An Examination of the Quality of Narratives Produced by Children With Language Disorders

Teresa Ukrainetz McFadden, and Ronald B. Gillam

Lang Speech Hear Serv Sch 1996;27;48-56

An Examination of the Quality of Narratives Produced by Children With Language Disorders

Teresa Ukrainetz McFadden
Ronald B. Gillam
University of Texas at Austin

ABSTRACT: A team of regular and special educators used a holistic scoring procedure to rate the overall quality of spoken and written narratives produced by students with language disorders and their age-, language-, and reading-matched peers. Students with language disorders earned significantly lower holistic scores than their age-matched peers. However, their holistic scores were similar to the scores earned by their language- and reading-matched peers. Correlations between holistic scores and structural measures of language revealed that quality judgments were moderately related to textual-level measures of form and content but were unrelated to sentence-level measures of form and content. Holistic scoring is shown to have clinical and research utility as a means for socially validating the effects of language disorders on storytelling. Clinicians who want to influence the overall quality of their students' stories may wish to focus their intervention on textual-level narrative features.

KEY WORDS: narrative, language disorder, discourse, assessment, holistic scoring

Researchers examining the narratives of students with language disorders have focused primarily on structural aspects of complexity. In comparison to narratives produced by their age peers, narratives produced by students with language disorders have been shown to be less complex, with deficits in sentence grammar, story grammar, and cohesion (Gillam, 1989; Gillam & Johnston, 1992; Graybeal, 1981; Liles, 1985, 1987; Merritt & Liles, 1987). It has been assumed, but not specifically demonstrated, that the structural deficits in these students' stories are apparent to listeners and readers and negatively influence audience judgments of quality.

Narrative quality derives from more than the appropriate organization of cohesively linked and grammatically well-formed sequences of propositions. Narratives are produced within specific contexts for present or implied audiences (Bakhtin, 1986), and they present the author's perspective of the events that are being recounted (Labov, 1972; Polanyi, 1989). Narrators achieve suspense, mystery, curiosity, and emotional involvement through skillful manipulation of language and textual organization (Brewer, 1985). Structural analyses do not directly capture aspects of narrative composition such as charm, interest, clarity, vigor, honesty, appropriateness, freshness, subtlety, or depth.

Holistic scoring is an approach to narrative analysis that takes into consideration the sum of quantifiable elements of story such as grammar, vocabulary, and episodic organization, as well as less quantifiable elements like charm, interest, and clarity. Holistic scorers think about stories as wholes and rank them in accordance with hierarchically ordered descriptions. Holistic scoring is frequently used in the evaluation of writing composition, producing more reliable results than individual instructor grading (Diederich, 1974). Methods of holistic scoring vary across authors (e.g., Diederich; Kirby & Liner, 1981; Myers, 1981), but the essential aspects involve grouping a collection of stories into quality-based categories, selecting stories that best exemplify the categories (called anchors), collaborating on descriptions of the commonalties among anchors (called rubrics), and assigning a single score to each of the remaining stories in the collection based on a comparison with the anchors and rubrics.

Holistic scoring can be an important adjunct to analytic measures of performance. For example, Daiute (Daiute, 1986, 1989; Daiute & Dalton, 1988) has used holistic scoring as well as analytic measures such as story length, number of story elements, use of details, and presence of dialogue to study the effects of collaboration and computers on narrative composition. Daiute found that overall quality increased when students wrote collaboratively rather than individually. Scardamalia and Bereiter (1985) and Woodruff, Bereiter, and Scardamalia (1981) evaluated the efficacy of various writing guidance techniques on opinion essay writing. These authors used holistic scoring in addition to more focused measures that indexed reflective thought and quality of revisions. Their results indicated that writing instruction positively influenced students' use of planning and reflection, but it did not influence the overall quality of their compositions.

Holistic scoring also has been used to compare narratives produced by two groups of children with language disorders who received different types of instructional programs. Gillam, McFadden, and van Kleeck (1995) compared the effects of whole language and language skills curricula on the narrative performance of 9- to 12-year-old children with language disorders. On three measures of language content (number of ideas per T-unit, number of problems that were resolved, and percentage of episodes that were embedded within other episodes), values for students in the whole language group were higher than values for students in the language skills group for spoken stories. On three measures of language form (number of morphemes per T-unit, percent of grammatically acceptable T-units, and percent of temporal, causal, and conditional relationships that were marked by connectives), values for students in the language skills group were higher than values for students in the whole language group for both spoken and written stories.

These same stories were analyzed according to a holistic scoring procedure. Fifty percent of the whole language group's spoken narratives and 25% of their written narratives were judged to be "good" or "strong" stories (ratings of 3 or 4 out of 4). Stories that were produced by students in the language skills group did not fare as well; none of their spoken stories and only 12.5% of their written stories received "good" or "strong" ratings. Stories produced by students in the whole language group were judged to be superior in terms of overall quality even though they were inferior in terms of structural analyses of language form.

The use of both structural analyses and holistic scoring enables investigators to measure the effects of language interventions on the overall quality and structural complexity of narratives. However, little is known concerning how analytic measures of complexity relate to judgments of discourse quality. This is a significant issue given that the findings of Scardamalia and Bereiter (1985) and Gillam, McFadden, and van Kleeck (1995) suggest a degree of dissociation between language complexity and overall quality of stories.

Two studies have specifically investigated those aspects of language that affect judgments of overall quality. Nold and Freedman (1977) correlated holistic scoring results with the outcomes of 17 analytic measures such as words per T-unit, subordinate clauses per T-unit, number of possessive nouns and pronouns, number of modals, number of adverbs of time, and number of gerunds. Results indicated that sophistication in modification, especially in the use of final modifiers (e.g., "He wrote creatively"), was positively correlated with judgments of overall quality. The use of "be," "have," and modals as auxiliaries were negatively related to judgments of overall quality. Nold and Freedman also found that shorter stories tended to receive lower holistic scores than longer stories.

Stein and Policastro (1984) studied the relationship between episode structure and narrative quality. These authors constructed texts that were missing various episodic elements. Teachers and second-grade students were asked to rate the texts on a 7-point continuum that ranged from "not a story" to "a story" to "a good story." Results showed a strong positive relationship between the number of episodic elements that were included in a text and the goodness rankings of both groups of raters.

Nold and Freedman's (1977) study provides some information on structural predictors of the quality of expository essays. However, the factors involved in judgments concerning freshman essays may be different from the factors that affect the quality of children's imaginative narratives. In addition, Nold and Freeman measured only sentential complexity. It would seem likely that judgments of quality could also be affected by factors such as textual organization and episodic complexity. Stein and Policastro's (1984) study implicates episodic complexity in decisions concerning narrative quality. However, their contrived, episodically-controlled stories were far removed from authentic texts that contain a host of other factors (like charm, clarity, and humor) that influence whether stories are appealing or unappealing to an audience.

In this study, we examined the overall quality of spoken and written narratives produced by students with language disorders and their age-, language-, and reading-matched controls. In addition, we examined the relationship between judgments of holistic quality and analytic measures of sentential and textual complexity in the domains of language form and content. The specific research questions were:

- Does the overall quality of spoken and written narratives produced by students with language disorders differ from that of their age-matched, language-matched, and reading-matched peers?
- What is the relationship between holistic judgments of quality and structural analyses of form and content at sentential and textual levels of discourse?

METHOD

Participants

The subjects in this study have been previously described in Gillam (1989) and Gillam and Johnston (1992). Forty school-age students participated, including 10 students with language disorders between the ages of 9:0 and 11:7 (years:months), with a mean age of 10:7. Each of these students was matched with three same-sex students with no history of speech, language, or hearing disorders according to age, spoken language ability, and reading ability. The age-matched group ranged in age from 9:1 to 11:7, with a mean age of 10:7. The spoken language- and reading-matched groups ranged in age from 7:7 to 8:9 (mean of 7:11) and 6:8 to 8:9 (mean of 7:9), respectively. There were seven boys and three girls in each group. All the students had normal hearing and vision and came from monolingual English-speaking homes. Profile information for the four groups is presented in Table 1.

The language disordered (LD) group. Students in the LD group had average or above average nonverbal intelligence together with significant deficits in spoken language and reading. These students had been diagnosed as learning disabled by school district special education personnel according to Wyoming state guidelines. These students' nonverbal cognitive abilities, as indicated by their performance on the Test of Nonverbal Intelligence (TONI) (Brown, Sherbenou, & Johnsen, 1982) were well within normal limits (TONI quotient x = 103.6; range = 93–129). Their verbal abilities were well below age expectations, as indicated by a mean verbal cluster quotient of 75.4 (range = 72–80) on the verbal aptitude composite of the Detroit Test of Learning Aptitude–2 (DTLA–2) (Hammill, 1985). Discrepancies between verbal (DTLA–2) and nonverbal (TONI) quotients ranged between 16 and 52 points, with a mean of 28.2 points. Reading performance, as indicated by age percentile scores on the Reading Recognition subtest of the Peabody Individual Achievement Test (PIAT, Dunn & Markwardt, 1970) ranged from the 2nd to the 12th percentile, with a mean percentile value of 6.3.

The age-matched (AGE-M) group. The AGE-M group consisted of students whose birth dates fell within ±3 months from the birth date of a member of the LD group. All students in the AGE-M group were estimated to be functioning within the middle two quartiles of their class by their teachers. These students were given the reading recognition subtest of the PIAT, the TONI, and the Sentence Imitation subtest of the DTLA–2 for descriptive purposes.

The language-matched (LANG-M) group. A language age was computed for each child with language disorders by multiplying the verbal quotient obtained on the DTLA–2 and the child's chronological age (in months). This product was converted to an age:months value. Each language-matched child met three criteria: their chronological age was within ±3 months of the language age for a child in the LD group, they obtained a Sentence Imitation score on the DTLA–2 that was within ±2 points of the score obtained by their match in the LD group, and their teachers indicated that they functioned within the middle two quartiles of their class. After selection, language-matched controls were given the PIAT Reading Recognition subtest and the TONI for descriptive purposes.

Table 1. Mean ages and test scores for participants in the four matched groups.

	Scores					
Groups	*CA*	*G*	*LA*	*TONI Q*	*MA*	*PIAT R*
LD	10:7	4.0	7:10	103.6	10:9	26.6
AGE-M	10:7	4.4	11:5*	99.5	10:7	49.8
LANG-M	7:11	2.1	8:2*	101.4	8:2	37.7
READ-M	7:9	1.7	7·9*	99.0	7:6	26.7

Note. CA = chronological age; G = grade in school, LA = language age from five verbal subtests on the DTLA–2 (*language age computed from four verbal subtests on the DTLA–2); TONI Q = Test of Nonverbal Intelligence quotient; MA = mental age based on TONI scores; PIAT R= raw score on the Reading Recognition subtest of the Peabody Individual Achievement Test.

The reading-matched (READ-M) group. Reading-level controls were selected on the basis of raw scores on the Reading Recognition subtest of the PIAT that were within ±2 points of the score of a student in the LD group. After selection, reading-matched controls were given the DTLA–2 Sentence Imitation subtest and the TONI for descriptive purposes. Chronological ages and language ages were not used in the matching equation for students in the READ-M in order to permit freedom for variation among the three groups on the matching variables. Similarity between the READ-M and LANG-M groups on the Sentence Imitation task was incidental to the matching design of the study, and probably occurred as a function of the relationship between language and reading abilities.

Procedures

Each of the 40 participants produced two spoken stories (n = 80) and two written stories (n = 80) based on picture stimuli. Participants were shown sets of three 7" x 10" color pictures. Each set of three pictures contained a nature picture, an outdoor action picture, and a portrait. Picture sets were changed for each narrative in order to prevent participants from using the same picture cue more than once. Students were asked to select a picture from the set.

In the spoken narrative condition, students were given the following oral instructions:

> I want you to take a couple of minutes to make up a story that has to do with the picture. Try to make your story as long as you can. Make sure your story has a beginning, some things that happen, and an ending. Put things in your story like where it takes place, the names of the people or animals in it, what happens, and why it happens. Now, take some time to think of a story.

In the writing condition, students received the same instructions, but were also told, "this time I want you to write your story. I can't help you with the writing. Just write your story so that you can read it back to me." This procedure elicited self-generated stories in a manner that enabled the students to have a degree of ownership of their story topics. This story elicitation context was also very familiar to these students because it was quite similar to the story creation tasks they routinely encountered in their elementary school classrooms.

Spoken stories were tape recorded. After writing their stories, students were asked to read their stories aloud. These readings were tape-recorded for use during data analysis in the event that a child's handwriting or spelling made it difficult to determine what had been written.

Two presentation sequences were used to control for possible task order effects: spoken narrative before written in the first session and written before spoken in the second, or the reverse order. Each student in the LD group and his or her three matches received the same sequence, with sequences counterbalanced across the four groups.

Spoken and written language samples were segmented into T-units (an independent clause plus its associated dependent clauses) (Hunt, 1970) and were transcribed according to Systematic Analysis of Language Transcripts (SALT) (Miller & Chapman, 1984) conventions. Spelling errors and mazes (false starts, repetitions, and reformulations) were removed from the transcripts in order to prevent obvious identification of the stories as either spoken or written.

Language Analysis

The holistic scoring procedure described in Myers (1981) was used to assess the overall quality of the narratives. After reading a random selection of the narratives (25 of the 160), the first author, as group leader, created six quality categories with associated descriptions for each category. She then selected two to three stories (anchors) that she considered to be representative of the tentative categories. A team of scorers consisting of one speech-language pathologist, two primary grade teachers, and one college English instructor independently rated the tentative anchors without knowledge of the group leader's rankings.

Narratives that the group of scorers had rated differently from the group leader were discussed and re-rated. In choosing rankings that would minimize the possibility of disagreement, the scorers reached consensus on four quality categories and accompanying rubrics and anchors for each category. Category one stories (weak) consisted of descriptions and poorly organized, uncaptivating stories. Category two stories (adequate) consisted of stories that took one of four forms: (1) an event recount, without a central climax; (2) a bare-bones narrative, with no elaboration; (3) a narrative without an ending; or (4) a confusing narrative with strong descriptive segments. Category three stories (Good) were captivating stories that contained problems and resolutions. These narratives may have had some organizational difficulties. Category four stories (strong) were easily understood narratives with a clear, integrated story line, elaboration, interesting word choices, and some captivating features such as a climax, an ending twist, or a compelling personal voice. Examples of the anchor stories representing each of the four categories are presented in the Appendix.

Narratives that had not been selected to be anchors were randomly assigned to two members of the scoring team who independently rated them according to the four quality categories. The scorers were blind to the group membership of the child who produced each narrative. Following Myers procedure (1981), narratives that differed by one point (e.g., received a "1" by one scorer and a "2" by another) were considered to fall between categories (better than "1" but not quite "2") and were not considered to be disagreements. Narrative scores that differed by more than one point (e.g., received a "1" by one scorer and a "3" by another) were considered to represent true disagreements concerning quality. On this basis, 5 of 160, or 3%, of the corpus were rescored by the group leader, who adjusted one of the discrepant scores. Narratives that had been used as anchors were recorded as having received two scores in the category they exemplified (e.g., an anchor story for the weak quality category received two "1s" as scores).

Gillam (1989) and Gillam and Johnston (1992) had evaluated the linguistic complexity of a subset of the stories that were used in the holistic analysis. They defined complexity in two ways: by amount and by nature of organization. Their analysis system included measures of language form and language content at both sentential and textual levels of discourse.

The measures of language form were:

- *sentence level amount*—morphemes per T-unit (MLT-u), e.g., "I wanted to go/" (5 morphemes).
- *text level amount*—number of T-units per story (T-units), e.g., "I saw a dog/ and I ran after him/ and I grabbed him and hugged him." (3 T-units).
- *sentence level organization*—percent of grammatically acceptable complex T-units (%complex), e.g., "He went down at store because him mom's there." (a grammatically unacceptable, complex T-unit).
- *text level organization*—number of connectives (e.g., because, so, then) per T-unit (connectives), e.g., "John called after I left/ but it didn't matter because I knew he wasn't home/" (1.5 connectives per T-unit).

The measures of content were:

- *sentence level amount*—number of idea units (predicates plus affiliated arguments) per T-unit (propositions). This measure concerns semantically defined units of activities, relations, and states. In the utterance, "They are sawing the trees," "sawing" is the predicate, "they" is an agent argument, and "trees" is a patient argument (1 proposition per T-unit).
- *text level amount*—number of plot units or constituents per story (constituents). This is a more fine-grained approach than the familiar story grammar analysis, and was developed by Sutton-Smith, Botvin, & Mahony (1976). Plot units are thematic elements such as participants, location, villainy, lack, departure, plan, and attack. The story, "Jan was afraid of the mean bear/ she ran to her house/ then the bear went away /and Jan wasn't afraid anymore/" contains 7 constituents (underlined).
- *sentence level organization*—number of predicate types per T-unit (predicate types). Utterances consist of the main (or nuclear) predicate plus adverbial, embedded, and associated predicates. The utterance, "Yesterday she taught the little boy to put coins in his piggy bank/" contains all four predicate types. "Teach" is the nuclear predicate, "put in" is an embedded predicate, "yesterday" is an adverbial predicate, and "little" is an associated predicate (4 predicate types per T-unit).
- *text level organization*—percent of constituents that are expressed as problem-resolution pairs (dyads). In the preceding bear story, 28.6% of the constituents (two of seven) are involved in the problem-resolution dyad, "afraid/wasn't afraid."

RESULTS

The first question of interest concerned potential group differences in the overall quality of spoken and written narratives. Recall that each of the 20 spoken and 20 written stories produced by each group of students received two holistic quality scores (a total of 80 scores per group). Because holistic scores are ordinal, nonparametric statistics were used to test the questions of interest.

Picture selection could have influenced the type of story that was generated which, in turn, could have influenced listener judgments of quality. For example, children might have created different kinds of stories concerning portraits than they created concerning outdoor action pictures, and listeners might have preferred action-based stories over portrait-based stories. Across the 160 stories, students chose nature pictures (49% spoken, 39% written) and outdoor-action pictures (35% spoken, 50% written) more often than portraits (16% spoken, 11% written). The distribution of picture selection choices was remarkably similar across the four groups of students. Kruskal-Wallis one way analysis of variance by ranks tests (Feldman & Gagnon, 1986) were computed in order to determine whether stories generated from the three types of pictures differed with respect to their holistic scores. There were no significant differences between the scores assigned to stories generated from the three types of picture stimuli for either spoken (H = 2.511, df = 2, p = .285) or written (H = .576, df = 2, p = .7497) modalities. If different kinds of stories were generated from the three types of picture cues, these differences did not significantly affect judgments of overall quality.

To assess potential modality differences, χ^2 analyses were computed on the distributions of spoken versus written holistic scores. Table 2 presents each group's spoken and written scores for each of the four categories. A comparison of spoken and written distributions across groups did not reach significance (χ^2 = 1.901, df = 3, p = .5919), nor did separate comparisons of spoken and written distributions within each of the four groups. For further analyses, holistic scores were collapsed across modalities to produce a distribution of category scores for each group. A χ^2 analysis yielded a significant group difference for the distribution of quality scores (χ^2 = 81.971, df = 9, p <.0001). Follow-up analyses revealed a significant difference between the LD group and the AGE-M group (χ^2 = 48.911, df = 3, p < .0001) but not between the LD, LANG-M, and READ-M groups (χ^2 = 5.442, df = 6, p = .4884).

Examination of the distribution of scores in Table 2 reveals a relatively high concentration of scores in the weak and adequate categories for students in the LD group (83%), the READ-M group (74%), and the LANG-M group (81%). In contrast, students in the AGE-M group showed the greatest concentration of scores in the good and strong categories (71%). These findings indicate that students in the LD group, like those in the reading- and language-matched groups, produced spoken and written stories that were judged to be significantly lower in overall quality than the scores produced by students in the age-matched group.

Table 2. Number of scores in each quality category as a function of group membership and mode.

	Quality			
Group	*Weak*	*Adequate*	*Good*	*Strong*
Spoken				
LD	13	18	7	2
AGE-M	1	10	21	8
READ-M	6	22	10	2
LANG-M	17	18	5	0
TOTAL	37	68	43	12
Written				
LD	10	25	5	0
AGE-M	4	8	18	10
READ-M	8	23	8	1
LANG-M	7	23	9	1
TOTAL	29	79	40	12

Note. Each of the 80 spoken and 80 written narratives received two scores (one from each scorer), for a total of 320 scores.

The second question of interest concerned potential relationships between holistic scores and previously reported measures of language form and content (Gillam, 1989; Gillam & Johnston, 1992). The previous χ^2 analyses indicated that the AGE-M group was drawn from a different population than the LD, LANG-M, and READ-M groups. Correlation coefficients are influenced by the way populations are established and by their sampling frequencies (Wickens, 1989). Because a coefficient based on data from two separate populations would not reflect any true population characteristic, data from the AGE-M group was excluded from the correlations. Holistic scores and data from structural analyses were not collapsed across modalities because Gillam and Gillam and Johnston found different patterns of structural complexity for spoken versus written narratives.

To be consistent with the earlier studies, only scores from students' "longest" spoken and written narratives (based on the number of story constituents) were used for this analysis. The two holistic scores given to each narrative were averaged to provide a single quality score that was correlated with each of the eight structural measures.

Spearman rho correlation results are presented in Table 3. The results show a consistent relationship between holistic scores and textual level measures of form and content. With the exception of the number of connectives in written stories, the textual measures of number of T-units per story, number of constituents per story, number of connectives per T-unit, and percent of dyadic constituents were moderately correlated with narrative quality in spoken and written modalities (ρ values between .42 and .60). None of the correlation coefficients for the sentence level measures (ranging from .01 to .21) reached significance. Thus, textual measures of language structure were related to judgments of overall quality, but sentential measures were not.

Table 3. Spearman rho correlations (corrected for ties) for holistic rankings and structural measures of language.

	ρ	
Structural category	*Spoken*	*Written*
Sentential level		
MLT-u	.064	.139
%Complex	.017	.127
Propositions	.078	.209
Predicate types	.091	.124
Textual level		
T-units	.483[a]	.488[a]
Connectives	.416[b]	.134
Constituents	.598[a]	.440[b]
Dyads	.463[a]	.474[a]

[a] p <.01 [b] p <.05

DISCUSSION

A holistic scoring protocol was used to assess the overall quality of spoken and written narratives produced by students with language disorders and three groups of matched controls. A high proportion of narratives produced by students in LD, READ-M, and LANG-M groups were judged to be in the weak or adequate categories. In contrast, a high proportion of the narratives produced by students in the AGE-M group were judged to be in the good or strong categories. These results are consistent with previously established findings of form and content deficits in narratives produced by students with language disorders (Gillam, 1989; Gillam & Johnston, 1992; Graybeal, 1981; Liles, 1985, 1987; Merritt & Liles, 1987).

Judgments of lower overall quality by a panel of educators who were blind to group membership demonstrate the clinical and ecological significance of the narrative difficulties of students with language disorders. Although definitions of language disorders are based, in part, on the social evaluation of a listener (e.g., Fey, 1986; Tomblin, 1983), and improvement in intervention should reasonably be expected to be perceptible to an informed audience, social validity assessments are uncommon in clinical practice (Campbell & Dollaghan, 1992). This study provides an example of one form of social validation: Students with language disorders produce narratives that are not only structurally less complex but, to an audience of educators, are noticeably lower in quality than those of their same age peers.

To examine the relationship between judgments of overall quality and structural measures of complexity, correlations were computed between holistic quality scores and previously obtained measures of textual and sentential complexity (Gillam, 1989; Gillam & Johnston, 1992). Judgments of the overall quality of narratives were related to textual-level measures of form and content but bore little relation to sentential-level measures. Longer stories with more story constituents (particularly problem-resolution units) were more likely to receive positive judgments of overall quality. However, the use of longer, syntactically complex utterances was not reliably associated with either positive or negative quality judgments. Although the number of significance tests performed in this study runs the risk of inflating the rate of positive findings (Stevens, 1992), the clear clustering of significant results within the textual measures of complexity contributes to our confidence in these findings.

A previous analysis of this corpus of narratives indicated that the group of students with language disorders evidenced lower performance than their age-matched peers on seven of eight measures of complexity (Gillam, 1989).[1] The results indicated that students with language disorders performed similarly to their younger reading- and language-matched peers. The current findings suggest that, despite their weaknesses at both sentential and textual levels of language, it was their performance on textual-level measures that was reliably associated with negative judgments of overall quality of the narratives produced by students with language disorders.

This finding is of some significance because intervention with school-age students has traditionally focused on sentence-level language. Further, debates concerning school-age language intervention often concern the relative emphasis on language form or content (Gillam, McFadden, & van Kleeck, 1995), not the relative emphasis on sentential or textual discourse levels. Certainly, listeners and readers consider both form and content when judging the quality of a story. However, the results of this study suggest that it is textual form and content rather than sentential form and content that may be the most salient features of overall quality. Clinicians who are interested in improving the quality of their students' stories might achieve more ecologically significant results by attending to such quantifiable textual-level elements of story as length and episodic organization, as well as such qualitative elements as charm, interest, subtlety, and clarity. Gillam (1995), Gillam, McFadden, and van Kleeck (1995), Norris and Hoffman (1993), and van Dongen and Westby (1986) have presented useful strategies for implementing textual approaches to narrative language intervention.

Our findings are applicable to imaginative narratives. Different relations between sentential and textual measures of language structure and overall judgments of quality may be obtained for other types of discourse such as expository essays or personal narratives. Further, there are numerous sentential features of language that were not measured by the structural scoring system reported herein. Different measures could have resulted in different relationships between sentential-level analyses and holistic judgements of quality. For this reason, we are not suggesting that clinicians cease to target sentential-level aspects of language in language intervention. However, our results suggest that

[1] A four-way repeated measures ANOVA yielded a significant main effect for group, which favored the age controls. Intercorrelations among the measures prevented independent statistical examination of each index. However, visual examination of the data revealed clear differences that favored the age-matched controls for seven of the eight structural measures (with the exception of number of connectives)

clinicians should pay careful attention to textual-level targets.

Holistic scoring is an infrequently used analysis tool in speech-language pathology. Its lack of use may be attributed in part to a lack of knowledge of methods used in writing evaluation. Additionally, the single case focus of the speech-language pathologist precludes immediate application of this group-based approach. However, the method is flexible and can be modified to fit individual situations (Gillam & McFadden, 1994). For example, a meeting among several speech-language pathologists or several educators within one school provides an opportunity to rate a corpus of narratives, to develop rubrics and anchors, and to establish inter-rater reliability. Speech-language pathologists can use these scoring standards to evaluate changes in the quality of spoken and written narratives produced by individual students.

Holistic scoring has potential as a reliable means of qualitative assessment (Gillam & McFadden, 1994). The scorer creates categories of quality that are specific to the corpus of texts being analyzed. This means that the scorer is not dependent on pre-selected standards and contexts of elicitation that may not be suited to the age, developmental level, or cultural background of the individuals being assessed. The applicability of holistic scoring to expository compositions, such as reports and opinion essays, may be of particular benefit because expository composition has less of a predictable structure than narration (Bereiter & Scardamalia, 1982) and few genre-specific ways of evaluating those products have been developed. By focusing on general impressions received from thinking about stories as wholes and attempting to collaboratively describe aspects of stories that are considered to be strong or weak, speech-language pathologists can work with other educators to arrive at reliable judgments of narrative quality and new insights about what makes a good story.

ACKNOWLEDGMENTS

The research reported in this paper was partially supported by a doctoral fellowship to the first author from the Social Sciences and Humanities Research Council of Canada and by a grant to the second author from the National Institutes of Deafness and Other Communication Disorders (1K08 DC00086-01).

The authors wish to thank David McFadden, Martha Garcia, Debbie Stahle, and Noel Stahle for serving on the holistic scoring team. Wayne Secord, Nancy Creaghead, Cheryl Scott, and one anonymous reviewer provided helpful comments concerning an earlier version of the manuscript.

REFERENCES

Bakhtin, M. M. (1986). *Speech genres and other late essays.* Austin, TX: University of Texas Press.

Bereiter, C., & Scardamalia, M. (1982). From conversation to composition: The role of instruction in a developmental process. In R. Glaser (Ed.), *Advances in instructional psychology, Vol. 2* (pp 1–64). Hillsdale, NJ: Lawrence Erlbaum.

Brewer, W. R. (1985). The story schema: Universal and culture-specific properties. In D.R. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language, and learning: The nature and consequences of reading and writing* (pp. 167–194). Cambridge, UK: Cambridge University Press.

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (1982). *Test of Nonverbal Intelligence.* Austin, TX: Pro-Ed.

Campbell, T. F., & Dollaghan, C. (1992). A method for obtaining listener judgments of spontaneously produced language: Social validation through direct magnitude estimation. *Topics in Language Disorders, 12*(2), 42–55.

Daiute, C. (1986). Physical and cognitive factors in revising: Insights from studies with computers. *Research in the Teaching of English, 20,* 141–159.

Daiute, C. (1989). Play as thought: Thinking strategies of young writers. *Harvard Educational Review, 59*(1), 1–23.

Daiute, C., & Dalton, B. (1988). "Let's brighten it up a bit": Collaboration and cognition in writing. In B.A. Rafoth & D.L. Rubin (Eds.), *The social construction of written communication* (pp. 249–272). Norwood, NJ: Ablex.

Diederich, P. B. (1974). *Measuring growth in English.* Urbana, IL: National Council of Teachers of English.

Dunn, L. M., & Markwardt, F. C. (1970). *Peabody Individual Achievement Test.* Circle Pines, MN: American Guidance Service.

Feldman, D. S., & Gagnon, J. (1986). *Statview 512+.* Calabasas, CA: BrainPower.

Fey, M. E. (1986). *Language intervention with young children.* Boston: College-Hill Press.

Gillam, R., McFadden, T. U., & van Kleeck, A. (1995). Improving the narrative abilities of children with language disorders: Whole language and language skills approaches. In M. Fey, J. Windsor, & J. Reichle (Eds.), *Communication intervention for school-age children* (pp. 145–182). Baltimore, MD: Paul H. Brookes.

Gillam, R. B. (1989). *An investigation of the oral language, reading, and written language competencies of language impaired and normally achieving school-age children* (Doctoral dissertation, Indiana University). University Microfilms, No. 9012207.

Gillam, R. B. (1995). Whole language principles at work in language intervention. In D.F. Tibbit (Ed.), *Language intervention: Beyond the primary grades* (pp. 219–256). Austin, TX: Pro-Ed.

Gillam, R. B., & Johnston, J. R. (1992). Spoken and written language relationships in language/learning impaired and normally achieving school-age children. *Journal of Speech and Hearing Research, 35,* 1303–1315.

Gillam, R. B., & McFadden, T. U. (1994). Redefining assessment as a holistic discovery process. *Journal of Childhood Communication Disorders, 16,* 36–40.

Graybeal, C. M. (1981) Memory for stories in language-impaired children. *Applied Psycholinguistics, 2,* 269–283.

Hammill, D. (1985). *Detroit Test of Learning Aptitude–2.* Austin, TX: Pro-Ed.

Hunt, K. (1970). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development, 35* (Serial No. 134).

Kirby, D., & Liner, T. (1981). *Inside out: Developmental strategies for teaching writing.* Montclair, NJ: Boynton-Cook.

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

Liles, B. Z. (1985). Cohesion in the narratives of normal and language disordered children. *Journal of Speech and Hearing Research, 28,* 123–133.

Liles, B. Z. (1987). Episode organization and cohesive conjunctives in narratives of children with and without language disorder. *Journal of Speech and Hearing Research, 30,* 185–196.

Merritt, D. D., & Liles, B. Z. (1987). Story grammar ability in children with and without language disorder: Story generation, story retelling, and story comprehension. *Journal of Speech and Hearing Research, 30,* 539–552.

Miller, J., & Chapman, R. (1984). *Systematic analysis of language transcripts.* Madison: University of Wisconsin.

Myers, M. (1981). *A procedure for writing assessment and holistic scor*ing. Urbana, IL: NCTE.

Nold, E. W., & Freedman, S. W. (1977). An analysis of readers' responses to essays. *Research in the Teaching of English, 11,* 164–174.

Norris, J., & Hoffman, P. (1993). *Whole language intervention for school-age children.* San Diego: Singular.

Polanyi, L. (1989). *Telling the American story: A structural and cultural analysis of conversational storytelling.* Cambridge, MA: MIT Press.

Scardamalia, M., & Bereiter, C. (1985). Development of dialectical processes in composition. In D.R. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language and learning: The nature and consequences of reading and writing* (pp. 307–332). Cambridge, MA: Cambridge University Press.

Stein, N. L., & Policastro, M. (1984). The concept of a story: A comparison between children's and teachers' viewpoints. In H. Mandl, N.L. Stein, & T. Trabasso (Eds.), *Learning and comprehension of text* (pp. 113–155). Hillsdale, NJ: Lawrence Erlbaum.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences (2nd ed.).* Hillsdale, NJ: Erlbaum.

Sutton-Smith, B., Botvin, G., & Mahony, D. (1976). Developmental structures in fantasy narratives. *Human Development, 19,* 1–13.

Tomblin, J. B. (1983). An examination of the concept of disorder in the study of language variation. *Proceedings from the Fourth Wisconsin Symposium on Research in Child Language Disorders.* Madison, WI: University Book Store.

van Dongen, R., & Westby, C. (1986). Building the narrative mode of thought through children's literature. *Topics in Language Disorders, 7*(1), 70–83.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Woodruff, E., Bereiter, C., & Scardamalia, M. (1981). On the road to computer assisted compositions. *Journal of Educational Technology Systems, 10,* 133-148.

Received June 28, 1994
Accepted January 20, 1995

Contact author: Ronald Gillam, University of Texas at Austin, Jesse H. Jones Communication Center, Austin, TX 78712-1809.

APPENDIX. EXAMPLES OF ANCHORS USED IN HOLISTIC SCORING

Weak Narrative

One day, three people were climbing a big big tree.
Their names are Sue, Susan, and Tom.
They have lots of fun.
I thought they're climbing that tree to see if there's other places they haven't been before, states, towns or even cities that they haven't been to.
Some cities they have traveled through.
And they're barefooted.
It'd be much better if they had socks and shoes on.
They don't want to climb too high up or else they could fall off and hurt themselves really badly.
One had shorts on.
That was Susan.
Tom had a red and white striped shirt.
Sue had shorts and a black and white striped shirt.

Adequate Narrative

It starts out like, I have him as a pet.
And it was goose time.
I meant, time to shoot gooses.
I forgot to put him inside so he was shot.
I started to cry.
And my mom bought me a new one and I was happy.

Good Narrative

Once upon a time in the state of Nebraska, there lived a famous family.
There was a man whose name is Rick and his wife whose name is Amanda, the two twin girls whose names Amanda and Ramona, and one little boy whose name is Rick Jr.
Well, Rick the father played in a famous band.
He played the drums.
Rick Jr. practiced on Rick's drums at the house.
He only made noise.
But Rick said it was music just to make Rick Jr. feel better.
When he was about ten, he quit school and practiced on the drums.
He figured he didn't have to be smart to enjoy and play the drums.
One day Rick Jr. got really sick.
He was almost going to die because he couldn't figure out what the cure was.
So he went back to school and found out what the cure was.
Then he started going back to school.
At the end of the school year he had really learned how to make music on the drums so his father let him play with

him in the band.
At the end of the one performance he said, thank you for letting me live and do what I want.
Thank you dad for making me realize what I should have done.
Thank you.

Strong Narrative
Julie and Sarah were best friends.
They both lived in Wyoming.
Their families were poor and worked hard to get money.
They picked corn and did many other hard jobs.
Both Julie and Sarah had to help.
One day Julie and Sarah were taking corn out of bags and putting it in baskets to take to market.
When they had finished their work, they decided to take a ride in one of the canoes sitting on the bank.
They slipped out when no one was watching and got into one of the canoes.
They started rowing and rowing until they were far out and could not be seen.
They were giggling and laughing.
And they did not notice storm clouds rolling in.
After awhile, they felt drops of rain coming down.
And they decided to go back.
They started rowing.
And soon the rain came pouring down.
And they could not see where they were going.
Julie pointed one way.
And she said, I think it's this way Sarah.
They began rowing the way Sarah had pointed.
The water was getting rough.
And they could not paddle much longer.
Finally, they could not row it.
And they huddled together for what seemed hours and hours.
Then the storm let up.
Julie looked around for the oars.
She could not find them.
Both girls frantically searched for them.
It's no use, cried Sarah, they're gone.
Julie could see now that they had gone the wrong way.
They could not decide what to do.
They decided to swim around for awhile to find the oars.
Then Julie called out, I found one.
And soon they found the other.
Now that they could see, they paddled toward home.
When they had gotten pretty far, both girls felt they could paddle no longer.
Then they saw a canoe ahead.
When they got closer, they saw their mammas and papas.
They rushed up to meet them and returned safely home.