Brigham Young University
BYU ScholarsArchive

All Theses and Dissertations

2010-03-19

Establishing Reliability of Reading Comprehension Ratings of Fifth-Grade Students' Oral Retellings

Laura Elizabeth Bernfeld
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

 Part of the Teacher Education and Professional Development Commons

BYU ScholarsArchive Citation

Bernfeld, Laura Elizabeth, "Establishing Reliability of Reading Comprehension Ratings of Fifth-Grade Students' Oral Retellings" (2010). *All Theses and Dissertations*. 2040.
https://scholarsarchive.byu.edu/etd/2040

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

Examining Reliability of Reading Comprehension Ratings of Fifth-Grade Students' Oral Retellings

L. Elizabeth Shirley Bernfeld

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Timothy G. Morrison, Chair
Bradley R. Wilcox, Member
Richard R Sudweeks, Member

Department of Teacher Education

Brigham Young University

April 2010

Copyright © 2010 L. Elizabeth Bernfeld

All Rights Reserved

ABSTRACT

EXAMINING RELIABILITY OF READING COMPREHENSION RATINGS OF FIFTH-GRADE STUDENTS' ORAL RETELLINGS

L. Elizabeth Shirley Bernfeld

Department of Teacher Education

Master of Arts

The purpose of this study was to rate the oral retellings of fifth-grade students to determine to what degree passages, raters, and rating occasions affect those ratings, and to identify what combination of those elements will produce reliable retelling ratings. Thirty-six fourth-grade students read and orally retold three contemporary realistic fiction passages. Two raters rated these retellings on two separate occasions using the Reader Retelling Rating Scale. These ratings were analyzed quantitatively using generalizability software. Two research questions were answered by the generalizability (G) and decision (D) studies. The G study answers the first question regarding the percentages of the total variation that can be attributed to the students, the raters, the rating occasions, the passages, and interactions among these factors. The G study found that the largest sources variation were the students, the passages, and the student-by-passage interaction. The D study answered the second question about how many raters, rating occasions, and passages would be needed to obtain a reliability coefficient for similar students in another setting. To obtain high reliability coefficients, retellings of a minimum of four (preferably six) passages should be rated by at least two raters on one occasion.

Keywords: Reading Comprehension, Oral Retelling

ACKNOWLEDGEMENTS

I could not have completed a research study and thesis project such as this on my own. Many individuals have contributed to my completion of the graduate program, research study, and thesis. I wish to express my appreciation to them.

I wish to thank my dear husband, Luke. I am so very grateful for his confidence in my abilities, even when I doubted myself. He has been perfectly supportive, with his unfailing patience and his listening ear. His love and support have been priceless.

I owe many thanks to Dr. Timothy Morrison, the chair of my graduate committee. His wisdom and patience in guiding my progress have been invaluable. He has been an ideal mentor through this process. I could not have succeeded without him. I would also like to thank the members of my graduate committee for their expertise and suggestions. Thank you to Dr. Richard Sudweeks for his vast knowledge in measurement and analysis. I appreciate his willingness to answer my many questions and to explain the intricacies of generalizability analysis. I am grateful to Dr. Brad Wilcox for his encouraging words and suggestions that result in better writing.

Finally, thank you to the graduate faculty in the department of Teacher Education at Brigham Young University. My outlook on education and life will never be the same. I appreciate the students who participated in my study, and their teachers for lending them to me for a short time. I am grateful for family and friends for their constant encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES AND FIGURES

Table Page

Figure

CHAPTER 1

INTRODUCTION

In elementary schools, reading instruction focuses on a variety of elements, including phonemic awareness, phonics, fluency, vocabulary, and comprehension (National Institute of Child Health and Human Development, 2000). Though each of these elements is essential in reading, most educators would agree that comprehension is the most essential outcome. Other elements usually serve as means to that end. Many suggest that comprehension is the purpose of reading (Bell & McCallum, 2008; Goodman, Watson, & Burke, 2005). Without comprehension, reading can be merely superficial word-calling. Since comprehension is so important, it is vital to help children to become better comprehenders.

As teachers teach children how to read and comprehend and researchers study the process, teachers must also be able to reliably assess students' reading comprehension which may be done in a number of ways. The most widespread method involves students reading passages and answering questions about the content of each (Afflerbach, 2007). This procedure is used in informal reading inventories (Afflerbach, 2007; Bell & McCallum 2008; Johnson, 1965; Morrow, 2005), criterion referenced tests (Bell & McCallum, 2008), and standardized tests (Afflerbach, 2007; Morrow, 2005). Though valuable information can be obtained about a student's reading comprehension through this method, other information is excluded, leaving teachers and researchers unable to gain a complete picture of the student's understanding.

The questions on the tests themselves may provide the student with information about the expected answers (Goodman, et al., 2005). Some questions may be poorly written or may only assess surface-level understanding (Afflerbach, 2007). Some questions may assess students' prior knowledge and not their comprehension of the text read. When questions are presented in a

multiple-choice format, students may simply guess the correct answer, when little or no comprehension has actually occurred.

Common alternatives to these question answering methods are the cloze procedure (Taylor, 1953) and retellings (Morrow, 1996). The *cloze procedure* involves students reading a passage in which words have been systematically omitted. Students then try to fill in the blanks with the exact words that are missing. The original intent of the cloze procedure was to match a specific student with a specific text, but is now more commonly used as an instructional tool. The cloze procedure has numerous limitations. Studies suggest that cloze ratings as an assessment tool measure within-sentence comprehension, or intrasentential comprehension, but not understanding across multiple sentences, or intersentential comprehension (Shanahan, Kamil, & Tobin, 1982).

Retelling, sometimes referred to as free-recall (Johnston, 1983), may provide rich information about students' understandings of written text. Retelling requires students to recall and then reconstruct their understandings of the text without being prompted in order to retell the story. The retelling process allows students to include more information about their comprehension than with other measures (Goodman, et al., 2005; Morrow, et al., 1986). However, there are limitations with this method as well. Students may have limited verbal or written communication skills and, consequently, they may be unable to completely communicate their understandings (Johnston, 1983). Retelling is not a natural process for children. But given instruction and practice retelling stories, students are able to retell stories more easily and with greater skill (Morrow, 1996). Common methods of judging the quality of retellings include rubrics and scales based on story grammars (Mandler & Johnson, 1977), which include major story elements and their sequence in passages. Examples of measures based on story grammars

are the Sense of Story Structure (Morrow, 2005) and the Reading Miscue Inventory: Construction of Meaning (Goodman, et al., 2005). Other scales, such as the Richness of Retellings scale (Irwin & Mitchell, 1983), rate retellings using additional factors, including connections students make between the text and their own experiences, inferences students make from the text, and students' ability to summarize.

Regardless of the instrument used, ratings of retellings encounter challenges. Rating retellings involves raters making decisions about the quality of each retelling. Many rating methods are quite subjective and reliability can be hard to establish.

Several potential sources of inconsistency in the ratings need to be considered when examining the reliability of retellings. These include the raters, the passages, and the rating occasions. The rating of a retelling requires one or more raters. If only one rater is involved, it is impossible to determine if the ratings would be consistent if a different rater were to rate the same retelling. When additional raters are involved, interrater reliability is an issue. This reliability between the raters would need to be examined. If an individual rated a retelling on any one occasion, intrarater reliability (reliability of scores by the same rater on more than one rating occasion) would also need to be examined.

Characteristics of passages can also affect ratings of students' reading comprehension. Passage content, structure, and length can influence retellings, and subsequently the ratings of those retellings. Students who are already familiar with the subject matter of particular passages or have experience with a particular text structure may score differently than students with limited knowledge or experience. Some passages may be too short to contain enough information for a complete assessment or may be so long that the student is overwhelmed with so much to

remember. For these reasons, passage effect should be taken into consideration when seeking to accurately assess students' reading comprehension.

As children's comprehension is assessed, they may experience a degree of fatigue while reading a very long passage or a number of different passages for an extended period. This fatigue could potentially impact their performance in an assessment. Teachers and researchers should find ways to gather enough information to make accurate assessments while not causing readers to become overly fatigued. Providing several separate opportunities for assessment (rating occasions) may decrease such fatigue.

One crucial purpose of comprehension assessment is to learn about students as readers and comprehenders, not to simply to see how many pieces of a particular text can be recalled. A retelling should be analyzed for more than story element correspondence to the original text, but as a text itself. Students' interpretations, inferences, and conclusions about texts they have read reveal a great deal about them as readers (Kalmbach, 1986).

Determining how to accurately gauge a child's comprehension using a retelling measure is difficult, considering the many factors that are potential problems. But, with the wealth of information that may be gained about students as readers from retellings, the challenge may be worth the effort. This study sought to establish what conditions needed to be in place in order for oral retelling to be an accurate, reliable measure of reading comprehension.

A similar study conducted by Sudweeks, Glissmeyer, Morrison, Wilcox, and Tanner (2004) investigated how adult English language learners comprehended expository passages through use of oral retellings. Story grammar protocols were used to measure the college-aged students' oral retellings. Oral retellings were used instead of written retellings because the English writing abilities of the participants were not as strong as their verbal abilities. The study

examined to what degree the passages, the raters, and the rating occasions affected the reliability of the ratings, and used generalizability theory to determine what combination of those factors were necessary to yield consistent, reliable ratings. The information from their study is valuable in understanding the comprehension of that population. However, no research has sought to determine how those factors might come into play when assessing the oral retellings of narrative passages from an elementary level, native English-speaking population.

Statement of Purpose

The purpose of this study was to rate the oral retellings of fifth-grade students to determine to what degree passages, raters, and rating occasions affect those ratings and to identify what combination of those elements will produce reliable retelling ratings.

Research Questions

Because this study sought to replicate the Sudweeks et al. (2004) study mentioned above, with a number of variations, the purpose and questions of the study were similar. This study sought to answer these questions:

1. What percentage of the total variance in ratings of oral retellings can be attributed to the differences in the passages sampled, inconsistencies between raters, inconsistencies within raters across rating occasions, interactions among these facets, and interactions with the object of measurement?
2. How many passages, raters, and rating occasions are needed in order to obtain a mean rating that provides a dependable estimate of how well an English-speaking elementary student would likely perform on other similar, but unobserved, passages rated by other similar raters on other similar occasions?

Limitations

Though the results of this study are not universally generalizable, they are generalizable to students in self-contained fifth-grade classrooms, who are of similar cultural and socio-economic status (SES) backgrounds. Additionally, these results are specific to the passage genre—in this case, to passages that are contemporary realistic fiction.

CHAPTER 2

LITERATURE REVIEW

Comprehension is essential to the reading process and it is crucial that teachers and researchers are able to accurately and reliably assess students' reading comprehension. Because retelling has the potential to provide rich information on what a student has comprehended while reading, retelling assessment is of particular interest. However, reliability of retelling assessment for elementary students reading narrative passages has not been studied. This study sought to determine to what degree the raters, the passages, and the rating occasions influence retelling ratings, and what factors needed to be in place to find accurate, reliable ratings.

It was important to first look at what the research has already discovered about reading comprehension assessment. Question- and-answer assessments and the cloze procedure have commonly been used to assess students' reading comprehension. This chapter examines these traditional comprehension assessments with their benefits and shortcomings, as well as retelling as an alternative comprehension assessment method.

Traditional Comprehension Assessment

Researchers and teachers use a variety of methods to assess students' reading comprehension. Traditionally, question-and-answer assessments have been the most prevalent (Johnston, 1997). Common alternatives to the question-answer method include the cloze procedure (Taylor, 1953) and retellings (Morrow, 1996).

Question-and-Answer Assessment

In question-and-answer assessments, students read a text and then are asked questions regarding it. The questions may be presented and answered orally or in written form. These

questions may be asked as part of informal reading inventories, standardized tests, criterion referenced tests, or in informal teacher assessment.

Numerous issues related to use of this method of reading comprehension assessment must be acknowledged. First, when questions are presented, students may obtain some information from them to prompt their responses. Because of this, it may be difficult to know what is being assessed: students' reading comprehension or the readers' ability to use information provided by the question (Goodman, et al., 2005; Johnston, 1983). Second, some questions may be poorly written and could communicate an erroneous message to the students regarding what is important in the passage. Such questions may be focused on unimportant information that assesses only surface-level, literal understanding (Guszak, 1967). Third, a response may underestimate students' true comprehension of the text passage. Students are often only able to present understanding about a text if there is a particular question asking about it. All additional understanding about a text may remain unexpressed (Johnston, 1983). Fourth, if students have prior knowledge relating to the topic of a text, they may be able to answer the questions that are presented without having to read or understand the text (Afflerbach, 2007). Finally, if the questions are presented in a multiple-choice format, students may simply guess the correct answer, giving the appearance of understanding when little comprehension may have occurred. It is not known if students provide the right answer for the right reason, or when the answer is wrong, why they answer incorrectly (Johnston, 1983). Overall, the question-and-answer method of reading comprehension assessment, while efficient, may provide a shallow or skewed view of students' reading comprehension.

The length of passage used in testing may also affect students' comprehension scores. Passages used in traditional types of assessment are often short. When they are short, more

passages can be used in the assessment, allowing for a variety of topics to be considered. An advantage to the use of several passages is that it decreases the affect of a student's limited prior knowledge related to a specific passage. However, the limited length of short passages allows for the creation of only a few questions. This small number of questions limits what information students can provide about their comprehension. When longer passages are used, fewer topics can be introduced, and prior knowledge about each passage can have a larger effect on the results. Longer passages allow for more good-quality questions on the passage content.

The questions derived from the text present another potential problem. The comprehension questions may vary in difficulty. Some questions may focus on literal or explicit information in a passage, while other questions might focus on inferential or implicit information (Guszak, 1967). Depending on the question, children may exhibit varying depths of understanding.

The Cloze Procedure

The cloze procedure (Taylor, 1953) is founded on the psychological notion of closure, making a language pattern whole again. This is where the procedure got its name. It was originally developed to assess reading comprehension, but is now most often used as an instructional method. The procedure involves manipulating a text in which words are systematically omitted (e.g., every fifth word or every seventh word), but the first and last sentences remain intact. The reader is required to replace the missing words. Readers' ability to replace words in the blanks with the exact missing words indicates how well they are able to comprehend the text.

One advantage of the cloze procedure is that it is similar to the experience a reader might encounter when reading a text with a number of difficult words. It can be administered in a group

setting and does not require students to answer comprehension questions. It is easy to prepare and to score. Still, the procedure has some disadvantages. There is evidence cloze texts do not require readers to comprehend beyond the immediate sentence. It is primarily sensitive to intrasentential comprehension, but lacks sensitivity to intersentential comprehension (Shanahan, et al., 1982). It appears cloze assesses comprehension on a low level with inferential comprehension not well assessed. Its unusual format may confuse students as well. Because only exact words are scored as correct, but synonyms are not (Johnston, 1983; 1997), students' ratings may not always reflect understanding. The measure has not been found to be reliable or valid in rating reading level below third grade. (Johnston, 1983; McKenna & Stahl, 2003).

Retelling as an Alternative Comprehension Assessment

Another method of reading comprehension assessment is that of retelling. Retelling can be used as an instructional strategy or as an assessment tool (Morrow, 1996). "Retellings are postreading or postlistening recalls in which readers or listeners tell what they remember from their reading or listening. Retelling can be oral or written." (p. 267). Retellings, whether completed orally or in writing, require students to communicate what they remember in their own words, not simply memorize the story (Morrow, 1996). According to Johnston (1983) retelling is "the most straightforward assessment . . . of the result of text/reader interaction" (p. 54).

Retelling as a measure of reading comprehension addresses some of the shortcomings of the question-answer and cloze assessment methods. Children's retellings reveal their comprehension in a holistic way, and this method is more advantageous that the more traditional piecemeal approach of the question-and-answer method to assessing children's reading comprehension (Irwin & Mitchell, 1983; Morrow, 1996; Morrow, et al., 1986). All of the

information in a retelling comes from the student without the prompting and information that questions may provide. Though general prompts may be given after a retelling to access further understanding, students are able to communicate their complete understandings about what is important in the text without the limitations of answering questions. Retelling leaves students free to express the depth of their understanding without boundaries. Lack of understanding is readily evident because no information is available to students to use as a crutch during the retellings. A researcher or teacher may be certain that what is expressed in a retelling is owned by the student, and not some outside source.

Retelling both requires and allows students to reconstruct understanding from a text within their own minds and then present information about the text, as they understand it (Goodman, et al., 2005; Morrow, 1996). "Retelling encourages both integration and personalization on content, helping children see how parts of the text interrelate and how they mesh with their own experiences," (Morrow, 1996, p. 268). Analysis of a child's retelling can reveal his or her ability for literal (remembering facts and details) and inferential (cause and effect relationships and sequencing of events) recall.

> Retellings can reveal a child's sense of story structure. For example, does a child's transcribed retelling include statements of setting, theme, plot episodes, and resolution? Through retelling, children also reveal their ability to make inferences as they organize, integrate, and classify information that is implied, but not expressed in the story. (Morrow, 1996, p. 276)

They may generalize, interpret feelings, or relate ideas to their own experiences (Irwin & Mitchell, 1983).

There is some discussion about whether prompts should be used during a child's retellings. Morrow (1996) suggested that prompting is appropriate during retelling instruction, but should not be used in an assessment situation. Literature since then (Bell & McCallum, 2008; Goodman, et al., 2005), including Morrow's own work (2005) has agreed that general prompts may be used after a free recall to probe for additional understanding. This procedure allows evaluators to gain insight into what was really understood and what was not. Morrow (1988; 1996) gave suggestions of general prompts that may be used at various stages of a retelling to determine understanding without providing the student with information. Before a retelling, a teacher might say, "Would you retell the story as if you were telling it to a friend who has never heard it before?" If a student has trouble getting started, the teacher might prompt, "Once upon a time," or "Once there was . . ." If a child stops before the end of the story, the teacher might ask, "What comes next?" or "Then what happened?" If a child cannot retell the story, the teacher may prompt the retelling step-by-step, with a number of questions:

> Who was the story about? When did the story happen? Where did the story happen? What was the main character's problem in the story? How did he (or she) try to solve the problem? What did he (or she) do first (second, next)? How was the problem solved? How did the story end? (Morrow, 1996, p. 270)

It should be noted that retelling orally and in writing are not easy, natural processes (Morrow, 2005). Students need to be taught how to retell and given practice retelling (Morrow, 1996). With training and practice, the quality and ease of student retellings will increase (Morrow, 1985, Morrow, et al., 1986). Johnston (1983) raised a good question when he asked, "What of the reader who can answer any question on what he has read, but if asked for a free recall has great difficulty producing an organized response?" (p. 57). This is one such situation in

which instruction and practice in retelling (and, perhaps, general prompting after a retelling) are essential.

Types of Retellings

Retellings may be expressed in writing and orally. The general benefits and limitations of retellings were discussed previously. In addition to these benefits and limitations, written and oral retellings each have unique advantages and drawbacks.

Written. Written retellings, also called rewritings (Morrow, 2005), require students to communicate their understanding and recall of a text through writing. This method has several benefits. Written language is more formal than oral language. Because an immediate response is not required, students have more time to think and may produce better-organized retellings. When writing, they have the opportunity to revise their retellings by adding and deleting information. Written retellings give students an opportunity to retell without the teacher being immediately and directly involved (Goodman, et al., 2005). Additionally, written retellings "offer the same benefits of enhancing the interpretation of text and give practice through a different communication modality" (Morrow, 1996, p. 276).

There are drawbacks and limitations to written retellings. Retelling in writing may be a challenge for students who have difficulty expressing their ideas in writing. Young children and English language learners are often unable to communicate their retellings in writing. Writing requires more time and effort. Students may be unmotivated to take the time to put into writing all of the ideas they comprehended. Their complete understanding may not be evident because of their shortcomings in writing (Johnston, 1983), fatigue, boredom, or impatience. Oral retelling provides a way around these problems.

Oral. Oral retellings require students to vocalize to a teacher or researcher their recall and comprehension of a written text. There are many benefits of oral retelling. Oral retelling does not require writing ability, but allows students to communicate in a modality that requires less time and effort for many. Students are able to use language that is familiar to them. Oral retellings are spontaneous and do not take a great deal of the students' time.

There are challenges in using oral retellings to measure reading comprehension. These challenges must be addressed for oral retellings to be a viable measure of students' reading comprehension. Because of the spontaneous nature of oral retellings, they may be less thought out and less organized. Oral retellings may also be briefer. Students may have a tendency to summarize, rather than to express their complete understandings (Morrow, 1985). Instruction and practice in oral retelling may address these limitations. When students are taught how retellings are constructed and given experience in retelling orally, the quality and ease of their retellings will improve (Morrow, 1985, Morrow, et al., 1986).

Assessment of the Retellings

When students retell passages they have read and those retellings are rated for comprehension, students are engaged in a form of performance assessment. As with any performance assessment, rater-mediated judgments are necessary to complete the ratings. A number of factors are involved in making such ratings and a variety of assessment instruments may be used to aid raters in making judgments.

Factors in rating retellings. In order to score or evaluate retellings, raters must make decisions about the quality of the retelling. They have either a student's live retelling, the recording of a student's oral retelling, a transcript of the oral retelling, or a written retelling. The

raters, passages, and rating occasions are all factors that must be taken into account when assessing reading comprehension.

The individuals rating the retelling (raters) should have training and practice using the rating instrument. When more than one person is rating the retellings, interrater reliability (reliability between the raters) should be established. This ensures that, regardless of which rater is rating the retellings, the retelling ratings will be reliable.

The passages a student reads affect comprehension. Students' interest, prior knowledge, and reading skill will vary depending on the passage. It is important to carefully select passages that will control, as much as possible, for these variables. In this study, control for these variables will be done by carefully selecting the passages based on readability formulas, text length, topic, and genre.

Retelling ratings should be consistent, whether a retelling is scored today or on another occasion. Therefore, with training and practice, a rater should be able to produce high intrarater reliability by having very close to the same score on a retelling on more than one rating occasion.

Retelling instruments. Finding an instrument to use to assist with scoring the retellings is often a concern. Three well-known instruments with contrasting perspectives have been used to assess oral retellings. They are the Sense of Story Structure (Morrow, 2005), The Reading Miscue Inventory: Construction of Meaning (Goodman, et al., 2005), and the Richness of Retellings scale (Irwin & Mitchell, 1983).

Morrow's (2005) Sense of Story Structure examines students' understandings of story grammar. Story grammars emphasize components of a narrative text, including setting, characters, problem or goal, episodes leading to solving the problem or meeting the goal, and the resolution of the problem. Ratings on the story structure protocol reveal the percentage of the

story elements recalled by the student as described and weighted in the protocol. An advantage of this protocol is that it is relatively easy to use. However, the ratings may not reveal the degree to which each element was referred to or explained by the students. It does not describe the depth of readers' comprehension and does not reveal ideas that are outside the story grammar frame, such as inferences, opinions, or associations (Wilson, Martens, Arya, & Jin, 2007). Kalmbach (1986) suggested that recall is only part of a retelling. Irwin and Mitchell (1983) asserted that inferences, connections, and conclusions in a retelling should also be scored.

Goodman et al.'s (2005) Reading Miscue Inventory (RMI) uses a retelling guide that is divided into character analysis and story events. The character names and descriptions are assigned point values in the character analysis section. In the story events section, story episodes are assigned more or fewer points depending on the number of major and minor events in each episode and how important they are to the overall story. A benefit of the RMI is that it awards points, to a degree, for more extensive understandings of the characters and plot episodes. Its rating, however, does not reflect readers' opinions, inferences, insights, and connections. An additional difficulty with this protocol is in the preparation of the retelling guide. Because each guide is unique to the passage, agreement on how many points to award characters and events can be a challenge. It is unclear how to decide how character or event importance in the story should translate into the point values (Wilson, et al., 2007).

Irwin and Mitchell's (1983) Richness of Retellings scale looks at retellings in a holistic manner. They suggested that assessments that "assign points to reflect the relative importance of various elements of retelling" (p. 391) are unable to "capture the interrelationships of all the individual factors [and] the individuality of a student's point of view" (p. 392). This rating guide distinguishes five separate levels of retellings, based on the ratings of eight characteristics:

generalizations beyond the text, a thesis (summarizing) statement, major points, supporting details, supplementations, coherence, completeness, and comprehensibility. Once a rater has taken into account all of the characteristics listed above, he or she assigns the student an overall score of 1 to 5, based on the depth and richness of the retelling as a whole. The Richness of Retellings scale has numerous benefits. It allows for the unique nature of retellings by individual students. It values the ability to recall elements from the text, but also the way in which readers make inferences, connect ideas to themselves and to life, draw conclusions, and engage in higher-level thinking in relation to the text. A drawback to this rating protocol is that, because each student receives an overall score between 1 and 5, it does not reveal details about what was and was not understood by the reader (Wilson, et al., 2007).

The researchers who developed the Sense of Story Structure (Morrow, 2005), The Reading Miscue Inventory: Construction of Meaning (Goodman, et al., 2005), and the Richness of Retellings scale (Irwin & Mitchell, 1983) did not provide information on the reliability and validity of scores obtained using these assessment instruments when they were introduced into the literature. Furthermore, no research has been found to establish the reliability or validity of any of these instruments. Because reading comprehension assessment is critical and oral retelling is a valuable way to assess comprehension, establishing the reliability and validity of a tool to score oral retellings is an important undertaking. This study sought to accomplish this objective by examining the sources of variability in rating retellings using the Reader Retelling Rating Scale, and determining what combination of raters, passages, and rating occasions are needed to achieve reliable scores.

Others who have studied how to rate oral retellings of students include Burton (2008) and Sudweeks, et al. (2004). Burton’s (2008) generalizability study examined the variability in the

ratings of fourth-grade students' oral retellings of expository texts. It investigated how much variability could be attributed to the students, the passages, the day of test administration, the raters, the rating occasions, and interactions among these factors. She found that to obtain the highest reliability coefficients teachers should have students read and retell at least two passages across two days, with at least two individuals rating the retellings.

Sudweeks, et al. (2004) investigated how adult English language learners comprehended expository passages. The oral retellings were rated using a story grammar protocol. The facets of the study were passages, the raters, and the rating occasions. They investigated how these facets affected the reliability of the ratings, and used generalizability theory procedures to determine what combination of those factors were necessary to yield consistent, reliable ratings. Using generalizability software, the Sudweeks, et al. (2004) study that found that readers should retell at least four, but preferably six, passages to obtain high reliability coefficients.

The results from these generalizability studies offer some information about their student populations and the text genres the students read. Though both of these studies contain similarities to the current study, it is unclear how their results would generalize to different populations and different settings. This study seeks to gain information about the rating of fifth-grade students' oral retellings of narrative passages.

CHAPTER 3

METHODOLOGY

Research has not yet investigated what conditions need to be in place in order for oral retelling to be an accurate, reliable measure of reading comprehension for elementary-age students reading narrative passages. For this reason, this study sought to rate the oral retellings of narrative passages given by fifth-grade students, to determine to what degree passages, raters, and rating occasions affected those ratings, and to identify what combination of those elements produced reliable retelling ratings.

Participants

Participants included 36 students from three self-contained fifth-grade classes from a school in Utah County, Utah. This K-6 school's student population of 965 students was approximately 95% white, 2% Asian, 1% Latino, and 2% other. The school's socio-economic status was primarily middle to upper-middle class. There were 12 English language learners receiving services at the school and 11% of the students in the school received free and reduced lunch.

Fifth-grade students were appropriate for this study because, by the fifth grade, most students are beyond the decoding stage and are essentially fluent readers. Data from students who were reading below fourth grade level were not scored or included in the results of this study. Only data from students reading at a fourth-grade reading level or higher were scored and analyzed. This was to help ensure that the participants' comprehension was not diminished by decoding struggles. The reading level of the students was assessed using the Developmental Reading Assessment (Beaver, 2006) conducted by the classroom teachers at the beginning of the school year.

The students this school and school district are accustomed to retelling a text after reading because of past school experience. As first- and second-grade students, they were required to orally retell stories to their classroom teachers for the reading benchmark tests that they took three times each year. Then, as third- and fourth-grade students, they were required to compose written retellings of the stories they read for their twice-yearly reading benchmark test. Because students were already familiar with and accustomed to retelling passages, the researchers in this study did not provide additional practice opportunities or instruction on retelling for the study participants.

Passages

Each student read three narrative passages that were taken directly from *Power Reading Pak 4-B* (Cole & Larkin, 2002) part of the Power Reading program. The titles of the three passages were “The Skate Park Stranger,” by Olivia Cole (482 words, Flesch-Kincade level 4.3), “Finding Freddie,” by Barbara Larkin (441 words, Flesch-Kincade level 4.4), and “The Bike Race,” by Barbara Larkin (520 words, Flesch-Kincade level 4.8). This set of leveled stories is one of several sets available to all teachers at the school being studied.

The content of the passages differed in topic, but were all from the contemporary realistic fiction genre. They were written at reading levels ranging from 4.3 to 4.8. The level of the passages was measured by the Flesch-Kincaid readability formula. Passages at this level were selected to ensure that the students who were reading at a fifth-grade reading level would be less likely to have their comprehension hampered by having to focus on decoding text that was too difficult for them.

Each passage was 400-600 words long. Passages of this length were long enough to allow for a self-contained, stand-alone story, but short enough that the reader could read each passage within a few minutes without becoming overwhelmed with too much information.

Instrument

Because rating retellings is a rater-mediated process, the rater is the real instrument. However, to promote uniformity among multiple raters, an additional aid is needed. For the purposes of this study, the Reader Retelling Rating Scale was the aid developed to guide in rating the retellings (see Appendix A).

This measure is the researchers' adaptation of Morrow's (1988) Reader Retelling Profile, developed from Irwin and Mitchell's (1983) Richness of Retellings scale. This scale was chosen because it not only rates the reader's recall of the story events, but it also takes into account the reader's background knowledge and deeper connections. Exact directions to the administrator/rater and students are printed with the assessment instrument (see Appendix A).

Two researchers practiced rating retellings by other fifth-grade students using a version of the rating instrument prior to the beginning of the study that had six items. The original six items included a number of observations about the reader: (a) includes information that is directly stated in the text, (b) infers information directly or indirectly from the text, (c) provides relevant content and concepts, (d) indicates reader's attempts to make summary statements or generalizations, (e) indicates reader's ability to organize or compose the retelling, and (f) demonstrates appropriate use of language conventions, sentence structure, and vocabulary.

During the rating practice sessions, the raters determined that two of the items on the original rating rubric were somewhat redundant. They felt that item b (inferring information from the text) and item d (making summary statements or generalizations) were closely related to item

a (information stated in the text). All three items were related to processing and stating content from the passage. The decision was made to collapse these items into one, reworded item. The final wording of the combined item stated that the student includes information that is directly stated or inferred/summarized from the text.

Additionally, item f, dealing with the participants' use of language, was deleted. The researchers felt that, because the rating instrument's purpose is to make decisions about reading comprehension, the language-related item was not directly related to the rating of reading comprehension. After these modifications were made to the rating instrument, three rating items remained, one dealing with the content, one with the relevance, and one with the organization of the students' retellings. The content item specifically scored readers' abilities to include information that was directly stated in or was inferred/summarized from the text in their retellings. The relevance item dealt with the degree to which the readers provided relevant content and concepts. The organization item investigated the readers' abilities to organize or compose their retellings. Together, the three items make up the items in the Reader Retelling Rating Scale. This instrument went beyond the story grammar by scoring students summaries and inferences. Additionally, the relevance item required students to make decisions about what was important in the story, not including every story detail in their retellings, and scored students' ability to do so.

Raters using this instrument assigned a rating of one to four to each of these three items on the scale. A rating of four on an item suggests a high level of proficiency in the retelling. This means that during the retelling, approximately 90% or more of the important story content was retold or summarized, 90% or more of the information that was included was relevant to the content of the passage, or all story events were told in sequence. A rating of three indicated a

moderate level of proficiency. This means that approximately 75 to 89% of the important story content was included in the retelling, 75 to 89% of the retelling included relevant information, or the retelling contained one story event told out of sequence. A rating of two indicated a low level of proficiency. The raters assigned this score if approximately 50 to74% of the important content was included in the retelling, 50 to 74% of the information included in the retelling was relevant, or if the retelling contained two or three events out of sequence. A rating of one indicated a very low level—less than approximately 50% of the content was included or was considered relevant, and four or more events were told out of sequence. The item could also receive a score of one if an item (content, relevance, or organization) was not addressed in the retelling. The percentages indicated above were agreed upon by the raters during the practice rating sessions and were used as a guide in rating the retellings.

Procedures

Students were asked to silently read and orally retell three passages to the researcher. The passages were read on one occasion. The stories were presented to the students in different orders. This counterbalancing was done to control for presentation effect. Students were individually introduced to the first passage and asked to read it silently to themselves. They then were then prompted by the researcher to retell what they recalled from the passage to the researcher without having access to the text. These retellings were audio recorded for later rating by the researchers. Students were then introduced to a second passage. They read it to themselves and then retold the story to the researcher. This retelling was also audio recorded for later rating. This procedure was repeated with a third passage. While the researcher was prepared to provide general prompts (e.g., “tell me more,” or “what happened next,”) to encourage

expanded retellings, no retelling session required their use. The same researcher collected all of the retellings by the students.

As the raters practiced rating, it became apparent that they also needed to agree upon the elements and events of each story that should be included in a complete retelling of the story. Because the raters were each going to be rating a total of 108 retellings on both rating occasions, they created a list of story elements/events for reference during the rating sessions (see Appendix B). They practiced rating retellings by fifth-grade students who met the criteria for participation in the study, but who were not selected as study participants. The researchers continued to practice until they had established 90% agreement on each practice passage. The ratings obtained during the practice sessions were used for the purpose of establishing interrater reliability and were not included in the study results.

The researchers then rated the retellings of all 36 participants retelling three passages each (n=108) on one rating occasion. While both raters were in the same room and listened to the recordings at the same time, each researcher rated the retellings independent of the other. Ratings of all 108 retellings again occurred on a separate occasion several days later, following the same procedures. The order of the presentation of the retellings to the raters was varied on each rating occasion. The data were then used to calculate interrater and intrarater reliability.

Data Analysis

The statistical analysis procedure used to examine the retelling ratings was based on generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001). Generalizability theory allows researchers to differentiate among multiple sources of error in estimating reliability. This is in contrast to classical test theory, which decomposes a participant's score into the true score and an undifferentiated error term. While classical theory's

concept of error term cannot differentiate among the multiple sources of error, generalizability theory is able to decompose the total error variance and attribute it to multiple sources of variance (called facets), the object of measurement, and interactions among those factors, with only a residual error value left to unidentified sources.

The generalizability study (G study) in this research used the retelling ratings collected to estimate the percent of variance associated with each source of variability in the ratings. The decision study (D study) extrapolated the data from the G study and investigated how changing the number of raters, rating occasions, and passages would produce high reliability ratings.

The design of this study was a three-facet, fully-crossed design. This fully-crossed design means that both raters rated each students' retelling of each passage on both rating occasions. After student retellings were rated by the researchers/raters, those ratings were analyzed quantitatively, using the G and D studies of generalizability theory to answer the two research questions. The GENOVA (Crick & Brennan, 1982) computer software was used to run both the G and D studies.

Students (S) were the object of measurement in this study. The facets were identified as the raters (R), the rating occasions (O), and the passages (P). The G study analyzed what percent of the variability could be attributed to the object of measurement and the three facets, as well as the amount of variance that could be attributed to interactions among these factors. The interactions included the student-by-occasion interaction (S x O), the student-by-passage interaction (S x P), the student-by-rater interaction (S x R), the occasion-by-passage interaction (O x P), the occasion-by-rater interaction (O x R), the passage-by-rater interaction (P x R), the student-by-occasion-by-passage interaction (S x O x P), the student-by-occasion-by-rater interaction (S x O x R), the student-by-passage-by-rater interaction (S x P x R), and the occasion-

by-passage-by-rater interaction (O x P x R). The G study provided a direct answer to the first research question.

After the G study had calculated the estimated variance components for each facet, the object of measurement, and possible interactions, the D study calculated reliability coefficients for a variety of combinations of factors. For example, the D study calculated the change in reliability if two passages were used, versus four passages or six passages, or if one rater was used, versus two, three, or four raters. Reliability coefficients were obtained for a number of different combinations of factors in this fully-crossed design. Once analyzed, results gave insight into the optimum number of raters, rating occasions, and passages that are needed in order to obtain accurate ratings of a student's oral retelling, in the context of a fully-crossed design. This D study answered the second research question.

CHAPTER 4

RESULTS

The purpose of this study was to find the optimal combinations of passages, raters, and rating occasions to be able to reliably and accurately assess fifth-grade students' reading comprehension of contemporary realistic fiction text. As stated in Chapter 1, the specific research questions were the following:

1. What percentage of the total variance in ratings of oral retellings can be attributed to the differences in the passages sampled, inconsistencies between raters, inconsistencies within raters across rating occasions, interactions among these facets, and interactions with the object of measurement?
2. How many passages, raters, and rating occasions are needed in order to obtain a mean rating that provides a dependable estimate of how well an English-speaking elementary student would likely perform on other similar, but unobserved, passages rated by other similar raters on other similar occasions?

G Study Results

The purpose of the generalizability study (G study) in this research was to identify, out of all the possible sources of variance, the degree to which the respective factors—the student, the rating occasion, the rater, and the passage—contributed to the variation in the ratings. The G study answers the first of the research questions driving this study.

The results of the G study are summarized in Table 1. The rows in the table identify the possible sources for variation in the total passage ratings that can be estimated from a three-facet, fully-crossed design where students are the object of measurement. The columns of the table show the estimated variance component for each factor and interaction of factors, the percent of

Table 1

Variability in Individual Passage Rating Totals Accounted for by Each Source of Variation and Their Interactions

Source of Variation	Degrees of Freedom	Estimated Variance Component	Percent of Total Variation	Standard Error
Students	35	.9256	31.81	.2842
Occasions	1	.0404	1.39	.0451
Passages	2	.5284	18.16	.4179
Raters	1	.0298	1.02	.0588
S x O	35	.0000	0.00	.0458
S x P	70	.4762	16.36	.1266
S x R	35	.0000	0.00	.0584
O x P	2	.0076	0.26	.0114
O x R	1	.0129	0.44	.0171
P x R	2	.0743	2.56	.0606
S x O x P	70	.0804	2.76	.0539
S x O x R	35	.1515	5.21	.0671
S x P x R	70	.2127	7.31	.0731
O x P x R	2	.0002	0.00	.0076
Residual	70	.3702	12.72	.0617

Note: The negative variance component estimates were set to zero following Brennan's (1992; 2001) guideline.

the total variation that can be attributed to each source of variance, and the standard error for each of the estimated variance components.

The estimated variance component for students (S) was the largest (.9256), accounting for nearly 32% of the total variance. The variance components for the passages (P) (approximately 18%) and student-by-passage interaction (S x P) (approximately 16%) were also relatively large when compared to the other factors. The total variance in the students, the passages, and the student-by-passage interaction accounted for approximately 66% of the total variation in the retelling ratings.

Students

The variance component for the students was larger than any other factor interaction in the study. It accounted for nearly 32% of the total variation. Because students were the population of interest and the researchers sought to make inferences about them, students were the object of measurement in this study. The high percentage of variation resulting from the students was to be expected because the researchers assumed there would be differences in the reading and retelling skills in the individual students. Just as one would expect different students to perform differently on a series of given tasks, one would expect their retelling ratings to differ.

Passages

The passages accounted for 18% of the total variation. Passages, by nature, tend to be different. There is variability among the overall means of the three passages. This was not surprising. Though care was taken to find several passages with similar readability from the same genre, inherent differences in the passages still existed. These differences may have included differences in text structures, vocabulary, and concept load, for example.

Student-by-Passage (S x P) Interaction

The student-by-passage interaction accounted for approximately 16% of the total variation. This means that the relative ordering of the students (as determined by the mean passage ratings) was not the same from one passage to the next. This interaction may be due to the way different students interacted with the passages, thereby producing variation in their retellings. The content of each passage may have been more or less familiar to the individual students. Students' prior knowledge or experience with certain topics over others will impact their passage comprehension. Similarly, the text structure and vocabulary may have been easier or more difficult for various students. This provides evidence that a teacher should not judge a student's comprehension based on a single passage.

Raters and Occasions

The fact that the rater variance component only accounted for approximately 1% of the total variation suggests that there was solid interrater reliability. Though there were two different raters, little variation between their ratings was noted. The variance component for the rating occasion also accounted for approximately 1% of the total variation. This low percentage indicates consistent intrarater reliability. Regardless of when the retelling was rated by an individual rater, the ratings were consistent.

Occasion-by-Rater (O x R) Interaction

The occasion-by-rater interaction accounted for less than one-half of one percent of the total variance. This negligible value indicates that the mean student ratings obtained were consistent between raters.

Results of the estimated variance components for two interactions—S x O and S x R—were reported by the GENOVA software (Crick & Brennan, 1982) as zero. These variance

components may have actually had a negative value, but were automatically set to zero by Brennan's (1992; 2001) rule. This rule sets the negative estimates to a value of zero, but uses the original negative estimates in computing other variance components. Brennan's rule allows for unbiased results in estimating the other variance components.

D Study Results

This study was a three-facet, fully-crossed design. The facets included passages, raters, and rating occasions. The object of measurement was the students. This fully-crossed design required that every rater scored every retelling of each passage, on every occasion. In this study, that means that two raters rated 108 retellings each on two different occasions, for a grand total of 216 retellings each. Because it is not always possible for two raters to be willing to rate so many passages, a D study was conducted to determine the reliability coefficients for all sources of variance using other configurations of raters, rating occasions, and passages. For example, how would the reliability coefficients change with only one rater, or with three? What number of passages is necessary to obtain the best reliability? Do the retellings need to be rated on multiple occasions? If so, how many? The results of a D study can give researchers and educators information to help them determine how varying the combination of factors to create more economical or practical configurations will affect the error variance.

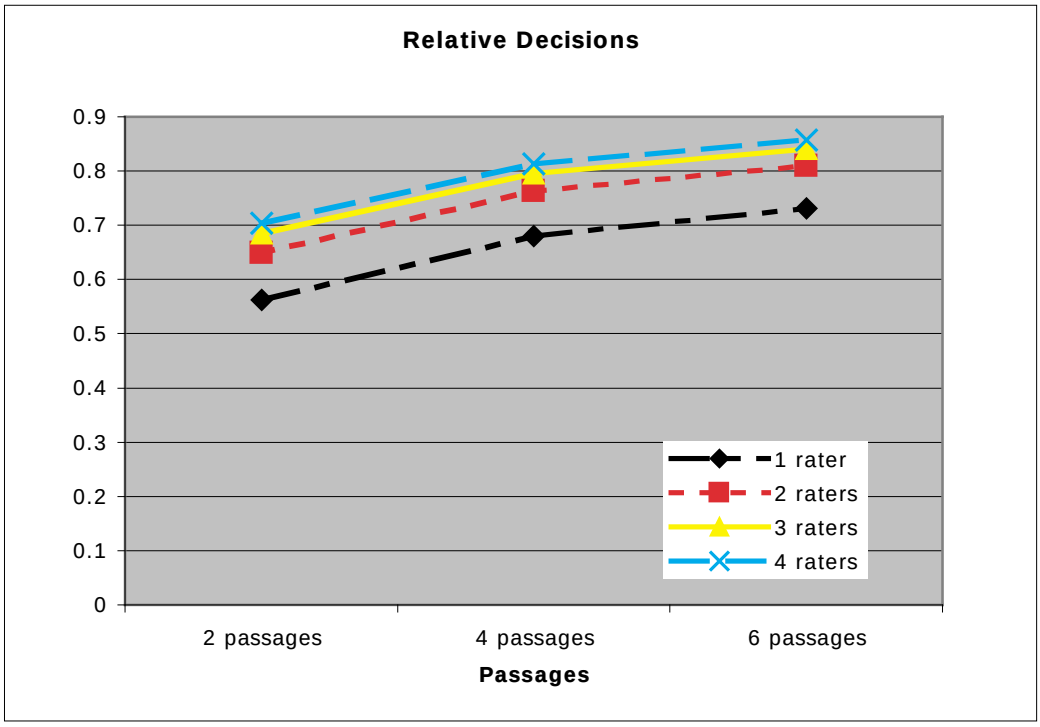
Essentially, the D study in this research answers the second research question in the context of a fully-crossed design. This question asks how many passages, raters, and rating occasions are needed in order to obtain a mean rating that provides a dependable estimate of how well an English-speaking elementary student would likely perform on other similar, but unobserved, passages rated by other similar raters on other similar occasions.

Figure 1 illustrates how increasing the number of raters or the number of passages students retell affects the coefficients for making relative and absolute decisions. Relative decisions, illustrated on the left side of Figure 1 are made using the G-coefficient. These numbers would appropriately indicate reliability when students' ratings are being compared to one another. Absolute decisions, on the other hand, are based on the Phi-coefficient values. The values for absolute decisions can be found on the right side of Figure 1. These values should be considered when students are being compared against a pre-determined standard.

When examining the lines indicating the number of raters in the relative decisions figure, there is a sizable increase in the reliability coefficients with two raters (.6494 for two passages, .7629 for four passages, and .8101 for six passages) compared to one rater (.5621, .6796, and .7306 for two, four and six passages respectively). The increase is less when using two to three raters (.6849, .7953, .8405), and even less benefit when moving from three to four raters (.7041, .8126, .8566). This same trend held true in the figure for absolute decisions.

Likewise, the reliability coefficients increase significantly when four passages were rated, rather than two passages. There is also an increase in the reliability coefficients when going from four passages to six, but the increase is less dramatic than when moving from two passages to four. These changes indicate that the benefits to increasing the number of passages level off somewhat.

In many classroom and research situations, it is only feasible for one individual to rate a student on one occasion. The D study found that, when making absolute decisions, if a student retells two passages that are rated by one rater on one occasion, the reliability coefficient is .4548. If the student retells four passages, the reliability jumps to .5793. If the student retells six passages, the reliability is .6375.



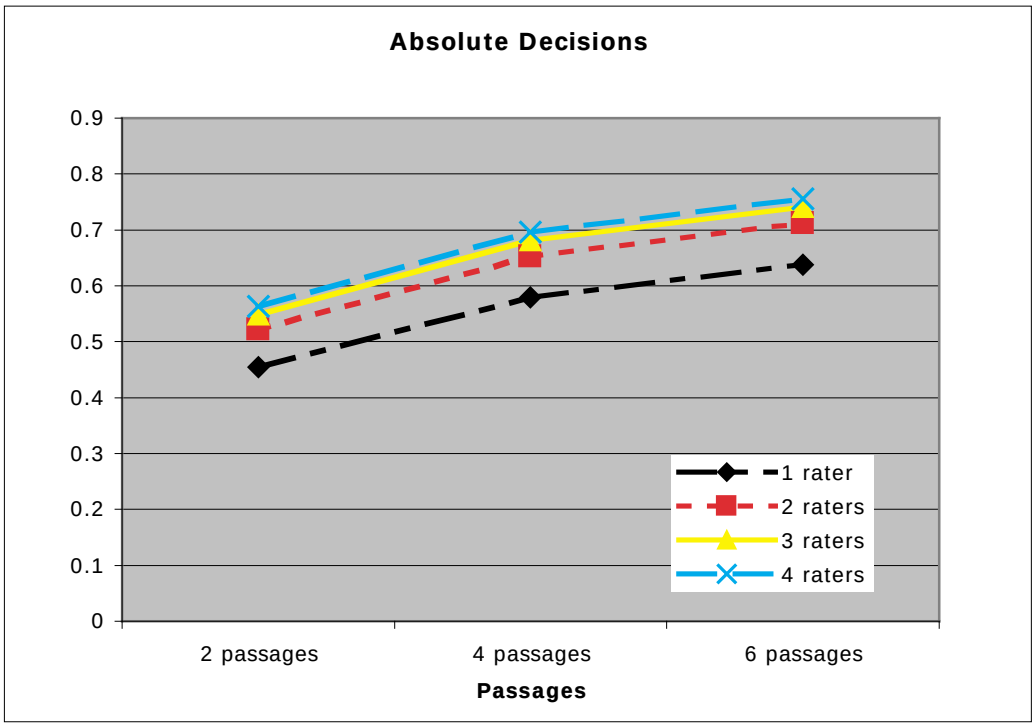


Figure 1. Reliability for Relative and Absolute Decisions by Number of Passages and Number of Raters

Summary

Much information is gained in answer to the two research questions through the G and D studies. The G study answers the first question regarding the percentages of the total variation that can be attributed to the students, the raters, the rating occasions, the passages, and interactions among these factors. The G study found that the largest percentage of variation was from the students (32%), the passages (18%), and the student-by-passage interaction (16%). The percent of residual factors was about 13%. All the other facets and interactions had small percentages that, combined, accounted for the remaining 21 percent of the total variance.

The D study answered the question about how many raters, rating occasions, and passages would be needed to obtain a reliability coefficient for similar students in another setting. To obtain high reliability coefficients, retellings of a minimum of four (preferably six) passages should be rated by at least two raters on one occasion.

CHAPTER 5

DISCUSSION

The purposes of this study were to identify the percent of total variation that can be attributed to a number of factors in the rating of oral retellings, and to make recommendations about how those factors can be manipulated to achieve high reliability coefficients. The researchers in this study sought to achieve these objectives by rating oral retellings using the Reader Retelling Rating Scale and analyzing the ratings using Generalizability software. This study provided much information about using oral retelling to measure the reading comprehension of fifth-grade students who read narrative passages, but further research should be conducted to answer similar questions about different populations in varying testing circumstances.

Reflections on Findings

A number of insights may be gained when reflecting on the findings of this study. The Reader Retelling Rating Scale may be used as an instrument to reliably assess readers' oral retellings. The sources of variability in the retelling scores were identified and this information was useful in determining the reliability of the scores if the assessment conditions were varied.

The Use of the Reader Retelling Rating Scale

Trained raters can use this Reader Retelling Rating Scale to obtain reliable scores of fifth-grade students' oral retellings of narrative texts. The researchers in this study created and used the Reader Retelling Rating Scale to aid their ratings of students' retellings. When practicing using the rating scale, it quickly became evident that training is necessary to establish interrater and intrarater reliability. If researchers or educators wish to use the Reader Retelling Rating Scale to rate retellings, training and practical experience is recommended to establish reliability.

While training using the rating scale, the researchers in this study found it helpful to agree on a list of story elements and events. For the purposes of this study, the list was created for later reference (see Appendix B). A similar list of story elements and events should be agreed upon by raters if the Reader Retelling Rating Scale is going to be reliably used with other retellings in the future.

Sources of Variability

The G study answered the first research question that dealt with the sources of variation. In this study, the largest sources of variance were the students, the passages, and the student-by-passage interaction. These three accounted for a little more than 66% of the total variance. Raters and rating occasions totaled only about 2% of the total variability. Residual factors accounted for about 13% of the total variability. Sudweeks et al. (2004) also found students to be the greatest source of variation, followed by passages, and the student-by-passage interaction. Burton (2008) likewise found that students account for the greatest percent of the variation. However, in her study, the passages and the student-by-passage interaction accounted for very little of the variation. In her study, more variability was a result of passage-by-rater interactions.

Potential Variation of Assessment Conditions

The answer to the second research question is provided by results of the D study. Including additional raters is beneficial. When at least two raters are used to rate retellings, the relative increase in benefit is the greatest. Three or four raters may be used, but the relative increased benefit is less than using two raters, compared to using only one.

When examining the number of passages that should be used to obtain acceptable reliability coefficients, there is a notable advantage in rating the retellings of four passages over two. There is an additional benefit to rating six passages, but the increase in the reliability

coefficient is not as great. These results are consistent with the findings and recommendations of Sudweeks, et al. (2004).

The results of the D study indicate a slight benefit to including a second rating occasion compared to using a single rating occasion, though the benefit is small. Using additional passages has greater affect on reliability than adding additional rating occasions. Comparatively, it is much more beneficial (and more realistic in a classroom setting) to rate retellings of additional passages than to rate retellings of fewer passages on multiple occasions. The reliability coefficients for a single rating occasion seem sufficient.

In the context of this fully-crossed design, the greatest increase in reliability coefficients can be found when two raters rate the retellings from a minimum of four passages on a single rating occasion. If it is feasible to rate retellings from six passages, the reliability coefficients are even better. A minimum of 4 passages should be retold and rated for the highest relative benefit.

Recommendations for Future Research

The research questions in this study were answered in the context of a three facet, fully-crossed design. However, it is not always feasible outside of a research setting to use this design because of limitations in time and limited availability of multiple raters. Future studies could investigate the effects of these same sources of variability in more feasible designs. What if some of the facets of this study were nested? A design in which retellings of different passages were nested in raters would mean that some raters would rate some of the retellings, while different raters would rate the remainder of the retellings. This nested design would be less taxing on each rater. How would reliability coefficients be affected if the design was nested in this or other ways? Could researchers create a design that produces high reliability while being more feasible to use in classroom and research settings?

This study utilized the Reader Retelling Rating Scale that is designed specifically for use with retellings of contemporary realistic fiction passages. Not all passages read by elementary students in a classroom setting are fiction. Development of a rating scale similar to the Reader Retelling Rating Scale, but that is created for use with nonfiction text would be beneficial. After such a scale is created, study is needed to establish whether it can be used to produce reliable ratings.

Results of this study provide evidence that researchers can use the Reader Retelling Rating Scale to obtain reliable scores of fifth-grade students' of oral retellings of fiction passages. When students from this population read a minimum of four passages that are rated by at least two raters on one rating occasion, estimated reliability coefficients are high. These results are consistent with the findings of Sudweeks, et al. (2004). Future research may provide additional information on the rating of oral retellings to measure reading comprehension.

REFERENCES

Afflerbach, P. (2007). *Understanding and using reading assessment, k-12.* Newark, DE: International Reading Association.

Beaver, J. M. (2006). *Developmental reading assessment* (2nd ed). Parsippany, NJ: Celebration Press.

Bell, S. M., & McCallum, R. S. (2008). *Handbook of reading assessment.* Boston: Allyn & Bacon.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Burton, R. C. (2008). *Oral retelling as a measure of reading comprehension: The generalizability of ratings of elementary school students reading expository texts*. Unpublished master's thesis, Brigham Young University, Provo, Utah.

Cole, O., & Larkin, B. (2002). *Power reading power pak 4-b.* Syosset, NY: National Reading Styles Institute.

Crick, J. E., & Brennan, R.L. (1982). *GENOVA: A generalized analysis of variance system* (FORTRAN IV computer program and manual). Dorchester, MA: Computer Facilities, University of Massachusetts at Boston.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability of scores and profiles.* New York: Holt, Rinehart & Winston.

Goodman, Y. M., Watson, D., & Burke, C. (2005). *Reading miscue inventory: From evaluation to instruction.* Katonah, NY: Richard C. Owen.

Guszak, F. J. (1967). Teacher questioning and reading. *The Reading Teacher, 21,* 227-234.

Irwin, P. A., & Mitchell, J. N. (1983). A procedure for assessing the richness of retellings. *Journal of Reading, 26,* 391-396.

Johnson, M. S. (1965). *Informal reading inventories.* Newark, DE: International Reading Association.

Johnston, P. H. (1983). *Reading comprehension assessment: A cognitive basis*. Newark, DE: International Reading Association.

Johnston, P. H. (1997). *Knowing literacy: Constructive literacy assessment.* York, ME: Stenhouse Publishers.

Kalmbach, J. R. (1986). Getting at the point of retellings. *Journal of Reading, 29,* 326-333.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9,* 111-151.

Martens, P., Arya, P., Wilson, P., & Jin, L. (2007). Text structures, readings, and retellings: An exploration of two texts. *Literacy Teaching and Learning, 11*(2), 49-64.

McKenna, M. C., & Stahl, S. A. (2003). *Assessment for reading instruction.* New York: Guilford Press.

Morrow, L. M. (1985). Retelling stories: A strategy for improving children's comprehension, concept of story structure and oral language complexity. *Elementary School Journal, 85,* 647-661.

Morrow, L. M. (1988). Retelling stories as a diagnostic tool. In S. M. Glazer, L. W. Searfoss, & L. M. Gentile (Eds.) *Reexamining reading diagnosis: New trends and procedures.* (pp. 128-149). Newark, DE: International Reading Association.

Morrow, L. M. (1996). Story retelling: A discussion strategy to develop and assess comprehension. In L. B. Gambrell & J. F. Almasi (Eds.), *Lively discussions!: Fostering engaged reading* (pp. 265-285). Newark, DE: International Reading Association.

Morrow, L. M. (2005). *Literacy development in the early years: Helping children read and write* (5th ed). Needham Heights, MA: Allyn & Bacon.

Morrow, L. M., Gambrell, L., Kapinus, B., Koskinen, P. S., Marshall, N., & Mitchell, J. N. (1986). Retelling: A strategy for reading instruction and assessment. In J. Niles (Ed.), *Thirty-fifth yearbook of the National Reading Conference.* Rochester, NY: National Reading Conference.

National Institute of Child Health and Human Development. (2000). *Report of the national reading panel: Teaching children to read* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*, 229-255.

Sudweeks, R. R., Glissmeyer, C. B., Morrison, T. G., Wilcox, B. R., & Tanner, M. W. (2004). Establishing reliable procedures for rating ELL students' reading comprehension using oral retellings. *Reading Research and Instruction, 43*(2), 65-86.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.

Wilson, P. G., Martens, P., Arya, P., & Jin, L. (2007). The anatomy of retelling ratings: What these ratings do (and don't) reveal about readers' understandings of texts. In D. W. Rowe, R. T. Jimenez, D. L. Compton, D. K. Dickinson, Y. Kim, K. M. Leander, & V. J. Risko

(Eds.), *56th yearbook of the National Reading Conference* (pp. 362-376). Oakcreek, WI: National Reading Conference.

APPENDIXES

Appendix A

Rating Instrument and Administration Protocol

Reader Retelling Rating Scale

Student _____ Rater _____ Date _____

Introduction: The title of this story is ____. Have you ever read or heard this story before? (If yes: How familiar are you with this story?) What do you already know about __(topic of story)__? What do you think might happen in this story?

Please read this story silently to yourself. After you finish reading, I will ask you to retell the story to me.

Initial prompt: Now that you have read this story, I'd like to have you retell it as if you were a storyteller, telling this story to someone who has never read or heard it before.

	1	2	3	4
1. Includes information that is directly stated or inferred/summarized from the text. *(Content)*				
2. Provides relevant content and concepts. *(Relevance)*				
3. Demonstrates ability to organize or compose the retelling. *(Organization)*				

Intermediate prompts: (indicate which prompts are used. Use only if student is unable to continue the retelling.)

Once there was…
What comes next?
Then what happened?
Who was the story about?
When did the story happen?
Where did the story happen?
What was the main character's problem in the story?
How did he (or she) try to solve the problem?
What did he (or she) do first (second, next)?
How was the problem solved?
How did the story end?

Follow-up prompt: Can you tell me anything else about this story?

Appendix B

Story Element/Event Lists Used for Rating

Student Name: _____ Title: <u>Finding Freddie</u> Date:_____Rater:_____
Story Elements:
Setting- at home, outside
Characters-Merilee, Freddie, Mom
Event 1- Merilee has to tend her little brother, Freddie, while their mom was gone. Merilee didn't want to.
Event 2- Merilee told Freddie to play inside with his fire engine while she beaded a necklace.
Event 3- After a while, she realized that Freddie was gone.
Event 4- The door was open and it was storming outside.
Event 5- Merilee went outside to look for Freddie.
Event 6- She was afraid that if Freddie was hiding in the culvert, it could fill with water from the storm and he could drown.
Event 7- After she heard his voice coming from the culvert, she dragged him out and took him home.
Event 8- Merilee told Freddie that he could have died, he didn't understand, and she explained what it means.

Student Name: _____ Title: <u>The Bike Race</u> Date:_____Rater:_____
Story Elements:
Setting- home, store, race
Characters- Jillian, Mark, Mom, older girl
Event 1- Jillian received a bike that had been used by other members of her family
Event 2- She was disappointed because it was in poor condition and didn't have the features of a racing bike.
Event 3- Jillian wanted to win the 4th of July bike race.
Event 4- Mark showed off his racing bike and teased her.
Event 5- Jillian got her money and went to the store.
Event 6- She bought items to improve her bike, and fixed it up.
Event 7- She trained for the race.
Event 8- During the race, Mark looked back to mock her and crashed.
Event 9- She took 2nd place to an older girl, but decided that 2nd place wasn't bad.

Student Name: _____ Title: Skate-Park Stranger Date:_____Rater:_____

Story Elements:

Setting- Skate park, home

Characters- Miranda, Jamie, other boys

Event 1- Jamie's little sister, Miranda, wanted to skateboard at the skate park.

Event 2- The boys teased her because she is a girl and wouldn't let her skate.

Event 3- Miranda had an idea and went home.

Event 4- She disguised herself as a boy.

Event 5- She returned to the skate-park, she was allowed to skate because the boys didn't recognize her.

Event 6- She performed some difficult skateboard tricks, making them look easy. The boys were impressed.

Event 7- She removed her disguise and the boys recognized her.

Appendix C

Participant Consent Form

Establishing Reliable Assessments of Fifth-grade Students' Oral Retellings
Consent to be a Research Subject

Introduction

The purpose of this research study is to determine how to best use children's oral retellings of what they have read as a measure of their reading comprehension. It is being conducted by L. Elizabeth Shirley Bernfeld, a graduate student at Brigham Young University, and Dr. Timothy G. Morrison, a faculty member at Brigham Young University. You were selected as a participant because you are a fifth-grade student who reads well.

Procedures

You will be asked to read three stories and, following each story, retell it as if you were telling the story to someone who has not read or heard it before. Your retelling will be recorded so the researchers can listen to them again later. The time to read and retell each story will be about 15 minutes, or about 45 minutes total. We will ask you to read and retell at your school, where your teacher will release you to participate.

Risks

There are no known risks for participation in this study. You may feel a little nervous in the retellings, but you may also enjoy reading three stories you may not have read before.

Benefits

There are no direct benefits to you. However, it is hoped that through your participation, researchers will learn more about how to gain accurate information about students' reading comprehension.

Participation and Confidentiality

Participation in this research is voluntary. You have the right to refuse to participate and the right to withdraw at any time without any jeopardy or penalty. Strict confidentiality will be maintained. No individual identifying information will be disclosed. This means that only the researchers will hear my answers, unless my parents want a copy. Where possible, all identifying references will be removed and replaced by control numbers.

Questions about the Research

If you have any questions regarding this research project, you may contact :
Liz (Shirley) Bernfeld, Deerfield Elementary
4353 W. Harvey Blvd., Cedar Hills, Utah 84062
(801)796-3141

Questions about your Rights as Research Participants

If you have questions regarding your rights as a research participant, you may contact:
Dr. Christopher Dromey, Chair of the BYU Institutional Review Board
(801)422-6461
Christopher_dromey@byu.edu

I have read, understood, and received a copy of the above consent. Now that I know about the study and what it means, I agree to be part of the study. I know that I don't have to be part of the study if I don't want to, and I can quit at any time without any penalty.

Student name _____ Date _____

Student Signature _____

I hereby give my voluntary consent for my child to participate in this research study.

Parent/Guardian Name _____ Date _____

Parent/Guardian Signature _____

I certify that this study and the procedures involved have been explained to _____ in terms he/she could understand and that he/she freely assented to participate in this study.

Signature of Person Obtaining Consent _____ Date/Time _____