HAMMILL INSTITUTE ON DISABILITIES

Communication Disorders Quarterly
2017, Vol. 38(2) 96–106
© Hammill Institute on Disabilities 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1525740116651442
journals.sagepub.com/home/cdq
SAGE

# Monitoring Indicators of Scholarly Language: A Progress-Monitoring Instrument for Measuring Narrative Discourse Skills

**Sandra Laing Gillam, PhD[1], Ronald B. Gillam, PhD[1], Jamison D. Fargo, PhD[1], Abbie Olszewski, PhD[2], and Hugo Segura, MsC-SLP[3]**

**Abstract**

The purpose of this study was to assess the basic psychometric properties of a progress-monitoring tool designed to measure narrative discourse skills in school-age children with language impairments (LI). A sample of 109 children with LI between the ages of 5 years 7 months and 9 years 9 months completed the *Test of Narrative Language* (TNL). The stories told in response to the alien picture prompt were transcribed and scored according to the TNL manual criteria and the criteria established for scoring the progress-monitoring tool, Monitoring Indicators of Scholarly Language (MISL). The MISL total score demonstrated acceptable levels of internal consistency reliability, inter-rater reliability, and construct validity for use as a progress-monitoring tool for specific aspects of narrative proficiency. The MISL holds promise as a tool for tracking growth in narrative language proficiency that may be taught as part of an intervention program to support the Common Core Standards related to literacy.

**Keywords**

assessment, narrative(s), literacy, speech-language pathologists (SLPs), elementary school, age, language

Speech-language pathologists (SLPs) are increasingly being called upon to provide evidence that their intervention efforts result in positive educational outcomes for students in school-based settings (American Speech-Language-Hearing Association [ASHA], 2000). This involves the provision of educationally relevant instruction and authentic documentation of student outcomes through a process called progress-monitoring (S. Gillam & Gillam, 2006; S. Gillam & Justice, 2010). The information obtained through progress-monitoring is used to inform clinical decisions about methods and procedures, dosage, service delivery, and to communicate accurate and consistent information about a child's progress to others (D. Paul & Hasselkus, 2004; Sutherland Cornett, 2006; Warren, Fey, & Yoder, 2007). Ideally, these tools should possess some basic psychometric properties such as inter-rater reliability, internal consistency reliability, and construct validity if SLPs are to have some degree of confidence in their ability to capture differences in performance as a result of intervention (American Institutes for Research, 2015a).

One of the roles and responsibilities of SLPs employed in educational settings is to design and implement intervention programs that target the language underpinnings that are foundational to curricular content related to literacy development. Then, they should monitor how well students respond to the instruction (ASHA, 2001; Ehren & Whitmire, 2009). According to Common Core State Standards (CCSS-ELA.Literacy.W.3.3), school-age children must be able to "compose narratives to develop real or imagined experiences or events using effective technique, well chosen details, and well-structured event sequences" (CCSS; National Governors Association and Council of Chief State School Officers, 2011). Component language skills that may be taught in support of this overarching discourse-level goal may include teaching students to "ask and answer questions about key details in text (CCSS.ELA-Literacy.RL.1.1)," "retell stories including key details (CCSS.ELA-Literacy.RL.1.2)," and to "describe the overall structure of a story, including how the beginning introduces the story and the ending

[1]Utah State University, Logan, USA
[2]University of Nevada, Reno, USA
[3]Universidad Autonoma de Chile, Talca

**Corresponding Author:**
Sandra Laing Gillam, Communicative Disorders and Deaf Education, Emma Eccles Jones Early Childhood Education and Research Center, Utah State University, 2610 Old Main Hill, Logan, UT 84322, USA.
Email: sandi.gillam@usu.edu

concludes the action (CCSS.ELA-Literacy.RL.2.5)." The authors designed a progress-monitoring tool to measure growth in the ability to generate fictional stories consistent with standards outlined in the CCCS (National Governors Association and Council of Chief State School Officers, 2011). A brief list of the reading and writing anchor standards that define what students should understand and be able to accomplish by the end of Grade 3 that are directly measured on the progress-monitoring tool described in this article (Monitoring Indicators of Scholarly Language [MISL]; S. Gillam, Gillam, & Laing, 2012) is provided in the Online Supplemental Material A. The purpose of this study was to evaluate the psychometric properties of a progress-monitoring tool called MISL (S. Gillam et al., 2012).

## Measuring Key Components of Narrative Discourse

In addition to measuring skills that are related to the Common Core, a narrative progress-monitoring tool should contain items that are consistent with models of narration. Narratives are generally characterized according to macrostructure and microstructure components. *Macrostructure* is usually defined as a setting plus one or more episodes (Stein, 1988; Stein & Glenn, 1979, in Squires et al., 2004). A *setting* is a reference to the time or place that the story occurred. Children may use fairly simple setting references, such as "outside" or "in the rain," or more specific, sophisticated setting elements such as "Central Park" or "Washington, D.C." A basic episode consists of an *initiating event* (IE), which is an incident that motivates actions by the main character(s)' goal-directed actions known as *attempts*, and a *consequence* (or outcome) that is related to both the IE and the actions. By 8 years of age, typically developing children tell complex narratives that contain complicating actions (occurrences that interfere with the goal-directed actions of characters) and/or multiple IEs with associated actions and consequences (Berman, 1988). For story coherence, it is important that the temporal and causal relationships between the IE, character actions related to the IE, and the consequences of those actions are clear to the listener. In fact, the amount of information one can retrieve for use in answering questions and composing retells is related to the number of causal relationships contained in a story (van den Broek, Linzie, Fletcher, & Marsolek, 2000; White, van den Broek, & Kendeou, 2007).

Narrative microstructure consists of the words and sentences that comprise a story. A critical part of narrative development during the school-age years relates to the increased use of literate or scholarly microstructure forms, sometimes referred to as literate language structures (Greenhalgh & Strong, 2001; R. Paul, 1995; Westby, 1985). Important aspects of literate language include coordinating and subordinating conjunctions (*for, and, nor, but, or, yet, so*), adverbs (*suddenly, again, now*), and elaborated noun phrases (*the big green monster*). Other literate language features include metacognitive verbs such as *think, believe*, and *decide* that refer to acts of thinking or feeling, and metalinguistic verbs such as *tell, yell*, and *argue* that refer to acts of speaking (Westby, 2005).

Measures of microstructure summarize relevant aspects of linguistic proficiency and have been used to differentiate between typically developing children and children with delayed or impaired language abilities (Justice, 2006; Liles, Duffy, Merritt, & Purcell, 1995). Conjunctions, adverbs, elaborated noun phrases, and metacognitive and metalinguistic verbs appear less frequently in the narratives of children with language impairments (LIs) than their typically developing peers (Greenhalgh & Strong, 2001). A progress-monitoring tool known as the Index of Narrative Microstructure (INMIS; Justice et al., 2006) was designed to assess narrative microstructure in children aged 5 to 12. The measure yields information about language productivity (word output, lexical diversity, T-unit output) and complexity (syntactic organization). Scores on two factors (productivity and complexity) may be compared against field test reference data based on age or grade level.

Some narrative measures have been developed to examine aspects of both macrostructural and microstructural aspects of narratives produced by school-age children (Heilmann, Miller, Nockerts, & Dunaway, 2010). For example, the Narrative Scoring Scheme (NSS; Heilmann et al., 2010) incorporates a Likert-type scale scoring approach for coding story elements related to introduction (setting, characters), character development (main character, supporting characters, first person), mental states (feelings), referencing (unambiguous pronouns), conflict resolution (clearly stated), cohesion (logical order, smooth transitions), and conclusion (story has clear ending). Story elements are coded as *proficient* (score of 5), *emerging* (score of 3), or *minimal/immature* (score of 1). Normative databases using the NSS to score self-generated stories and retells generated from wordless picture books are included in the Systematic Analysis of Language Transcripts manual (Miller, Andriacchi, & Nockerts, 2011).

The *Index of Narrative Complexity* (INC) was also developed for measuring macrostructure and microstructural elements of narration in school-age children (Petersen, Gillam, & Gillam, 2008). The INC contains scales to measure macrostructure components (character, setting, IE, internal response, plan, attempt, consequence) and microstructure features (coordinated and subordinated conjunctions, adverbs, metacognitive and metalinguistic verbs, and elaborated noun phrases) of self-generated stories and retells. We revised the INC into a measure, called MISL, which was designed to track the range of progress from the production of simple descriptions produced by very young children to more sophisticated multi-episode narratives produced by children in the upper elementary grades (the MISL rubric is available as Online Supplemental Material B).

The MISL is primarily used for assessing self-generated narratives elicited in response to sequenced pictures and single scene prompts, but it has also been used to track progress in story retelling. In the next sections, we describe the psychometric properties that we report for the MISL including estimates of reliability and construct validity.

## Characteristics of Psychometrically Sound Progress-Monitoring Tools

A progress-monitoring tool should yield reliable scores for measuring the component skills that correspond to success in a particular domain (American Institutes for Research, 2015a). According to The National Center on Intensive Intervention technical review committee, progress-monitoring tools should contain estimates of reliability and construct validity (American Institutes for Research, 2015b).

Reliability estimates for performance-level scores may include internal consistency reliability and inter-rater reliability. Internal consistency reliability refers to the extent to which responses to the items on a scale correlate with one another. Typically, internal consistency reliability is measured using a statistic called Cronbach's alpha. Inter-rater reliability refers to the degree to which different raters reach the same conclusions in scoring. To demonstrate minimum reliability, reliability coefficients should be equal to or greater than .70 (Nunnally & Bernstein, 1994).

In addition to being reliable, progress-monitoring tools should be valid (Briesch & Volpe, 2007; Lueger & Barkham, 2010; Overington & Ionita, 2012). One measure of validity is construct validity, which is an accumulation of evidence indicating that scores from an instrument measure what the instrument is intended to measure. A confirmatory factor analysis (CFA) may be conducted to establish this construct. In CFA, examiners create factor structures that test whether hypotheses made about the measure correspond to a theoretical notion. For example, if a clinician wished to measure narrative discourse skills, the tool should be composed of items known to reflect knowledge of narrative macrostructure and microstructure. The purpose of this study was to assess the inter-rater reliability, the internal consistency reliability, and the construct validity of the MISL. Our research questions were as follows:

> **Research Question 1:** To what extent do two raters who score narratives independently agree on the values that are assigned to the MISL items (inter-rater reliability)?
>
> **Research Question 2:** To what extent do the items on the MISL correlate with each other (internal consistency reliability)?
>
> **Research Question 3:** Are there two multiple dimensions (macrostructure and microstructure) underlying the items on the MISL (construct validity)?

## Method

The participants were 109 children (69 males and 40 females) with identified LIs between the ages of 5 years 7 months and 9 years 9 months. These participants were recruited as part of a series of studies to examine the outcomes of language and narrative instructional approaches. Consistent with the EpiSLI model (Tomblin et al., 1997), children were determined to have LI if they displayed standard scores at or below 81 on two or more composite scores from the *Test of Language Development–Primary–3rd Edition* (TOLD-P-3; Newcomer & Hammill, 1997) or a composite score below 82 on the *Comprehensive Evaluation of Language Fundamentals–4* (CELF-4; Semel, Wiig, & Secord, 2004) or the *Comprehensive Assessment of Spoken Language* (CASL; Carrow-Woolfolk, 1999). None of the participants presented with hearing, visual, or gross neurological impairments, oral-structural anomalies, or emotional/social disorders, but they all demonstrated average to above average nonverbal reasoning skills as measured by the *Brief Kaufman Intelligence Test* (K-BIT-2: Kaufman & Kaufman, 1990) or the *Universal Nonverbal Intelligence Test* (UNIT: Bracken & McCallum, 1998). Ninety-two of the children were from Texas, and 17 were from Utah. Their demographic characteristics are shown in Table 1.

### Procedures

Trained research assistants or certified SLPs administered the *Test of Narrative Language* (TNL; Gillam & Pearson, 2004) to all the participants before their respective intervention programs began (pre-test). All the assistants were graduate students in speech-language pathology programs under the direct supervision of certified SLPs. Training was provided by the first and second authors to all the research team involved in conducting these assessments. The TNL is a standardized test designed to assess narrative comprehension (NC) and production in children between the ages of 5 and 12. The TNL utilizes three successively more difficult contexts to assess narrative production proficiency. The first context is a scripted narrative. Children were asked to answer questions about the story and to retell it. In the second context, children listened to a story that corresponded to a series of five sequenced pictures. They answered questions about the story they heard and then generated their own story that corresponded to a novel set of five sequenced pictures. The prompts for the third narrative context were single scene pictures depicting fictional events. Children listened to a story about a dragon guarding a treasure and answered questions about it. Then, children were asked to generate a story that

**Table 1.** Demographic Characteristics of Study Samples.

| Demographics | Children with language impairments *M* (*SD*) |
|---|---|
| Gender | |
| Male | 69 |
| Female | 40 |
| Variable | |
| Race and ethnicity | |
| White, not Hispanic | 26 |
| White, Hispanic | 26 (12 bilingual, English was first language) |
| African American | 21 |
| American Indian | 24 |
| Asian | 12 |
| *Test of Narrative Language* (NLAI) | 76.67 (11.99) |
| *Comprehensive Assessment of Spoken Language* | 75.91 (11.10) |
| *Comprehensive Evaluation of Language Fundamentals* | 73.00 (8.66) |
| Nonverbal intelligence quotient | 95.54 (8.33) |

*Note.* Nonverbal Intelligence Quotient—UNIT or K-BIT. NLAI = *Narrative Language Ability Index*; UNIT = *Universal Nonverbal Intelligence Test*; K-BIT = *Kaufman Brief Intelligence Test.*

corresponded to a novel scene depicting an alien family landing in a park. The TNL yields an overall *Narrative Language Ability Index* (NLAI) as well as composite scores for NC and oral narration (ON). MISL scoring was conducted on the narratives generated while children looked at the novel scene depicting an alien family landing in the park.

## Transcription

The stories told in response to the alien picture prompt were digitally recorded and transcribed according to Systematic Analysis of Language Transcription (SALT) conventions (Miller & Chapman, 2004). Narratives were transcribed verbatim with the inclusion of both child and examiner utterances when applicable. Two research assistants who did not administer the TNL and who were unaware of the purpose of the research project segmented transcripts into communication units (C-units; Loban, 1976) that consisted of an independent main clause and any phrases or clause(s) subordinated to it. Utterances were also coded for the presence of mazes (reformulations, reduplications, and false starts). Accuracy of the transcription and coding process was reviewed by examining 30% of the written transcripts. Percentage of agreement between primary and secondary transcribers/coders was 98% for C-unit segmentation and 95% for mazes.

*MISL description and scoring procedures.* The MISL has a Macrostructure subscale and a Microstructure subscale whose scores are combined to reflect an overall narrative proficiency score (total MISL score). The Macrostructure subscale consists of seven story elements (character, setting, initiating event, internal response, plan, action, and consequence). Definitions for these story elements and examples for each are provided in Table 2. Scores of 0 are interpreted as evidence that a story does not contain elements that constitute a basic episode. Accounts that earn scores of 0 may contain simple descriptions of objects or actions (*There is a tree. They are running.*). Scores of 1 indicate that a story has an emerging episodic structure (*There is a boy. He's at the table eating.*). Scores of 2 are taken as evidence that a story contains the necessary elements to constitute a basic episode (*The boy is eating breakfast and then he is going to school. He likes school, so he is hurrying to finish. He ran to school after breakfast.*), and scores of 3 indicate that a story is complex and elaborated (*John and Bill are brothers. They are hurrying to eat breakfast before school. They love going to State Middle School so they are hurrying. All of a sudden, they knocked their cereal bowls over and milk went everywhere. They decided to clean it all up and grab breakfast bars instead. They ate their breakfast bars as they ran to school. They got there just before the bell rang. They were glad they'd gotten to eat breakfast and that they'd made it to school on time.*).

The scoring system for character and setting is similar such that items related to the use of character earn a score of 0 if no reference to a character is made; a score of 1 if an ambiguous reference is stated (*the boy, in the park*); a score of 2 if a specific name is used (*Mark, Central Park*); and a score of 3 if two or more specific references are indicated in the story (*Mark and Mary, Central Park and California*). Therefore, *Mary and Mark walked through Central Park in California* would receive a score of 3 for character and 3 for setting. Recall that scores of 3 are interpreted as evidence that a story is complex and elaborated. The scoring procedures for IE, internal response, plan, action, and consequence is based on whether there is clear evidence that the elements are causally linked and is anchored at a score of 2. (See Online Supplemental Material C for more detail regarding macrostructure scoring.)

There are seven items on the microstructure scale: five items that relate to literate language, a grammaticality item, and a tense item. Nippold (1998) used the term *literate lexicon* to refer to words that are "important for the literate activities of reading, writing, listening to lectures, talking about language and thought, and mastering school curriculum" (p. 21). More recently, R. Paul (2007) wrote that literate language is "the style used in written communication and is typically more complex and less related to the physical context than the language of ordinary conversation" (p. 394). There are five specific linguistic forms that are identified as literate language on the Microstructure subscale of the MISL (Benson, 2009): coordinating conjunctions (*for, and, nor, but, or, yet, so*), subordinating conjunctions (*so, that, because*), adverbs (*quickly, slowly, fast*), metacognitive

**Table 2.** Macrostructure Subscale: Story Elements, Definitions, and Scoring Criteria.

| Story element | Definition | 0 Not present | 1 Emerging | 2 Mastery | 3 Elaborated knowledge |
|---|---|---|---|---|---|
| Character | An agent who performs an action | No main character is included, or uses ambiguous pronouns | Includes at least one main character using non-specific labels with a determiner (the, a) | Includes at least one main character using a proper name | Includes more than one main character using proper names |
| Setting | Information about location or time | No reference to location or time | Reference to a general place or time (not necessarily related to story) | Reference to a specific place or time in the story (related to story)<br>Clear causal connection indicated by use of causal adverbs (e.g., because, so) | Reference to place denoted using proper name, or reference to specific time |
| IE | Event(s) that motivate characters to take action | Series of descriptions, no indication of | Event stated, does not motivate action | One event stated that motivates action | Two or more events that motivate separate actions (complex episode story) |
| Internal response | Feelings stated about the IE. Must be made by the character taking the actions related to the IE | No feelings stated | Feelings stated, but not clearly related to IE | Feelings stated that are clearly related to IE | More than one instance of feelings stated that are clearly related to IE |
| Plan | Thoughts stated by characters related to a decision to take action | No statement provided about the character's plan to take action | Statements about plans to take action, not related to IE | One statement about a plan to take action that is related to IE | Multiple statements about plans to take action that are related to the IE |
| Attempt | Actions taken by characters motivated by IE | No actions are taken by a character | Use of action verb(s) in descriptive sentences. No clear link to an IE | Use of action verbs in sentences clearly linked to an IE | The inclusion of a complicating action that impedes actions taken in response to IE |
| Consequence | End result of characters actions in relation to the IE | No clear "ending" or resolution related to an IE | Outcome of action linked to another action, not to IE | One outcome of action, related to IE | Two or more outcomes, related to IE |

*Note.* IE = initiating event.

verbs (*thought, planned, decided, said, yelled*), and elaborated noun phrases (*the girl, the happy girl, the sweet happy girl*). The grammaticality item relates to grammatical errors such as improper use of pronouns, lack of subject–verb agreement, or tense and inflection errors. For example, the utterance, *Her went home*, would be judged as ungrammatical because it contains a pronoun use error. The tense item assesses whether sentences produced in students' stories contain changes from present to past or future tense or reflect consistent use of one tense. For example, *Yesterday, she* ***walked*** *home. She* ***runs*** *all the way there. She* ***will walk*** *home yesterday*, would be scored as two tense changes. Stories that contained three or more grammatical or tense errors earned scores of 0 in each category. A score of 3 was given for each item if the story contained no grammatical errors or tense changes. Table 3 contains the literate language structures, definitions, and scoring criteria for the Microstructure subscale items.

## *Inter-Rater Reliability*

Two research assistants (coders) who were trained in the use of the MISL and blind to group assignment and the purpose of the study independently scored all the stories produced by participants. The coders had previously participated in an hour-long training to learn how to use the MISL rubric to score macrostructure and microstructure for stories not included in this study. During preliminary training, coders were asked to score four or five stories with the first author, and to ask clarifying questions. The first author discussed scoring scenarios with them and answered their questions about scoring the stories according to the rubric. The coders were cleared to begin scoring stories for this project after they had attained 90% or higher inter-rater reliability with the first author on five consecutive stories.

The procedure for scoring the stories used in this project was as follows: The coders were asked to score 10 stories that were selected randomly from the total corpus of transcripts and then meet to calculate their levels of agreement. Care was taken to select stories from children at each age level (5- to 6-year-olds, 7- to 8-year-olds, 9- to 10-year-olds). Discrepancies were resolved through consensus and confirmed by the first author who made the final decision on scoring. Then, coders were instructed to score 10 additional stories and to meet again to calculate their agreement scores. This procedure of coding 10 stories, meeting to

**Table 3.** Microstructure Subscale: Literate Language Structures, Definitions, and Scoring Criteria.

| Literate language structure | Definition | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| CC | Coordinates clauses | No CCs in story | One CC in story | Two different CCs in story | Three or more different CCs in story |
| SC | Joins a subordinate clause to a main clause | No SCs in story | One SC in story | Two different SCs in story | Three or more different SCs in story |
| MCV | Verbs that refer to passive states of cognition, thinking, knowing, or perception about the world | No MCVs in story | One MCV in story | Two different MCVs in story | Three or more MCVs in story |
| Adverbs | Adjectives used to modify the verb | No adverbs in story | One adverb in story | Two different adverbs in story | Three or more adverbs in story |
| ENP | Modifiers + noun including articles, possessives, determiners, wh-words, adjectives | No ENPs in story | One ENP in story | Two different ENPs in story | Three or more different ENPs in story |
| Grammaticality[a] | The extent to which an utterance is grammatically correct | ≥3 grammatical errors in story | Two grammatical errors in story | One grammatical error in story | No grammatical errors in story |
| Tense[a] | The extent to which the speaker maintains the same tense throughout the story | ≥3 tense changes in story | Two tense changes in story | One tense change in story | No tense change in story |

*Note.* CC = coordinating conjunctions; SC = subordinating conjunctions; MCV = metacognitive, metalinguistic verbs; ENP = elaborated noun phrases.
[a]Grammaticality and tense should not be included in total scores (see validity and reliability results).

resolve discrepancies, and oversight by the first author was incorporated to control for coder drift (S. Gillam, Olszewski, Fargo, & Gillam, 2014). Coder drift is a phenomenon in which reliability decreases over time due to a lack of calibration. Inter-rater reliability percentages were calculated for 20 stories (20%) that had been scored independently by the two raters. To obtain the percentages, the total number of items that the raters agreed on was divided by the total number of items in each subtest and for the total index, then multiplied by 100. The final inter-rater reliability percentages are presented in Table 4. Scores for inter-rater reliability are discussed in the following "Results" section.

## Results

### Inter-Rater Reliability

The first research question was, "To what extent do two raters who score narratives independently agree on the values that are assigned to the MISL items?" We wanted to know whether the MISL possessed reasonable inter-rater reliability to be useful in measuring narrative discourse skills. As can be seen in Table 4, the inter-rater reliability scores for items and subscales ranged from 90% to 100%. For the Macrostructure subscale, inter-rater reliability ranged from 92% (consequence) to 100% (character) and for microstructure, 90% (elaborated noun phrases) to 100% (coordinating conjunctions). The inter-rater reliability scores for each item, the total score, and the macrostructure and microstructure scores were 90% or higher, indicating acceptable levels of coder reliability. These data represent scores for students who range in age from 5 years 7 months to 9 years 9 months.

### Internal Consistency Reliability

The second research question was, "To what extent are the items on the MISL internally reliable?" Total, subscale, and item-level descriptive statistics for the MISL are presented in Table 5. Reliability coefficients at or greater than .70 were considered acceptable (Nunnally & Bernstein, 1994). Preliminary analyses suggested that the measure we used to calculate internal consistency (Cronbach's α) for the MISL significantly improved with the removal of the grammaticality and tense items. The Cronbach's alpha improved from .67 to .79 for the total instrument, and from .36 to .67 for the microstructure scale after removal of these two items. In summary, scores obtained from the MISL demonstrated acceptable levels of internal consistency reliability for the total instrument (α = .79) and the Macrostructure subscale (α = .71), but were slightly lower for the Microstructure subscale (α = .67).

### Construct Validity

The third research question was, "Are there two multiple dimensions (macrostructure and microstructure) underlying the items on the MISL?" Construct validity was evaluated

**Table 4.** Inter-Rater Reliability Scores for Each Macrostructure and Microstructure Element.

| Macrostructure element | % agreement | Microstructure element | % agreement |
|---|---|---|---|
| Character | 100 | Coordinating conjunctions | 100 |
| Setting | 96 | Subordinating conjunctions | 96 |
| IE | 96 | Metacognitive, metalinguistic verbs | 92 |
| Internal response | 98 | Elaborated noun phrases | 90 |
| Plan | 98 | | |
| Action | 96 | | |
| Consequence | 92 | | |
| Macrostructure total | 97 | Microstructure total | 93 |
| | | MISL total | 95 |

*Note.* IE = initiating event.

**Table 5.** Item, Reliability, and Validity Statistics for MISL Items and Subscales.

| | | | Confirmatory factor analysis | | |
|---|---|---|---|---|---|
| MISL items | *M* (*SD*) | Cronbach's α | $R^2$ | *SE* $R^2$ | *p* value $R^2$ |
| MISL total | 11.37 (5.24) | .79 | | | |
| Macro subscale total items | 6.59 (3.40) | .71 | | | |
| Character | 1.14 (0.58) | | .23 | .12 | .043* |
| Setting | 0.73 (0.63) | | .27 | .08 | .001** |
| IE | 1.42 (0.97) | | .93 | .05 | <.001*** |
| Internal response | 0.38 (0.78) | | .04 | .04 | .336 |
| Plan | 0.74 (0.83) | | .14 | .07 | .050 |
| Attempt | 1.58 (0.77) | | .95 | .05 | <.001*** |
| Consequence | 0.60 (1.0) | | .58 | .08 | <.001*** |
| Micro subscale total items | 4.85 (2.51) | .67 | | | |
| Coordinated conjunction | 1.53 (0.85) | | .79 | .14 | <.001*** |
| Subordinated conjunction | 0.32 (0.58) | | .22 | .10 | .018* |
| Metacognitive and metalinguistic verbs | 1.07 (0.95) | | .53 | .12 | <.001*** |
| Adverbs | 0.45 (0.71) | | .22 | .10 | .027* |
| Elaborated noun phrases | 1.51 (0.65) | | .15 | .08 | .055 |
| Grammaticality[a] | 0.97 (1.05) | | | | |
| Tense[a] | 1.60 (1.20) | | | | |

*Note.* Mann–Whitney *U* test was used to calculate. MISL = Monitoring Indicators of Scholarly Language; IE = initiating event.
[a]Items not included in reliability and validity calculations.
*$p < .05$. **$p < .01$. ***$p < .001$.

by conducting a CFA that assessed the extent to which items within each subscale (i.e., Macrostructure or Microstructure) correlated, forming a construct or latent variable. The fit of the CFA was estimated by comparing the observed correlation structure with that obtained through model fitting. Model fitting involved determining how well the proposed theoretical model (narrative macrostructure and microstructure) captured the covariance between all the items in the model. If the correlations were low, the results of the CFA would indicate a poor fit, prompting the removal of items.

We conducted a full information CFA with a weighted least square parameter estimator (WLSMV) due to the presence of categorical data to assess the degree of fit between the item properties and the measurement model. Two latent variables (i.e., macrostructure and microstructure) were allowed to covary in this model. Latent variables were not directly observed, but were "inferred" from the variables that were directly observed (component items and subscales of the MISL). The following guidelines were used for identifying the characteristics of an "adequately fitting" CFA: composite reliability estimates ≥.70 for each latent variable (Fornell & Larcker, 1981; Hatcher, 1994, p. 339); a chi-square statistic to degrees of freedom (*df*) ratio ≤ 2 (Hatcher, 1994, p. 339); a comparative fit index (CFI) and a Tucker–Lewis index (TLI) ≥ .95 (Hu & Bentler, 1999); a root mean square error of approximation (RMSEA) ≤ .06 (Hu & Bentler, 1999); and a weighted root mean square residual (WRMR) ≤ .90 (Yu & Muthén, 2002). After removing items related to grammaticality and tense from the Microstructure subscale, the CFA measurement model consisting of two latent factors (Macrostructure and Microstructure subscales) demonstrated an overall model fit with $\chi^2(df = 53) = 81.27$, $p = .008$, $\chi^2/df$ ratio = 1.53; CFI = .99; TLI = .98; RMSEA = .06; and the average WRMR = .82.

Estimates of variance accounted for by each item (from the latent variable) in the form of $R^2$ (variance explained by the model), their standard errors, and $p$ values are presented in Table 5. A $p < .05$ was judged to be significant. As shown in the table, items measuring setting, IE, attempt, consequence, coordinating conjunctions, and metacognitive/metalinguistic verbs were highly significant at $p = .01$; and, items measuring character, subordinating conjunctions, and elaborated noun phrases were moderately significant at $p \leq .043$. Items that were not significant included internal response ($p = .336$) and plan ($p = .05$). Nonsignificance for internal response and plan reflects floor effects for these elements because the children who participated in this study rarely included them in their stories. Aside from a slightly larger RMSEA, results of the CFA measurement model indicated adequate model fit to support the construct validity of the MISL instrument.

## Discussion

The purpose of this study was to assess the inter-rater reliability, internal consistency, and construct validity of the MISL. Our first question was, to what extent do two raters who score narratives independently agree on the values that are assigned to the MISL items (inter-rater reliability)? Our second question was, to what extent are the items on the MISL internally reliable as measured using Cronbach's alpha (≥.70; internal consistency)? Our final question was, are there two multiple dimensions (macrostructure and microstructure) underlying the items on the MISL (construct validity)?

### *Inter-Rater Reliability*

Recall that inter-rater reliability is the extent to which two raters agree on how to score individual items. This construct is important for a progress-monitoring tool because a determination of progress can only be trustworthy to the extent that another professional would have obtained the same scores. We found relatively high levels of inter-rater reliability (90%–100%) across all of the items on the MISL rubric. One potential reason for this high degree of inter-rater reliability was the rigorous training and support the coders received as they learned to use the rubric. Recall that coders were asked to independently score four or five stories on their own and then turn them in with any questions they had to the first author. The coders reported that as they became more familiar with the rubric, their independent scoring time decreased by as much as 50% depending on the length or complexity of the narrative. Rapid and accurate scoring was related to the amount of experience the coders had with the rubric, meetings among coders to discuss their scores (5–10 min per meeting), and group discussion of discrepancies.

School-based SLPs may not have the luxury of meeting with other trained professionals after they score every five to 10 narratives to refine and calibrate their skills. Also, they will not be able to meet regularly with research staff to obtain final judgments on scoring discrepancies. Therefore, inter-rater reliability estimates among clinicians in authentic contexts may be somewhat lower than those reported here, at least initially. It is expected, however, that clinicians will increase their scoring proficiency and their scoring reliability as they become more familiar with the tool and how to use it to monitor narrative proficiency with their own students.

One additional consideration when using the MISL rubric in authentic, school-based settings is whether or not to orthographically transcribe narratives before attempting to score them. Recall that the stories in this study were orthographically transcribed before they were scored. School-based practitioners may not have the time and resources necessary to use this process to score every narrative obtained from students on their caseload. One way to reduce the amount of transcription that may be necessary for reliable scoring is to digitally record stories told by students and then take abbreviated notes while replaying them. These notes may be used during the scoring process. The use of audio-recordings to score narratives has been shown to have adequate inter-rater reliability using procedures outlined in the manual for the TNL. However, more research is necessary to determine whether the MISL may be scored reliably using a similar method.

The most important way to achieve sufficient inter-rater reliability using the MISL rubric is to adhere to the operational definitions of the items included in the measure. The definitions contained in this article, and the examples provided in the supplemental materials should assist clinicians in achieving sufficient inter-rater reliability to use the rubric for the purpose of progress-monitoring in school-based settings.

### *Internal Consistency Reliability*

Internal consistency represents the homogeneity of the items that have been selected to measure a particular construct. The MISL rubric was intended to measure narrative proficiency (the construct of interest). Toward that end, the items that were included on the rubric were selected because they have been shown to contribute to narrative skill. Initially, it was thought that grammar and tense may be important items to include in the measurement of narrative proficiency, however, the analysis suggested otherwise. The overall internal consistency reliability of the MISL was sufficient (Cronbach's α = .79) only after the removal of the two items related grammatical acceptability and tense change. The data in this study suggest that grammar and tense, although important linguistic skills, may not be critical contributors to overall narrative competence.

Recall that the internal consistency of the Macrostructure and Microstructure subscales was minimally acceptable when measured independently, particularly the Microstructure subscale (α = .67). However, the total MISL score, with a Cronbach's alpha of .79, may be a more meaningful measurement of narrative proficiency than either scale used in isolation. For statistical reasons, we recommend that clinicians base global decisions about intervention progress on the total score as a reliable indicator of change rather than the macrostructure or microstructure scores separately. That is not to suggest that clinicians should not utilize each of the individual subscale scores to monitor mastery of each of these important skills. We have used individual scores to make decisions about specific targets for intervention sessions and feel that this is a useful tool for planning.

### *Construct Validity*

In combination, the nature of the relationships among item scores on the MISL was consistent with the theory that narratives are comprised of macrostructure and microstructure components. The macrostructure items included on the MISL that are consistent with theory were setting, IE, internal response, plan, attempt, and consequence (Stein & Glenn, 1979). The MISL also included an additional element, character, because many narrative intervention programs often include instruction on this component. Character was shown to load or be consistent with the other macrostructure items on the MISL.

The microstructure elements that were included on the MISL were coordinating and subordinating conjunctions, adverbs, elaborated noun phrases, metacognitive and metalinguistic verbs. The items related to grammaticality and tense were removed from the rubric because model fit statistics indicated that these items were inconsistent with the other microstructure items on the scale. For the type of coding that was used for the MISL, grammatical accuracy and consistency of tense did not correlate well with other aspects of macrostructure or microstructure. Clinicians who work on these aspects of language during intervention would want to use a means other than the MISL to monitor children's progress in these domains.

The data presented in this article suggest that a progress-monitoring tool designed to measure narrative proficiency may not be improved by adding measures of grammaticality or tense change. Our findings of lower internal consistency when items measuring grammar and tense were included were very important findings relevant to clinical practice. It is possible that grammaticality and tense are "distinct skills" that are separate from macrostructure and microstructure. If so, including grammar and tense in a narrative rubric may indicate a lack of progress in narrative skills when in fact progress is being made. If clinicians are targeting grammaticality in therapy, it may be important to acknowledge that fluctuations between grammaticality and narrative discourse proficiency may occur as students focus on learning difficult narrative discourse skills, although again, we do not provide data in this study to support this assertion. Tentative data suggest that grammaticality will improve after knowledge of narrative content has become more stable (Crotty & Gillam, 2015a, 2015b). Future research is needed to provide solid evidence for this hypothesis. What is important to note is that we are not saying to clinicians they should not work on tense and grammar. We are suggesting that these items may make a tool for measuring narrative proficiency less reliable in measuring macrostructure and microstructure relative to narrative production.

## Limitations

The MISL was designed to measure changes in a very specific set of macrostructure and microstructure features that have been documented to contribute to narrative proficiency and that are aligned with Common Core Curricular Standards (S. Gillam et al., 2014). If a clinician is not teaching these aspects of narrative macrostructure and/or microstructure in their narrative instruction, the MISL may not be as useful in documenting progress. In addition, the pilot studies we have conducted with versions of this progress-monitoring tool have included fairly small numbers of participants. Therefore, findings related to reliability and validity using larger samples could yield different results from those reported here. Finally, we calculated the psychometric properties based on only one elicitation context (spontaneous generation) using a specific prompt from the TNL. It is possible that findings may differ using different elicitation contexts (retelling) and prompts (sequenced pictures, story books). Future research may investigate the ways in which the MISL might be modified for use with various other elicitation contexts including sequenced scene pictures and retells.

## Summary and Clinical Implications

The purpose of progress-monitoring tools such as the MISL is to provide clinicians with information that can inform clinical decisions about the nature of narrative intervention needed to support children's ability to meet the language demands of the classroom curriculum. Valid and reliable outcome measures are crucial for progress-monitoring tools to be useful in driving systematic, data-based decisions about language instruction. Progress-monitoring for narrative discourse poses a unique challenge to researchers, educators, and clinicians. This is because measuring narrative proficiency requires tracking multiple sources of macrostructure and microstructure information in increasingly more demanding contexts (Petersen et al., 2008).

The data collected in this study suggest that the unified construct score (total MISL score) is the most valid measure for assessing narrative discourse progress using the MISL rubric. Neither the Macrostructure nor the Microstructure subscale on its own was sufficient to reflect the complexity of narrative discourse proficiency. These assumptions were drawn from the psychometric data reporting lower internal consistency scores for each of these scales when evaluated independently. This is not to say that the individual subscales (Macrostructure, Microstructure) are not informative for intervention planning. For example, clinicians may use data from the subscales to note macrostructure and microstructure features that are consistently absent from students' stories and target them explicitly during future sessions. When evaluating progress in response to narrative instruction, the total MISL score is the most well supported of the three scores that may be obtained using the rubric.

## Acknowledgments

The authors would like to thank Allison Hancock, Natalie Nelson, Julise Nelson, Sara Hegsted, Sara Hicken, Katie Squires, Shannon Davenport, and all of the undergraduate and graduate research assistants who administered tests and analyzed language samples. The authors also convey a special thank you to their former doctoral student Doug Petersen who contributed to an earlier version of this instrument.

## Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The Monitoring Indicators of Scholarly Language (MISL) is part of the Supporting Knowledge in Language and Literacy (SKILL) narrative intervention program. S.L.G. and R.B.G. receive royalties from the SKILL program. R.B.G. is also the author of the *Test of Narrative Language*, which was used in the study.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by a grant from the Institute for Educational Sciences, National Center for Special Education Research (Award Number R324A100063). The Supporting Knowledge in Language and Literacy (SKILL) program can be ordered at https://usuworks.usu.edu and includes progress-monitoring tools and video examples to support implementation.

## Supplemental Material

The online supplements are available at http://journals.sagepub.com/doi/suppl/10.1177/1525740116651442.

## References

American Institutes for Research. (2015a). *The essential components of RTI*. Available from www.rti4success.org

American Institutes for Research. (2015b). *Technical review committees process*. Retrieved from http://www.rti4success.org/technical-review-committees-process

American Speech-Language-Hearing Association. (2000). *Guidelines for the roles and responsibilities of the school-based speech-language pathologist*. Retrieved from www.asha.org/policy

American Speech-Language-Hearing Association. (2001). *Roles and responsibilities of speech-language pathologists with respect to reading and writing in children and adolescents* [Position statement]. Retrieved from www.asha.org/policy

Benson, S. (2009). Understanding literate language: Developmental and clinical issues. *Contemporary Issues in Communication Science and Disorders*, *36*, 174–178.

Berman, R. (1988). On the ability to relate events in narrative. *Discourse Processes*, *11*, 469–497.

Bracken, B. A., & McCallum, R. S. (1998). *The Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside.

Briesch, R., & Volpe, R. (2007). Selecting progress monitoring tools for evaluating social behavior. *School Psychology Forum*, *1*, 59-74.

Carrow-Woolfolk, E. (1999). *Comprehensive Assessment of Spoken Language*. Circle Pines, MN: American Guidance service.

Crotty, B., & Gillam, S. (2015a, June 3). *Content and form in the narratives of children with Autism Spectrum Disorder in two elicitation contexts: Implications for Assessment and Instruction*. Poster presentation presented to the Symposium on Child Language Disorders, Madison, WI.

Crotty, B., & Gillam, S. (2015b, November). *Content and form in the narratives of children with Autism Spectrum Disorder in two elicitation contexts: Implications for Assessment and Instruction*. Technical session presented to the American Speech Language and Hearing Convention, Denver, CO.

Ehren, B., & Whitmire, K. (2009). Speech-language pathologists as primary contributors to response to intervention at the secondary level. *Seminars in Speech and Language*, *30*, 90–104.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*, 39–50.

Gillam, R. B., & Pearson, N. (2004). *Test of narrative language*. Austin, TX: Pro-Ed.

Gillam, S., & Gillam, R. (2006). Making evidence-based decisions about child language intervention in schools. *Language, Speech, and Hearing Services in Schools*, *37*, 1–12.

Gillam, S., Gillam, R., & Laing, C. (2012). *Supporting knowledge in language and literacy (SKILL): A narrative intervention program*. Logan: Utah State University.

Gillam, S., & Justice, L. (2010). Progress monitoring tools for SLPs in response to intervention (RTI): Primary grades. *The ASHA Leader, 9/21 Issue: Feature*. Retrieved from http://leader.pubs.asha.org/article.aspx?articleid=2291687

Gillam, S., Olszewski, A., Fargo, J., & Gillam, R. (2014). Classroom-based narrative and vocabulary instruction: Results of an early-stage, nonrandomized comparison study. *Language, Speech, and Hearing Services in Schools*, *45*, 204–219.

Greenhalgh, K., & Strong, C. (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*, 114–125.

Hatcher, L. (1994). *A step-by-step approach to using the SAS® system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute.

Heilmann, J., Miller, J., Nockerts, A., & Dunaway, C. (2010). Properties of the narrative scoring scheme using narrative retells in young school-age children. *American Journal of Speech-Language Pathology*, *19*, 154–166.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Justice, L. M. (Ed.). (2006). *Clinical approaches to emergent literacy intervention*. San Diego, CA: Plural Publishing.

Justice, L. M., Bowles, R., Eisenberg, S. L., Kaderavek, J. N., Ukrainetz, T. A., & Gillam, R. B. (2006). The index of narrative micro-structure (INMIS): A clinical tool for analyzing school-aged children's narrative performance. *American Journal of Speech-Language Pathology*, *15*, 177–191.

Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman Brief Intelligence Test*. Circle Pines, MN: American Guidance Service.

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

Loban, W. (1976). *Language development: Kindergarten through grade twelve*. Urbana, IL: National Council of Teachers of English.

Lueger, R., & Barkham, M. (2010). Using benchmarks and benchmarking to improve quality of practice and service. In M. Barkham, G. E. Hardy, & J. Mellor-Clark (Eds.), *Practice-based evidence: A guide for psychological therapies* (pp. 223-256). Hoboken, NJ: Wiley-Blackwell.

Miller, J. F., Andriacchi, K., & Nockerts, A. (2011). *Assessing language production using SALT Software: A clinician's guide to language sample analysis*. Middleton, WI: SALT Software.

Miller, J. F., & Chapman, R. S. (2004). Systematic Analysis of Language Transcripts (SALT, v8*.0)* [Computer software and manual]. Madison: Language Analysis Laboratory, Waisman Center, University of Wisconsin–Madison.

National Governors Association and Council of Chief State School Officers. (2011, October 27). *Common Core State Standards initiative*. Available from www.corestandards.org

Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development-Primary: Third edition*. Austin, TX: Pro-Ed.

Nippold, M. (1998). Later language development: The school-age and adolescent years (2nd ed.). Austin, TX: Pro-Ed.

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Paul, D., & Hasselkus, A. (2004). *Clinical record-keeping in speech-language pathology for healthcare and third-party payers*. Rockville, MD: American Speech-Language-Hearing Association.

Paul, R. (1995). *Language disorders from infancy through adolescence: Assessment and intervention*. St. Louis, MO: Mosby.

Paul, R. (2007). *Language disorders from infancy through adolescence: Assessment and intervention* (3rd ed.). St Louis, MO: Mosby.

Petersen, D., Gillam, S., & Gillam, R. (2008). Emerging procedures in narrative assessment: The index of narrative complexity. *Topics in Language Disorders*, *28*, 111–126.

Semel, E., Wiig, E., & Secord, W. (2004). *Clinical evaluation of language fundamentals (CELF-4)*. Austin, TX: Pro-Ed.

Squires, K. E., Lugo-Neris, M., Pena, E., Bedore, L., Bohman, T., & Gillam, R. (2014). Story retelling by bilingual children with language impairments and typically developing controls. *International Journal of Communication Disorders*, *49*, 60–74.

Stein, N. L. (1988). The development of children's storytelling skill. In M. B. Franklin & S. S. Barten (Eds.), *Child language: A reader* (pp. 282–297). New York, NY: Oxford University Press.

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

Sutherland Cornett, B. (2006, September 5). Clinical documentation in speech-language pathology: Essential information for successful practice. *The ASHA Leader*. Retrieved from http://leader.pubs.asha.org/article.aspx?articleid=2278198

Tomblin, B., Records, N., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, E. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, *40*, 1245–1260.

van den Broek, P., Linzie, B., Fletcher, C., & Marsolek, C. (2000). The role of causal discourse structure in narrative writing. *Memory & Cognition*, *28*, 711–721.

Warren, S., Fey, M., & Yoder, P. (2007). Differential treatment intensity research: A missing link to creating optimally effective communication intervention. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*, 70–77.

Westby, C. E. (1985). Learning to talk—Talking to learn: Oral literate language differences. In C. S. Simon (Ed.), *Communication skills and classroom success: Therapy methodologies for language-learning disabled students* (pp. 181–213). San Diego, CA: College-Hill.

Westby, C. E. (2005). Assessing and facilitating text comprehension problems. In H. Catts & A. Kamhi (Eds.), *Language and reading disabilities* (pp. 157–232). Boston, MA: Allyn & Bacon.

White, M., van den Broek, P., & Kendeou, P. (2007, April). *Comprehension and basic language skills predict future reading ability: A cross-sectional study of young children*. Symposium paper presentation at the Society for Research on Child Developments Biennial Conference, Boston, MA.

Yu, C.-Y., & Muthén, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.