invariance of narrative skill scores across the four elicitation books was assessed using a differential item functioning analysis, described under Study 2 below.

Development of Item Pool

The NAP-2 was designed so that the narrative elicited from a child is scored from video using event-based frequency scoring, in which the scorer (who need not be the same individual who elicited the narrative) identifies occurrences of specific indicators that reflect an aspect of narrative skill (e.g., use of an infinitive). As in the original NAP (Justice et al., 2010), the assessor scores the frequency of occurrence for a series of individual items based on a rating scale of 0 (*no occurrence*), 1 (*one occurrence*), 2 (*two occurrences*), or 3+ (*three or more occurrences*). This event-based frequency scoring eliminates the need for transcription of children's narratives. The original NAP scored the frequency of occurrence for 18 items representative of a narrative's microstructure, such as use of copula "be" verbs and irregular past tense verbs. The NAP-2 was developed to provide a more comprehensive, holistic representation of children's narratives, representing both

macro- and microstructural features, with items identified empirically through a development process.

To develop the NAP-2 item pool, the research team, which included individuals with expertise in education, developmental psychology, psychometrics, and speech-language pathology, identified methods of scoring narratives used in previous research studies and published assessments and adapted the items (i.e., the specific methods used to score the narrative) to the NAP-2 event-based scoring approach. Our goal for the initial item pool was to take a broad, comprehensive approach to scoring the narrative by including all previous scoring methods that could be adapted to the NAP-2 as well as by identifying gaps in the scoring methods and creating additional items to fill the gaps. Thus, the initial item pool included all items from the original NAP (a total of 18 items; see Supplemental Material S1) and additional items as identified in a comprehensive analysis of extant research reports that describe measurable features of children's linguistic output within narrative or conversational contexts (e.g., Eisenberg et al., 2008; Huttenlocher et al., 2002; Justice et al., 2008; Liles et al., 1995; Peña et al., 2006; Petersen et al., 2008; Price et al., 2006). To identify these items, two graduate student research assistants independently searched scholarly journals on ProQuest/PsycInfo using search terms related to narrative assessment (e.g., narrative language assessment, microstructure, macrostructure, story grammar). We also conducted a web search using Google for commercially available language assessments. The searches had no restrictions on year published. From these assessments, we added any additional items to the initial item pool that could be adapted to the NAP-2 event-based scoring approach. The item pool was then reviewed for comprehensiveness by three experts who have extensive research experience and publication records related to children's narrative skill. Experts were asked to identify gaps in the item pool; based on the expert feedback, we added additional items to address narrative artfulness (e.g., inclusion of character name). The final item pool consisted of 60 items, listed in Table 1 and described in detail in Supplemental Material S2. Through procedures described shortly, this item pool was subsequently pruned to arrive at 20 empirically derived items featured in the NAP-2.

Study 2: Validation and Structure of Narrative Skill

The aim of Study 2 was to identify the structure of narrative skill, examine how the items form a coherent measurement system, and identify a smaller set of items that maintained sufficient reliability and validity. The latter goal was desirable in order to reduce the time needed to score the NAP-2 and thereby improve usability. To achieve this, we collected two narratives using the NAP-2 elicitation materials from 470 children ages 3–6 years and scored each narrative on all 60 items in the initial item pool. We considered two methods of dividing the full sample. First, we divided the narratives into two subsamples, a calibration subsample and a validation subsample, with the two narratives elicited from each child randomly assigned to one of the two subsamples. Thus, the calibration and validation subsamples included the same children, but with one narrative from a particular child in the calibration subsample and the other narrative from the same child in the validation subsample.[1] This allowed us to take a cross-validation approach by considering whether the findings from the calibration sample were cross-validated with the validation sample, leading to much stronger conclusions. This approach also minimized concerns with local dependence while maintaining large subsample sizes for the analyses. Second, we divided the sample into four samples based on the book used for the narrative elicitation. This allowed us to consider differences in item functioning across books.

We first provide descriptive findings for the entire sample. We then report results of an EFA used to identify the number of dimensions of narrative skill that the NAP-2 measures and to identify items that do not load strongly on any dimension. Finally, we used a Rasch measurement approach to examine item functioning and select a final set of items for the NAP-2.The Rasch approach offers a strong method for offering evidence of construct validity (Baghaei, 2008). Throughout, we repeated each analysis on the calibration, validation, and book subsamples to consider replicability of our results. We report detailed results on the calibration sample and note instances in which results from other subsamples differed.

Participants

Participants were 470 children (M_{age} = 59.2 months, SD = 12.1 months, range: 36–83 months) recruited from preschools and kindergartens from regions surrounding two research sites in the midwestern United States. Children were recruited primarily through communication with area schools and other educational organizations. Eligibility criteria were restricted solely to age (children were to be between the ages of 36 and 83 months [3;0–6;11 years; months]), language background (children were to communicate fluently in English), and no medical or developmental conditions that would moderately or greatly interfere with the ability of the child to provide a narrative; all exclusion criteria were based on parent or other primary caregiver report. Thus, the sample included both typically developing children (n = 421) and children with disabilities that did not substantially interfere with their ability to provide a narrative (n = 49). Some parents specified a type of disability; these included a language disability (n = 16), vision disability (n = 16), speech disability (n = 4), cerebral palsy (n = 4), unspecified physical disability (n = 2), learning disability (n = 2), hearing disability (n = 1), and disability related to attention (n = 1). Children were primarily

[1]Here, "subsample" refers to a subsample of the entire sample of narratives rather than a subsample of the children in the sample.

Table 1. Children's responses to initial Narrative Assessment Protocol-2 item pool (proportion per categorical response).

Dichotomously scored items				
Item	**No**	**Yes**		
Title[a]	95.9	4.1		
Abstract	99.3	0.7		
Conventional opening[a]	81.3	18.7		
Establish setting	30.0	70.0		
Establish overall goal	30.3	69.7		
Completion of overall goal	32.3	67.7		
Resolution of overall goal	25.4	74.6		
Conventional ending[a]	58.7	41.3		
Coda	98.4	1.6		
Polytomously scored items				
	0	**1**	**2**	**3+**
Interrogative: Tag questions	99.7	0.2	0.0	0.1
Interrogative: Yes/no questions	97.3	2.3	0.5	0.0
Interrogative: *Wh*-questions[a]	90.2	6.6	2.1	1.1
Prepositional phrase	27.5	20.8	18.2	33.5
Compound sentence	37.7	18.4	13.5	30.4
Complex sentence: Infinitive form	44.5	20.3	16.0	19.2
Complex sentence: Let form	97.8	2.2	0.0	0.0
Complex sentence: Coordinated form	61.4	18.7	11.4	8.5
Complex sentence: Subordinated form	49.0	20.9	14.3	15.8
Pluralized noun	18.1	17.8	22.6	41.6
Elaborated noun phrase with 1 adjective	56.0	21.8	13.1	9.1
Elaborated noun phrase with demonstrative and quantifier determiners	38.7	19.9	14.7	26.7
Elaborated noun phrase 2[a]	82.8	14.2	2.5	0.5
Elaborated noun phrase with possessive determiner	26.4	12.1	8.8	52.7
Postnoun modifier	94.4	4.7	0.7	0.2
Possessive form	96.4	2.7	0.7	0.2
Compound word	22.9	13.8	14.7	48.6
Tier 2 noun	26.0	13.9	18.8	41.3
Tier 2 adjective[a]	97.8	2.1	0.1	0.0
Pronoun error (reverse scored)	95.1	2.6	0.3	1.9
Auxiliary verb + main verb	27.7	22.5	20.3	29.4
Modal verb	44.5	28.6	12.9	14.0
Copula	26.7	18.8	18.0	36.4
Regular past tense verb	19.0	10.0	10.4	60.6
Irregular past tense verb	16.4	8.0	10.6	65.0
Negative verb form	44.0	26.4	15.6	14.0
Tier 2 verb[a]	78.0	12.9	6.7	2.3
Tier 2 adverb[a]	79.0	13.3	5.2	2.5
Place adverb	44.9	23.0	17.8	14.3
Time adverb	75.1	17.8	5.1	2.0
Manner or degree adverb	34.9	22.9	18.0	24.2
Conjoined adverbial phrase[a]	91.7	7.3	0.7	0.3
Verb morphology errors (reverse scored)	59.7	23.6	9.0	7.6
Character reference[a]	64.9	15.6	7.3	12.2
Temporal ordering[a]	86.0	9.4	3.6	1.0
Emotion reference[a]	76.3	18.9	2.8	2.0
Onomatopoeia	82.2	13.9	2.7	1.2
Stress	84.5	8.9	4.1	2.5
Elongations[a]	83.8	10.3	3.4	2.6
Repetition	74.1	18.1	5.4	2.5
Similes and metaphors[a]	93.7	5.4	0.9	0.0
Gratuitous terms[a]	68.5	17.2	6.0	8.3
Time reference[a]	63.2	26.7	6.8	3.2
Place reference	21.9	15.9	16.4	45.8
Direct quote with carrier	56.5	19.0	11.9	12.6
Direct quote without carrier	67.9	16.6	5.9	9.5
Indirect quote	85.4	11.1	2.7	0.9
Subgoal[a]	15.8	11.7	22.0	50.6
Subproblem[a]	13.1	12.0	18.9	56.0
Subsolution[a]	15.2	9.6	19.5	55.7
Subresolution[a]	13.1	10.3	18.4	58.1

[a]Items in the final version of the Narrative Assessment Protocol-2.

White (72%), Black (13%), or multiracial (10%); 2% were Asian, and the remaining 3% were marked as "other" or not reported by the caregiver. English was the primary language spoken by 96% of the children, according to the caregiver report. Caregiver-reported education included less than high school degree (3%), high school graduation (28%), 2-year degree (6%), 4-year degree (30%), master's degree (22%), and PhD (10%); the remaining 1% reported "other" or did not respond. Primary caregivers gave informed consent, and all children provided assent before any data collection session.

Procedure

The primary procedure for the initial NAP-2 validation study involved collecting two narratives from the children during a maximum 30-min one-on-one session with project staff. Children were assessed at schools, at home, in university laboratories, or at other community locations when necessary (e.g., libraries), based on caregiver preference. Prior to the session, caregivers filled out a demographic questionnaire and gave their permission for their children's participation, and children gave verbal assent to participate at the start of each session.

During the assessment session, trained research assistants elicited two separate narratives from each child, using two of the four story scripts developed in Study 1. The stories were randomly assigned to each child, and within a given session, these were administered in a randomly assigned order. To elicit the narrative, the research assistant read the book script to the child in page-by-page coordination with the wordless picture book and then handed the wordless picture book to the child and said, "Now it is your turn to tell me a make-believe story using the pictures in this book." Children's narratives were video-recorded for later scoring and analysis within a laboratory setting.

Narrative Scoring

Each child's videotaped narrative was scored by trained, reliable coders for frequency of occurrence for each of the initial 60 items. Most items were coded for frequency of occurrence, with a range of 0 to 3+, such that more than three occurrences were considered ceiling, although some items were coded as absent/present (0/1; e.g., presence of a title). Each coder was assigned only a subset of the 60 items to code, in order to minimize the amount of training needed to achieve reliable scoring of individual items. The items were divided into five sets of conceptually similar items (i.e., nouns and noun modifiers, verbs and verb modifiers, sentence complexity, storytelling conventions, story grammar), and each coder was trained to reliability on two or three item sets.

Prior to conducting any scoring, the coders were trained using a set of 18 master-coded videos that were coded for all items independently by three experts. Training involved reading sections of the *Syntax Handbook* (Justice & Ezell, 2002) to become familiar with key terminology related to the item set (e.g., verb, noun, clause, phrase). Then, each assistant read the item descriptions (for the subset of items to which they were assigned) and watched two videos with accompanying master-coded transcripts to familiarize themselves with the scoring process. Next, they practiced scoring five videos, comparing their scores to the master codes. Finally, research assistants scored three videos and had to reach 85% exact agreement with the master codes to achieve coding reliability; those who were unable to reach the criterion could repeat with a new set of three videos, repeating the reliability assessment up to a total of four times. All research assistants were able to achieve coding reliability within this approach.

Following training, the coders scored each narrative on the set of items on which they had trained. Scoring was conducted using videos collected from the children, and in determining the scores, the coders could pause or rewind the video as many times as necessary. To assess reliability of the scoring procedures, 20% of the narratives were randomly selected and double-coded. Overall, interrater agreement was high: The average exact agreement was .84, close to the trained criterion level of .85.

Results

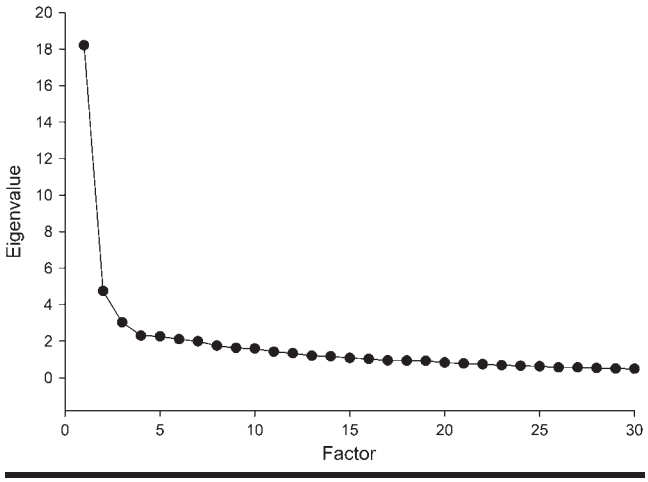
Table 1 provides the proportions of responses for each item in the initial pool. Across the 60 items, 12 had a highly skewed response pattern, in which at least 90% of responses were in a single category: title, abstract, coda, conjoined adverbial phrase, similes and metaphors, post-noun modifier, possessive form, Tier 2 adjective, pronoun error, interrogative tag, interrogative yes/no, and interrogative *wh*–.

EFA

EFA was used to examine the factor structure characterizing the 60 items in the initial set. The purpose of the analysis was to examine the nature of narrative skill, specifically its dimensionality (or factor structure), as well as to select items from among the larger, initial item pool of 60 items.

This analysis used the calibration sample and treated the item responses as categorical. (We also conducted an identical analysis, treating item responses as continuous; the results were similar to those we report here.) The pattern of eigenvalues supported a single factor, as shown in the scree plot in Figure 1. The one factor solution fit poorly (CFI = .85, TLI = .84, root-mean-square error of approximation [RMSEA] = .05), whereas the two-factor solution fit somewhat better ($\Delta\chi^2 = 974$, $\Delta df = 57$, CFI = .89, TLI = .89, RMSEA = .05). However, geomin rotated factor loadings indicated that the second factor was almost entirely driven by one item, abstract ($\lambda = -1.61$), and to a lesser extent several other items almost all rarely observed in children's narratives: coda ($\lambda = -.51$), onomatopoeia ($\lambda = .55$),

Figure 1. Scree plot for exploratory factor analysis (all 60 items).



similes and metaphors (.53), direct quote with carrier (.58), post–noun modifier (.51), Tier 2 adjective (−.59), pronoun error (.57), and complex sentence with subordinating clause (.60). Thus, the second factor appears to be primarily a difficulty factor. Dropping the abstract item yielded similar results, but with coda having a very strong loading (1.22). Continuing this process led to dropping two additional rare items: coda and Tier 2 adjective. Because of the consistency of results, we then dropped all items with proportions of 0 scores greater than .95: title, possessives, pronoun errors, and complex sentence let form (conclusions from EFA can be sensitive to highly skewed indicators; e.g., McDonald, 1965). This led to a substantially smaller second eigenvalue (3.08) and adequate to good fit for the one-factor model (CFI = .91, TLI = .90, RMSEA = .05). The two-factor model continued to fit significantly but only slightly better, likely due to the large sample size ($\Delta\chi^2 = 414$, $\Delta df = 50$, CFI = .93, TLI = .93, RMSEA = .04); furthermore, almost all items loaded strongly on the first factor ($\lambda > .40$), with those items loading strongly on the second factor not cohering theoretically (onomatopoeia, repetition, direct quotation with carrier, pronoun error, complex sentence subordinated form).

The factor analytic process was replicated with the validation sample and the four book samples, yielding similar results and clear evidence of a single factor. Overall, we concluded that a single factor best described narrative skill as measured by the NAP-2.

Rasch Analysis

Rasch analysis was subsequently used to further examine construct validity and to reduce the number of items in the item pool. To ensure that the full range of narrative ability could be measured, we considered all items, including those with extreme proportions identified in the factor analyses. The items were analyzed with the partial credit model (Masters, 1982), an extension of the Rasch (1960/1980) model for items scored with two or more ordered categories. The model is as follows:

$$\ln \text{odds}(X_{ni} = x \,|\, X_{ni} = x \text{ or } x-1) = \theta_n - \beta_{ix}, \quad (1)$$

where ln odds($X_{ni} = x \mid X_{ni} = x$ or $x - 1$) is the log odds that child n's response to item i is scored in category x, given that the score is either x or $x - 1$; θ_n is child n's level of narrative skill; and β_{ix} is the category threshold for item i and category x, similar to an item difficulty. This model has theoretical advantages such that fit to the model provides evidence of construct validity (Bond & Fox, 2001; Embretson, 1983; Fisher, 1994; Rost, 2001). It also has important practical advantages, including relatively modest sample size requirements (Linacre, 1994).

For our purposes, the model was estimated with the computer program Facets (Linacre, 2014). Consistent with typical practice with Rasch analysis, we examined the fit associated with each item by using two standard Rasch-based fit statistics, mean square outfit and mean square infit (Linacre, 2002), which have an expected value of 1 when an item fits the model. We identified misfitting items as having outfit or infit outside 0.6 to 1.4 bounds, based on Wright and Linacre's (1994) rating scale criteria. Based on these criteria, nine items had substantial misfit in the calibration sample: abstract (outfit = 2.30), repetition (outfit = 1.64), direct quotation without carrier (outfit = 1.68), manner adverb (outfit = 1.45), postnoun modifier (outfit = 1.82), compound word (outfit = 1.43), Tier 2 noun (outfit = 1.78), pronoun error (outfit = 3.86), and interrogative: tag questions (outfit = 2.12). Iteratively dropping misfitting items and repeating the analysis yielded six more misfitting items: coda (outfit = 1.45), onomatopoeia (outfit = 1.42), stress (outfit = 1.56), indirect quotation (outfit = 1.45), elaborated noun phrase (ENP) with demonstrative and quantifier determiners (outfit = 1.41), and possessive form (outfit = 1.43). In the validation sample, no additional items misfit, although repetition and all the iteratively dropped items that misfit in the calibration sample did not misfit in the validation sample. For the book samples, because of the higher likelihood of error from multiple assessments of fit with smaller sample sizes, we considered an item as misfitting if the item had fit statistics outside the bounds on at least two books. This yielded the same set of misfitting items except interrogative: tag questions, coda, onomatopoeia, ENP with demonstrative and quantifier determiners, and possessive form. To be conservative, we chose to exclude all 15 items identified as misfitting from further analyses.

Next, we considered differential item functioning (DIF) across the four books used to elicit narrative samples in the NAP-2 with both the calibration and validation samples, so as to ensure that items selected for the final version of the tool were invariant across books. There was some evidence of differences in overall difficulty across books within both the calibration sample (largest difference = .10, test of equality: $\chi^2 = 13.0$, $df = 3$, $p < .01$) and the

validation sample (largest difference = .21, test of equality: $\chi^2 = 65.0$, $df = 3$, $p < .01$); however, the differences were small in magnitude, and the largest differences were associated with a different pair of books in the calibration sample than in the validation sample. We set two criteria for identifying potential DIF relative to the item difficulty of the other three books combined: statistical significance ($p < .05$) and, consistent with typical practice in a Rasch measurement framework, a difference in item difficulty of at least 0.5 logits, which is considered large enough to impact ability estimates (e.g., Linacre, 2016). Only 11 out of 180 Item × Book interactions displayed DIF by these criteria, and there was no consistency in which items or books displayed DIF. Repeating the DIF analysis with the validation sample yielded 11 Item × Book interactions displaying DIF. Between the two samples, the DIF overlapped for six items: overall solution, plural noun, ENP possessive form, prepositional phrases, ENP demonstrative form, and complex sentence let form. These items were removed from further analyses. Finally, we combined the calibration and validation samples to perform a maximally powered DIF analysis using the entire sample. One additional item, ENP with one adjective, showed substantial and statistically significant DIF, so this item was deleted, leading to a final pool of 38 items.

Next, we considered DIF across gender separately for each book, combining the calibration and validation samples to increase power. Overall, boys scored slightly better than girls (difference = .08, test of equality: $\chi^2 = 20.1$, $df = 1$, $p < .01$), a difference that was consistent across all four books (range of difference: .04–.14). Only six out of 152 Item × Gender interactions displayed DIF, and there was no consistency in the number of items with DIF by book (0, 1, 2, or 3), nor which items displayed DIF. Thus, we concluded that there was no evidence of DIF associated with gender and continued with the pool of 38 items.

As a final step, we iteratively removed the individual items that had the lowest interrater agreement among the remaining items until we reached the smallest number of items for which reliability of the full item pool remained above .80, as estimated with the Rasch framework. This led to 20 items remaining in the final version of the NAP-2 with an estimated reliability of .81; the items are highlighted in Table 1, along with category frequencies. The NAP-2 spans a wide range of narrative skill levels, as it includes items that occurred relatively seldom in children's narratives, such as statement of the title and use of similes and metaphors, as well as items that occurred at relatively high levels, including references to time and references to characters. Mean scores for four age groupings are shown in Table 2. Rasch-based narrative skill estimates were positively correlated with age ($r = .35$, $p < .01$), and for nearly all items, there were gradual age-related changes, indicating that the NAP-2 captured age-related changes in narrative skill.

General Discussion

This article presents the development (Study 1) and validity evidence (Study 2) of a new tool for the assessment of narrative skill for children between the ages of 3 and 6 years. The NAP-2 joins a small group of newer generation narrative assessments that lessen the burden of transcription. The NAP-2 offers several key features that make it potentially an important part of the assessment toolkit for SLPs, educators, and researchers. First, the NAP-2 requires little time to administer: less than 10 min to collect the narrative sample and, in our experience, less than 10 min to complete scoring, although such speedy scoring does require training and experience in the coding system. Administration time is roughly equivalent or shorter than other narrative assessments. Second, the NAP-2 is cost-effective, as all materials for training, administration, and scoring are available free of charge online. Third, the NAP-2 uses a traditional narrative elicitation approach, although this precludes real-time scoring as children can produce many different forms of narratives with no constraint on instances of scored items (e.g., any Tier 2 verb counts even if it was not part of the original script); our experiences indicate that video recording with the option to pause and rewind is necessary for reliable coding, limiting use of the NAP-2 to those with access to video equipment. We note, however, that such high-quality video recording capacity is part of many technologies available in educational settings (e.g., laptops), so this constraint is unlikely to be a substantial barrier. Future research may establish that the NAP-2 can be scored from audiotape, potentially reducing but not eliminating the equipment challenge. Finally, the NAP-2 provides a broad assessment of narrative skill including microstructural and macrostructural features of children's narratives, has strong evidence of content validity through the development process described in Study 1, and has strong evidence of construct validity through a rigorous development process and item analysis using Rasch measurement techniques described in Study 2. The resulting tool, with its 20 items, provides the field an easy-to-use and scalable tool to examine children's narrative skill.

Narrative Skill As Measured by the NAP-2

One of the challenges with assessing narrative skill is the lack of agreement regarding what constitutes narrative skill and the nature of the construct of narrative skill. In our comprehensive literature review of existing narrative assessments, which involved 60 initial items, we found a wide variety of approaches to scoring narratives, with sometimes little or no overlap in item content across existing assessments. Thus, our finding that narrative skill as measured by the NAP-2 consisted of a single factor is particularly noteworthy; although we found some statistical evidence for a second factor, there was no coherence in content of items with loadings on the second factor. Prior work examining the structure of narrative skill in young children has relied on conceptual distinctions among various aspects of narrative, such as microstructural aspects (e.g. Justice et al., 2006), macrostructure (Stein & Glenn, 1979), and narrative quality (Ukrainetz et al., 2005). Furthermore, some narrative assessments provide separate scores for

Table 2. Children's mean performance on the Narrative Assessment Protocol-2 final item set by age groupings.

No.	Item	Age groupings			
		3 years (*n* = 104)	4 years (*n* = 145)	5 years (*n* = 124)	6 years (*n* = 97)
1[a]	Title	0.01	0.01	0.04	0.09
2[a]	Conventional opening	0.09	0.15	0.21	0.30
3	Character reference	0.31	0.57	0.82	0.91
4	Temporal ordering	0.15	0.17	0.24	0.26
5	Emotion reference	0.24	0.32	0.36	0.36
6	Elongations	0.34	0.20	0.25	0.19
7	Similes and metaphors	0.02	0.09	0.11	0.06
8	Gratuitous terms	0.48	0.53	0.56	0.78
9	Time reference	0.27	0.44	0.55	0.74
10	Tier 2 verb	0.13	0.29	0.48	0.53
11	Tier 2 adverb	0.14	0.42	0.42	0.36
12	Tier 2 adjective				
13	Interrogative: *Wh*-questions	0.07	0.08	0.20	0.34
14	Conjoined adverbial phrase	0.07	0.12	0.08	0.08
15	Elaborated Noun Phrase 2	0.09	0.19	0.26	0.36
16	Subgoal	1.41	1.93	2.32	2.65
17	Subproblem	1.70	2.10	2.40	2.61
18	Subsolution	1.50	2.07	2.46	2.67
19	Subresolution	1.60	2.11	2.52	2.71
20[a]	Conventional ending	0.38	0.36	0.46	0.50
	Rasch narrative skill estimate	18.96	19.81	20.37	20.93

Note. Rasch narrative skill estimate is the logit narrative skill estimate rescaled to a mean = 20, *SD* = 2 scale.
[a]These items are scored as absent (0)/present (1); all others are coded for frequency of occurrence (0, 1, 2, 3+).

different aspects of narratives, such as separately scoring story grammar and language complexity, without empirical support for such a distinction (e.g., Renfrew Bus Story; Renfrew, 1969). Our empirical data suggest that such conceptual distinctions between different aspects of narrative do not reflect distinctions in the narratives that young children produce, and in particular, there may be no meaningful distinction between microstructural narrative skill and macrostructural narrative skill for young children. Further research is needed to confirm this counter-theoretical finding.

The finding that narrative skill forms a unidimensional construct coincides with other research in prekindergarten and kindergarten, which indicates that a unidimensional model of grammar, vocabulary, and discourse describes children's language skills as well as, or better, than multidimensional models (Language and Reading Research Consortium, 2015). The unidimensional nature of language in prekindergarten has led some researchers to recommend that researchers and educators streamline assessment batteries for this age group so as to reduce redundant information (Anthony et al., 2014). Indeed, the final 20 items of the NAP-2 does this while still representing a variety of narrative features thought to be key for capturing children's language development in the extant literature. Story grammar elements (subgoals, problems, solutions, resolutions) are included in four items (Stein & Glenn, 1979). Storytelling conventions (title, conventional opening and ending, orientation to time), deemed to be important in the high-point analysis work of Hughes et al. (1997), are included in four items. Evaluative aspects that lend color and artfulness to narratives (elongations, similes and metaphors, gratuitous terms, emotional state references) are represented in four items (Peterson & McCabe, 1983). Lexical diversity and microstructural aspects of narrative (Tier 2 adjectives, Tier 2 verbs, Tier 2 adverbs, ENPs, conjoined adverbials) are captured in five items. Two items represent information regarding character perspectives (character references, *wh*-questions), and one item represents conjunctive cohesion (temporal ordering).

Thus, it is evident that items on the NAP-2 represent a broad and comprehensive range of narrative features that work together to provide structural coherence, internal cohesion, and richness to spoken narratives. Nonetheless, further research is needed to understand the nature of the construct of narrative skill. In particular, we encourage research considering how the narrative skill, as measured by the NAP-2, is related to narrative skill as measured by other contemporary assessments such as the TNL-2 and CUBED NLM, as well as more traditional approaches involving transcription.

Development of Narrative Skill

The NAP-2 was designed to be appropriate for children ages 3–6 years, a time when children's language skills are developing rapidly (Curenton & Justice, 2004). Consistent with this development, we found that the scores on the NAP-2 and the frequency of occurrence of almost all items increased with age. This pattern of age-related differences in narrative structures is consistent with prior work, showing

that children are rapidly acquiring and consolidating information about how stories are organized (Khan et al., 2016; Trabasso et al., 1992). Simultaneously, children are also acquiring linguistically specific knowledge to express temporal relations, mark references, and create interclausal connectivity in their narratives (e.g., Hickmann, 2004). Thus, the NAP-2 not only captures variability across children in narrative skill but also offers the potential to capture within-child developmental changes, although longitudinal studies are needed to confirm this potential.

Elongations was the one item that did not increase in frequency across age. Elongations did meet item selection criteria using exploratory factor analysis and Rasch analysis, as our item selection process did not consider developmental progressions. Thus, it is possible that this unexpected finding is simply the result of chance with a relatively rare item and not scientifically meaningful. However, it is possible that the use of elongations, which might be considered a form of narrative artfulness, decreases with age, perhaps as children employ a wider variety of artfulness techniques such as similes, metaphors, and gratuitous terms. To our knowledge, no previous research has considered the nature of narrative development at this level of detail.

Limitations

Two important limitations should be noted. First, the sample of children who participated in Study 2 was primarily monolingual, English-speaking children. However, the number of English language learners is increasing rapidly within the United States (NCES, 2009), and the nature of language development appears to be somewhat different for children who are bilingual (Bedore & Peña, 2008). It is unknown whether our conclusions about the nature of narrative skill or the final selection of items for the NAP-2 would be the same if a more linguistically diverse sample of children had been included. Research on the nature of narrative skill in different cultural and linguistic contexts is quite limited, so an important future direction is examination of the functioning of the NAP-2 and consideration of the construct of narrative skill with diverse populations.

Second, the event-based scoring method used with the NAP-2 limits the ability to identify all possible ways to score narratives. In particular, event-based scoring without transcripts does not allow for consideration of items or scoring methods that require the full narrative. For example, a common form of narrative scoring involves counting the total number of words or total number of different words (e.g., Justice et al., 2006). Some prior research suggests that such measures of productivity may reflect a different aspect of narrative skill than the measures of complexity captured in the NAP-2 (Justice et al., 2006).

Using the NAP-2

Currently, the NAP-2 offers an accessible tool for researchers, educators, and allied professionals to measure a 3- to 6-year-old child's narrative skill. To maximize usefulness of the NAP-2 for this purpose, we have created a website, https://www.narrativeassessment.com, which includes all elicitation materials, along with training and online scoring. End users must complete the training and meet the same scoring reliability as described in Study 2 (85% exact agreement across three videos). Preliminary studies indicate that end users, including early childhood educators and SLPs, can reliably score the NAP-2, although further research is needed to confirm this finding. The NAP-2 therefore has the potential to provide information about a child's narrative skill in a very simple and easy-to-use task format.

Results from these studies are promising with respect to the psychometric caliber of the NAP-2, but before recommending it for broad usage, we encourage further research in two key potential areas that the NAP-2 might be used. First, the NAP-2 may serve to augment comprehensive language evaluations to identify areas of strength and needs for children, such as determining whether discourse-level skills are an area of need for children. Current consensus statements argue for the use of multicomponent tools in determining the presence of language impairment in young children (Bishop et al., 2016); many tools used in the process may be complementary to those used for diagnostic purposes, namely, standardized tools with high degrees of sensitivity and specificity. Given the feasibility of NAP-2 for inclusion in multicomponent batteries, we argue that it can be used as an important mechanism for assessing children's narrative-based skills. Relatedly, we also propose that the NAP-2 has potential to serve as a diagnostic measure for language impairment, in line with other assessments of narrative skill (e.g., Gillam & Pearson, 2017; Pankratz et al., 2007; Peña et al., 2006). However, additional research is needed not only to identify diagnostic cutoffs and criteria such as sensitivity and specificity but also to determine if the items of the NAP-2 work differently for children with disabilities such as language impairments. For example, we found that pronoun errors misfit in the Rasch measurement approach; it is possible that such errors would offer important information for diagnosing language impairments even if they are not informative for measuring narrative skill in general. Because we have provided the NAP-2 freely to the community of researchers, there is great potential for the ongoing evaluation of the tool's diagnostic potential for numerous subgroups of children at risk for narrative concerns.

Second, the NAP-2 also may be used for formative and summative purposes so as to examine the efficacy of programs designed to improve narrative skill, at the individual child level and at the group level. For instance, a kindergarten teacher may utilize the NAP-2 to examine narrative performance for each child in her classroom and to provide a snapshot of classroom-level growth over an academic year. Given the increasing prominence of narrative promotion within educational standards and practice guides (e.g., Foorman et al., 2016), professionals will have interest in determining the effects of specific practices on students' narrative development. We offer norm-referenced

interpretations of NAP-2 scores on our website based on the sample in Study 2, acknowledging that a U.S. population–based representative sample (such as offered by the TNL-2) would provide more valid benchmarks. By making the tool itself openly available for utilization by professionals, the NAP-2 has the potential to impact large numbers of educators at very low costs. While further research is needed to identify and validate bench marks with the NAP-2, including benchmarks for specific populations such as children with disabilities, its open-source availability helps to ensure that such research can be readily pursued by members of the research and practice community.

Conclusion

The NAP-2 offers a psychometrically sound and easy-to-use assessment of narrative skill for children ages 3–6 years and, with further validation research, may augment comprehensive language evaluations, potentially serve in a diagnostic capacity, or be used for formative and summative purposes. The NAP-2 uses an event-based frequency approach to eliminate transcription, reducing the time required for scoring, and it is available freely online for use by SLPs, educational practitioners, and researchers. The NAP-2 has good evidence of internal structure validity through a rigorous Rasch measurement item analysis, which indicated that narrative skill, as measured by the NAP-2, forms a unidimensional construct. In light of these features and psychometric evidence, we conclude that the NAP-2 is a promising approach to better understand children's narrative skill.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A110293 awarded to Michigan State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anthony, J. L., Davis, C., Williams, J. M., & Anthony, T. I. (2014). Preschoolers' oral language abilities: A multi-level examination of dimensionality. *Learning and Individual Differences, 35,* 56–61. https://doi.org/10.1016/j.lindif.2014.07.004

Baghaei, P. (2008). The Rasch model as a construct validity tool. *Rasch Measurement Transactions, 22*(1), 1145–1146.

Bedore, L. M., & Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism, 11*(1), 1–29. https://doi.org/10.2167/beb392.0

Berman, R. A. (1995). Narrative competence and storytelling performance: How children tell stories in different contexts. *Journal of Narrative and Life History, 5*(4), 285–313. https://doi.org/10.1075/jnlh.5.4.01nar

Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & Catalise Consortium. (2016). CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLOS ONE, 11*(7), e0158753.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Erlbaum.

Botvin, G. J., & Sutton-Smith, B. (1977). The development of structural complexity in children's fantasy narratives. *Developmental Psychology, 13*(4), 377–388. https://doi.org/10.1037/0012-1649.13.4.377

Boudreau, D. M., & Hedberg, N. L. (1999). A comparison of early literacy skills in children with specific language impairment and their typically developing peers. *American Journal of Speech-Language Pathology, 8*(3), 249–260. https://doi.org/10.1044/1058-0360.0803.249

Calkins, L., Ehrenworth, M., & Lehman, C. (2012). *Pathways to the common core*. Heinemann.

Curenton, S. M., & Justice, L. M. (2004). African American and Caucasian preschoolers' use of decontextualized language: Literate language features in oral narratives. *Language, Speech, and Hearing Services in Schools, 35*(3), 240–253. https://doi.org/10.1044/0161-1461(2004/023)

De Villiers, P., & Burns, F. (2003, November). *Assessing narrative skills in children*. Paper presented at the American Speech-Language-Hearing Association Annual Convention, Chicago, IL, United States.

Deese, J. (1983). *Thought into speech: Psychology of a language* Prentice Hall.

Eisenberg, S. L., Ukrainetz, T. A., Hsu, J. R., Kaderavek, J. N., Justice, L. M., & Gillam, R. B. (2008). Noun phrase elaboration in children's spoken stories. *Language, Speech, and Hearing Services in Schools, 39*(2), 145–157. https://doi.org/10.1044/0161-1461(2008/014)

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197. https://doi.org/10.1037//0033-2909.93.1.179

Engel, S. (1995). *The stories children tell*. W. H. Freeman.

Fisher, W. P., Jr. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. II, pp. 36–72). Ablex Publishing.

Foorman, B., Beyler, N., Borradaile, K., Coyne, M., Denton, C. A., Dimino, J., Furgeson, J., Hayes, L., Henke, J., Justice, L., Keating, B., Lewis, W., Sattar, S., Streke, A., Wagner, R., & Wissel, S. (2016). *Foundational skills to support reading for understanding in kindergarten through 3rd grade* (NCEE 2016-4008). National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. http://whatworks.ed.gov

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., Bohnacker, U., & Walters, J. (2012). *Multilingual Assessment Instrument for Narratives (MAIN)*. ZAS.

Gardner-Neblett, N., & Iruka, I. U. (2015). Oral narrative skills: Explaining the language-emergent literacy link by race/ethnicity and SES. *Developmental Psychology, 51*(7), 889–904. https://doi.org/10.1037/a0039274

Gardner-Neblett, N., Pungello, E. P., & Iruka, I. U. (2012). Oral narrative skills: Implications for the reading development of African American children. *Child Development Perspectives, 6*(3), 218–224. https://doi.org/10.1111/j.1750-8606.2011.00225.x

Gavin, W. J., & Giles, L. (1996). Sample size effects on temporal reliability of language sample measures of preschool children.

Journal of Speech and Hearing Research, 39(6), 1258–1262. https://doi.org/10.1044/jshr.3906.1258

Gillam, R. B., & Pearson, N. (2017). *Test of Narrative Language–Second Edition*. Pro-Ed.

Govindarajan, K., & Paradis, J. (2019). Narrative abilities of bilingual children with and without developmental language disorder (SLI): Differentiation and the role of age and input factors. *Journal of Communication Disorders, 77,* 1–16. https://doi.org/10.1016/j.jcomdis.2018.10.001

Griffin, T. M., Hemphill, L., Camp, L., & Wolf, D. P. (2004). Oral discourse in the preschool years and later literacy skills. *First Language, 24*(2), 123–147. https://doi.org/10.1177/0142723704042369

Hickmann, M. (2004). Coherence, cohesion, and context: Some comparative perspectives in narrative development. In S. Strömqvist & L. Verhoeven (Eds.), *Relating events in narrative: Typological and contextual perspectives* (pp. 281–306). Erlbaum.

Hughes, D. L., McGillivray, L., & Schmidek, M. (1997). *Guide to narrative language: Procedures for assessment*. Thinking Publications.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology, 45*(3), 337–374. https://doi.org/10.1016/s0010-0285(02)00500-5

Hux, K., Morris-Friehe, M., & Sanger, D. D. (1993). Language sampling practices: A survey of nine states. *Language, Speech, and Hearing Services in Schools, 24*(2), 84–91. https://doi.org/10.1044/0161-1461.2402.84

Justice, L. M., Bowles, R. P., Kaderavek, J. N., Ukrainetz, T. A., Eisenberg, S. L., & Gillam, R. B. (2006). The Index of Narrative Microstructure: A clinical tool for analyzing school-age children's narrative performances. *American Journal of Speech-Language Pathology, 15*(2), 177–191. https://doi.org/10.1044/1058-0360(2006/017)

Justice, L. M., Bowles, R. P., Pence, K., & Gosse, C. (2010). A scalable tool for assessing children's language abilities within a narrative context: The NAP (Narrative Assessment Protocol). *Early Childhood Research Quarterly, 25*(2), 218–234. https://doi.org/10.1016/j.ecresq.2009.11.002

Justice, L. M., & Ezell, H. K. (2002). *The syntax handbook: Everything you learned about syntax…but forgot*. Pro-Ed.

Justice, L. M., Mashburn, A., Pence, K. L., & Wiggins, A. (2008). Experimental evaluation of a preschool language curriculum: Influence on children's expressive language skills. *Journal of Speech, Language, and Hearing Research, 51*(4), 983–1001. https://doi.org/10.1044/1092-4388(2008/072)

Khan, K. S., Gugiu, M. R., Justice, L. M., Bowles, R. P., Skibbe, L. E., & Piasta, S. B. (2016). Age-related progressions in story structure in young children's narratives. *Journal of Speech, Language, and Hearing Research, 59*(6), 1395–1408. https://doi.org/10.1044/2016_JSLHR-L-15-0275

Labov, W. (1972). *Language in the inner city*. University of Pennsylvania Press.

Laing, S. P., & Kamhi, A. (2003). Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Language, Speech, and Hearing Services in Schools, 34*(1), 44–55. https://doi.org/10.1044/0161-1461(2003/005)

Language and Reading Research Consortium. (2015). The dimensionality of language ability in young children. *Child Development, 86*(6), 1948–1965. https://doi.org/10.1111/cdev.12450

Language Dynamics Groups. (2019). *CUBED validity*. http://www.languagedynamicsgroup.com/products/cubed/cubed-validity/

Liles, B. Z., Duffy, R. J., Merritt, D. D., & Purcell, S. L. (1995). Measurement of narrative discourse ability in children with language disorders. *Journal of Speech and Hearing Research, 38*(2), 415–425. https://doi.org/10.1044/jshr.3802.415

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16,* 878.

Linacre, J. M. (2014). Facets computer program for many-facet Rasch measurement [Computer program]. Winsteps.

Linacre, J. M. (2016). *Winsteps Rasch measurement computer program user's guide*. Winsteps.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

McCabe, A., & Peterson, C. (1984). What makes a good story. *Journal of Psycholinguistic Research, 13*(6), 457–480. https://doi.org/10.1007/bf01068179

McCabe, A., & Rollins, P. (1994). Assessment of preschool narrative skills. *American Journal of Speech-Language Pathology, 3*(1), 45–56. https://doi.org/10.1044/1058-0360.0301.45

McDonald, R. P. (1965). Difficulty factors and non-linear factor analysis. *British Journal of Mathematical and Statistical Psychology, 18*(1), 11–23. https://doi.org/10.1111/j.2044-8317.1965.tb00690.x

Merritt, D. D., & Liles, B. Z. (1989). Narrative analysis: Clinical applications of story generation and story retelling. *Journal of Speech and Hearing Disorders, 54*(3), 438–447. https://doi.org/10.1044/jshd.5403.438

Muñoz, M. L., Gillam, R. B., Peña, E. B., & Gulley-Faehnle, A. (2003). Measures of language development in fictional narratives of Latino children. *Language, Speech, and Hearing Services in Schools, 34*(4), 332–342. https://doi.org/10.1044/0161-1461(2003/027)

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects.*

NCES. (2009). *The condition of education 2009*.

Neuman, S. B., & Roskos, K. (2005). The state of state prekindergarten standards. *Early Childhood Research Quarterly, 20*(2), 125–145.

Nicolopoulou, A., & Trapp, S. (2018). Narrative interventions for children with language disorders: A review of practices and findings. In A. Bar-On & D. Ravid (Eds.), *Handbook of communication disorders* (pp. 357–386). De Gruyter Mouton.

Pankratz, M. E., Plante, E., Vance, R., & Insalaco, D. M. (2007). The diagnostic and predictive validity of the Renfrew bus story. *Language, Speech, and Hearing Services in Schools, 38*(4), 390–399. https://doi.org/10.1044/0161-1461(2007/040)

Pavelko, S. L., Owens, R. E., Jr., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools, 47*(3), 246–258. https://doi.org/10.1044/2016_LSHSS-15-0044

Peña, E. D., Gillam, R. B., Malek, M., Ruiz-Felter, R., Resendiz, M., Fiestas, C., & Sabel, T. (2006). Dynamic assessment of school-age children's narrative ability: An experimental investigation of classification accuracy. *Journal of Speech, Language, and Hearing Research, 49*(5), 1037–1057. https://doi.org/10.1044/1092-4388(2006/074)

Pesco, D., & Kay-Raining Bird, E. (2016). Perspectives on bilingual children's narratives elicited with the Multilingual Assessment Instrument for Narratives. *Applied Psycholinguistics, 37*(1), 1–9. https://doi.org/10.1017/S0142716415000387

Petersen, D. B. (2011). A systematic review of narrative-based language intervention with childrenwho have language impairment. *Communication Disorders Quarterly, 32*(4), 207–220. https://doi.org/10.1177/1525740109353937

Petersen, D. B., Gillam, S. L., & Gillam, R. (2008). Emerging procedures in narrative assessment: The Index of Narrative Complexity. *Topics in Language Disorders, 28*(2), 115–130. https://doi.org/10.1097/01.tld.0000318933.46925.86

Petersen, D. B., & Spencer, T. D. (2016a). *CUBED*. Language Dynamics Group. http://www.languagedynamicsgroup.com

Petersen, D. B., & Spencer, T. D. (2016b). Using narrative intervention to accelerate canonical story grammar and complex language growth in culturally diverse preschoolers. *Topics in Language Disorders, 36*(1), 6–19. https://doi.org/10.1097/tld.0000000000000078

Peterson, C., & McCabe, A. (1983). *Three ways of looking at a child's narrative: A psycholinguistic analysis*. Plenum.

Price, J. R., Roberts, J. E., & Jackson, S. C. (2006). Structural development of the fictional narratives of African American preschoolers. *Language, Speech, and Hearing Services in Schools, 37*(3), 178–190. https://doi.org/10.1044/0161-1461(2006/020)

Rasch, G. (1980). *Probabilistic models for some intelligence and attainments tests* (Expanded edition). University of Chicago Press. (Original work published 1960)

Reese, E., Suggate, S., Long, J., & Schaughency, E. (2010). Children's oral narrative and reading skills in the first 3 years of reading instruction. *Reading and Writing, 23*(6), 627–644. https://doi.org/10.1007/s11145-009-9175-9

Renfrew, C. (1969). *The bus story: A test of continuous speech*.

Rost, J. (2001). The growing family of Rasch models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory*. Springer-Verlag.

Schraeder, T., Quinn, M., Stockman, I. J., & Miller, J. (1999). Authentic assessment as an approach to preschool speech-language screening. *American Journal of Speech-Language Pathology, 8*(3), 195–200. https://doi.org/10.1044/1058-0360.0803.195

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *Advances in discourse processes* (Vol. 2): New directions in discourse processing (pp. 53–120). Ablex.

Tilstra, J., & McMaster, K. (2007). Productivity, fluency, and grammaticality measures from narratives: Potential indicators of language proficiency. *Communication Disorders Quarterly, 29*(1), 43–53. https://doi.org/10.1177/1525740108314866

Trabasso, T., Stein, N. L., Rodkin, P. C., Munger, M. P., & Baughn, C. R. (1992). Knowledge of goals and plans in the online narration of events. *Cognitive Development, 7*(2), 133–170. https://doi.org/10.1016/0885-2014(92)90009-g

U.S. Department of Health and Human Services, Administration for Children and Families, & Office of Head Start. (2015). *Head Start early learning outcomes framework: Ages birth to five*.

Ukrainetz, T. A., Justice, L. M., Kaderavek, J. N., Eisenberg, S. l., & Gillam, R. B. (2005). The development of expressive elaboration in fictional narratives. *Journal of Speech, Language, and Hearing Research, 48*(6), 1363–1377. https://doi.org/10.1044/1092-4388(2005/095)

Washington, J. A., & Craig, H. K. (2004). A language screening protocol for use with young African American children in urban settings. *American Journal of Speech-Language Pathology, 13*(4), 329–340. https://doi.org/10.1044/1058-0360(2004/033)

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370.

Zevenbergen, A. A., Whitehurst, G. J., & Zevenbergen, J. A. (2003). Effects of a shared-reading intervention on the inclusion of evaluative devices in narratives of children from low-income families. *Journal of Applied Developmental Psychology, 24*(1), 1–15. https://doi.org/10.1016/s0193-3973(03)00021-2

Appendix A

Final Item Set for the Narrative Assessment Protocol-2

No.	Item	Example
1[a]	Title	Raccoon makes lemonade.
2[a]	Conventional opening	Once upon a time…
3	Character reference	Rita went to the garage.
4	Temporal ordering	First, she found her bike.
5	Emotion reference	She was excited.
6	Elongations	It took a looooooong time.
7	Similes and metaphors	Her eyes as got as big as tomatoes.
8	Gratuitous terms	She was really frustrated.
9	Time reference	It was morning.
10	Tier 2 verb	Rita sprinted into the garage.
11	Tier 2 adverb	She suddenly got on her bike.
12	Tier 2 adjective	She got her beautiful bow.
13	Interrogative *wh*- questions	"Why are you going inside?"
14	Conjoined adverbial phrase	She quickly and carefully got on her bike.
15	Elaborated noun phrase	The cute furry rabbit
16	Subgoal	He opened the fridge to get the pitcher.
17	Subproblem	The pitcher was empty.
18	Subsolution	He sets up a ladder to reach it.
19	Subresolution	He pumps up the tire.
20[a]	Conventional ending	The end.

[a]These items are scored as absent (0)/present (1); all others are coded for frequency of occurrence (0, 1, 2, 3+). The complete coding catalog, with descriptions and additional examples for each item, is available for download at https://www.narrativeassessment.com.

Appendix B

Sample Narrative Assessment Protocol-2 Script: Raccoon Makes Lemonade

1. One hot, summer day, Rachel Raccoon and her three best friends were playing soccer. The sun was beating down, and everyone was getting hot and sweaty. Soon, everyone was thirsty.
2. Suddenly, Rachel had an idea. “Let’s have some lemonade!” she exclaimed.
3. While her friends finished the game, Rachel dashed inside and swung open the fridge. There, on the top shelf, was a pitcher of lemonade.
4. Rachel reached for the lemonade, but it was too high. She couldn’t reach it! “How am I going to get that down?” she asked herself.
5. Rachel was feeling frustrated, until she spotted a step stool under the counter. “Aha!” she thought. Rachel grabbed the heavy stool and put it in front of the fridge. Thud!
6. Rachel climbed on top of the stool and carefully removed the pitcher from the fridge.
7. Rachel took a big gulp and tasted the drink.
8. “Bleck!” she cried, with puckered lips. “We can’t drink that, it’s way too sour!” But Rachel had a plan.
9. First, Rachel got the sugar bag from the cupboard, and then she put three heaping spoonfuls into the lemonade.
10. Rachel tried the lemonade a second time, and this time it tasted yummy and sweet! “Perfect!” she said happily.
11. Very carefully, Rachel carried the pitcher outside so she could share it with her friends. She announced that the lemonade was ready. “Yay!” everyone jumped and cheered.
12. They quickly ran over for their drinks, but there was something missing. No cups! How were they going to drink the lemonade?
13. Rachel felt silly—of course they needed cups! They couldn’t drink lemonade out of their hands! While her friends waited, Rachel ran inside quick as a flash to get cups.
14. After just a moment, Rachel returned with a tray and four cups, grinning happily. The cups clinked against one another —Clink! Clink!
15. Rachel very slowly and carefully poured the lemonade into each cup. She made sure that everyone got the same amount so that no one felt sad.
16. Finally, it was time to drink the lemonade. The four friends raised their cups and gulped down the sweet, refreshing lemonade. “Thank you, Rachel!” everyone yelled. The end.