

Journal of Educational Psychology

Improving Oral and Written Narration and Reading Comprehension of Children at-Risk for Language and Literacy Difficulties: Results of a Randomized Clinical Trial

Sandra Laing Gillam, Sharon Vaughn, Greg Roberts, Philip Capin, Anna-Maria Fall, Megan Israelsen-Augenstein, Sarai Holbrook, Rebekah Wada, Allison Hancock, Carly Fox, Jordan Dille, Beula M. Magimairaj, and Ronald B. Gillam

Online First Publication, September 8, 2022. <http://dx.doi.org/10.1037/edu0000766>

CITATION

Gillam, S. L., Vaughn, S., Roberts, G., Capin, P., Fall, A.-M., Israelsen-Augenstein, M., Holbrook, S., Wada, R., Hancock, A., Fox, C., Dille, J., Magimairaj, B. M., & Gillam, R. B. (2022, September 8). Improving Oral and Written Narration and Reading Comprehension of Children at-Risk for Language and Literacy Difficulties: Results of a Randomized Clinical Trial. *Journal of Educational Psychology*. Advance online publication. <http://dx.doi.org/10.1037/edu0000766>

Improving Oral and Written Narration and Reading Comprehension of Children at-Risk for Language and Literacy Difficulties: Results of a Randomized Clinical Trial

Sandra Laing Gillam¹, Sharon Vaughn², Greg Roberts², Philip Capin², Anna-Maria Fall², Megan Israelsen-Augenstein¹, Sarai Holbrook¹, Rebekah Wada¹, Allison Hancock¹, Carly Fox¹, Jordan Dille², Beula M. Magimairaj¹, and Ronald B. Gillam¹


¹ Department of Communicative Disorders and Deaf Education, Utah State University


² Meadows Center for Preventing Educational Risk, University of Texas at Austin

Narration has been shown to be a foundational skill for literacy development in school-age children. Elementary teachers routinely conduct classroom lessons that focus on reading decoding and comprehension, but they rarely provide instruction in oral narration (Hall et al., 2021). This multisite randomized controlled trial was designed to rigorously evaluate the efficacy of the *Supporting Knowledge of Language and Literacy (SKILL)* intervention program for improving oral narrative comprehension and production. Three hundred fifty-seven students who were at-risk for language and literacy difficulties in Grades 1–4 in 13 schools across seven school districts were randomly assigned to the *SKILL* treatment condition or a business as usual (BAU) control condition. *SKILL* was provided to small groups of two to four students in 36 thirty-minute lessons across a 3-month period. Multilevel modeling with students nested within teachers and teachers nested within schools revealed students who received the *SKILL* treatment significantly outperformed students in the BAU condition on measures of oral narrative comprehension and production immediately after treatment. Oral narrative production for the *SKILL* treatment group remained significantly more advanced at follow-up testing conducted 5 months after intervention ended. Improvements in oral narration generalized to a measure of written narration at posttest and the treatment advantage was maintained at follow-up. Grade level did not moderate effects for oral narration, but it did for reading comprehension, with a higher impact for students in grades 3 and 4.

Educational Impact and Implications Statement

Understanding and composing stories is a particularly important developmental milestone for school-age students but is not often a focus of instruction by classroom teachers. This study evaluated the impact of a comprehensive narrative instructional program on oral and written educational outcomes for a large sample of school-age students who were at-risk for language and literacy difficulties. Students were randomized to the treatment group or to a business-as-usual control group. Storytelling instruction was provided to small groups of two to four students in 36 thirty-minute lessons delivered across a 3-month period by trained teachers and special educators. The students who received the instruction significantly outperformed the students who did not on measures of

Sharon Vaughn  <https://orcid.org/0000-0001-8305-5549>


Philip Capin  <https://orcid.org/0000-0003-4955-9879>


Megan Israelsen-Augenstein  <https://orcid.org/0000-0003-0560-5351>

Rebekah Wada  <https://orcid.org/0000-0002-8654-007X>

Allison Hancock  <https://orcid.org/0000-0003-3529-3811>

Carly Fox  <https://orcid.org/0000-0003-1344-4965>

Jordan Dille  <https://orcid.org/0000-0002-5110-8973>

Ronald B. Gillam  <https://orcid.org/0000-0003-0106-1908>

This study was funded by Grant R305A170111 from the National Center for Education Research. Ronald B. Gillam receives royalties from the sale of the Test of Narrative Language, which was administered to the participants. Ronald B. and Sandra Laing Gillam receive royalties from the sale of the SKILL intervention. No other authors have any conflicts of interests to disclose.

David Francis and Laura Justice, who served as consultants to this project, provided valuable advice about methods, analyses, and data interpretation. This study could not have been completed without the dedication of the team of special educators, classroom teachers, and speech-language pathologists who provided the intervention and the team of teachers, speech-language pathologists, graduate students, and undergraduate students who collected and analyzed the data. We are particularly grateful to the school administrators and teachers in the participating schools in Texas and Utah. Last, but not least, we thank the children and their families for their commitment to this study.

Correspondence concerning this article should be addressed to Ronald B. Gillam, Emma Eccles Jones Early Childhood Education and Research Center, 2610 Old Main Hill, Logan, UT 84322, United States. Email: ron.gillam@usu.edu

storytelling and story comprehension immediately after treatment. The storytelling gains were maintained at follow-up testing 5 months after the instruction ended. Children in the treatment group also made greater gains on a measure of story writing at posttest and follow-up. This study highlights the importance of strengthening oral language skills to support the development of academic skills such as story writing.

Keywords: at-risk students, language intervention, oral narration, reading comprehension, written narration

Perhaps the most significant role of elementary teachers is to help students develop reading and writing skills that allow them to access print for enjoyment and lifelong learning (Gambrell et al., 2011). An important, and often overlooked, aspect of literacy instruction involves activities designed to foster the development of oral language proficiency beyond the development of vocabulary knowledge (Hall et al., 2021; Language and Reading Research Consortium [LARRC], 2015). The development of strong oral language skills in the primary grades provides a solid foundation for cognitive, academic and social growth in later school years and into adulthood (Kamhi & Catts, 2012; Kim et al., 2020). Difficulty with the development of oral narration places students at a learning disadvantage even after controlling for general cognitive abilities, memory, phonological skills, and mother education (Babayigit et al., 2021; Catts et al., 1999).

Narratives consist of a series of interrelated sentences that provide information about real or imagined events (Curenton & Justice, 2004). Most often, narratives are described as consisting of a macrostructure (a set of story grammar elements that represent the structure of episodes) and a microstructure (the words and sentences that form the story; Mandler & Johnson, 1977; Stein and Glenn, 1979). The macrostructure in children's stories (initiating events, internal responses, plans, attempts, consequences, and reactions) are remarkably similar across languages and cultural groups (Berman & Slobin, 1994; McCabe & Bliss, 2005; Squires et al., 2014). However, the way story elements are sequenced—as well as the specific vocabulary and sentence structures used—often vary as a function of linguistic and sociocultural differences (Champion, 2003; Gillam et al., 2012; Price et al., 2006). In addition, as children progress through school, they acquire “literate language,” which is decontextualized, complex language used to convey specific meanings often through the use of conjunctions, elaborated noun phrases, mental and linguistic verbs, and adverbs and clausal embedding (Greenhalgh & Strong, 2001). In general, children in their early elementary school years produce short oral narratives that contain the basic elements of an episode (initiating events, goal-directed actions of characters, and the consequence of those actions). Their stories often do not include language that clearly specifies temporal, causal, and logical relationships, nor does it include richness and elaboration that contribute to higher quality narratives (Berman, 1988). Children develop broader knowledge of narrative discourse organization as well as the complex vocabulary and grammatical skills necessary to create high quality complex narratives as they progress through school (Berman, 1988; Wells, 2009). In fact, the development of narrative proficiency is well-represented in the curricular standards and

guidelines for English Language Arts (ELA) instruction (National Governors Association Center for Best Practices, 2010).

Many children who are at risk for language and literacy difficulties struggle with these expectations, in part because their insufficient oral language development does not support their access to classroom instruction (August & Shanahan, 2006; Rand Reading Study Group, 2002; Scott & Windsor, 2000). Children with limited language skills and/or reading comprehension problems frequently create oral and written stories that are shorter, less complex and contain more grammatical errors than stories produced by their typically developing peers (Allen et al., 2012; August & Shanahan, 2006; Cain, 2003; Fey et al., 2004; Tsimpli et al., 2016).

Narrative Intervention Research

Because of the importance of narrative proficiency for successful educational outcomes and the potential impact of poor narrative skills on literacy outcomes, researchers and educators have conducted several studies investigating the efficacy of various instructional approaches to improve narrative proficiency. Many of these investigations have been summarized in recently published systematic reviews and meta-analyses of narrative interventions, the vast majority of which were small-scale, nonrandomized studies conducted with preschool and kindergarten-age children (Favot et al., 2021; Nicolopoulou & Trapp, 2018; Pesco & Gagné, 2017; Petersen, 2011; Pico et al., 2021; Rogde et al., 2019). The intervention procedures most associated with positive narrative production and/or comprehension outcomes included the use of authentic literature, verbal scaffolding, story grammar instruction with icons or cue cards, instruction in the use of causal and temporal relationships among events, and scaffolded instruction that supports story retelling and story generation. Effect sizes for measures of comprehension or production of story macrostructure in these small-scale studies varied widely from .16 to 1.57. Similarly, effect sizes for measures of narrative microstructure, which typically assessed sentence length or complexity, ranged from .77 to 1.33. There were no specific intervention procedures that yielded consistently better outcomes than others.

Relatively few interventions that have been studied to date have targeted both narrative macrostructure and *literate language* (Rubin et al., 2000; Tannen, 1982) aspects of microstructure, which includes the types of sentences, often used in literature, that contain conjunctions, elaborated noun phrases, mental and linguistic verbs, adverbs, and clausal embedding. Literate language typically develops in the early elementary grades when classroom instructional contexts provide formal exposure to the types of

language structures that are needed to communicate complex concepts. Literate language is important because it leads to the use of language to monitor, reflect, reason, and plan. However, such language is frequently decontextualized, which may be particularly challenging for children who are at risk for language and literacy difficulties (Greenhalgh & Strong, 2001).

We know of three RCTs to date that studied the outcomes of narrative interventions with school-age, at-risk or language impaired children that targeted narrative macrostructure and literate language skills. Gillam et al. (2008) compared the language and auditory processing outcomes of 216, school-age children (Grades 1–4) with developmental language disorder who were randomly assigned to one of four conditions. The primary focus of the study was on *Fast ForWord-Language* (Tallal, 2013), but one of the active comparison conditions was a multicomponent intervention that incorporated vocabulary, phonological awareness, grammatical morphology, syntax and narrative targets. It was the only treatment that addressed narrative proficiency. Students in all four conditions made meaningful language-related gains over time. The group-level samples were small (~54 per group), and the authors only reported standardized differences from pretest to posttest within groups. They did not standardize mean differences across subsets of treatment conditions, making it difficult to evaluate the sample-independent magnitude of treatment differences.

The LARRC (2015) assessed the language and reading outcomes of a multicomponent, language-focused intervention called *Let's Know!* A total of 160 elementary classrooms in four geographic regions with 938 typically developing and at-risk children were randomly assigned to the treatment or business-as-usual control conditions. The authors reported significant, large-sized treatment effects on measures of vocabulary and comprehension monitoring strategies for 1st, 2nd, and 3rd grade students. Only the 3rd graders in the treatment group outperformed controls on a measure of oral narrative comprehension. The vocabulary and comprehension monitoring measures were described as very proximal to the intervention, including words and strategies taught only to treated students. There were no differences on distal measures or on standardized language measures.

Finally, Petersen et al. (2022) conducted an effectiveness study that contained a cluster randomized component in which they examined narrative outcomes of 686 kindergarten students from 28 classrooms across four school districts. Teachers provided whole-classroom narrative instruction twice each week for 15–20 minutes. A secondary, quasi-experimental aspect of the study involved small sets of matched students from the original treatment and control groups who were identified as “at risk” because they scored below designated benchmarks on researcher designed measures of narrative ability and/or were being served on existing IEPs at the outset of the study. In the cluster-randomized efficacy study, the authors reported statistically significant results favoring the treatment classrooms on two researcher-developed measures, oral personal story generation (effect size = .21) and oral narrative retell (effect size = .49), that were closely aligned with the intervention. There were also statistically significant group differences, with small effect sizes, favoring the treatment group for experimenter-designed measures of written language (effect size = .19) and expository language (effect size = .16). The authors did not indicate that the measures were administered by blinded examiners, and they did not provide information about the psychometric

validity and reliability of their outcome measures, making these promising results difficult to interpret.

Although nonrandomized studies with relatively small samples have suggested that narrative intervention has the potential to result in positive, short-term gains in narrative abilities for children with diverse learning needs in elementary grades, these treatment effects lack stability given the low statistical power of the studies. One potential reason for the highly variable outcomes may relate to the nature of instructional activities used to target narration. In the smaller studies, narrative skills were taught explicitly and directly and often served as the sole target of instruction, whereas the treatment conditions in the larger experimental studies included activities focused on narration and other language targets. The current study was designed to address some of these issues by comparing the language and literacy outcomes of the manualized *Supporting Knowledge in Language and Literacy* program (*SKILL*; Gillam, Gillam, et al., 2018) that targeted narrative macrostructure and literature language aspects of microstructure to business-as-usual classroom practices.

Theoretical Underpinnings of SKILL

The development of the *SKILL* intervention program was informed by three cognitive models that address how information is stored, activated, and used in comprehension and production. These include the Adaptive Control of Thought (ACT-R) model of cognition (Anderson, 1993; Anderson & Lebiere, 2014), the Embedded Processes model of working memory (Cowan, 1999, 2014), and the Construction-Integration (CI) model of text comprehension and production (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983). In all three models, discourse comprehension and production are characterized as multilevel cognitive processes (Graesser, 2007; Wilkinson & Son, 2011) that draw heavily on object knowledge, event knowledge, and language knowledge within associative networks in long-term memory (Federmeier & Kutas, 1999; Lewis et al., 2006; Was & Woltz, 2007). According to ACT-R, knowledge of story structure is generated from episodic encoding operations in which information from semantic and syntactic language systems is integrated into multilevel representations (chunks) that support performance in narrative comprehension and production tasks (Bower, 2008). Chunks become increasingly large and complex as a function of experience listening to and telling more complex stories. Cowan (2014; Cowan et al., 2021) has suggested that cognitive strategies (i.e., *schemas*) are used to organize information in LTM from experiences. According to the Embedded Processes model of WM, schemas assist in recoding or regrouping chunks of information in ways that support their retrieval during comprehension and production activities (i.e., listening to, reading or telling stories).

Kintsch's CI model suggests that comprehension and production of discourse requires the coordinated use of construction and integration processes. First, surface level information (e.g., language microstructure) is used to create propositions or meaningful “chunks” that form a textbase which represents the intended meaning of the discourse. The textbase includes the microstructure as well as the macrostructure which includes a hierarchical organization of the concepts and information contained in the text. Integration is said to occur when the textbase is integrated with the listener (or reader's) knowledge, goals, motivations, and purposes for listening (or reading) when a “situation model” or mental

representation is created. The situation model includes chunks of information that include the characters, settings, action and events that are explicitly mentioned or inferentially suggested in the text (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983).

The three phases of the *SKILL* program were designed to transfer our theoretical orientation into a set of learning and practice activities designed to increase organizational frameworks in episodic memory resulting in the ability to comprehend and produce more complex stories.

Feasibility Studies

Three studies were conducted to assess the feasibility of conducting an RCT with the *SKILL* narrative intervention program. Gillam and colleagues (2014) evaluated the impact of the *SKILL* program in a limited sampling design study with two 1st grade classrooms. Twenty-one children in the treatment classroom received 30 minutes of whole-class *SKILL* intervention three times per week for six weeks as a supplement to the typical classroom language arts curriculum. Twenty-two children in the comparison classroom participated in the language arts curriculum with no *SKILL* instruction. The pre–post effect sizes on a narrative language sample were three times larger for children in the experimental classroom ($d = .82$) than children in the BAU class ($d = .21$). Further, there were differential effects favoring the high-risk children ($d = 1.0$) over the low-risk children ($d = .34$). These results provided preliminary evidence for the use of *SKILL* to improve narrative outcomes for children who are at-risk for language and literacy difficulties.

A small-scale RCT of an early version of the *SKILL* program was conducted with 20 elementary-age children with developmental language disorders (DLD; R. Gillam et al., 2017). Ten children were randomly assigned to receive *SKILL* instruction, and 10 continued to receive traditional speech and language services. Children were seen in groups of three for 35–40 minutes per session, three times each week for 6 weeks. The *SKILL* instruction resulted in statistically significant gains on a standardized measure of narrative proficiency (Test of Narrative Language, TNL; Gillam & Pearson, 2004), yielding a large Cohen's d effect size of 1.45 for group differences at posttest.

Finally, a single-case, multiple-baseline study was conducted to assess the *SKILL* intervention's impact on individual children with DLD (Gillam, Olszewski, et al., 2018). Six children participated in the investigation with two children remaining in baseline across the entire study. The four participants who received the intervention made significant positive changes on measures of narrative productivity and complexity from baseline to treatment phases (Tau-U range across measures from .61 to .92), whereas the narrative skills of the two children who remained in baseline did not change from the first half to the second half of the study (Tau-U range across measures from $-.13$ to $.15$).

The purpose of the present article is to report the findings of a rigorously conducted RCT of narrative intervention using the *SKILL* curriculum. Our study directly addresses gaps in the literature by directly focusing on macrostructure and microstructure, examining immediate and long-term outcomes, using standardized measures that were not closely aligned with the *SKILL* program as well as researcher designed tasks that were closely assigned to the *SKILL* program as outcome measures, and including a large

population sample of at-risk children in 1st through 4th grades in the United States.

We asked the following research questions:

1. What is the effect of the *SKILL* program on overall oral narrative proficiency, oral narrative comprehension, and oral narrative production in at-risk students in 1st through 4th grade? To what extent is the effect maintained five-months post treatment?
2. What is the effect of the *SKILL* program on literacy outcomes (written narration and reading comprehension)? To what extent are significant effects maintained five-months post treatment?
3. To what extent is the effect of the *SKILL* program moderated by students' grade level classification (1st/2nd grade versus 3rd/4th grade)?

Method

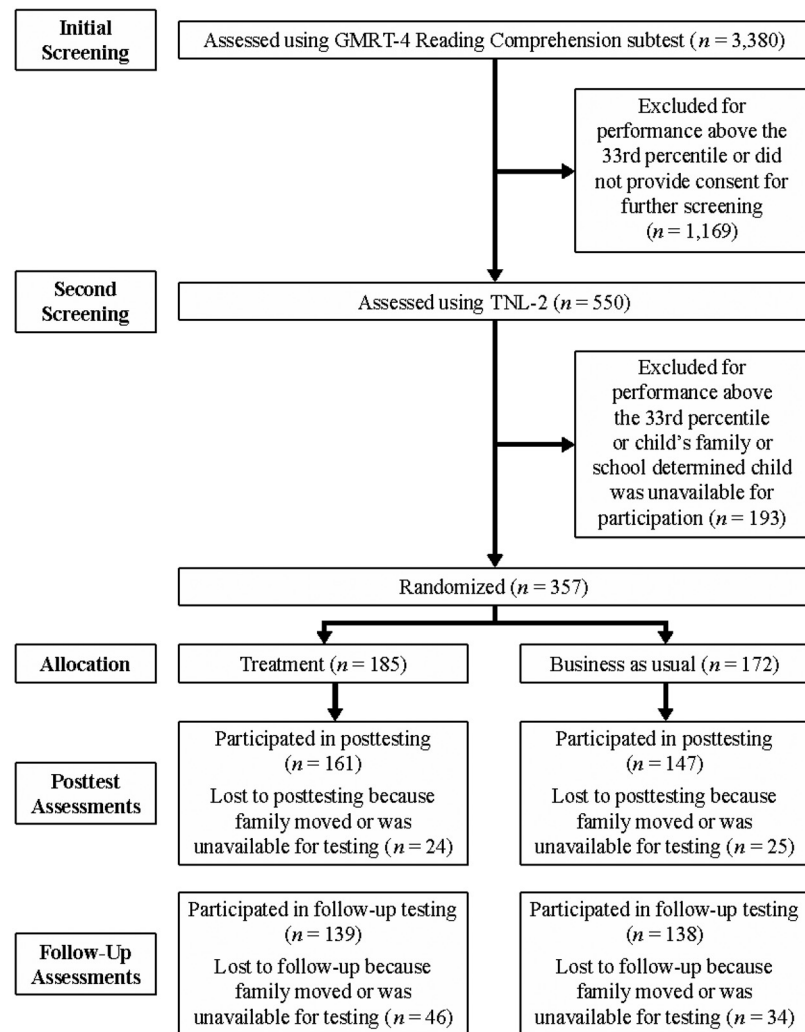
Study Design

At-risk children in 133 classrooms within 14 schools in two states in the Western United States were randomly assigned to a treatment group or a business-as-usual (BAU) control group in a randomized controlled trial to rigorously evaluate the efficacy of the *SKILL* narrative intervention program. The study, which was conducted in three yearly cohorts between September 2017 and October 2021, was approved in an Institutional Authorization Agreement by the Institutional Review Boards at Utah State University and the University of Texas at Austin. The research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local statutory requirements. All participants and their parents gave written informed consent to participate.

Participants

Our research team used a multiple screening procedure to identify students in Grades 1 through 4 with language and literacy difficulties (LLD) for participation. As shown in the CONSORT Flow Diagram (see Figure 1), we initially screened all students ($n = 3,380$) in Grades 1–4 at 14 participating elementary schools using the Gates-MacGinitie Reading Test – 4th edition reading comprehension subtest (GMRT; MacGinitie et al., 2002). Students who scored at or below the 33rd percentile on the GMRT test and whose families consented for participation advanced to the second phase in the process ($n = 550$). These students were assessed using the Test of Narrative Language – 2 (TNL-2; Gillam & Pearson, 2017). Like previous studies (Bowyer-Crane et al., 2008; Coyne et al., 2013; Simmons et al., 2011), students who scored at or below the 33rd percentile on both measures were considered to be at-risk for language and literacy difficulties and were invited to serve as participants in the study. We used the 33rd percentile as our cut-off for inclusion as “at-risk” rather than the more common 25th percentile because we were interested in including participants representing a wider range of abilities in reading comprehension and oral narration. This

Figure 1
CONSORT Flow Diagram



allowed us to better estimate the functional relationships between oral narration, written narration, and reading comprehension (Barnes et al., 2016; Preacher et al., 2005).

A total of 357 students over 3 years met the qualifying screening criteria and received parent consent for participation. Randomization to treatment and control groups was conducted after screening was completed each year by an investigator with no clinical involvement in the trial. Assignment by a computerized random number generator was blocked by classroom to increase the likelihood that similar numbers of children in each classroom were assigned to the two conditions. We blocked on classroom because we expected between-classroom variance. A main purpose of blocking was to improve the statistical power of difference tests and the precision of parameter estimates by decreasing residual variance.

Of the 357 students who participated in this RCT, 57% were Latino, 32% were White, 4.5% were African American, and 1.4% were Asian. The majority of students were male (54%) and 41% of the sample were from bilingual backgrounds (i.e., parents reported a

home language other than English). Thirty-seven percent of participating students were receiving special education services. The most reported disability categories for the participants were speech language impairments (19% of total sample) and learning disabilities (16%). Table 1 presents demographic data on the participants by condition. There were no significant differences at pretest between students randomized to the *SKILL* treatment or BAU conditions on key demographic variables (i.e., sex, ethnicity, grade level, age, special education status, and bilingual status).

We paid particular attention to attrition because the data collected at posttest and follow-up for Cohort 3 were collected via videoconference after pandemic-related school closures. Elsewhere we describe methods for collecting these data (Magimairaj et al., 2022). Here we describe rates of overall and differential attrition, their threats to internal validity, and related analytic adjustments. Overall attrition ranged from 12% to 14.6% at posttest and from 22% to 23% at follow-up across outcomes. Rates of attrition were comparable across treatment conditions. At posttest, the difference in attrition rates for *SKILL* and BAU ranged from

.1% to 1.7% across outcomes. At follow-up, differential attrition ranged from 4% to 5.7%. At posttest and follow-up, the threat of attrition-related bias is tolerable under cautious assumptions outlined by the What Works Clearinghouse (WWC, 2020).

In Table 1, we summarize the baseline characteristics for nonattending students in the *SKILL* intervention and in the BAU. Effect sizes (ES) indicate equivalence (or at least tolerable levels of nonequivalence) at posttest and at follow-up (Hedges g ranged from .02 to .21). For effects between .05 and .25, WWC (2020) recommends including known correlates of the outcome of interest in the analytic models to balance or rebalance groups. Missing data were handled using full information maximum likelihood estimation.

Description of the Intervention

The manualized *SKILL* program consists of three phases: *Teaching Story Structure and Causal Language*; *Teaching Strategies for Creating a Situation Model*; and *Teaching Strategies for integration into Long-Term Memory*. Two language intervention techniques, *parallel story production* and *vertical structuring* are used extensively during all three phases of instruction. In parallel story construction, instructors first model a story, then help children “coconstruct” a similar story using scaffolding techniques. In vertical structuring, instructors use scaffolding techniques to help students “connect” short, simple sentences to form complex sentences. The primary purpose of these activities is to teach children to use complex sentences (microstructure) to communicate causal connections between story grammar elements (macrostructure). Each phase ends with a series of literature-based activities designed to help students transfer new skills into authentic contexts they will encounter in school.

Phase I, *Teaching Story Structure and Causal Language*, contains 20 lessons about story construction and organization based on Stein and Glenn’s (1979) notion of episode structure. Verbal (key words) and visual cues (icons), wordless picture books, and graphic organizers are used in activities that target the construction of complete, single episode stories that contain initiating events, internal responses, plans, attempts, consequences, and reactions that are causally related. Students have multiple opportunities to practice using the concepts they are learning in a variety of contexts including retelling, parallel story development, pictographic planning, story discussion, and answering literal and inferential questions.

Phase II, *Teaching Strategies for Creating a Situation Model*, contains 18 lessons focused on elaborating on basic episodes and creating schemas for organizing story propositions into situation models. The activities are similar to those in Phase I, with the addition of lessons that use advanced planning strategies to help students add complicating actions and dialogue to their stories. New icons (e.g., dialogue, plan again) and more elaborate graphic organizers are used to demonstrate and give students opportunities to practice comprehending and producing stories containing multiple episodes with more complex vocabulary and syntax. The lessons target cohesion and use of clausal complements, metacognitive and metalinguistic verbs, elaborated noun phrases, and adverbs.

Phase III, *Teaching Strategies for integration into Long-Term Memory*, contains 12 lessons that offer students multiple opportunities to critically evaluate stories they hear and create. Instruction occurs almost entirely in children’s literature and authentic discourse contexts with the addition of metacognitive activities designed to provide opportunities for students to engage in

episodic encoding operations that facilitate their ability to integrate semantic and syntactic knowledge into chunks that are used to support comprehension and production. Toward this end, students are taught to use a rubric that contains the macrostructure and microstructure features they have learned and practiced in earlier phases. This phase ends with students creating and retelling their own multi-episodic stories and participating in authentic story comprehension activities that mirror those they will encounter in the classroom (e.g., listening to novel stories, retelling them, and answering comprehension questions about them).

Tutors

Twenty-four tutors (total) provided the *SKILL* treatment across the 3 years of the study. All tutors had a bachelor’s or master’s degree in education or an education-related discipline (e.g., speech-language pathology) and all had worked in an education setting previously for a minimum of 3 years. All tutors were hired, trained, and supervised by the research team. Each year, the lead author of the *SKILL* program provided 8 hours of professional development about implementing the manualized *SKILL* program to the tutors at all sites before intervention began.

Fidelity of Intervention

Intervention sessions were audio- and video-recorded digitally and were uploaded onto a secure server. An intervention observation checklist (IOC) was created for each lesson. A member of the research team observed and scored every session for the number of critical aspects of the lesson that were taught. A second research assistant independently rated fidelity for 20% of the lessons. The percentage of critical aspects that were taught in each lesson, averaged across all the lessons, was 90.5%, 98.4%, and 98.4% for years 1, 2 and 3, respectively and 95.8% overall. If fidelity fell below 85% for any lesson, the observer contacted the interventionist to review the IOC and discuss aspects of the lesson that were missed. In most cases, when an interventionist failed to include all the critical concepts in a lesson, it was because they ran out of time during a session. For example, the teacher may have started lesson 1 but was unable to finish it. It was almost never the case that entire sections of lessons were omitted for other reasons. However, when parts of lessons were missed, the interventionist was instructed to include them in the next session.

Outcome Measures

Testing sessions were conducted in quiet rooms in the schools the children were attending. All outcome measures were administered by trained examiners who were blind to group assignment.

Test of Narrative Language-2

The TNL-2 (Gillam & Pearson, 2017), a norm-referenced measure of narrative comprehension and production, was used to assess narrative language change before, immediately after, and five months after intervention ended. Children listened to stories, answered literal and inferential questions about them, and then told stories that were scored for content and complexity. The test included three types of stories: (a) a short restaurant script with a single problem, (b) single-episode personal narrative-like stories

Table 1
Covariate Balance Checking for the Analytic Sample at Baseline

Baseline variable	SKILL		BAU		Hedges <i>g</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Posttest analytic sample					
Demographic variables					
Female (%)	0.41	—	0.47	—	0.15
SES (%)	0.44	—	0.45	—	0.02
ELs (%)	0.45	—	0.37	—	0.21
SWDs (%)	0.42	—	0.39	—	0.08
Latino (%)	0.55	—	0.57	—	0.05
White (%)	0.34	—	0.32	—	0.05
Grade 1 (%)	0.16	—	0.18	—	0.08
Grade 2 (%)	0.27	—	0.26	—	0.04
Grade 3 (%)	0.31	—	0.30	—	0.01
Grade 4 (%)	0.26	—	0.26	—	0.00
Oral narrative outcomes					
MISL oral narrative	10.80	6.32	10.41	5.72	0.06
TNL production	6.54	1.82	6.29	2.15	0.12
TNL comprehension	6.30	2.23	6.26	2.19	0.02
Literacy outcomes					
MISL written narrative	8.33	5.31	8.18	5.60	0.03
GMRT	82.01	9.45	82.67	9.31	0.07
Follow-up analytic sample					
Demographic variables					
Female (%)	0.44	—	0.47	—	0.08
SES (%)	0.42	—	0.46	—	0.09
ELs (%)	0.45	—	0.38	—	0.19
SWDs (%)	0.41	—	0.38	—	0.08
Latino (%)	0.52	—	0.54	—	0.05
White (%)	0.37	—	0.34	—	0.11
Grade 1 (%)	0.16	—	0.20	—	0.16
Grade 2 (%)	0.29	—	0.23	—	0.21
Grade 3 (%)	0.27	—	0.29	—	0.06
Grade 4 (%)	0.27	—	0.28	—	0.02
Oral narrative outcomes					
MISL oral narrative	11.11	6.43	10.54	5.67	0.09
TNL production	6.60	1.87	6.37	2.14	0.11
TNL comprehension	6.27	2.23	6.34	2.23	0.03
Literacy outcomes					
MISL written narrative	8.37	5.37	7.98	5.36	0.07
GMRT	81.84	9.44	82.99	9.23	0.12

Note. SKILL = Supporting Knowledge of Language and Literacy; BAU = business as usual; SES = socioeconomic status; EL = English learner; SWD = students with disabilities; MISL = Monitoring Indicators of Scholarly Language; TNL = Test of Narrative Language; GMRT = Gates-MacGinitie Reading Test.

related to common school experiences, and (c) multi-episode fictional narratives. There were 47 items across the three comprehension tasks and 88 items across the three production tasks. Internal consistency reliability for the participants in this study ranged from .86 to .92 across the three testing periods with a mean of .89.

Gates-MacGinitie Reading Test

The GMRT-4 (MacGinitie et al., 2002) reading comprehension subtest is a group administered, timed assessment designed to access a student's reading comprehension abilities. Students were given 45 minutes to silently read expository and narrative passages and to mark their answers to multiple choice comprehension questions on test booklets. The total number of items in the booklets varied from 39 to 48 across the five levels that were administered in this study. Internal consistency reliability for the GMRT-RC ranged from .81 to .84 across the three testing periods with a mean of .82.

Spontaneous Oral and Written Narrative Samples

We elicited spontaneous narratives by asking children to create oral and written stories that corresponded to different single-scene prompts. The examiners said, "I am going to show you a picture. I want you to make up a story using this picture. Tell (Write) the best story you can. Start when you are ready." The icons that were used in the intervention were available on the table, but students were not provided with any explanation of what the icons were or how to use them during the testing session. Examiners did not prompt the students to elaborate their stories. Digital recorders were used to record oral stories as they were being told as well as written stories that were "read" aloud to the examiner after they were composed.

Research assistants who were blind to participant group assignment transcribed the stories (oral and written) according to Systematic Analysis of Language Transcripts conventions (Miller et al., 2019). The stories were segmented into communication units (Loban, 1976) that consisted of an independent main clause and any

phrases or clause(s) subordinated to it. Each transcript was checked by a second research assistant for spelling, mazing, morpheme segmentation, and utterance segmentation. All transcription disagreements were resolved as the two transcribers listened to the digital recording together for a third time. As a check on the accuracy of the original transcription and coding process, 20% of the transcripts were transcribed independently by a second transcriber. Percentage of agreement between primary and secondary transcribers was 97.54% for communication-unit segmentation and 94.12% for the identification of mazes, indicating stable and reliable transcription.

A narrative scoring rubric called, Monitoring Indicators of Scholarly Language (MISL; S. L. Gillam et al., 2017), was used to evaluate the complexity of microstructure and macrostructure aspects of narration. Oral narratives were scored using the finalized transcripts of each student's digital recording. Written narratives were scored from the students' original written products. For illegible written stories, we scored the transcribed version of what the child read aloud. This occurred for 15% (163 of 1,071) of the written narratives.

Each oral and written narrative was coded for seven aspects of narrative macrostructure (e.g., character, setting, initiating event, internal response, plan, attempts, and consequence) and six aspects of narrative microstructure (coordinating conjunctions, subordinating conjunctions, adverbs, mental verbs, linguistic verbs, and elaborated noun phrases). Each item was scored according to a four-point scale that reflected whether an element was absent (score of 0), emerging (score of 1), present (score of 2), or elaborated (score of 3). The dependent variable was the total of the 13 macrostructure and microstructure items. Internal consistency reliability values were acceptable for the total instrument ($\alpha = .86$), the macrostructure subscale ($\alpha = .75$) and the microstructure subscale ($\alpha = .77$).

Training was conducted to ensure that blinded scorers were reliable across all 14 MISL items. Scorers were trained on 10 stories and then independently coded 10 additional stories not associated with the study. This process was continued until scorers were 95% reliable with one of the authors (a gold-standard rater) on item-by-item scoring across three or more stories. Scorers met to discuss disagreements with the gold-standard rater after every 20 stories were scored to control for coder drift and to recalibrate; 95% or higher reliability was maintained throughout the scoring process. If reliability fell below that value, the author who served as the gold standard rater met individually with the scorer to review the scoring protocol. The final interrater reliability, calculated on 20% of the transcripts, which were randomly selected across all 3 years, was 95.7% for oral narratives and 93.2% for written narratives. Clearly, our goal for having a stable score across blinded raters was met.

Data Analysis Plan

We fit multilevel models to estimate the effects of the *SKILL* intervention on oral narrative and literacy outcomes. Multilevel models accounted for dependencies in nested data by estimating residual components (random effects, errors, etc.) at each level and by partitioning total variance into its level-specific component parts (Hox et al., 2017). Note that each outcome was modeled according to the earlier-discussed concerns with attrition and balance across groups at posttest. Because attrition varied across outcomes and because the correlation of covariate and outcome differed by measure, models were specific to each outcome. Further, for all outcomes, four-level models (Raudenbush & Bryk, 2002) were fit

initially to evaluate the source(s) of significant clustering, as measured by the intraclass correlation (ICC). There was no tutor-level clustering for any outcomes with the exception of the MISL oral narrative at posttest (ICC = .11) and for the GMRT comprehension at follow-up (ICC = .24). Also, teacher-level differences were nominal for all outcomes except for GMRT comprehension.

Our first set of research questions involved the main effect of the *SKILL* intervention on oral narrative production and oral narrative comprehension outcomes at posttest and follow-up, controlling for pretest scores on these measures. These questions were evaluated according to the models (Equations 1–6) in Table 2. Again, these models reflected a concern with pretest balance at treatment's end and at follow-up and with managing nested data. The second set of research questions addressed the main effect of *SKILL* intervention on written narration and reading comprehension at posttest and follow-up. These questions were evaluated by the multilevel models in Equations 7–10. As before, these models reflected equivalence- and clustering-related concerns. The third research question considered the extent to which *SKILL*'s effect differs on average for students who were in 1st or 2nd grade versus students who were in 3rd or 4th grade. To answer this research question, we expanded the earlier models to include a main effect for grade level and a conditional effect representing grade's potential moderation of treatment's effect (i.e., its interaction with treatment condition).

Per recommendations of the WWC (2020), we fit separate models for each dependent variable and adjusted for false discovery rates (the probability of type I error) using the Benjamini-Hochberg correction (Thissen et al., 2002). Hedges' g effect sizes were calculated per the following equation (WWC, 2020):

$$g = \frac{\gamma}{\sqrt{\frac{(n1-1)S_1^2 + (n2-1)S_2^2}{(n1+n2-2)}}} \quad (1)$$

where γ is the coefficient for the intervention's effect; $n1$ and $n2$ are student sample sizes in the intervention and treatment groups; and S_1^2 and S_2^2 are the student-level unadjusted posttest standard deviations for the intervention group and the comparison group, respectively.

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. All data, analysis code and research materials are available at (provide URL here). All analyses were run with *lme4* package (1.1–27.1, Bates et al., 2015) in R (R Core Team, 2020). Two-way interactions were decomposed, and contrasts were computed using the "emmeans" package (1.7.0, Lenth et al., 2020) in R. This study's design and its analysis were not preregistered.

Measurement Invariance

The COVID-19 pandemic occurred at the end of our third cohort year, and we administered the posttest and follow-up batteries online using the Zoom platform. To evaluate the comparability of TNL-2 scores at posttest and follow-up in Cohort 3 to those collected in person during Cohorts 1 and 2, we fit measurement-invariance models at the passage level (Magimairaj et al., 2022). Technically, measurement invariance is an item-level analysis. It considers the extent to which items

Table 2
Equations Describing the Models for Each Outcome of Interest

1. MISL oral narrative at posttest	$Y_{ik} = \gamma_{000} + \gamma_{100}(\text{Pretest})_{ik} + \gamma_{200}(\text{SKILL})_{ik} + \gamma_{300}(\text{Female})_{ik} + \gamma_{400}(\text{SWD})_{ik} + \gamma_{500}(\text{EL})_{ik} + \gamma_{600}(\text{Gradelevel})_{ik} + u_{00k} + u_{0ik} + e_{ik}$
2. TNL production at posttest	$Y_{ik} = \gamma_{00} + \gamma_{10}(\text{Pretest})_{ik} + \gamma_{20}(\text{SKILL})_{ik} + \gamma_{30}(\text{Female})_{ik} + \gamma_{40}(\text{SWD})_{ik} + \gamma_{50}(\text{EL})_{ik} + \gamma_{60}(\text{Gradelevel})_{ik} + u_{0k} + e_{ik}$
3. TNL comprehension at posttest	$Y_{ik} = \gamma_{00} + \gamma_{10}(\text{Pretest})_{ik} + \gamma_{20}(\text{SKILL})_{ik} + \gamma_{30}(\text{Female})_{ik} + \gamma_{40}(\text{SWD})_{ik} + \gamma_{50}(\text{EL})_{ik} + \gamma_{60}(\text{Gradelevel})_{ik} + u_{0k} + e_{ik}$
4. MISL oral narrative at follow-up	$Y_{ik} = \gamma_{00} + \gamma_{10}(\text{Pretest})_{ik} + \gamma_{20}(\text{SKILL})_{ik} + \gamma_{30}(\text{Female})_{ik} + \gamma_{40}(\text{SWD})_{ik} + \gamma_{50}(\text{EL})_{ik} + \gamma_{60}(\text{Gradelevel})_{ik} + u_{0k} + e_{ik}$
5. TNL production at follow-up	$Y = \beta_0 + \beta_1(\text{Pretest}) + \beta_2(\text{SKILL}) + \beta_3(\text{female}) + \beta_4(\text{SWD}) + \beta_5(\text{EL}) + \beta_6(\text{Grade level}) + \varepsilon_i$
6. TNL comprehension at follow-up	$Y_{ij} = \beta_0 + \beta_1(\text{Pretest}) + \beta_2(\text{SKILL}) + \beta_3(\text{female}) + \beta_4(\text{SWD}) + \beta_5(\text{EL}) + \beta_6(\text{Grade level}) + \varepsilon_i$
7. MISL written narrative at posttest	$Y_{ik} = \gamma_{00} + \gamma_{10}(\text{Pretest})_{ik} + \gamma_{20}(\text{SKILL})_{ik} + \gamma_{30}(\text{Female})_{ik} + \gamma_{40}(\text{SWD})_{ik} + \gamma_{50}(\text{EL})_{ik} + \gamma_{60}(\text{Gradelevel})_{ik} + u_{0k} + e_{ik}$
8. GMRT comprehension at posttest	$Y_{ij} = \gamma_{00} + \gamma_{10}(\text{Pretest})_{ij} + \gamma_{20}(\text{SKILL})_{ij} + \gamma_{30}(\text{Female})_{ij} + \gamma_{40}(\text{SWD})_{ij} + \gamma_{50}(\text{EL})_{ij} + \gamma_{60}(\text{Gradelevel})_{ij} + u_{0j} + e_{ij}$
9. MISL written narrative at follow-up	$Y_{ik} = \gamma_{00} + \gamma_{10}(\text{Pretest})_{ik} + \gamma_{20}(\text{SKILL})_{ik} + \gamma_{30}(\text{Female})_{ik} + \gamma_{40}(\text{SWD})_{ik} + \gamma_{50}(\text{EL})_{ik} + \gamma_{60}(\text{Gradelevel})_{ik} + u_{0k} + e_{ik}$
10. GMRT comprehension at follow-up	$Y_{ij} = \gamma_{000} + \gamma_{100}(\text{Pretest})_{ij} + \gamma_{200}(\text{SKILL})_{ij} + \gamma_{300}(\text{Female})_{ij} + \gamma_{400}(\text{SWD})_{ij} + \gamma_{500}(\text{EL})_{ij} + \gamma_{600}(\text{Gradelevel})_{ij} + u_{00k} + u_{0ik} + e_{ik}$

Note. Here, i represents students, t represents tutors, j represents teachers, and k represents schools. Parameters γ_{000} , γ_{00} and β_0 are the student-level mean outcome for each measure in the study; Pretest_{ik} , Pretest_{itk} , Pretest_{ij} , and Pretest_{ij} are student level pretest score for each outcome centered around its grand mean (Enders & Tofghi, 2007); SKILL_{ik} , SKILL_{itk} , SKILL_{ij} , and SKILL_{ij} are student-level dummy-coded variable representing condition, where SKILL intervention is coded as 1 and BAU is coded as 0; Female_{ik} , Female_{itk} , Female_{ij} , and Female_{ij} is student's gender with male coded as 0 and female coded as 1; SWD_{ik} , SWD_{itk} , SWD_{ij} , and SWD_{ij} is disability status with the non-SWD group coded as 0 and SWD coded as 1; EL_{ik} , EL_{itk} , EL_{ij} , and EL_{ij} is EL status with non-ELs coded as 0 and ELs coded as 1; Gradelevel_{ik} , Gradelevel_{itk} , Gradelevel_{ij} , and Gradelevel_{ij} is student's grade level with first and second graders coded 1 and third and fourth graders coded 0; and residuals e_{ij} , u_{0j} and u_{00k} are Level 1, Level 2, and Level 3 random effects, respectively. SKILL = Supporting Knowledge of Language and Literacy; EL = English learner; SWD = students with disabilities; MISL = Monitoring Indicators of Scholarly Language; TNL = Test of Narrative Language; GMRT = Gates-MacGinitie Reading Test.

“behave” similarly across occasions, across groups, or across occasions by groups. The TNL-2 presented test-takers with multiple tasks per passage, which means that items were nested in passages. Testlet models could be fit to account for this dependence. However, the available sample was not adequate to model item-level invariance using these approaches based on common heuristics (e.g., five cases per estimable parameter). Instead, we modeled passage as the primary outcome in our confirmatory models, with the latent *comprehension* variable predicting performance on the scores across three passages and the latent *production* variable predicting performance on scores across another three passages. We fit several models (see the Appendix) to evaluate the absolute fit of the measurement models on each occasion and to evaluate the relative fit of each occasion in the context of other occasions across cohorts.

First, we fit a CFA for the TNL-2 at pretest using data combined across cohorts. These represented a very good fit with the data ($\chi^2 = 10.24$, $df = 7$, CFI = .99, TLI = .99, RMSEA = .04 90% CI [.00, .08]). The posttest and follow-up data fit the data similarly well ($\chi^2 = 80.02$, $df = 7$, CFI = .99, TLI = .99, RMSEA = .02 90% CI [.00, .08] and $\chi^2 = 17.78$, $df = 7$, CFI = .99, TLI = .99, RMSEA = .04 90% CI [.00, .08]), suggesting that TNL-2 data, when modeled at the passage level, aligned very strongly with the measure's a priori factor structure at each occasion. Further, because the posttest and follow-up data from Cohort 3 were included in these models, the findings suggest that differences in the mode of collection did not compromise model fit on each

occasion. When the pathways from *comprehension* and *production* to the posttest and follow-up occasions in Cohort 3 data were constrained, the overall fits were not different from the full model, suggesting that data collected online, and data collected in person were comparably useful in generating factor scores for the two latent constructs. To evaluate relative fit over occasions and cohorts, we fit the entire measurement model (pretest, posttest, and follow up), under the assumption that posttest and follow-up data in Cohort 3 (the data that were collected on line) did not differ (i.e., provided a nondifferent indication of performance on the latent comprehension and production constructs) from the posttest and follow-up data collected from students in Cohorts 1 and 2 (data collected face to face). The model fit was more than adequate ($\chi^2 = 164.52$, $df = 100$, CFI = .97, TLI = .97, RMSEA = .04 90% CI [.03-.04]). Indeed, given its relative complexity, the fit is very good. Based on methods typically used for evaluating invariance, we found partial scalar invariance on each occasion when compared with the prior occasion (post to pre, follow up to post). Further, constraints on the Cohort 3 data at posttest and follow-up did not change the patterns of invariance at the passage level.

Results

Research Question 1

Our first research question concerned the immediate effect of the *SKILL* program on overall *oral narrative abilities* and whether

the effect was maintained 5-months after treatment. Table 3 summarizes observed means and standard deviations at pretest, posttest, and follow-up for the *SKILL* (treatment) and BAU conditions. All variables were distributed normally based on estimates of skewness and kurtosis. There were no outlying values.

Students randomized to the *SKILL* intervention significantly outperformed students in BAU condition on all three oral narrative outcomes at posttest (see Table 4). Students in the *SKILL* treatment condition scored 4.16 points higher on the researcher designed MISL oral narrative rubric (total score) at posttest than students in the BAU condition ($\gamma_{200} = 4.16, p < .01$; Hedges' $g = .61$). On the TNL-2, the norm-referenced measure of narration, students randomized to receive *SKILL* intervention significantly outperformed students in the BAU on the oral narration production subtest ($\gamma_{20} = .84, p < .01$; Hedges' $g = .36$) and the oral narrative comprehension subtest ($\gamma_{20} = .48, p = .04$; Hedges' $g = .20$).

At the 5-month follow-up, students who received the *SKILL* intervention maintained their oral narrative advantage as measured by the MISL oral narrative rubric compared with students in BAU ($\gamma_{20} = 4.09, p < .01$; Hedges' $g = .63$). However, there were no significant group differences on the norm-referenced measure of narration (TNL-2) that included both production and comprehension components at follow-up, although the effect sizes were non-trivial (Hedge's $g = .21$ and $.11$, respectively).

Research Question 2

Our second research question concerned the potential effects of *SKILL* on literacy outcomes (written narration and reading comprehension) and the extent to which any significant effects were maintained at the 5-month follow-up testing.

Students randomized to the *SKILL* intervention significantly outperformed students randomized to the BAU condition on their writing sample scored using the MISL rubric at posttest ($\gamma_{20} = 2.16, p < .01$; Hedges' $g = .34$) and at follow up ($\gamma_{20} = 2.02, p < .01$; Hedges' $g = .35$).

Scores on the norm-referenced measure of reading comprehension (see Table 5) did not differ between groups at posttest (Hedges' $g = -.02$) or at follow-up, although the effect size (Hedges' $g = .16$) was in the low-average range.

Research Question 3

Our third research question concerned the extent to which outcomes were moderated by students' grade-level group (1st/2nd grade vs. 3rd/4th grade). Tables 6 and 7 summarize the moderating effect of grade level on oral language and literacy outcomes. *SKILL*'s effect on oral and written language did not vary reliably across participants, regardless of grade-level. However, grade-level group did moderate performance on the standardized reading comprehension measure (Figure 2). Specifically, on the GMRT comprehension subtest, the negative and statistically significant interaction term indicated that *SKILL* benefited 3rd/4th graders significantly more (Hedges $g = .26$) than students in the 1st/2nd grade-level group ($g = -.34$).

Discussion

Narration is a complex, often decontextualized, discourse-level skill requiring story structure, semantics, morphology, syntax, pragmatics, and adjusting communication for the listener (Johnston, 2008). Children who possess strong narrative language abilities in early grades tend to do better in comprehension and production activities associated with reading and writing in later grades (LARRC, 2015; Phillips et al., 2021). Even though State Curricular Standards contain multiple objectives for comprehension and production of narratives in K–5th grades, recent observational evidence suggests that explicit instruction in oral narrative proficiency is rarely addressed in the classroom (Hall et al., 2021). Such instruction may be especially beneficial for children who are at-risk for language and literacy difficulties, which includes diverse learners with a wide range of learning characteristics (Pico et al., 2021).

Table 3
Pretest, Posttest, and Follow-Up Means and Standard Deviations for the Outcome Measures

Outcome	Pretest			Posttest			Follow-Up		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Oral narrative outcomes									
MISL oral narrative									
SKILL	185	10.63	6.30	162	15.17	7.34	140	16.09	6.69
BAU	172	10.58	5.85	151	11.04	6.10	138	12.04	6.26
TNL production									
SKILL	185	6.50	1.89	158	8.12	2.35	140	8.16	2.36
BAU	172	6.30	2.12	147	7.24	2.26	137	7.67	2.57
TNL comprehension									
SKILL	185	6.28	2.21	159	7.84	2.21	141	8.38	2.62
BAU	172	6.24	2.24	148	7.41	2.58	138	8.30	2.53
Literacy outcomes									
MISL written narrative									
SKILL	185	8.16	5.18	162	10.76	6.89	139	11.48	5.96
BAU	172	8.18	5.67	152	8.70	5.71	139	9.47	5.55
GMRT									
SKILL	185	82.16	9.31	157	85.53	11.58	137	85.53	10.12
BAU	172	82.85	9.16	149	86.34	11.33	137	84.67	10.79

Note. SKILL = Supporting Knowledge of Language and Literacy; BAU = business as usual; MISL = Monitoring Indicators of Scholarly Language; TNL = Test of Narrative Language; GMRT = Gates-MacGinitie Reading Test.

Table 4*Estimating the Main Effect of Intervention on Oral Language Outcomes at Posttest and Follow-Up*

Effect	MISL oral narrative				TNL production				TNL comprehension			
	Estimate	SE	<i>p</i>	ES (<i>g</i>)	Estimate	SE	<i>p</i>	ES (<i>g</i>)	Estimate	SE	<i>p</i>	ES (<i>g</i>)
Fixed effects												
Posttest												
Intercept	12.83	0.80	0.00		7.65	0.33	0.00		7.78	0.31	0.00	
Pretest	0.51	0.05	0.00		0.36	0.06	0.00		0.54	0.05	0.00	
SKILL	4.16	0.73	0.00	0.61	0.84	0.24	0.00	0.36	0.48	0.23	0.04	0.20
Female	1.06	0.64	0.10		0.57	0.25	0.02		0.43	0.24	0.07	
SWDs	-0.21	0.67	0.75		-0.70	0.26	0.01		-0.38	0.25	0.12	
ELs	-1.59	0.67	0.02		-0.36	0.26	0.16		-0.49	0.25	0.05	
Grade level	-3.12	0.67	0.00		-0.24	0.25	0.33		-0.37	0.24	0.12	
Follow-up												
Intercept	13.21	0.81	0.00		8.32	0.32	0.00		8.86	0.30	0.00	
Pretest	0.49	0.06	0.00		0.37	0.07	0.00		0.59	0.06	0.00	
SKILL	4.09	0.65	0.00	0.63	0.51	0.27	0.06	0.21	0.29	0.26	0.27	0.11
Female	1.53	0.66	0.02		0.67	0.28	0.02		0.63	0.26	0.02	
SWDs	-1.18	0.69	0.09		-1.19	0.29	0.00		-0.89	0.27	0.00	
ELs	-0.76	0.70	0.28		-0.41	0.28	0.15		-0.70	0.27	0.01	
Grade level	-2.66	0.69	0.00		-0.81	0.28	0.00		-0.77	0.27	0.00	
Random effects												
Posttest												
Student-level	26.21	0.84			4.22	0.93			3.95	0.95		
Tutor-level	3.96	0.13										
School-level	0.94	0.03			0.31	0.07			0.23	0.05		
Follow-up												
Student-level	28.01	0.98										
School-level	0.79	0.02										

Note. (ES) *g* = Hedges *g* effect size; ICC = Intraclass Correlation Coefficient; SKILL = Supporting Knowledge of Language and Literacy; EL = English learners; SWD = students with disabilities; MISL = Monitoring Indicators of Scholarly Language; TNL = Test of Narrative Language.

A multisite randomized controlled trial was conducted to rigorously evaluate a manualized narrative intervention program (*SKILL*) that was designed to improve narrative comprehension and production in at-risk children. Our first research question asked whether children randomized to the *SKILL* program improved in their *oral narrative abilities* as compared with students in a BAU condition. Findings revealed that *SKILL* instruction had a significant, immediate effect on narrative ability as assessed by norm-referenced (TNL-2) and experimenter designed (MISL) measures. The *SKILL* effect on narrative comprehension, as assessed by the TNL-2, was significant, but was smaller ($g = .20$) than the effect on narrative production (TNL-2; $g = .36$) and the MISL ($g = .61$). To interpret our findings, we consulted the Lipsey et al. (2012) meta-analysis of randomized studies of educational interventions published from 1995 through 2012 that included low performing and at-risk students. Lipsey and colleagues found that the mean effect size for standardized, norm-referenced tests used as outcome measures that were narrow in scope (similar to the TNL-2) for elementary grade children was $.25$ ($SD = .42$). The mean effect size for researcher developed tests of outcomes for elementary grade children (similar to the MISL) was $.40$ ($SD = .55$). With reference to the Lipsey meta-analysis of effect sizes in educational research, we observed an average-sized effect of *SKILL* on narrative comprehension as measured by the TNL-2 administered immediately after intervention. Our effect sizes for oral narration, as measured by norm-referenced (TNL-2) and researcher developed measures (MISL) immediately after intervention, were above average.

Recall that the *SKILL* intervention was conducted in elementary schools with small groups of children. Lipsey et al. (2012) also examined effect sizes from randomized studies of education with respect to intervention contexts. They reported a mean effect size for small group instruction of $.26$ ($SD = .40$). Our effect size ($g = .20$) for narrative comprehension was $.15$ standard deviations smaller than the average effect size of randomized educational studies of small group interventions as reported by Lipsey, but our narrative production gains for the TNL-2 ($g = .36$) measure were $.25$ standard deviations larger than average and the gains as measured by the MISL ($g = .61$) were $.875$ standard deviations larger than the average effect size.

A recent systematic review (Rogde et al., 2019) examined the impact of intervention studies conducted in school settings to support language and reading skills of preschool and school-age children. Similar to our study, participants in the RCTs and quasi-experimental studies that were reviewed included second language learners, children with developmental language delays, and children who were otherwise at risk for language and reading difficulties. These previously conducted studies of language interventions (e.g., vocabulary, grammar, story grammar) tended to yield small, immediate effects ($g = .16$). For a subset of the narrative intervention studies ($n = 13$) that used investigator developed measures, the overall effect size was larger ($g = .42$). Our RCT of the *SKILL* intervention yielded effects that were somewhat larger than those of other language interventions in general as well as other narrative interventions, specifically. Based on comparisons to the outcomes of randomized studies of educational interventions in general, small-group

Table 5
Estimating the Main Effect of Intervention on Literacy Outcomes at Posttest and Follow-Up

Effect	MISL written narrative				GMRT comprehension			
	Estimate	SE	<i>p</i>	ES (<i>g</i>)	Estimate	SE	<i>p</i>	ES (<i>g</i>)
Fixed effects								
Posttest								
Intercept	9.43	0.68	0.00		86.72	1.50	0.00	
Pretest	0.66	0.06	0.00		0.34	0.07	0.00	
SKILL	2.16	0.56	0.00	0.34	-0.18	1.13	0.88	-0.02
Female	0.70	0.57	0.22		0.43	1.20	0.72	
SWDs	0.03	0.59	0.96		-2.65	1.27	0.04	
ELs	-0.58	0.58	0.32		-3.46	1.28	0.01	
Grade level	-1.98	0.63	0.00		4.50	1.42	0.00	
Follow-up								
Intercept	10.85	0.70	0.00		85.64	1.53	0.00	
Pretest	0.48	0.06	0.00		0.18	0.07	0.01	
SKILL	2.02	0.54	0.00	0.35	1.67	1.27	0.19	0.16
Female	1.21	0.56	0.03		0.97	1.26	0.44	
SWDs	-0.85	0.59	0.15		-2.03	1.38	0.14	
ELs	-0.23	0.59	0.70		-2.72	1.32	0.04	
Grade level	-3.14	0.64	0.00		0.28	1.40	0.84	
Random effects								
	Variance	ICC			Variance	ICC		
Posttest								
Student-level	23.63	0.99			87.75	0.78		
Teacher-level					25.11	0.22		
School-level	0.13	0.01						
Follow-up								
Student-level	19.84	0.96			65.29	0.60		
Tutor-level					32.21	0.30		
Teacher-level					10.61	0.10		
School-level	0.84	0.04						

Note. (ES) *g* = Hedges *g* effect size; ICC = Intraclass Correlation Coefficient; SKILL = Supporting Knowledge of Language and Literacy; EL = English learner; SWD = students with disabilities; MISL = Monitoring Indicators of Scholarly Language; GMRT = Gates-MacGinitie Reading Test.

interventions, and other language and narrative interventions, the results of our RCT of the *SKILL* intervention suggest that intervention with the *SKILL* program resulted in average to above-average benefits for children in the early elementary grades who were at-risk for language and literacy difficulties immediately after intervention.

Most narrative language intervention studies have reported significant immediate impacts but have not assessed maintenance over time (e.g., Dodd et al., 2011; Gillam et al., 2015; Petersen, 2011; Rogde et al., 2019; Schoenbrodt et al., 2003; Spencer & Slocum, 2010; Spencer et al., 2013; Spencer et al., 2015; Stringfield et al., 2011). Our results showed that the significant effects of *SKILL* were maintained five-months post treatment for oral narration as measured using a researcher-designed assessment that was closely aligned with the treatment (the MISL). However, maintenance of gains on our norm-referenced measure (TNL-2) was not observed. There are several possible reasons for this finding. First, the TNL-2 includes three narrative production contexts in which children are required to either retell or create stories that are “like” the story modeled for them. This places constraints on the structure (retell, creating stories from multiple or single pictures), vocabulary and sentence structures that may be used in their responses. The stories that were scored with the MISL were generated in response to general picture cues. That production context may have given students more creative license to craft their own narratives and engage in “expressive elaboration,” a term coined by Ukrainetz and colleagues

(2005) to describe the artful aspect of storytelling that goes beyond basic content and story grammar structure. Further, students in the *SKILL* treatment condition were very familiar with that elicitation context because it was used repeatedly in the *SKILL* lessons. This may have made the task more comfortable, familiar, and perhaps more motivating than the TNL-2 testing format, particularly after a significant time lapse, which, in this case, included a summer break.

Our second research question asked whether the effects of *SKILL* training generalized to literacy measures (written narration and reading comprehension) administered immediately after treatment and at follow-up five months after treatment ended. Children who received the *SKILL* training evidenced significantly greater gains on the researcher designed measure of written narration (MISL) than students in the BAU condition immediately after the intervention ($g = .34$) and at the 5-month follow-up assessment ($g = .35$). This finding was particularly compelling because the follow-up period included a summer break, which is historically associated with at least some loss in academic skills for many students who are at risk for language and literacy difficulties. Studies of the impacts of short-term school closures (e.g., summer, weather related, absenteeism) on learning loss show that achievement scores decline when students are out of school (Campbell et al., 2009; White et al., 2007). Summer loss disproportionately affects students from low socioeconomic backgrounds and students with language and literacy difficulties (Cooper et al., 1996; Menard & Wilson, 2014; Quinn & Polikoff, 2017). Our findings

Table 6
Estimating the Moderating Effect of Grade Level on Oral Language Outcomes at Posttest and Follow-Up

Effect	MISL oral narrative			TNL production			TNL comprehension		
	Estimate	SE	p	Estimate	SE	p	Estimate	SE	p
Fixed effects									
Posttest									
Intercept	12.63	0.86	0.00	7.50	0.35	0.00	7.57	0.33	0.00
Pretest	0.50	0.05	0.00	0.35	0.06	0.00	0.54	0.05	0.00
SKILL	4.52	0.92	0.00	1.09	0.32	0.00	0.84	0.31	0.01
Female	1.03	0.64	0.11	0.55	0.25	0.03	0.41	0.24	0.09
SWDs	-0.16	0.67	0.81	-0.67	0.26	0.01	-0.33	0.25	0.18
ELs	-1.57	0.67	0.02	-0.35	0.26	0.18	-0.47	0.25	0.06
Grade level	-2.69	0.95	0.01	0.07	0.36	0.84	0.07	0.35	0.84
Grade level × SKILL	-0.81	1.29	0.53	-0.59	0.48	0.23	-0.83	0.47	0.08
Follow-up									
Intercept	13.17	0.87	0.00	8.14	0.34	0.00	8.75	0.33	0.00
Pretest	0.49	0.06	0.00	0.37	0.07	0.00	0.59	0.06	0.00
SKILL	4.15	0.87	0.00	0.83	0.37	0.02	0.49	0.35	0.16
Female	1.52	0.66	0.02	0.65	0.28	0.02	0.62	0.26	0.02
SWDs	-1.17	0.70	0.09	-1.14	0.29	0.00	-0.86	0.28	0.00
ELs	-0.76	0.70	0.28	-0.39	0.28	0.17	-0.70	0.27	0.01
Grade level	-2.58	0.96	0.01	-0.43	0.40	0.28	-0.53	0.38	0.17
Grade level × SKILL	-0.14	1.31	0.92	-0.71	0.55	0.20	-0.46	0.52	0.38
Random effects									
	Variance	ICC		Variance	ICC		Variance	ICC	
Posttest									
Student-level	26.25	0.84		4.21	0.93		3.92	0.94	
Tutor-level	4.01	0.13							
School-level	0.90	0.03		0.30	0.07		0.23	0.06	
Follow-up									
Student-level	28.12	0.97							
School-level	0.79	0.03							

Note. ICC = Intraclass Correlation Coefficient; SKILL = Supporting Knowledge of Language and Literacy; EL = English learners; SWD = students with disabilities; MISL = Monitoring Indicators of Scholarly Language; TNL = Test of Narrative Language.

suggest that the *SKILL* intervention may minimize disparities in the education of at-risk children by improving oral and written language abilities that are maintained over time periods during which no instruction is occurring.

Our third research question asked whether the effects of the *SKILL* intervention were moderated by grade level. We grouped 1st and 2nd graders and 3rd and 4th graders because of similarities in academic expectations for these groups. We found statistically significant outcomes of the *SKILL* intervention on oral narrative production and written narrative composition regardless of grade level. However, grade-level did affect performance on the standardized reading comprehension measure. Children in 3rd and 4th grade demonstrated small, significant gains on a standardized measure of reading comprehension (the GMRT) even though reading ability was not directly targeted during any of the lessons. The ability to generalize from instruction on oral narrative comprehension and production to reading comprehension may relate to the fact that the older students had better developed word reading allowing them to access more complex text and thus use the narrative comprehension skills they acquired within text comprehension. Our findings are consistent with those reported by LARRC (2015) whose multicomponent language intervention program was found to indirectly impact reading comprehension for their 3rd grade students but not their 1st or 2nd graders. It is possible that the relatively lower reading comprehension scores of 1st and 2nd

grade students in the treatment condition compared with those in the BAU could be a function of the factors that influence reading comprehension in these very early grades. As we know from the simple view of reading (Gough & Tunmer, 1986), word reading plays a more significant role in reading comprehension in the early grades and thus the full effects on reading comprehension of a treatment addressing oral linguistic comprehension may be unrealized until grades 3 and above.

This RCT demonstrated that a relatively short (3 month), multi-component, manualized narrative intervention program that directly focused on narrative macrostructure and microstructure led to statistically and practically significant gains on oral narrative abilities directly, and written narrative abilities indirectly for at-risk students in Grades 1–4. Importantly, the significant gains on oral and written narration were sustained over a 5-month period that spanned the summer break, in which children did not receive formal instruction. Additionally, the *SKILL* intervention was associated with significant gains in reading comprehension for at-risk children in 3rd and 4th grades even though students were never asked to read texts during instruction.

Limitations and Future Directions

Although this study was conducted according to WWC standards and included an adequately sized sample across four grade-levels, there were limitations. First, to generalize our results to the kinds of

Table 7
Estimating the Moderating Effect of Grade Level on Literacy Outcomes at Posttest and Follow-Up

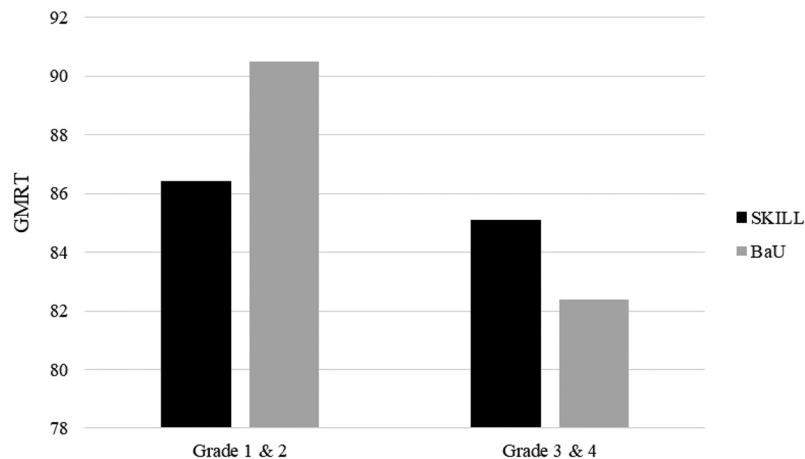
Effect	MISL written narrative			GMRT comprehension		
	Estimate	SE	<i>p</i>	Estimate	SE	<i>p</i>
Fixed effects						
Posttest						
Intercept	9.15	0.73	0.00	84.93	1.60	0.00
Pretest	0.66	0.06	0.00	0.35	0.07	0.00
SKILL	2.66	0.74	0.00	2.73	1.47	0.06
Female	0.67	0.57	0.24	0.29	1.19	0.80
SWDs	0.09	0.60	0.88	-2.12	1.27	0.10
ELs	-0.56	0.59	0.34	-3.20	1.26	0.01
Grade level	-1.37	0.87	0.11	8.13	1.85	0.00
Grade level × SKILL	-1.14	1.13	0.31	-6.84	2.26	0.00
Follow-up						
Intercept	10.62	0.75	0.00	84.90	1.64	0.00
Pretest	0.48	0.06	0.00	0.18	0.07	0.01
SKILL	2.44	0.73	0.00	3.03	1.70	0.08
Female	1.19	0.56	0.04	0.91	1.26	0.47
SWDs	-0.78	0.59	0.19	-1.78	1.39	0.20
ELs	-0.20	0.59	0.73	-2.64	1.32	0.05
Grade level	-2.65	0.86	0.00	1.83	1.89	0.33
Grade level × SKILL	-0.95	1.10	0.39	-3.07	2.56	0.23
Random effects						
	Variance	ICC		Variance	ICC	
Posttest						
Student-level	23.64	0.99		85.23	0.78	
Teacher-level				24.65	0.22	
School-level	0.11	0.01				
Follow-up						
Student-level	19.86	0.96		66.20	0.61	
Tutor-level				31.62	0.29	
Teacher-level				9.92	0.09	
School-level	0.84	0.04				

Note. SKILL = Supporting Knowledge of Language and Literacy; EL = English learners; SWD = students with disabilities; MISL = Monitoring Indicators of Scholarly Language; GMRT = Gates-MacGinitie Reading Test.

students who comprise most elementary school classrooms, we included all children who were screened who met our qualification criteria of performance at or below the 33rd percentile on standardized measures of language and reading comprehension. Our sample contained monolingual and bilingual children with mild to severe

language and literacy difficulties, and children represented a variety of disability categories (e.g., learning disabilities or speech and language impairment). We do not know the relative effects of the SKILL treatment for subsamples of children (e.g., learning disabled, severely impaired) because we did not stratify our randomization on

Figure 2
Grade Level as a Moderator of SKILL Effect on GMRT at Posttest



these variables, and the sample size for each possible subgroup would be relatively small. Our future research plan is to examine closely whether there are differential impacts of instruction on specific populations including students who are bilingual or students from low socioeconomic backgrounds who, because of reduced opportunities, may profit from systematic and supplemental multi-component language instruction.

Unfortunately, the COVID-19 pandemic occurred at the end of the third cohort year, which necessitated a transfer to online posttesting and follow-up testing. The TNL-2, which was our primary outcome variable, was not normed using online procedures. To assess the validity of online testing with the TNL-2, we conducted confirmatory factor analyses in which we compared the fit the TNL-2 at pretest, posttest and follow-up to the measure's a priori factor model (derived from in-person testing of the entire normative sample). The model fits were very similar to the full model at each occasion. We also found that the data that were collected online for the children in Cohort 3 did not differ from the face-to-face posttest and follow-up data that were collected from students in Cohorts 1 and 2 (Magimairaj et al., 2022). Our analyses, while suggesting that the TNL-2 operated similarly whether administered online or to face-to-face, suggest comparability of measurement across modes; they do not confirm such. Further, to the extent that changes in mode may have compromised the measures' comparability across platforms, bias in estimates of effect were not in favor of the treatment. That said, invariance evaluated using the above-described method does not necessarily preclude the possibility of an effect attributable to changes in the mode of test administration, whether considered at the item or at the passage level. It is conceivable that the changeover to zoom had the same effect (i.e., statistically nondifferent) on all students' responses to all items (or to all passages if items are aggregated) at later test occasions, such that interitem or interpassage correlations remained the same even in the presence of a "zoom effect." This possibility, if true, would attenuate the measures' reliability, resulting in underestimates of SKILL's effects. However, randomization should have mitigated the possibility of differential group effects of the COVID-19 pandemic on final data collection for Cohort 3.

Summary and Clinical Implications

It has been well-established that knowledge and use of multiple domains of language contribute to the acquisition of advanced literacy development (Hall et al., 2021; Phillips et al., 2021). Recent studies suggest that targeting multiple language skills (i.e., text structure, vocabulary, syntax, inferential skills, comprehension monitoring) rather than singular and/or modular abilities (i.e., vocabulary) may result in comprehensive and sustained changes in language in a relatively short amount of time (Clarke et al., 2010; Phillips et al., 2021). The SKILL intervention addressed multiple foundational language skills in authentic contexts. Importantly, these language skills were taught within a predictable organizational framework (text structure) that was introduced before direct instruction in decontextualized, literate language skills. This approach yielded practically and statistically significant improvements in narrative language and generalized to measures of written narration performance for students randomized to the narrative instruction group in our study. These results suggest that SKILL is beneficial for students with language and literacy difficulties.

References

- Allen, M. M., Ukrainetz, T. A., & Carswell, A. L. (2012). The narrative language performance of three types of at-risk first-grade readers. *Language, Speech & Hearing Services in Schools, 43*(2), 205–221. [https://doi.org/10.1044/0161-1461\(2011/11-0024\)](https://doi.org/10.1044/0161-1461(2011/11-0024))
- Anderson, J. R. (1993). *The Rules of the Mind*. Erlbaum.
- Anderson, J. R., & Lebiere, C. J. (2014). *The anatomic components of thought*. Psychology Press. <https://doi.org/10.4324/9781315805696>
- August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners (Report of the National Literacy Panel on Language Minority Children and Youth)*. Erlbaum.
- Babayigit, S., Roulstone, S., & Wren, Y. (2021). Linguistic comprehension and narrative skills predict reading ability: A 9-year longitudinal study. *The British Journal of Educational Psychology, 91*(1), 148–168. <https://doi.org/10.1111/bjep.12353>
- Barnes, M. A., Stuebing, K., Fletcher, J. M., Barth, A., & Francis, D. (2016). Cognitive difficulties in struggling comprehenders and their relation to reading comprehension: A comparison of group selection and regression-based models. *Journal of Research on Educational Effectiveness, 9*(2), 153–172. <https://doi.org/10.1080/19345747.2015.1111482>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Berman, R. A. (1988). On the ability to relate events in narrative. *Discourse Processes, 11*(4), 469–497. <https://doi.org/10.1080/01638538809544714>
- Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Erlbaum.
- Bower, G. H. (2008). The evolution of a cognitive psychologist: A journey from simple behaviors to complex mental acts. *Annual Review of Psychology, 59*(1), 1–27. <https://doi.org/10.1146/annurev.psych.59.103006.093722>
- Bowyer-Crane, C., Snowling, M. J., Duff, F. J., Fieldsend, E., Carroll, J. M., Miles, J., Götz, K., & Hulme, C. (2008). Improving early language and literacy skills: Differential effects of an oral language versus a phonology with reading intervention. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 49*(4), 422–432. <https://doi.org/10.1111/j.1469-7610.2007.01849.x>
- Cain, K. (2003). Text comprehension and its relation to coherence and cohesion in children's fictional narratives. *British Journal of Developmental Psychology, 21*(3), 335–351. <https://doi.org/10.1348/02615100332277739>
- Campbell, V. A., Gilyard, J. A., Sinclair, L., Sternberg, T., & Kailes, J. I. (2009). Preparing for and responding to pandemic influenza: Implications for people with disabilities. *American Journal of Public Health, 99*(S2), S294–S300. <https://doi.org/10.2105/AJPH.2009.162677>
- Catts, H., Fey, M., Zhang, X., & Tomblin, B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading, 3*(4), 331–361. https://doi.org/10.1207/s1532799xssr0304_2
- Champion, T. B. (2003). *Understanding storytelling among African American children: A journey from Africa to America*. Erlbaum.
- Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading comprehension difficulties: A randomised controlled trial. *Psychological Science, 21*(8), 1106–1116. <https://doi.org/10.1177/0956797610375449>
- Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greenhouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*(3), 227–268. <https://doi.org/10.3102/00346543066003227>
- Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge University Press. <https://doi.org/10.1017/CBO9781139174909.006>

- Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, 26(2), 197–223. <https://doi.org/10.1007/s10648-013-9246-y>
- Cowan, N., Morey, C. C., & Naveh-Benjamin, N. (2021). An embedded-processes approach to working memory: How is it distinct from other approaches and to what ends? In R. H. Logie, V. Camos, & N. Cowan (Eds.), *Working memory: State of the science* (pp. 44–84). Oxford University Press.
- Coyne, M. D., Little, M., Rawlinson, D. A., Simmons, D., Kwok, O. M., Kim, M., Simmons, L., Hagan-Burke, S., & Civetelli, C. (2013). Replicating the impact of a supplemental beginning reading intervention: The role of instructional context. *Journal of Research on Educational Effectiveness*, 6(1), 1–23. <https://doi.org/10.1080/19345747.2012.706694>
- Curenton, S. M., & Justice, L. M. (2004). African American and Caucasian Preschoolers' Use of Decontextualized Language: Literate Language Features in Oral Narratives. *Language, Speech & Hearing Services in Schools*. Advance online publication. [https://doi.org/10.1044/0161-1461\(2004\)023](https://doi.org/10.1044/0161-1461(2004)023)
- Dodd, J. L., Ocampo, A., & Kennedy, K. S. (2011). Perspective taking through narratives: An intervention for students with ASD. *Communication Disorders Quarterly*, 33(1), 23–33. <https://doi.org/10.1177/1525740110395014>
- Enders, C. K., & Tofghi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological methods*, 12(2), 121–138. <https://doi.org/10.1037/1082-989x.12.2.121>
- Favot, K., Carter, M., & Stephenson, J. (2021). The effects of oral narrative intervention on the narratives of children with language disorder: A systematic literature review. *Journal of Developmental and Physical Disabilities*, 33(4), 489–536. <https://doi.org/10.1007/s10882-020-09763-9>
- Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4), 469–495. <https://doi.org/10.1006/jmla.1999.2660>
- Fey, M. E., Catts, H. W., Proctor-Williams, K., Tomblin, J. B., & Zhang, X. (2004). Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research*, 47(6), 1301–1318. [https://doi.org/10.1044/1092-4388\(2004\)098](https://doi.org/10.1044/1092-4388(2004)098)
- Gambrell, L., Malloy, J., & Mazzoni, S. (2011). Evidence-based best practices in comprehensive literacy instruction. In L. Morrow & L. Gambrell (Eds.), *Best practices in literacy instruction* (4th ed., pp. 18–21). Guilford Press.
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, Reading, and Reading Disability. *Remedial and Special Education*, 7(1), 6–10. <https://doi.org/10.1177/074193258600700104>
- Gillam, R. B., Gillam, S. L., & Fey, M. E. (2017). Supporting knowledge in language and literacy (SKILL): A narrative-based language intervention. In R. McCauley, M. E. Fey, & R. B. Gillam (Eds.), *Treatment of language disorders* (2nd ed., pp. 389–420). Brookes.
- Gillam, S. L., Gillam, R. B., & Laing, C. (2018). *SKILL narrative: Supporting knowledge in language and literacy* (3rd ed.). Utah State University.
- Gillam, S. L., Hartzheim, D., Studenka, B., Simonsmeier, V., & Gillam, R. (2015). Narrative intervention for children with autism spectrum disorder (ASD). *Journal of Speech, Language, and Hearing Research*, 58(3), 920–933. https://doi.org/10.1044/2015_JSLHR-L-14-0295
- Gillam, R. B., Loeb, D. F., Hoffman, L. M., Bohman, T., Champlin, C. A., Thibodeau, L., Widen, J., Brandel, J., & Friel-Patti, S. (2008). The efficacy of Fast ForWord Language intervention in school-age children with language impairment: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research*, 51(1), 97–119. [https://doi.org/10.1044/1092-4388\(2008\)007](https://doi.org/10.1044/1092-4388(2008)007)
- Gillam, S. L., Olszewski, A., Fargo, J., & Gillam, R. B. (2014). Classroom-based narrative and vocabulary instruction: Results of an early-stage, non-randomized comparison study. *Language, Speech, and Hearing Services in Schools*, 45(3), 204–219. https://doi.org/10.1044/2014_LSHSS-13-0008
- Gillam, S. L., Olszewski, A., Squires, K., Wolfe, K., Slocum, T., & Gillam, R. B. (2018). Improving narrative production in children with language disorders: An early-stage efficacy study of a narrative intervention program. *Language, Speech, and Hearing Services in Schools*, 49(2), 197–212. https://doi.org/10.1044/2017_LSHSS-17-0047
- Gillam, R. B., & Pearson, N. A. (2004). *Test of narrative language: Examiner's manual*. Pro-Ed, Inc.
- Gillam, R. B., & Pearson, N. A. (2017). *Test of narrative language* (2nd ed.). Pro-Ed, Inc.
- Gillam, S. L., Fargo, J. D., Petersen, D. B., & Clark, M. (2012). Assessment of structure dependent narrative features in modeled contexts: African American and European American children. *English Linguistics Research*, 1(1), 1–12. <https://doi.org/10.5430/elr.v1n1p1>
- Gillam, S. L., Gillam, R. B., Fargo, J. D., Olszewski, A., & Segura, H. (2017). Monitoring indicators of scholarly language: A progress-monitoring instrument for measuring narrative discourse skills. *Communication Disorders Quarterly*, 38(2), 96–106. <https://doi.org/10.1177/1525740116651442>
- Graesser, A. C. (2007). *An introduction to strategic reading and comprehension*. Psychology Press.
- Greenhalgh, K. S., & Strong, C. J. (2001). Literate language features in spoken narratives of children with typical language and children with language impairments. *Language, Speech, and Hearing Services in Schools*, 32(2), 114–125. [https://doi.org/10.1044/0161-1461\(2001\)010](https://doi.org/10.1044/0161-1461(2001)010)
- Hall, C., Capin, P., Vaughn, S., Gillam, S. L., Wada, R., Fall, A.-M., Roberts, G., Dille, J. T., & Gillam, R. (2021). Narrative instruction in elementary classrooms: An observation study. *The Elementary School Journal*, 121(3), 454–483. <https://doi.org/10.1086/712416>
- Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315650982>
- Johnston, J. R. (2008). Narratives: Twenty-five years later. *Topics in Language Disorders*, 28(2), 93–98. <https://doi.org/10.1097/01.TLD.0000318931.08807.01>
- Kamhi, A. G., & Catts, H. W. (2012). *Language and reading disabilities*. Pearson.
- Kim, Y.-S. G., Petscher, Y., Uccelli, P., & Kelcy, B. (2020). Academic language and listening comprehension—Two sides of the same coin? An empirical examination of their dimensionality, relations to reading comprehension, and assessment modality. *Journal of Educational Psychology*, 112(7), 1267–1287. <https://doi.org/10.1037/edu0000430>
- Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363–394. <https://doi.org/10.1037/0033-295X.85.5.363>
- Language and Reading Research Consortium. (2015). Learning to read: Should we keep things simple? *Reading Research Quarterly*, 50(2), 151–169. <https://doi.org/10.1002/rrq.99>
- Lenth, R. V., Buerkner, P., Love, J., Riebl, H., & Singmann, H. (2020). Emmeans: Tests in linear mixed effects models. R package version 1.7.
- Lewis, R. L., Vasisht, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10), 447–454. <https://doi.org/10.1016/j.tics.2006.08.007>
- Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education into more readily interpretable forms*. National Center for Special Education Research.
- Loban, W. (1976). *Language development: Kindergarten through grade twelve*. National Council of Teachers of English.
- MacGinitie, W. H., MacGinities, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2002). *Gates-MacGinities reading tests: Technical manual, Forms S & T*. Rolling Meadows.

- Magimairaj, B. M., Capin, P., Gillam, S. L., Vaughn, S., Roberts, G., Fall, A.-M., & Gillam, R. (2022). Online administration of the Test of Narrative Language-Second Edition: Psychometrics and considerations for remote assessment. *Language Speech and Hearing Services in Schools*, 53(2), 404–416. https://doi.org/10.1044/2021_LSHSS-21-00129
- Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 9(1), 111–151. [https://doi.org/10.1016/0010-0285\(77\)90006-8](https://doi.org/10.1016/0010-0285(77)90006-8)
- McCabe, A., & Bliss, L. S. (2005). Narratives from Spanish-speaking children with impaired and typical language development. *Imagination, Cognition and Personality*, 24(4), 331–346. <https://doi.org/10.2190/CJQ8-8C9G-05LG-0C2M>
- Menard, J. J., & Wilson, A. M. (2014). Summer learning loss among elementary school children with reading disabilities. *Exceptionality Education International*, 23(1), 72–85. <https://doi.org/10.5206/eei.v23i1.7705>
- Miller, J., Andriacchi, K., & Nockerts, K. (2019). *Assessing language production using SALT software: A clinician's guide to language sample analysis* (3rd ed.). SALT Software, LLC.
- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). Reaching higher. http://www.corestandards.org/assets/CommonCoreReport_6.10.pdf
- Nicolopoulou, A., & Trapp, S. (2018). Narrative interventions for children with language disorders: A review of practices and findings. In A. Baron & D. Ravid (Eds.), *Handbook of communication disorders* (pp. 357–386). De Gruyter Mouton. <https://doi.org/10.1515/9781614514909-018>
- Pesco, D., & Gagné, A. (2017). Scaffolding narrative skills: A meta-analysis of instruction in early childhood settings. *Early Education and Development*, 28(7), 773–793. <https://doi.org/10.1080/10409289.2015.1060800>
- Petersen, D. B. (2011). A systematic review of narrative-based language intervention with children who have language impairment. *Communication Disorders Quarterly*, 32(4), 207–220. <https://doi.org/10.1177/1525740109353937>
- Petersen, D. B., Staskowski, M., Spencer, T. D., Foster, M. E., & Brough, M. P. (2022). The effects of a multitiered system of language support on kindergarten oral and written language: A large-scale randomized controlled trial. *Language, Speech, and Hearing Services in Schools*, 53(1), 44–68. https://doi.org/10.1044/2021_LSHSS-20-00162
- Phillips, B. M., Kim, Y. S. G., Lonigan, C. J., Connor, C. M., Clancy, J., & Al Otaiba, S. (2021). Supporting language and literacy development with intensive small-group interventions: An early childhood efficacy study. *Early Childhood Research Quarterly*, 57(4), 75–88. <https://doi.org/10.1016/j.ecresq.2021.05.004>
- Pico, D. L., Hessling Prah, A., Biel, C. H., Peterson, A. K., Biel, E. J., Woods, C., & Contesse, V. A. (2021). Interventions designed to improve narrative language in school-age children: A systematic review with meta-analyses. *Language, Speech, and Hearing Services in Schools*, 52(4), 1109–1126. https://doi.org/10.1044/2021_LSHSS-20-00160
- Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, 10(2), 178–192. <https://doi.org/10.1037/1082-989X.10.2.178>
- 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\(2006/020\)](https://doi.org/10.1044/0161-1461(2006/020))
- Quinn, D., & Polikoff, M. (2017). *Summer learning loss: What is it, and what can we do about it*. Brookings Institution.
- R Core Team. (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. <https://www.r-project.org><https://www.r-project.org>
- Rand Reading Study Group. (2002). *Rand report*. Rand.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.
- Rogde, K., Hagen, A., Melby-Lervag, M., & Lervag, A. (2019). The effect of linguistic comprehension instruction on generalized language and reading comprehension skills: A systematic review. *Campbell Systematic Reviews*. Advance online publication. <https://doi.org/10.1002/cl2.1059>
- Rubin, D. L., Hafer, T., & Arata, K. (2000). Reading and listening to oral-based versus literate-based discourse. *Communication Education*, 49(2), 121–133. <https://doi.org/10.1080/03634520009379200>
- Schoenbrodt, L., Kerins, M., & Gesell, J. (2003). Using narrative language intervention as a tool to increase communicative competence in Spanish-speaking children. *Language, Culture and Curriculum*, 16(1), 48–59. <https://doi.org/10.1080/07908310308666656>
- Scott, C. M., & Windsor, J. (2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research*, 43(2), 324–339. <https://doi.org/10.1044/jslhr.4302.324>
- Simmons, D. C., Coyne, M. D., Hagan-Burke, S., Kwok, O., Simmons, L. E., Johnson, C., Zou, Y., Taylor, A. B., McAlenney, A. L., Ruby, M., & Crevecoeur, Y. C. (2011). Effects of supplemental reading interventions in authentic contexts: A comparison of kindergarteners' response. *Exceptional Children*, 77(2), 207–228. <https://doi.org/10.1177/001440291107700204>
- 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(3), 178–199. <https://doi.org/10.1177/1053815110379124>
- Spencer, T. D., Kajian, M., Petersen, D. B., & Bilyk, N. (2013). Effects of an individualized narrative intervention on children's storytelling and comprehension skills. *Journal of Early Intervention*, 35(3), 243–269. <https://doi.org/10.1177/1053815114540002>
- 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(2), 196–217. <https://doi.org/10.1177/1476718X13515419>
- Squires, K. E., Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., Bohman, T. M., & Gillam, R. B. (2014). Story retelling by bilingual children with language impairments and typically developing controls. *International Journal of Language & Communication Disorders*, 49(1), 60–74. <https://doi.org/10.1111/1460-6984.12044>
- Stein, N. L., & Glenn, C. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *New directions in discourse processing* (Vol. 2, pp. 53–120). Ablex.
- Stringfield, S. G., Luscre, D., & Gast, D. L. (2011). Effects of a Story Map on accelerated reader postreading test scores in students with high-functioning autism. *Focus on Autism and Other Developmental Disabilities*, 26(4), 218–229. <https://doi.org/10.1177/1088357611423543>
- Tallal, P. (2013). Fast ForWord®: The birth of the neurocognitive training revolution. *Progress in Brain Research*, 207(2), 175–207. <https://doi.org/10.1016/B978-0-444-63327-9.00006-0>
- Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language*, 1(1), 1–21. <https://doi.org/10.2307/413530>
- Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, 27(1), 77–83. <https://doi.org/10.3102/10769986027001077>
- Tsimpli, I. M., Peristeri, E., & Andreou, M. (2016). Narrative production in monolingual and bilingual children with specific language impairment. *Applied Psycholinguistics*, 37(1), 195–216. <https://doi.org/10.1017/S0142716415000478>
- Ukrainetz, T. A., Justice, L. M., Kaderavek, J. N., Eisenberg, S. L., Gillam, R. B., & Harm, H. M. (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\(2005/095\)](https://doi.org/10.1044/1092-4388(2005/095))
- van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.

Was, C. A., & Woltz, D. J. (2007). Reexamining the relationship between working memory and comprehension: The role of available long-term memory. *Journal of Memory and Language*, 56(1), 86–102. <https://doi.org/10.1016/j.jml.2006.07.008>

Wells, G. (2009). *The meaning makers: Learning to talk and talking to learn* (2nd ed.). Multilingual Matters. <https://doi.org/10.21832/9781847692009>

What Works Clearinghouse. (2020). *What Works Clearinghouse Standards Handbook, Version 4.1*. U.S. Department of Education, Institute of

Education Sciences, National Center for Education Evaluation and Regional Assistance. <https://ies.ed.gov/ncee/wwc/handbooks>

White, G. W., Fox, M. H., Rooney, C., & Cahill, A. (2007). *Assessing the impact of hurricane Katrina on persons with disabilities*. Research and Training Center on Independent Living, University of Kansas. http://www.rtcil.org/products/NIDRR_FinalKatrinaReport.pdf

Wilkinson, I. A. G., & Son, E. H. (2011). *Handbook of reading research, handbook IV*. Routledge.

Appendix

Measurement Invariance Analysis

We provide the following tables in support of the measurement invariance analyses. The fundamental question for those analyses was “to what extent do the patterns of responses differ or not differ across in-person and online testing modes at pretest, posttest, and follow-up.” We report details for data at posttest, follow-up, and for data combined across pretest, posttest, and follow-up. Fit indices are reported across test administration mode for different levels of measurement invariance, including configural invariance, metric invariance, and scalar invariance.

Configural invariance concerns the number of factors and patterns of factor-indicator relationships. Measurement across groups or across occasions is invariant at the configural level if there are no differences in the number of factors or patterns of factor-indicator relationships. Metric invariance means that factor loadings are equal (or nondifferent) across groups or occasions. Scalar invariance indicates that mean differences in the latent construct capture all mean differences in the shared variance of the items. Scalar invariance is tested by constraining the item intercepts to be equivalent in the two groups. A final type of invariance is residual invariance, which means that the sum of specific variance (variance of the item that is not shared with the factor), and error variance (measurement error) is similar across groups. Although a required component for full factorial invariance, testing of residual invariance is not a prerequisite for testing mean differences because the residuals

are not part of the latent factor, so invariance of the item residuals is inconsequential to interpretation of latent mean differences. For this reason and because tests for residual invariance only make sense in the presence of full scalar invariance, we did not fit residual invariance models.

Table A1 presents fit indices across measurement models for the TNL administered online and in-person at posttest. The results indicate partial scalar invariance across modes of administration (online vs. in person)—one of the intercepts varied across online and in-person testing—which means that the measurement models for online and in person were nondifferent when the intercept for McDonald’s story was allowed to vary. Note that the relative fit indices (CFI and TLI) are supportive of full invariance; however, they should be considered in light of the statistically significant χ^2 value. Table A2 summarizes the fit indices for levels of invariance across administration modes at follow-up. Again, partial scalar in the TNL as supported. In this case, the intercepts for McDonald’s story, Late for School, and Aliens were allowed to vary. Finally, Table A3 presents fit indices for models using data at all test occasions (pretest, posttest, and follow-up). As before, we found scalar variance when the intercept for McDonald’s story at posttest, McDonald’s story at follow-up, Late for School at follow-up, Aliens at follow-up, and Treasure at follow-up were allowed to vary (i.e., partial scalar invariance).

Table A1

Fit Indices for Measurement Invariance Models for Online and In-Person Testing on TNL Posttest Data

Model	χ^2	<i>df</i>	CFI	TLI	RMSEA	$\Delta\chi^2$	Δdf	<i>p</i> value
Configural invariance	20.21	14	0.99	0.97	0.05			
Metric invariance	22.93	18	0.99	0.99	0.04	2.719	4	0.61
Scalar invariance	32.24	22	0.98	0.98	0.06	9.311	4	0.05
Partial scalar 1	28.55	21	0.99	0.98	0.05	5.62	3	0.13

(Appendix continues)

Table A2*Fit Indices for Measurement Invariance Models for Online and In-Person Testing on TNL Follow-Up Data*

Model	χ^2	<i>df</i>	CFI	TLI	RMSEA	$\Delta\chi^2$	Δdf	<i>p</i> value
Configural invariance	23.66	14	0.98	0.96	0.07			
Metric invariance	29.17	18	0.98	0.96	0.07	5.51	4	0.24
Scalar invariance	74.60	22	0.90	0.86	0.13	47.24	6	0.00
Partial scalar 1	58.63	21	0.93	0.90	0.11	29.46	3	0.00
Partial scalar 2	48.25	20	0.95	0.92	0.10	19.07	2	0.00
Partial scalar 3	32.12	19	0.98	0.96	0.07	2.96	1	0.09

Table A3*Fit Indices for Measurement Invariance Models for Online and In-Person Testing on TNL Pretest, Posttest, and Follow-Up Data*

Model	χ^2	<i>df</i>	CFI	TLI	RMSEA	$\Delta\chi^2$	Δdf	<i>p</i> value
Configural invariance	290.76	200	0.96	0.94	0.05			
Metric invariance	302.00	212	0.96	0.95	0.05			
Scalar invariance	363.66	224	0.94	0.92	0.06	61.67	12	0.00
Partial scalar 1	348.68	223	0.95	0.93	0.06	46.68	11	0.00
Partial scalar 2	345.00	222	0.95	0.93	0.06	43.01	10	0.00
Partial scalar 3	322.21	221	0.96	0.94	0.05	20.21	9	0.02
Partial scalar 4	319.38	220	0.96	0.94	0.06	17.39	8	0.03
Partial scalar 5	313.90	219	0.96	0.94	0.05	11.90	7	0.10

Received October 7, 2021
Revision received June 1, 2022
Accepted June 3, 2022 ■