1990

The Generalizability of Teachers' Judgements of Developing Story Structure in Children 4- to 6- Years of Age Within Three Pedagogical Beliefs of Retelling

Josephine C. Logan-Woods
Loyola University Chicago

THE GENERALIZABILITY OF TEACHERS' JUDGMENTS OF DEVELOPING STORY STRUCTURE IN CHILDREN 4- TO 6-YEARS OF AGE WITHIN THREE PEDAGOGICAL BELIEFS OF RETELLING

by

Josephine C. Logan-Woods

A Dissertation Submitted to the Faculty of the School of
Education of Loyola University, Chicago in Partial
Fulfillment of the Requirements for the
Degree of Doctor of Education

May

1990

Josephine C. Logan-Woods

Loyola University, Chicago

THE GENERALIZABILITY OF TEACHERS' JUDGMENTS OF DEVELOPING STORY STRUCTURE IN CHILDREN 4- TO 6-YEARS OF AGE WITHIN THREE PEDAGOGICAL BELIEFS OF RETELLING

ABSTRACT

Generalizability theory was used to investigate the degree to which a set of teachers' judgments generalized to the population of teachers who share the sample teachers' belief system regarding story retelling as an assessment procedure. Three teachers of young African American children in a large, urban, public school formulated judgments regarding their students' story schema. A checklist of eight story structure items was used to stimulate the teachers' judgments during the students' retelling of a familiar story. A generalizability (G) study and a decision (D) study were conducted. For purposes of analysis, teacher was conceptualized as a form of a test and judgments were considered scores on the test.

Indicies of generalizability were computed using variance components. The variance components were estimated from the expected mean squares for a split-plot, components of variance, ANOVA model. It can be generalized that teacher judgments are accurate, consistent and stable across the time of assessment. It can also be generalized that teachers who regard story retelling as an inference-based process tend to have higher means and lower

error variances than do teachers who view retelling as a structure-based process. The generalizability of teacher judgments can be increased if an informal observational checklist is regarded as a fixed facet.

ACKNOWLEDGEMENTS

The author is deeply indebted to Dr. Robert C. Cienkus whose support and belief sustained the author for many years. The author is especially thankful to Dr. Cienkus who revealed the wonders of theory. Sincere gratitude is expressed to Dr. Jack A. Kavanagh for his shared insights into the problems of educational research, insights which continue to challenge the author. The author is also grateful to Dr. Barney M. Berlin who provided opportunities for the author's professional growth.

Sincere gratitude is expressed to the "subjects" of the study who gave so freely of their time. Without these teachers, there would have been no study. Also gratefully acknowledged is the quality of instruction the author experienced in the classes of Drs. Robert Cienkus, Jack Kavanagh, Dianne Schiller, Todd Hoover, and Max Bailey from which grew the conceptualization of this study. The author gratefully acknowledges the encouragement and assistance of J. Burgis, J. Roberts, Y. Wersching, S. Nesbury, Dr. Kay Smith, and Matthew Creighton, S. J., who, each in their own way, enabled the completion of this document.

Most of all, the writer wishes to express her gratitude for the unfailing support, encouragement, and assistance from her family and friends, especially her husband, Marshall, her son, Lyle, and her study-buddies, Norvella Carter and Josephine Seaton Franklin.

VITA

The author, Josephine C. Logan-Woods is the daughter of Maurice and Marie Colombel. She was born October 27, 1932, in Chicago, Illinois.

She attended Chicago public schools from kindergarten through junior college.

Mrs. Logan-Woods received a Bachelor of Science in Music Education from Northern Illinois State Teachers College, May, 1954. She attended Chicago Teachers College for certification for teaching grades 3 - 8 in the Chicago Public Schools. She received the Masters of Science in the teaching of reading, June, 1971, from Chicago State College.

Mrs. Logan-Woods was a classroom teacher in the Chicago Public Schools from 1954 to 1971. From 1971 to 1977, she served as a consultant for ESEA reading labs and tutorial programs. In September of 1977, she was appointed principal of the Shields Elementary School. In August 1979, she was selected to be principal of the Betsy Ross Elementary School where she continues today.

Mrs. Logan-Woods was a 1989 receipient of the Whitman Corporation's Award for Excellence in Educational Management. She is the author of "Class Size: The Bridge Between Diagnostic Information and Actions for School Improvement," Education and Urban Society, February, 1989. She is on the editorial board of Illinois School Research and Development, the journal of the Illinois Association for Supervision and Curriculum Development.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

The need to improve student assessment is acknowledged not only widely in the literature (Haney and Madaus, 1989; Johnston, 1987; Madaus, 1988; Munby, 1982; Nickerson, 1989; Sanders, Hills, Merwin, Trice, and Dianda, 1989; Stiggens, 1988; Student assessment, 1989) but also by state legislatures (e.g. Illinois, Michigan, Georgia,) who have enacted statutes that mandate school district accountability through testing (States, 1989). However, most of the testing done in schools is done by individual teachers with tests they have constructed or selected to evaluate student progress in specific courses and for assigning grades (Nickerson, 1989). Stiggens and Bridgeford (1985) stated that "teachers purposely go beyond test scores and are intent on using observation-based modes of assessment to acquire information for decision making" (p. 272). The influence of this type of testing on student learning may be as great as, or greater than, that of large-scale evaluation programs (Farstrap, 1989/1990; Stiggens, 1988; Tyler, 1986).

A test is a statement of a theory of learning (Sergiovanni, 1989). When a test is selected, also selected is a specific and particular theory of learning. Concern has been raised that assessments of student performance are developmentally and cognitively appropriate (Bredekamp, 1987; Frederiksen, 1984; Nickerson, 1989; Sanders, Hills, Merwin, Trice, and Dianda, 1989; Stiggens, 1989; Tyler, 1986). This concern has been expressed in terms of a need for alternatives to standardized tests, alternatives that reflect the new understandings of the processes in learning and performing the intellectual tasks to be tested (Johnston, 1987; Haney and Madaus, 1989; Neil and Medina, 1989; Nickerson, 1989).

"Story retelling is a recent development of this type of performance assessment that is concerned with the form and quality of student's learning" (Clay, 1986, p. 769). Research has established retelling both as an instructional strategy and as a tool for learning how children use story structure (Applebee, 1978; Blank and Sheldon, 1971; Bower, 1976; Mandler, 1987; Morrow, Gambrell, Kapinus, Marshall, and Mitchell, 1986). The assessment process, according to Schmitt and O'Brien (1986), could be simply the use of "the story grammar as a checklist for assessing children's overall understanding of a story when they have been asked to summarize it" (p. 5).

Although Schmitt and O'Brien (1986) declared story retelling as the state of the art, the use of an observation checklist transforms teacher from decision maker, or user of evaluative information, to teacher-as-assessment instrument, or the generator of evaluative information (Hennerson, Morris, and Fitz-Gibbon, 1987). Involved are teacher judgments--judgments of the presence or absence of predetermined student behavior that gives evidence of the student's story schema, "an idealized internal representation of the parts of a typical story and the relationships among those parts" (Mandler and Johnson, 1977, p. 111). This is an assessment approach that can sometimes be influenced by factors such as domain knowledge, beliefs, and ideological perspectives (Nickerson, 1989). Not known is the dependability of a teacher's judgments when the judgments indicate the absence or presence of a student's cognitive attributes.

Assessment is part of the instructional process (Farstrup, 1989; Nitko, 1989; Tyler, 1986). "The most straightforward assessment (in terms of initial preparation) of the result of the text-reader interaction is a free recall" (Johnston, 1983, p. 54). However, warnings have been issued that "the ease of preparation of this type of measure of student performance is inversely proportional to the ease of interpretation" (Johnston, 1983, p. 54). A test that consists of informal teacher observations requires that the teacher hypothesize a theory of growth and development, generate the test construct, and validate the construct with the coding of performance (Clark, 1980;

Shavelson and Stern, 1981; Stenner, Smith III, and Burdick, 1983). The teacher functions as a qualitative researcher who must systematically search for confirming as well as disconfirming data and analyze negative performance (Kirk and Miller, 1986). Frederiksen (1984) stated that "there are problems of discovering what are the salient aspects of performance in carrying out a particular task and in identifying the cognitive process that it requires" (p.200). It has been argued that a teacher's definition of the situation affects his or her decision (Clark, 1980). Judgments are based on observations and observations entail the recording of the reaction of some entity to some stimulus, even if the only stimulus is the act of measurement. However, "the act of measurement is likely to produce changes" (Calfree, 1983, p. 61). Immediately, issues of reliability (consistency) and validity (completeness) surface.

Story retelling is a structure based process involving inference based responses and observations. Inference based structures assume that all the information for comprehension may not be explicitly stated. A structure based assessment of a student's comprehension of a story requires that the child's recall, or summarization, is processed "according to the expectations for story components specifically constructed or chosen for the occasion" (Schmitt and O'Brien, 1986, p. 2).

However, Stein and Glenn (1979) pointed out that a student may not remember the exact syntactic and semantic aspects of stories. The information may undergo blending, omissions, inventions of new detail, and similar transformations. Mandler and Johnson (1977) found that not all parts of a story are remembered equally well. Bower (1976) found that children who were not aware of story structure told fractured stories with various elements missing, unexplained, or out of sequential order. "It is possible that teachers could base their reconstruction not on what children actually recall, but rather "upon probalistic estimates of what could have occurred . . . ." (Bower, 1976, p. 54).

Research has revealed also issues of criterion-related validity in performance assessments (Stiggens, 1988; Stiggens and Bridgeford, 1985). Stiggens and Bridgeford (1985) investigated (1) patterns of test use, (2) teachers' concerns about assessment and reasons for these concerns, and (3) teachers use of structured performance assessments, focusing specifically on aspects of criterion-related validity. They found that (1) teachers tend to rely heavily on mental record keeping to store and retrieve information on student performance, and, (2) that teacher's concerns about assessment quality increase with grade level: the higher the grade level, the greater the tendency for teachers to write down criteria and inform students of them, plan scoring procedures, define levels of performance, and conduct blind ratings.

Heishema and Sulzby (1985), discovered that (1) the role the investigator takes during a session affected the child/investigator interaction, and (2) the knowledge of rapport, use of encouragements, and amount of wait time affected the child's degree of participation and the success of the child/investigator interaction. They also noted that the older children tended to suppress knowledge of the story because they perceived the situation as a test of the story. Beagles-Roos and Gat (1983) noted similar responses of students to the technique.

Morrow (1984) concluded that during a retelling, the child is actively engaged in an interaction with the teacher; Sulzby (1982) determined that the teacher's model of literacy had important implications; and Shavelson (1976) asserted that "evidence existed that the teacher may conflict perceptions of teaching ability with input of student behavior" (p.401).

The potential for inconsistencies in judgments exists given the teacher's beliefs about themselves and their roles as teachers, and beliefs about the aims of education and how to achieve them (Shavelson, 1976). Rumelhart (1981) noted the following about the assessment of story schema:

> Once we have determined that a particular schema accounts for some event we may not be able to determine which aspects of our beliefs are based on direct sensory information and which are merely consequences of our interpretations (p. 10).

In fact, during an observation, a teacher can label a student's performance as incorrect or substandard when it is really only a normal variation (Neil and Medina, 1989; Wiggins, 1989). Johnston (1987) articulated the problem with a problematical statement: "In this sense, when we are talking about refining the assessment instrument, making them more valid and reliable, we should be talking about teacher education, since the teacher is the instrument" (p. 350).

The task of explaining even so long-recognized an ability as teacher judgment-making in terms of a theory of information processing or other advanced concepts has barely begun. Bejar (1984), in a broad discussion of approaches to diagnosis, indicated that issues of knowledge representation and issues of representing the teacher must be addressed. According to Bejar (1984) "issues of knowledge representation . . . would parallel the content perspective, and issues of representing the [teacher] . . . would parallel the response consistency perspective" (p. 10).

Little research exists that investigated the dependability of teacher judgments when coding student performance. In fact, "almost no attention has been given to the nature or quality of observational assessment methods in the classroom" (Stiggens and Bridgeford, 1985, p. 271). Johnston (1987) pointed out that "reseachers have until recently come up with developmental models of literacy. . . ., it has been difficult to supply research-based help with informal observations . . . ." (p.350).

This lack in the research also has been blamed on the use of classical test theory as the framework for research in measurement (Brennan, 1983; Kane 1986; Hopkins, 1984; Shavelson, Webb, and Rowley, 1989). Classical test theory is customarily used to distinguish between persons, however, "it cannot estimate the separate sources of error simultaneously" (Shavelson, Webb, and Rowley, 1989, p. 924). As stated by Brennan (1989):

> Classical test theory postulates that an observed measurement can be decomposed into a 'true' score and a single source of random 'error'. As such, any single application of the classical test theory model cannot clearly differentiate among multiple sources of error" (p. 1).

"Generalizability theory recognizes multiple sources of error, estimates each source separately, and provides a mechanism for optimizing the reliability" (Shavelson, Webb, and Rowley, 1989, p. 923).

Researchers have been investigating the mental processes that are involved in teacher decision making, judgments, and planning. Organized around two models ---information processing and decision-making--- the researchers (Borko, Cone, Russo, and Shavelson, 1979;

Byers and Evans, 1980; Clark and Peterson, 1986; Clark and Yinger, 1979; Peterson, Marx, and Clark, 1978; Shavelson, 1976; Shavelson, Cadwell, and Izu, 1977; Shavelson and Stern, 1981) have concluded that teachers integrate the large amount of information about students, teaching, and learning into judgments about the student's affective and behavioral states. This new information is used to make judgments. These same researchers have acknowledged the limitations of statistical models for purposes of determining the influence of teachers' thinking on preinstructional decisions. Yet, as articulated by Shavelson (1976), statistical models can serve as "an heuristic for examining teacher' decision making" (p.376). "A major contribution of generalizability theory is that it allows the researcher to pinpoint the sources of error (e.g., rater, occasion, or both". . . (Shavelson, Webb, and Burstein, 1986, p. 66).

Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 5). Such theoretical orientations, like belief systems and philosophical principles, serve to guide teachers when establishing expectations about pupil behaviors as well as other decisions related to classroom practices (House, Mathison and McTaggart, 1989; Munby, 1982; Smith and Shephard, 1989).

As a result, it has been suggested that measurement concerns include validity that comes from the inferences that teachers themselves draw from their own experiences because those inferences are primary influences on practices (House, Mathison and McTaggart, 1989). Consequently, if retellings are to be used for planning and implementing appropriate programs, the first question should be: How good is teacher judgment? Anderson (1985) would argue that a good interpretation "is complete and consistent" (p. 373). Completion implies congruency with theory, theory that is the result of the teacher's construction of reality, or, in other words, a theory that is the result of pedagogical concerns. Consistency, when the teacher is the measurement instrument, is a set of judgments that are in agreement with his/her beliefs about the uses of retelling as an indicator of a child's story schema.

TERMS

The purpose of this study is to determine the generalizability--the dependability--of teacher judgments of young children's story schema. Its framework is built on elements from schema theory, information theory, measurement theory, and generalizability theory. To contribute to the discussion, the following terms and concepts are defined.

Story schema. The knowledge that people have about how stories may be constructed is called a story schema (Page and Stewart, 1985; Poulsen, Kintsch, Kintsch, and premack, 1979; Rumelhart, 1981, 1977, 1975). The term "story schema" is used by story grammar researchers to (1) refer to an idealized internal representation of the parts of a typical story (Mandler and Johnson, 1977), (2) the relationships among those parts (Kintsch and Kozminsky, 1977), and (3) a set of expectations about the internal structure of stories which serves to facilitate both encoding and retrieval (Griffith, Ripich and Dastoli, 1986; Stein and Glenn, 1979).

Story grammar. Simple stories have a very definite structure identified in the research as its grammar. The grammar of a story "is a set of rules about how certain story structures combine in patterns to form a meaningful story [and] is similar to sentence grammar, which is a set of rules about how words combine in a pattern to form a meaningful sentence" (McGee, 1981, p. 428). Stein and Glenn (1979) viewed story grammar as a set of rules that describe how a story can be broken down into units and how these units are related to one another.

Pedagogy. Pedagogy is understood to refer to the discipline of education (Kindsvatter, Wilen, and Ishler, 1988). "Pedagogy, . . ., is defined as an extensive body of knowledge that incorporates those principles and practices that have been validated by research and scholarly scrutiny, or by the teachers' theoretical beliefs. . . ." (Kindsvatter, Wilen and Ishler, 1988, p. xx).

Judgment. Shulman and Elstein (1975) concluded that a judgment may be defined as the evaluation or categorizing of an object of thought. Tenbrink (1974) asserted that "this logically differentiates from productive thought in that nothing is produced. The material is merely judged; i.e., put into one category or another" (p. 14). "Judgments, unlike decisions, do not call for action. Instead, they are estimates of present condition or predictions of future problems" (Tenbrink, 1974, p. 1).

Information theory. Information theory holds that cognitive processes are involved in "the development of internal representations of problems, the organization of information in long-term memory for efficient retrieval, the acquisition of pattern cognition and automatic-processing skills, use of strategic and heuristic procedures in problem solving, and how to compensate for the limited capacity in working memory" (Frederiksen, 1984, p. 200). In order to handle an information overload, human beings integrate information into judgments about the other's affective and behavioral states using heuristics to formulate judgments (Shulman and Elstein, 1975).

Validity. It can be concluded that teacher judgment is theoretically "a construct and such, is a complex product of many determinants containing random error and systematic bias due to irrelevent components" (Cook and Campbell, 1979, p. 14). Therefore, in addition to statistical conclusion validity (variability and sampling error in units (c.f., Cook and Campbell, 1979), internal validity, construct validity, theoretical and criterion-related validity are of equal importance. Internal validity, the "sine qua non" of any study, rests "on the construct validity of the treatment implementation and measures" (Cook and Campbell, 1979, p. 14). Duffy and Anderson (1984) asserted that teacher judgments cannot be investigated in the laboratory; questions of ecological validity [or the multiple correlation between the cues and the judgments of teachers (Kamil,1984)], have surfaced. "That is, the observed differences in performances are correlated with, but not necessarily caused by differences in the treatment implementation and measures" (Kamil, 1984, p. 48). Kirk and Miller (1986) stated that in qualitative research "the issue of validity is a fundamental problem of theory" (p.21) . . . and "theoretical validity underlies discussions of both apparent and instrumental validity" (p.23). In addition, House, Mathison and McTaggart (1989) argued that the validity of the practitioner's "causal knowledge is critical . . . so far as the conduct and improvement of professional practices are concerned" (p.15).

Pedagogical belief system. According to Smith and Shephard (1989), a teacher's pedagogical belief system is:

> A theory of development and early learning that a teacher holds to be true, the actions that are necessary to promote learning with a degree of credulity,

> how learning is represented, and how that representation facilitates the use of knowledge in particular ways, in relation to other beliefs, values, and emotional attitudes, and in light of what consequences such a belief has in any actions taken (p. 307).

Irwin and Mitchell (1988) identified three theoretical orientations to, or pedagogical belief systems about, retelling. Retelling, according to Irwin and Mitchell, can indicate (1) comprehension of text information; (2) metacognitive awareness, strategy use, and involvement with text; and (3) facility with language and language development. A pedagogical belief system could be anyone of the positions defined by Irwin and Mitchell, or, the combination of any two at any degree of acceptance, all three at a moderate degree, or all three at a degree of full acceptance of each.

Reliability. Reliability in generalizability theory depends upon how accurately observed scores permit generalizations about a person's behavior in a defined universe of situations. Qualitative researchers (Kirk and Miller, 1986) assert that reliability depends essentially on explicitly described observational procedures. Recent research has enabled scoring of retellings in a fairly consistent way (Johnston, 1983). However, "accurate assessment depends on accurate analysis" (Morrow, 1988, p. 129). Research has revealed that when judges are asked to explain the variables considered in making a judgment, "they typically believe they make use of more variables than they actually do.

Reliability, or accuracy of judgments, then --like validity-- is meaningful only by reference to some theory" (Shulman and Elstein, 1975, p. 17).

The limitations in the ability of teachers to process all information in their environment has been accounted for by theorists (Gage and Needles, 1989; House, Mathison and McTaggert, 1989; Rumelhart, 1981; Schmitt and O'Brien, 1986); researchers (Hennerson, Morris, and Fitz-Gibbon, 1987; Pfaffenberger, 1988; Shavelson and Stern, 1981) and measurement specialists (Fine and Sandstrom, 1988; Kirk and Miller, 1986; Thorndike, 1973). Researchers (Deford, 1985; Kindsvatter, Wilen, and Ishler, 1988; Munby, 1982) have asserted that a well formed belief system is the most credible basis for rational teacher decisions.

Retelling assessments are based on the teacher's observations and professional judgments. Their purpose as an approach to educational assessment is to create a method that is linked to classroom practice and that is valid in terms of the knowledge of the way children learn. They are designed to reveal information regarding student skills and products. The essence of the problem in a study of judgment making lies in the analysis. This was succinctly stated by Hopkins (1984) who said that models must be used that allow for "inferences associated with the dependent variable when it is scores on a test or inventory." (p. 704). Generalizability theory recognizes these validity concerns and enables the researcher to address each one (Crocker and Algina, 1986).

Purposes of the Study

1. To establish that teachers' judgments are generalizable

2. To estimate the dependability of teachers' reconstructions of students' story schemas

3. To determine whether teachers reconstruct students' story retellings in a manner that reveals their stated pedagogical beliefs about retelling

4. To determine whether teachers discern the developmental aspects of story structure

CHAPTER II

REVIEW OF RELATED LITERATURE

Story retelling as a tool for performance assessment is a recent development of the type of informal performance assessment that is concerned with the form and quality of students' learning. Involved are teacher judgments--judgments of the presence or absence of predetermined student behavior that gives evidence of the student's story schema, an internalized version of the structure of a story (Mandler and Johnson, 1977). Such judgment-making renders the teacher to an evaluation instrument. Evaluation instruments are theory statements (Sergiovanni, 1989), and so, the act of measuring occurs in a theoretical context that influences judgment making. One of earliest references to an influencing variable within the context of informal performance assessments is in the scriptures where Jesus is quoted as having said:

> "First cast out the beam out of thine own eye, and then shalt thou see clearly to cast out the mote out of thy brother's eye" (Matthew 7:1-5).

Classical test theory would suggest that the "beam" is undifferentiated sources of measurement error, a component of any test score. Research by Shavelson (1976) and his colleagues in teacher decision-making (Borko, Cone, Russo, and Shavelson, 1979; Shavelson, Cadwell, and Izu, 1977; Shavelson and Russo, 1977; Shavelson and Stern, 1981; Shavelson, Webb, and Berstein, 1986) has compiled evidence that the "beam" can be identified, and thus, the dependability and the validity of teachers' judgments can be determined. The "beam," according to these researchers, might be the teachers' own theories of teaching or beliefs about teaching (Shavelson, Cadwell, and Izu, 1977).

The primary purpose of this review of literature is to bring together the findings of two separate and distinct bodies of research, teacher decision making and story grammar, that impact upon judgment making when story retelling is used as an informal indicator of student performance. The major conclusion of this review is stated in the words of Morris, Fitz-Gibbon and Lindheim (1987): "The outcomes that underlie [a] test should be of high priority" (p. 24).

This review is divided into three major sections: (1) the theoretical, research using judgments as data; (2) the practical, research of influences of teachers' judgments, and (3) the problematical, aspects of using retelling as a performance test. Within each section the decision rules for research inclusion are delineated, the studies are reviewed, and a table outlining the salient aspect of the studies is presented.

The Theoretical: Research Using Judgments as Data

Stiggens and Bridgeford (1985) defined performance assessments as the "observation and rating of student behavior and products in contexts when students actually demonstrate proficiency" (p. 273). This raises questions that denote measurement problems, e.g., content validity, knowledge representation, and response consistency. To determine if these assumptions regarding measurement problems when teachers make judgments are supported in the research, a library search was conducted for studies that identified "teacher" as the independent variable. In addition, a set of three determiners for including a study was generated. A study had to have the following listed components to be included in this section of the review.

1. The description of the methodology identified the cognitive processes used by the teacher(s).

2. The study included a description of the domain of knowledged accessed by the teacher in performing the task.

3. The statistical model used provided control for consistency of teacher responses.

Six studies from the body of research on teacher decision making were found that fit the criteria for inclusion. Table 1 outlines and summarizes the findings of the studies. It is organized under three aspects of a perfor-

Table 1

Summary of Research with Teacher as Independent Variable

Knowledge Domain	Cognitive Processes	Response Consistency
Pedagogical processes (Clark,et al., 1979)	Discrimination of the variations in each of eight recitation strategies	Full factorial with teacher as fourth independent variable
Pedagogical Processes (Margolis & Nicholas,1986)	Discrimination of semantic relationships and lexical interpretations in a survey	Teachers choices were used to generate a Likert-type scale.
Role of voluntary reading in raising standardized test scores (Morrow, 1985a)	Discrimination of semantic relationships and lexical interpretations	Full factorial design with teacher as third indep. variable
Pedagogy of higher learning (Peterson & Comeaux,1987)	Discrimination of classroom teach/ learning strategies at the higher levels	Triangulation: 1)free recall 2)video taped 3)structured interviews
Validity of various sources of data used when grouping children for instruction (Shake, 1986)	Ranking, ordering weighing one bit of information against another; discerning differences in tasks	Production data was compared against response data using a contingency table denoting percentages across four categories and four grades
Domain of standardized test items; personal knowledge of student's ability (Coladarci, 1986)	Discrimination of students' academic peformance	Correlation; r's ranging from .74 - .77

mance test: (1) knowledge necessary to perform the task (knowledge domain), (2) the salient aspects of the task (cognitive processes), 3) consistency in response (control for external influences).

Shake (1986) found significant differences in a study to determine the impact of a defined task. First, she asked teachers to generate a list, in order of importance, of four sources of data used when grouping children for reading. Then she compared the results to data gathered earlier when the teachers responded to the importance of four sources of data used when grouping children for reading. Shake found that knowledge of the task had greater importance when teachers' generated, or produced, the responses. Consistency across grade levels and within grade levels was determined using a continguency table denoting percentages in each category.

Clark, Marx, Stayrook, Gage, Peterson and Winne (1979) established task knowledge as a variable when, using a full factorial design with teacher as the fourth independent variable, they examined teacher performance with different groups of students. These researchers were able to denote teacher discrimination of the variations in each of eight recitation strategies. Yet, the researchers questioned their results on an ecological basis: "There's no way of knowing how the results would have differed if the balance between experimental control and representativeness had been different" (p. 550).

The influence of "knowledge" on teachers' judgments was revealed in a study by Peterson and Comeaux (1987) in which response consistency was established with the use of a video tape of teacher performance with 1) subsequent teacher verbal reactions and 2) structured interviews. Peterson and Comeaux (1987) found that a difference in teaching performance was related to judgments made regarding the utilization of underlying principles of classroom learning and teaching.

Margolis and Nicholas (1986) investigated teachers' perceptions of influences on choice of reading material. They asked teachers to (1) identify factors teachers perceived as influencing choices of reading methods and strategies, and (2) to identify on a Likert-type scale the same factors as having positive, negative or neutral effects on their choices of reading materials and methods. Teachers' responses to this task required, in addition to being in touch with their emotions, knowledge of (1) the semantic relationships inherent in the two instruments, (2) implied pedagogical processes, and (3) classroom management strategies.

Morrow (1985a) surveyed the "attitudes of teachers, principals and parents toward promoting voluntary reading in the elementary classroom as a type of reading activity to promote improved standarized test scores" (p. 116). Teachers in the study who ranked development of voluntary reading as most important revealed knowledge of the role of literacy events in the realization of desired standarized reading scores.

Shavelson and Stern (1981), in their review of four laboratory and classroom studies of judgments and diagnosis regarding reading, concluded that the interaction between the teacher and the student is determined by the teacher's memory and strategy.

The Practical: Influences on Teachers' Decisions

This section explores the notion that theories of teaching and learning influence teacher judgments. Shavelson (1976) developed a cognitive model of teachers' judgments and pedagogical decisions as a heuristic for organizing and conducting research on teaching. "The model posits that teaching is a process by which teachers make reasonable decisions with the intent of optimizing student outcomes" (Shavelson and Stern, 1981, p. 471). The model assumed that "teachers have available a large amount of information about their students from many sources. . . . In order to handle the information overload, teachers integrate this information into judgments about the students' cognitive, affective, and behavioral states. These judgments, in turn, are used in making pedagogical decisions" (Shavelson and Stern, 1981, pp. 471-473).

Shavelson, Cadwell, and Izu (1977) investigated the sensitivity of teachers to the reliablility of information in making judgments. Teachers were assessed at two different times and, in each instance, decisions "were influenced by other factors not measured" in the study (Shavelson, Cadwell, and Izu, 1977, p. 95). One possible explanation for this finding was given by the authors: The subjects own theories of teaching or beliefs about teaching "may have been the overriding factor in the decision" (Shavelson, Cadwell, and Izu, 1977, p. 95).

Duffy (1982), as a challenge to laboratory investigations of teacher decision making, asserted that "there's more to instructional decision making in reading than the 'empty classroom'" (p. 295). Byers and Evans (1980), using the lens modeling technique, established a procedure for determining response consistency, or the correlation between the research findings and teacher judgments where both the findings and the judgments are dichotomized variables. They sought to assess the judgmental accuracy of teachers as they predicted the reading interest for their students from the clues provided. They concluded that "teachers are highly individualized in the judgment patterns for this task . . ." (Byers and Evans, 1980, p. 16).

The research reported below advances the idea that teachers have identifiable belief systems that influence their practices. Table 2 summarizes the methodology used in the studies and the presence of evidence either for a belief system and/or the influence of a belief system.

Table 2

Summary of Research of Influence of Teacher Belief Systems

Study	Methodology		Evidence for	
	Qualitative	Quantitative	Belief system	Influence of belief
Deford(1985)	X		X	
Duffy & Anderson (1986)	X		X	
Kinzer & Carrick (1986)	X		X	X
Mangano & Allen (1986)	X		X	X
Moore (1986)	X		X	X
Powell (1986)	X		X	
Rupley & Logan (1986)		X		X
Smith & Shephard (1988)	X		X	X
Taylor & Garcia (1987)	X		X	X

Duffy and Anderson (1984) investigated the beliefs and conceptions about reading held by classroom teachers when measured outside the classroom context and when actually teaching. Using a 45-item researcher designed instrument focused on two sets of beliefs, content centered or pupil centered, the researchers found that older, more experienced teachers tended towards more content centered conceptions, while younger, less experienced teachers tended toward pupil centered conceptions (natural language, interest, and integrated curriculum models. Thirty-seven teachers responded to the instrument which was designed to clarify their beliefs about reading. They then observed eight of the 37 teachers who manifested clear and categorical conceptions of reading. They found that four teachers employed practices which directly reflected their beliefs and four teachers exhibited practices that departed from their belief systems. Data from the researchers' follow-up three year field study revealed that observed teachers possessed a variety of beliefs about reading. The researchers concluded that the relation between teacher practices and their reported beliefs about how reading takes place were not strong.

However, according to DeFord (1985), who validated the construct of theoretical orientation in reading instruction, "teachers evaluate information in terms of their theoretical orientation and then act accordingly" (pg. 352). She found that when asked to perform a set of tasks, teachers of known orientation (established by having first chosen statements that they agreed with) exhibited response patterns more similar to one another than to teachers with other known orientations.

Mangano and Allen (1986) investigated the impact of theoretical orientation about language arts on instructional practices. They questioned if teachers' belief systems about language arts influenced teacher-pupil interactions. Two teachers, one skills-oriented and other whole language-oriented, responded to a structured interview, kept journal records and allowed observers to record interaction patterns during their writing instruction. They found that teachers and pupils appear to interact differently based on teachers' beliefs about language arts instruction.

Rupley and Logan, in separate studies (1986; 1985), explored the relationships between teacher reading beliefs and their reported use of questioning and engagement strategies. Two theoretically oriented scenarios were used to determine teacher beliefs. One was student-centered and the other was content centered. It was hypothesized that teachers who hold a student-centered belief about reading would most likely not perceive themselves as being the center of control in maintaining student engagement, but would view student's interests as a major factor that would encourage students' engagement in reading. A significant negative correlation was found. Teachers who held content oriented beliefs were not likely to be associated with engagement and questioning strategies (Rupley and Logan, 1986); and teachers who held student centered beliefs were not likely to value instruction focused on decoding (Rupley and Logan, 1985). The authors concluded that the lack of a significant relationship between content-centered reading beliefs and engagement scenerios based on teacher effectiveness research reflected the conflicts between teacher beliefs and "strategies that are not representative of their teaching" (Rupley and Logan, 1986, p. 168). Another confict noted by the authors was "regarding the nonvariance in instructional strategies for content material and the effective teaching research which would require that teachers give attention to individual student's needs and adjust their instruction appropriately" (Rupley and Logan, 1986, p. 168).

> Teachers who are student-centered are not likely to be associated with engagement and questioning strategies that are teacher directed. However, teachers who hold content-centered beliefs do not appear to be associated with such strategies either. It maybe that neither group of teachers is familiar with such strategies and, as a result, are not associated with their use in reading instruction. (Rupley and Logan, 1986, p. 169).

Taylor and Garcia (1987) studied "three teachers: what they said and what they did" (p. 17). The teachers wrote definitions of reading and were taped while teaching. Miscue feedback and stated beliefs were examined to determine if teacher's feedback to miscues matched their stated beliefs about the reading process. The authors reported that two of the teachers gave feedback to miscues that matched their stated beliefs while the third teacher's feedback was inconsistent with stated beliefs.

Moore (1986) conducted a study that compared reading education students' instructional beliefs and instructional practices. Instructional beliefs categories were goals for self, goals for children, and how the reading process works. Instructional practices categories were (1) practices planned for future use, (2) those which had been observed, (3) those which they had implemented, and (4) those that were hypothetical or ideal instructional practices. "When belief and practice statements were congruent, there was determined to be a point of integration. However, when there was incon- gruence between stated beliefs and practices, there was determined to be a point of conflict" (Moore, 1986, p. 145). With the exception of one student, all students showed the same point of conflict: "they frequently described instructional practices which could only be classed as isolated skills-based instruction" (Moore, 1986, p. 145). However, there were "several points of integration between instructional beliefs and instructional practices both within each subjects' journal and across all eight journals" (Moore, 1986, p. 145). The most commonly held points of integration across all eight journals were three concepts: (1) learning should be interesting and motivating, (2) learning should be relevant, and (3) children should be actively engaged in the learning process. Moore (1986) concluded that the students lacking consistency in beliefs and practices "had difficulty articulating justifications; they confessed to being confused and wondered how to reconcile the discrepancies" (p. 146).

Clinical interviews with teachers, participant observations, analysis of documents, and interviews with parents revealed to Smith and Shephard (1988) that teachers' beliefs about the development of school readiness could be described and ordered within seven categories related to beliefs about the nature of child development. Smith and Shephard (1988) were concerned with teachers' beliefs regarding development and early learning and philosophy of retention. They found that teacher belief systems influenced teacher retention policies that ranged from the teacher could intervene and promote learning to a posture that held learning had to wait for development to occur.

Kinzer and Carrick (1986) investigated teacher beliefs about how reading takes place and develops. First, teachers were asked to chose five statements from fifteen that were more valid and important for a teacher to know about reading development. Second, teachers were asked to read three sets of lesson plans reflecting three views of how reading takes place and develops. The researchers used a Chi-square analysis and found significant statistical relationships between teacher beliefs and their choices of lesson plans. The results indicated that the teachers had different belief systems and that teachers are more influenced by practical considerations (how reading develops) than by theoretical considerations (how reading takes place). However, Kinzer and Carrick (1986) pointed out that theoretical aspects of teacher's belief systems may be more influential when interpreting student responses, interpreting a miscue, or when interpreting the acceptability of a student's response to a comprehension question.

Powell (1986) conducted a unique study of teacher belief systems and pedagogy. The study was unique in that the sample consisted of two school districts. Each district had well-developed, written text book selection procedures. Teachers were surveyed to determine their textbook selection policies. This study revealed that (1) pedagogy, along with publishers politics and the people involved, (2) pilot tryouts of text book programs, and (3) the physical appearance of the program were the major perceived influences. Within the pedagogy category "primary teachers' beliefs about the early stages of reading were influential. . . ," (Powell, 1986, p. 150-151).

Ray, Lee, and Stansell (1986) predicted that commitment to theory and help in implementation of the indicated processes would bring about changes in practices. They succinctly stated their findings as follows:

> The teacher in the study had not implemented changes where they counted -- in her theory of . . . instruction . . . . Finally we understood that the theory we thought she had discarded was very much intact and still governing her teaching (p. 158).

As summed by the researchers, "teachers may, in fact, include myriad new techniques without altering their underlying theoretical orientation at all" (pg. 154).

The Problematical: Review of Research Using Retelling

The purpose of this section is not to review the large body of story grammar research, but to focus on the problems of validity and of reliability when the teacher functions as the test instrument. First, retelling as a technique for conducting informal performance assesments of a student's comprehension of text information is described. Second, a theoretical perspective, shaped by research findings, is presented. Next, the issue of knowledge representation when story retelling is used for evaluating students' story schema is addressed. This research is outlined in a table and is briefly referenced.

Story Retelling

Story retelling depends upon an internalized grammar, or the structure, of simple stories.

> In story retelling, . . . an individual recalls orally a story after having listened to it . . . . Retelling for assessment is carried out without prompts, props, or use of text/story. The text/story [is] not . . . discussed with the child after reading/listening and before retelling. . . . (Morrow, 1988, p. 128).

The child is simply asked to retell the story/test as if telling it to a friend who has never heard it before.

In a quantitative assessment, the readers or listeners are told to retell all they can remember from the text.

> Prior to retelling, the teacher will have parsed the story/text into units to be assessed (e.g., propositions, idea units, elements of story structure). The protocol of the reader's listener's retelling is then parsed into identical units and compared with the text units. The match between protocol units and text units represents the reader's/listener's comprehension score. (Morrow, 1988, p. 131).

Theoretical Perspective of Story Retelling

Several assumptions have been made concerning the analysis of stories. The first assumption is that story material has some kind of internal structure much like sentences, a schema (Rumelhart, 1975). The second assumption is that stories can be described in terms of a hierarchical network of categories and the logical relations that exist between these categories (Mandler and Johnson, 1977; Stein and Glenn, 1979; Rumelhart, 1975). It is further assumed that this network constitutes a schema that is used to process information concerning stories. Research has suggested that "the story framework--the set of basic constituents, or its grammar, or syntax--serves as a set of retrieval cues to prompt recall of the items of information filling their slots" (Bower, 1976, p. 523). However, more recent research (Mandler, 1989) has revealed that the network consists of only three elements: goal paths, episodes, and attempts.

Additional research conducted with the listener as the focus has led to the conclusion that there exists at least two listener organizations for stories (Page and Stewart, 1985). The first is structure based and is concerned with the stereotypical rules of a story or of the events that occur in a story. These rules, or story grammar, are the elements which describe the structural units of a story and the order in which these units typically appear (Stein and Glenn, 1979; Mandler and Johnson, 1977; Thorndyke, 1977; Rumelhart, 1975).

The second listener organization can be described as problem-solving and inference-based. The listener can focus (1) on a propositional analysis (Applebee, 1978; Kintsch and Kozminsky, 1977); (2) on the manner in which different parts of the story are tied to preceeding or subsequent parts (Black and Bower, 1980; and/or (3) on the series of actions a main character must complete in order to attain a goal (Black and Bower, 1980; Griffith, Ripich and Dastoli, 1986). How these events or actions are tied to the preceding or subsequent parts, the cohesion of the story, has been identified as the salient feature of story comprehension

(Griffith, et al., 1986; Stein and Glenn, 1979). The salient features of story comprehension are outlined in Table 3.

Text information. The studies outlined in Table 3 revealed that researchers have established that comprehension of text information differs with age. Comprehension was found (1) to increase in complexity with age, and (2), older children recalled more than younger children. Applebee (1978) determined that one-half of the stories told by child- ren at age 5 and one-half years were a focused chain in which characters were not repeated. The chain consisted of "and then . . . " type narratives focused on the continuing adventures of the main character. The move from this primitive type of narrative to a narrative built upon a theme seems to be the greatest in the development of a child's story schema.

Text structure. According to the researchers represented in Table 3, a story grammar consists of a series of constituents representing different levels of generality. The assumption has been made in the research that all of the units and sequences of units identified by story grammar research are reflected in story schemas (Mandler, 1984). It has since been determined that the basic constituents of a working story schema are three: (1) a goal, or a description of an internal state, (2) a setting, or a description of actions, and (3) episodes, a description of the begin-

Table 3

Summary of Story Grammar Research

Study	Comprehension	Language fluency	Self-regulated	Development of text
Applebee (1978)				X
Beagles-Roos&Gat(1983)			X	
Black & Bower (1980)	X			
Blank & Sheldon(1971)		X	X	
Bower (1976)			X	
Brown (1975)	X			
Gambrell,et al.(1985)		X		
Griffith,et al.(1986)	X			X
Grinnell (1984)				X
Mandler&Johnson (1977)				X
Meyer (1984)			X	
Morrow (1978)		X		
Morrow (1984)	X		X	
Morrow (1985b)		X		
Morrow, et al.(1986)			X	
Pickert & Chase (1978)	X	X		
Poulsen,et al.(1979)			X	
Rose, et al. (1984)		X		
Stein & Glenn (1979)	X		X	X
Thorndike (1977)	X			X

ning, the development, and the ending (Mandler, 1984). These three aspects of story schema have been found to have the greatest influence on recall of a story. Applebee (1978) earlier determined that stories based on these constituents made up over one-half of the stories of children at age 5. In addition, the task of retelling a story has been found to present difficulties for the young child. Implicit knowledge has been found to guide process- ing. Summarization and story generation appears to be influenced by consciously accessible knowledge (Mandle, 1987).

It can be argued that the salient features of story comprehension also represent the knowledge domain of a teacher-as-a-measurement instrument for retelling. This knowledge, gleaned from story grammar research, was outlined in Table 3.

Domain of Knowledge in Story Retelling

Language use and development. Retelling has been proposed as an approach to evaluate children's language (Irvin and Mitchell, 1988; Morrow, et al., 1986; Pickert and Chase, 1978). The skills involved are: (1) comprehension (understanding of grammatical forms and vocabulary words), (2) organization (ability to integrate visual and auditory information and to recall sequences of events), and (3) expression (expressing the story in fluent, connected sentences, using correct grammar).

In story retelling, children must be able to organize information and recall it in a logical manner. However, reseach has concluded that:

1) Skills improve with age, but individual differences prevail.
2) The ability to reconstruct a sequence of events also requires organization.
3) Children must first be able to express themselves in fluent, grammatical speech.

Summary

The research reported in this chapter has provided information to support the proposition that a teacher's pedagogical belief system is the key to teacher judgement's when using retelling to assess students' story schema. The task in judgment making regarding a child's story schema is one of placing incoming information in one category or another with implications of the selected category being supported by a particular pedagogical outcome that is determined by the teacher's belief and knowledge of how the teaching/learning process best transpires. In each study reviewed, the teacher was revealed to be a qualitative researcher who had to systematically search for confirming as well as disconfirming data and had to analyze negative performance. If this behavior of discrimination is used as the standard for evaluating the validity of using teacher judgments as data, in can be argued that the standard was (1) established by the findings of qualitative studies, and (2) tested with quantitative analysis procedures. A statement could be made, based on the studies reviewed, that the use of judgments as data is a valid procedure to investigate the influences of teachers' belief systems on their judgments.

An analysis of the studies revealed that, in the area of evaluating student based information, teachers belief systems tend to influence: (1) the interpretation of a miscue, and (2) the acceptability of a student's response to a comprehension question. In addition, teacher belief systems tend to interact with pedagogical decisions.

The requirement [in judgment-making] is for specific information. . . . However, apart from the obvious problem that most [cognitive mediation] processes are not directly observable, there are many constraints which affect the nature of performance assessments" (Johnston, 1983, pp. 40-41).

As outlined in Table 3, the vast body of story grammar research has revealed the knowledge representation when teacher is to be the measuring instrument. A brief summary indicates that children's retellings are not only highly organized (Stein and Glenn, 1979), but also:

1. Denote development in text information (Applebee, 1978; Baggett, 1979; Griffith, et al., 1986; Grinnell,

1984; Mandler and Johnson, 1977; Stein and Glenn, 1979).

2. Represent use of metacognitive strategies to regulate and control thought (Beagles-Roos and Gat, 1983; Blank and Sheldon, 1971; Bower, 1976; Meyer, 1984; Morrow, 1984; Paulsen, Kintsch, Kintsch, and Premack, 1979)

3. Signal a relation between concept of story and reading comprehension (Beagles-Roos and Gat, 1983; Blank and Sheldon, 1971; Bower, 1976; Meyer, 1984)

4. Verify that the extent to which an item is recalled is highly stable over time and between grade levels (Stein and Glenn, 1979).

"The discovery of large variations in accuracy of judgments provides researchers with conditions for learning about how teachers use the information available to them to make judgments and decisions" (Byers and Evans, 1980, p. 3). Training for performance assesment would be highly desirable and would enhance generalizabilty of skills from theory to real life teaching situations. "Knowledge is not free-floating but is situated in activity" (Wineburg, 1989, p. 8). General principles of tests and measurement must be embodied in the coding of students' performance in a retelling. Ultimately, the techniques of retelling, what we call knowledge, will determine "the marketplace of ideas" (Wineburg, 1989, p. 9) and services, as well as determine policies for the performance assessments of young children.

CHAPTER III

METHOD

This chapter will discuss the application of generalizability theory and experimental design theory in a study of teacher-as-a-measurement instrument.

Conceptually, each teacher is perceived as a different form of a test designed to measure the shape and the content of a student's schema for story. At issue are the general problems of reliability and validity of the measurements of mental ability. Two additional issues surface given the use of a checklist to frame the measurements: the adequacy and the stability of the measurements.

Generalizability theory is used to address these issues. It is applicable to testing situations where the concerns are generated by the use of more than one form of a test and where there is more than one testing occasion. Experimental design theory is used to organize the data to be analyzed within the discipline of test theory-- (1) estimating the extent to which these problems influence the measurements taken in a given situation, and (2) devising methods to overcome or minimize these problems (Crocker and Algina, 1986).

As a consequence, the design of this study is two studies in one, a G-study and a D-study. The G-study prepares the measurements to be analyzed in the D-study and establishes that the data are generalizable. The D-study estimates the generalizability of the data-- the extent to which random error has influenced the measurements and the sources of a problem, the problems of reliability and validity.

Generalizability theory identifies the problem of random error in test forms and item formats as the conditions of measurement. It assumes that the conditions are crossed in the universe of admissable observations. Generalizability theory acknowledges the presence of random error with its emphasis on the use of variance components to estimate indicies of generalizability.

The G-study

Sample

The sample consists of three public school teachers who work in a large, urban, elementary school that serves Afro-American children from 3 years of age through the completion of the eighth grade. Each teacher volunteered to participate in this study and each teacher held a different view of the use of retelling. Two are classroom teachers in a child centered preschool program. One taught 4-year old children and was concerned with matters of text recall as revealed in language use. The other classroom teacher taught kindergarten and was exploring the whole language approach. The third teacher used retelling to determine the effectiveness of an ECIA Chapter I program for kindergarden children. Each teacher is experienced with the age groups of the children assessed. Each teacher assessed children individually using a checklist and formulated a judgment as to the absence or presence of eight elements of a story structure. Two teachers assessed twenty-four students each; one teacher assessed twenty-seven students. The age range of the students was 4 years, 10 months to 6 years, 5 months with a median age of 5 years, 11 months.

Methodology

A tape recording was made of the reading of the story "Goldilocks and the Three Bears." The assessment process began with the playing of the recording within the classroom setting. Each child was asked individually to recall the story during which time the teacher indicated on the checklist the presence of eight structural story elements in the child's retelling. A story protocol was prepared for each student and each teacher. A protocol consists of one of two possible judgments for each of eight elements of story structure: 0, not present; 1, present. A protocol is displayed in Table 4. Following the construction of protocols, each teacher completed "The Reader Retelling Profile" (Irwin and Mitchell, 1988), a Likert-type instrument designed to assess teacher pedagogical beliefs regarding story retelling.

Table 4

An Hypothetical Protocol

Story Elements	1	2	3	4	5	6	7	8
Teacher A_j, Child_k	0	1	0	1	0	1	1	0

The "Reader Retelling Profile" (Irwin and Mitchell, 1988), is a self-report instrument of teacher beliefs about retelling. It is intended to provide a qualitative assessment of students story schema.

Universe of Generalizations

In generalizability theory, the interest lies not in the sample but in the measurement conditions, the universe of generalizations. The conditions of measurement are labled as facets. The measurement conditions in the G-study consisted of two facets, the test form and the story structure check list. Each was considered to be a random variable with many sources of error.

Test form. Each teacher is operationalized as a form of a test. Therefore, each test form is a construct which is the product of three determinants. First, each test form represented a specific pedagogical belief system regarding the purpose of a story retelling. Second, each test form was identified on a continuum from a structure-based organization to an inference-organization. Third, each test form contained bias given the knowledge of each child for whom judgments were formed of the presence or absence of an aspect of story schema. Finally, the measurement occured in a classroom structured to reflect a specific theory of growth and development.

Item checklist. The story grammar item checklist represented the expectations of a story structure. However, the checklist represented only one of many checklists that could be constructed from the story grammar item domain. In addition, attending each item is the notion of development. First, the level of development corresponding to a given item may vary from child to child. Second, the task-demands of each item have within each a developmental sequence. Furthermore, a lack of language fluency may inhibit the desired, or expected articulation.

The G-Study Design

The focus of the G-study was to generate the data needed to estimate the variance components for the D-study. Three measurement concerns determined the choice of a design. The first concern was that each teacher was represented systematically as a random variable in the experiment. The second concern centered on the occassions of measurement. The third concern was representing the set of judgments. A split-plot treatment by items ANOVA with judgments nested within teachers was used. This procedure identified seven sources of measurement error and allowed generalizations to the population of judgments.

Available were 600 observations. Using the formula and procedures outlined by Kirk (1982, p. 145) for estimating sample size, the number of "subjects" needed at each of the three treatment levels in a split-plot p.q design (p=.01) was 12. Twelve protocols selected at random from the total number prepared by each teacher were used for judgments. A grand total of 288 judgments (subjects) comprised the study. The chances of correctly rejecting a null hypothesis for N=288 at p=.01 is slightly less than nine in ten.

The strength of a split-plot design lies in the measure of the effect of a constant due to an individual being in a nested condition of measurement. The applicable assumptions of the split plot design were three:

1. There are at least two treatments and each treatment has more than two levels.

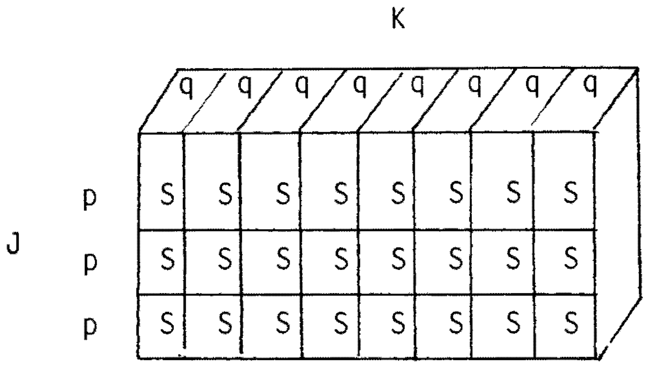
2. The protocols were selected randomly from the set of protocols which were prepared randomly.

3. Each effect is independent of every other effect.

These assumptions were met in this study. Facet P (teacher) was designated as a between-block treatment; facet I (check list) was designated as a within-block treatment. A random effects model of the ANOVA was used to analyze the data. The inferences made from the ANOVA procedure centered on the variance of the population of effects sampled by the researcher. Table 5. displays the design.

Table 5

Type Split-plot SPF-p.q 3.8 Design



Where p =levels of factor J (Teachers), p= 3
q =levels of factor K (Items), q= 8
n =sets of protocols S for each teacher, n= 12

The error terms were obtained using the procedures described by Kirk (1986). They are displayed in Table 6.

Table 6

Error Terms for a Split Plot p.q Design

SV	E(MS)	AET
J	$\sigma_e^2 + \sigma_{S(J)}^2 + nq\sigma_{JK}^2 + nq\sigma_J^2$	MS $_{S(J)}$
S(J)	$\sigma_e^2 + \sigma_{S(J)}^2 + n\sigma_J^2$	
K	$\sigma_e^2 + \sigma_{KS(J)}^2 + np\sigma_{JK}^2 + nq\sigma_K^2$	MS $_{KS(J)}$
JK	$\sigma_e^2 + \sigma_{KS(J)}^2 + n\sigma_{JK}^2$	MS $_{KS(J)}$
KS(J)	$\sigma_e^2 + \sigma_{KS(J)}^2$	

The major goal of the G-study was to determine if the data were generalizable. The goal had two components: (1) the establishment that the population of effects is not zero, and (2) to estimate the variance components needed for the D-study. These procedures enabled the researcher to infer that the data are generalizable.

The hypotheses were:

1. There are no experimental effects due to test forms.
2. There are no experimental effects due to check list.
3. There is no interaction of the forms of the test and the checklist.
4. Student's age does not influence teacher judgments.

Hypotheses 1-3 will be accepted or rejected at the .01 level at the .01 alpha level. Hypothesis number 4 is accepted if the F_{max} statistic is greater than the table value for the F_{max} at .01, 7,14 df.

The D-Study

Sample

The sample of the D-study is the set of estimated variance components.

Methodology

Two sets of equations were defined and developed. The first set is analogous to descriptive statistics, those needed to compute the average score, the expected observed score variance, and the error variance. The second set of equations were those needed to generate appropriate indicies of generalizability in terms of the estimated variance components for each index.

Statistical Design

In the universe of generalization, the n_i conditions of the random facet I (Items) are nested within the n_j conditions of random facet J (test forms). Both facets are crossed. The notation for this design is (j:p)xi which is read conditions of facet I are crossed in judgments and n_j conditions of facet J are nested within each judgment (p). The generalizations are made over the test forms and the occasions of measurement.

Descriptive statistics. The average observed score, the domain score, is the total score/N since the judgments are scored either 0 or 1. The expected observed score variance is the sum of the weighted variances of (1) residual error, (2) judgments in test forms plus test forms, and (3) interaction. The universe score variance is the variance of the judgments. The error variance, or the fluctuation or varying of measures due to chance, is estimated from the variance components of items and error.

Generalizability coefficients. Generalizability theory recognizes that there is potentially a large number of coefficients associated with a particular measurement. "Different generalizability coefficients can be defined according to how broadly or narrowly each decision maker proposes to generalize from a particular application of the measurement procedure" (Shavelson, Webb, and Rowley, 1989, p. 931). The following listed indicies of generalizability were identified in terms of the appropriate variance components for each index.

1. Coefficient of accuracy, the ratio of the universe score variance to the observed score variance. (Crocker and Algina, 1986, p. 159)

2. Coefficient of consistency, the ratio of the difference between the variances for judgments and error to the variance for judgments. (Rowley, 1976, p. 54)

3. Coefficient of construct validity, the ratio of the variance of test forms to the sum of the variances of test forms, interaction of forms and items, and error. (Hayes, 1981, p. 382)

4. Coefficient of criterion validity, the error variance, estimated from the variance components of items and error/ninj. (Crocker and Algina, 1986, p. 176)

5. Coefficient of content validity, the ratio of the variance of test forms to the sum of the variances of test forms and judgments in forms. (Crocker and Algina, 1986, p. 235)

Additional _analysis_. An evaluation of teachers' judgments over time, "steady state behavior" (Shavelson, Webb and Rowley, 1989, p. 926) was conducted. The determination of steady state behavior was derived from the relative size of the variance components for judgments in items and for interaction. In addition, a determination of the percent of variance due to judgments within groups was conducted using the F-max statistic. Of concern was the influence of error on judgments.

Summary

The above described procedures enabled the researcher to determine the degree to which the obtained judgments generalize to the universe of generalizations and to draw inferences regarding: (1) the adequacy of the D-study design, (2) the reliability and validity of teacher as a form of a test, (3) the content validity of a checklist, and (4) other difficulties such as (a) inadequate sample size, (b) a lack of reliability given the use of a checklist, and (c) criterion contamination.

The major goal for the D-study was:

> To determine the generalizability of teachers' judgments of developing story structure in children 4- to 6- years of age.

The major hypothesis was:

> The design will yield adequate generalizability.

Sub-hypotheses

5. The observed judgments do not predict the teachers' domain score.

6. The observed judgments are not consistent given the influence of random error from the measurement conditions.

7. There is no relationship between the content of the item format and the resulting judgments.

8. There is no similarity in judgments within teachers

9. The observed judgments are contaminated given the

influence of random error from the measurement conditions.

10. The observed judgments are not stable over time.

11. The observed judgments do not reflect the influence of the teachers' belief systems.

Hypotheses number 5-10 will be accepted or rejected on the relative size of the value of each generalizability coefficient which can range from .00 to .99. "The coefficients developed are descriptive statistics and do not depend upon any parametric assumptions about the distribution of errors" (Kane and Brennan, 1977, p. 290). Therefore, there are no statistical tests of significance for generalizability coefficients. Where appropriate, the magnitude of a relation will be determined by finding the square root of the obtained value. Hypothesis number 11 will be evaluated using the value of the Chi-square statistic.

Testing Instruments

Checklist for Retelling

The checklist from "Section III, Story Structure" of the "Early School Inventory-- Preliteracy (ESI-P)" (Nurse and McGauvran, 1987) is the first instrument. The ESI-P is designed to determine if (1) a child has acquired a concept of story, and (2) includes the basic story structures in the retelling of a familiar story. It is intended for use by teachers, pre-school through the first grade, as an aid for planning developmentally appropriate literacy experiences.

The specific objectives of the checklist is "to assess a child's ability to retell a familiar story using the conven- tional elements found in a story" (Nurse and McGauvran, 1987, p. 11). In addition to specific directions for administrating the assessment, in the test manual is a text version of "Goldilocks and the Three Bears". The manual for interpretation lists the following performance objectives as score criteria.

1) Beginning: The child can begin a story with "Once upon a time", or "One day," etc.

2) Setting: The child can tell where the story happened.

3) Characters: The child can name significant characters.

4) Sequence: The child can include at least three events in logical sequence.

5) Feelings: The child can describe at least one feeling of a character.

6) Description: The child can use descriptive words at least twice.

7) Conversation: The child can give at least one example of a character speaking.

8) Ending: The child can tell what happened at the end of the story.

The instructions include an alert that (1) the child may state an element that is not included in the list; (2) the content of the specific answer was not of importance, whether or not the retelling included the element was of importance; and 3) a teacher could use his/her own judgment as to whether or not the element is correct for the child's story.

Pedagogical Belief System Survey.

"The Reader Retelling Profile" (Irwin and Mitchell, 1988; permission granted for use) is a Likert-type instrument consisting of twelve propositions of abilities revealed during a child's retelling of a story. It was intended for use as a tool for a qualitative assessment of a story. Its objective is to document, in addition to story structure, the child's behaviors that supplement, provide coherence, and enhance completeness and comprehensibility of a story (Irwin and Mitchell, 1983). The authors provide for a categorical analysis of responses to the twelve propositions. The categories, or positions of beliefs are:

1. Retelling indicates the reader's comprehension of textual information.
2. Retelling indicates metacognitive awareness, strategy use, and involvement with text.
3. Retelling indicates facility with language and language development.

Irwin and Mitchell (1988) provided directions for a categorical analysis of the propositions that could be identified as components of one's pedagogical belief system. However, this instrument is not validated in research.

CHAPTER IV

RESULTS

This chapter discusses the results of the G-study and the D-study. The G-study analyzed the results of the ANOVA and estimated the components of variance. The D-study developed the universe score variance and the coefficients of generalizability for three different universes of generalization: (a) an infinite universe of teacher judgments, (b) an infinite universe of story structure items, and (c) an infinite universe of teachers.

The G-Study

The G-study was concerned with the generalizability of the data. Its purpose was to answer the question: Are the data generalizable. The data were determined to be generalizable because the estimated variance components were neither zero nor negative. The variance components were estimated from computed mean squares. The data were prepared for an analysis of variance procedure. Table 7 displays the means and standard deviations of the teachers' judgments. A random effects, teacher by items split-plot Anova was conducted. The main effects were used to test the hypotheses of no experimental effects.

Table 7

Means and Standard Deviations of Three Teachers' Judgments

	Mean	Standard Deviation
Teacher 1	33	.47
Teacher 2	84	.33
Teacher 3	24	.57

A summary of the procedure is displayed in Table 8.

Table 8

Split-plot Random Effects ANOVA Summary Table

SV	S.S.	df	MS	F
Between	36.125	35	1.032	
Teachers	13.5625	2	6.78125	9.918*
Judg.w.teacher	22.5625	33	.6837	
Within	32.75	252	.12996	
Items	6.0417	7	.8631	7.912*
ItemsXTeachers	2.6042	14	.1860	1.198
ItemsX(Judg.w.Tch)	24.1041	231	.1043	
Total	68.8775	287		

*p $<$.01

The effects were tested for significance. Effects of interaction were not significant, (12, 200) =2.27, $p > .01$. The treatment effects for teachers, F (2, 30) = 5.39, $p < .01$, and items, F (12, 200) = 2.27), $p < .01$, were significant. The hypotheses of no experimental effects due to teacher as a test form (hypothesis number 1) nor due to the check list (hypothesis number 3) were rejected at the .01 level. The hypothesis of no interaction of teacher judgments and the check list was accepted at the .01 level.

The homogeneity of the variances of teachers' judgments was tested with the F_{max} statistic. Table 9 displays the summary of the partition of the variance of judgments within teachers. The hypotheses of no treatment effects were rejected.

Table 9

Partition of Variance Within Teachers

SV	SS	df	MS
Judgments w.Tch.	22.56	p(n-1) = 33	.6837
Teacher 1	11.03	n-1 = 11	1.0028
Teacher 2	4.0	n-1 = 11	.3636
Teacher 3	7.53	n-1 = 11	.6849

The table values of F_{max} for 3 variances with 10 and

12 degrees of freedom are 5.85 and 4.91 respectively. The value of F_{max} obtained, 2.76, was less than the surrounding F_{max} tabled values. The hypothesis of homogeneity of variances was accepted. It was inferred that the age of the student did not infuence the judgments of the teacher. Hypothesis number 4 was accepted.

Equations based on the random effects ANOVA were defined for numerical estimates of the variance components. The equations and the value of the estimated variance components are displayed in Table 10.

Table 10

Estimated Variance Components for a (j:p)xi Design

SV	Estimated Variance Components	
Occasions (P)	σ^2_p =(MSbetween + MSwithin)/Sn_i	.009
Teachers (J)	σ^2_j =(MS_J - MS_{IJ} - $MS_{J:P}$)/$n_i n_j$	.246
Tch.w.Occ. (J:P)	$\sigma^2_{j:p}$ =(MS_{IP} + MS_J)/Sn_i	.077
Items (I)	σ^2_i =(MS_I - MS_{IP})/$n_i n_j$	.028
ItemsXtch's (IJ)	σ^2_{ij} =(MS_{IJ} - MS_{IP})/n_j	.027
I(J:P) (IP)	σ^2_{ip} =(MS_r)	.104

The equations in Table 10 are based on the mean squares of the split plot p-q design and the sample sizes from the G-study (Crocker and Algina, 1986).

The D-Study

The domain score, the universe score variance, and related descriptive statistice were developed. A set of equations were defined to compute the estimates of the average score, the observed score variance, the error variance, and the standard error of measurement. A summary of the descriptive statistics is presented in Table 11.

Table 11

Descriptive Statistics for the Generalizability of Teachers' Judgments

Universe of Generalization	
Domain score	.604
Universe score variance	.2518
Observed score variance	.2896
Error variance	.0302
SEM	.1738

As displayed in Table 11, in the universe of generalization, the domain score is the population score. The domain score is the proportion correct in the sample. The universe score variance is an estimate of the variance of judgments in the population. The observed score variance has as its counterpart the standard deviation of the judgments of teachers in the sample. That part of the observed score variance which is not universe score variance is the error variance in the population. It is what remains when all other sources of variance have been removed from the observed score variance. The standard error of measurement was used to form the confidence interval. The probability is .95 that the interval .264 to .945 includes all possible sample means of judgments in the population of teachers defined by the sample teachers.

Indicies of Generalizability

Using the estimated variance components and the sample sizes from the G-study, the estimated values of the indicies of generalizability were computed. These indicies are a function of the value of "rho," an estimate of the proportion of variance of the dependent variable (teachers' judgments) due to the presumed influences of the independent variables (test forms or the eight items of the checklist.) Table 12 displays the estimated rho values.

Coefficient of accuracy. The first rho coefficient computed was the coefficient of accuracy, an index of reliability. This coefficient is an estimate of how well the domain score generalizes to the universe score. The coefficient of accuracy is the ratio of the universe score variance to the observed score variance; its value was estimated to be .91. It can be inferred that all three sets of teacher judgments are highly reliable with 95 percent of the observed score variance of any one set of judgments predicted from any other set of judgments. It can also be inferred that 91 percent of the estimated observed score variance of teacher judgments will be attributable to variations in the judgments around the true judgement. Hypothesis number 5, which stated that the domain score cannot be predicted from the sample, is rejected.

Table 12

Estimated Rho Values of Coefficients of Generalizability

Coefficient	Rho
Coefficient of accuracy (p_i^2)	.91
Coefficient of consistency (p_{XX}^2)	.89
Coefficient of content validity (p)	.82
Construct validity (p_i)	.99

Coefficient of consistency. The coefficient of consistency is concerned with the influence of random, uncontrolled error on teachers' judgments. It is an index of reliability and provides an estimate of the lower bound of the percent of variance attributable to true score variance. This coefficient, p^2_{XX}, was estimated to have the value of .89. It is interpreted according to the value of its square root, .94. It can be inferred that at least 94 percent of the variance of teacher judgments is attributable to differences in the true judgments of the teacher. The hypothesis that the observed judgments are not consistent given the influence of random error is rejected (hypothesis number 6).

Coefficient of content validity. The validity of the content was tested within the theory that the item format is a different method of measurement. The value of rho was estimated at .82. It can be inferred that eighty-two percent of the variance in teachers' judgments is due to the influence of the random sample of items that form the content objectives of the checklist. It can be inferred that test content as represented by the items is valid. The hypothesis that there is no relationship between the item format and the resulting judgments is rejected (hypothesis number 7).

Coefficient of construct validity. The last rho coefficient to be computed was the coefficient of construct validity. The magnitude of this coefficient is interpreted as indication of the similarity of the judgments within each category used to identify each teacher as a form of a test constructed to assess students' story schema. The value of this coefficient was .9879, or rounded off, .99. It can be inferred that the judgments related to one test form were more similar in that form than to any other form. Hypothesis number 8 is rejected.

Three additional indicies of generalizability were computed: (1) criterion-related validity, (2) the stability of teacher judgments over time, and (3), the magnitude of the influence of teacher belief systems on teacher judgments.

Criterion-related validity. The error variance is defined as that part of the observed score variance that is not universe score variance. The logic of the error variance was extended to state that the error variance is an index of the extent to which random error impacts on the criterion of performance: the formation of each judgment as specified by the eight items of story structure. The error variance was estimated at .03. It can be inferred that no more than 3 hundreths of a percent of the variance in teacher judgements is unrelated to the specified criterion. The hypothesis that the observed judgments are contaminated given the influence of random error from the measurement conditions is rejected (hypothesis number 10).

Steady-state behavior. The determination of the consistency of teachers' judgments over time was determined by comparing the size of the variance component for judgments within teachers to the size of the variance component for interaction. Both are relatively small; changes in teachers' judgments over the occassions of twelve student assessments were ruled out. It can be inferred that teachers' judgments remain constant over observations.

Item bias. A chi-square analysis was used to determine the extent to which two selected items were biased towards the beliefs of the teachers. The hypothesis that the observed judgments do not reflect the influence of the teachers' belief systems was rejected (hypothesis number 11). This analysis required two separate procedures. First, an item analysis was conducted. Then, the procedures of Camilli's chi-square (Crocker & Algina, 1986) were followed. The summary of the item analysis is displayed in Table 13.

As displayed in Table 13, two items experienced a larger percent of variance, .24, relative to the average, .21, of the other six item variances. A judgment on an item may be subject to sources of variation other than the differences in responses from the students. Items 2 and 5 are unbiased if (1) the items are affected by the same source of variance in each sub-population of judgments and (2) among the teachers the distribution of irrelevant sources of variation is the same. Each item is affected by a different source of variation. This source of error was detected using the F-test for items against interaction in the analysis of variance (see Table 8). In addition, the partition of within variance (see Table 9) indicated that among teachers the distribution of irrelevant sources of variance was not the same. The chi-square analysis was conducted to estimate the magnitude of the amount of bias in the judgments of teachers for items 2 and 5. The value of chi-square (Camilli) was estimated to be .37 (without Yates correction; the chi-square value with Yates was spuriously high). It was inferred that 37% of the variance in the judgments of teachers for item 2 (the child can tell where the story happened) and Item 5 (the child can

Table 13

Summary of Item Responses

	Item							
	1	2	3	4	5	6	7	8
Item mean	.33	.61	.72	.72	.38	.69	.67	.72
Item Variance	.22	.24	.20	.20	.24	.21	.22	.20

describe the feelings of one character) is attributable to the teachers' belief systems. Chi-square obtained was significant (p= .01) The hypothesis that the observed judgments do not reflect the influence of the teacher's belief system was rejected (hypothesis number 11).

CHAPTER V

DISCUSSION

Chapter 5 will summarize the findings of this study, draw conclusions based on the findings, and, finally, make recommendations for utilization of the conclusions.

Two studies were conducted to determine the dependability of teacher judgments. The first was a generalizability (G) study which had as its purpose the estimation of variance components. The second study was a decision (D) study which had as its purpose the estimation of the generalizability coefficients. Generalizability coefficients are descriptive statistics and are not dependent upon parametric assumptions of distributions of error.

Two facets were identified in the conditions of measurement. The first facet was the teacher who was conceptualized as a form of a test. The second was the checklist composed of eight items of story structure. Each facet was treated as a random sample from the infinite universe of possible samples. Three concepts from experimental design theory were used to identify the model of analysis for both studies. First, the observations of each teacher were independent of any other teacher. Second, the number of observations per cell were more than one. Each teacher formed twelve judgments for each of the eight items of the checklist. Third, both independent variables were considered to be random variables. A random effects nested design was chosen to analyze the data in both studies. In this design judgments are nested within teachers and crossed with items; each teacher formed a different set of judgments and the same set of items were used to stimulate the judgments of each teacher.

Two questions were answered: (1) Are teacher judgments generalizable, and (2) are teacher judgments dependable given the assumed influence of an identified belief system. The first question was answered in the G-study. The estimates of the variance components led to the conclusion that teacher judgments are generalizable. The second question, are teacher judgments dependable, was answered with the finding that the age of the students did not influence the teacher's judgments, and with a set of "rho" coefficients that estimated the judgments to be reliable, valid, and stable.

Summary of the G-Study

The purpose of the G-study was to generate the components of variance. It was hypothesized that the G-study would yield appropriate generalizability. The data were determined to be generalizable given the magnitude of the variance components. Variance components for six sources of variation were estimated: Judgments (p), teachers (j), items (i), judgments within teachers (j:p), the two-way interaction (ij), and the two-way interaction confounded with error [(j:p)xi]. Variance components are considered to be an estimate of measurement error. The components of variance were neither zero nor negative. The magnitude of the variance components indicated the amount of measurement error attributable to each facet. "Even if an estimated variance component does not possess statistical significance, it is an unbiased estimate" (Kane and Brennan, 1976, p. 290).

A random effects, split-plot treatment by items ANOVA procedure was used to estimate the components of variance. The F test of column effects against interaction in the analysis of variance was significant. From this F test it was inferred that each item was affected by different sources of error. A partition of the within variance indicated that the age of the students did not influence teacher judgments.

Summary of the D-Study

The purpose of the D-study was to estimate the generalizability coefficients that characterize the dependability of the teacher's judgments. Five coefficients of generalizability were hypothesized to characterized the dependability of teachers judgements. A random effects (j:p)xi design was used to define the expressions for estimating the descriptive statistics of the universe of generalization.

The domain score, the universe score variance, the error variance, and the observed score variance were estimated. The error variance was use to determine the extent to which teacher judgments are related to the criterion as spec- fied by the items of the checklist. It was found that only 3% maximum error was due to random factors in the conditions of measurement. It was inferred that 97% of the judgments were criterion-related.

The estimated value of the coefficients of generalizability indicated that teachers' judgments are highly dependable. The value of the rho coefficients, the coefficients of generalizability, revealed the following characteristics of teachers' judgments:

1. Teacher judgments are very accurate. 95% of the variance in the universe of teacher judgments of students' schema for story was predicted from any other set of judgments.

2. Teacher judgments are consistent over the time of assessment. 94% of the observed judgments were free of the influence of random error.

3. Teacher judgments reflect the content of the universe of items of story structure. 82% of the observed score variance of judgments was related to the content of the check list.

4. Teacher judgments validate the construct "teacher as form of a test." 99% of the variance in judgments was related to the independent observations. The judgments of one form of the test were more similar within that form than to any other form.

An additional component of the dependability of teacher judgments was the influence of the teacher's belief system regarding retelling as an assessment strategy. It was hypothesized that teachers' judgments are constant over pobservations and that teachers' belief systems influence their judgments. These hypotheses were not rejected. The variance components and a chi-square analysis were used to evaluate the influence.

Teacher judgments were found to be stable over the time of the assessment. The changes in teachers' judgments over time were evaluated using the components of variance for judgments within teachers and interaction. The relative sizes of the variance components for within teacher judgments (.07) and interaction (.03) indicated "steady-state behavior" (Shavelson, Webb, and Rowley, 1989, p. 926) during the time of assessment.

Teacher judgments were found to be bias towards teacher held beliefs regarding the use of retelling as an assessment procedure. An item analysis led to the conclusion that judgments pertaining to item 2 and item 5 were biased towards teacher belief systems. The chi-square analysis estimated that 37% of the variance of teachers' judgments on these items may be due to true differences between the teachers, or differences in the teachers' beliefs regarding the uses of retelling as a measurement procedure.

Conclusions

Generalizability theory was used to address the issues of the dependability of teacher judgments. Three conditions of measurement defined the universe of admissible observations. Therefore, generalizations are to the infinite universe of (1) teachers who use a (2) checklist to form (3) judgments of students' schema for story. Teacher judgments were estimated to be dependable but biased towards a teacher's beliefs regarding retelling as an assessment procedure. These generalizations are applicable to all teachers who share the belief systems of the teachers in the sample and use a story structure item checklist to stimulate their judgments.

The checklist and the construct, teacher as measurement instrument, are both valid. However, a teacher as a form of a test is not parallel to another teacher as form of a test. The error variances of two of the three teachers were equal, but their mean scores were unequal. The third teacher had a higher mean and a smaller error variance than the other two. Crocker and Algina (1986) suggested that in this situation, the teacher with the higher true scores will more frequently respond correctly to the child's retelling.

Therefore, teachers who view story retelling as a function of development will listen within the inference-based structure and will tend to have a higher mean score for judgments and a lower error variance than teachers who view retelling as a structure-based procedure designed only to determine the deficits in a child's story schema. Differences in teachers' judgments on items that are theoretically a function of development will be attributable to the teacher's belief system regarding retelling as an assessment procedure.

Teacher judgments are not made independently of the content of a checklist and are not influenced by the age of the child. Teacher judgments are ecologically valid: they are not influenced by the conditions in the classroom, or other sources of criterion contamination, and they are constant over the time of observations. Therefore, it could be argued that the bias noted towards the developmental aspects of a story schema support the concept of intentional validity, or the validity of a teacher's causal knowledge when functioning as a measuring instrument.

Recommendations

The generalizability of teacher judgments can be increased if the checklist is regarded not as a random variable but as a fixed variable. When the items of the checklist are regarded as the only set of expectations for the structure of a child's retelling, the coefficient of reliability, or of prediction, increases to .95 (or, 97% of the variance in the population is predicted) and the error variance drops to .0155. In addition, when the checklist is regarded as a fixed facet, the issue of knowledge representation is resolved. The items of the fixed facet would then represent the content of the test, or the knowledge a teacher should possess when assessing a student for the purposes of planning appropriate instructional experiences.

Implications for Further Research

The question of the validity of the teacher as a test form is not answered satisfactorily for the researcher. This researcher wondered if the results of the effects of teacher as the test form would have been different had not well defined issues in behavioral research prevailed: issues of (1) sample size and power of the test, and (2) concern with a Type 1 error versus a Type 2 error. These issues prevailed given the overwhelming evidence of story grammar being a random variable.

Crocker and Algina (1986) discussed an alternate definition of true score and error. They suggested that from a pool of dichotomously scored items, two or more test forms may be constructed by drawing items randomly from this pool. "Such randomly parallel test forms need not have equal means or equal variances, nor do the items have to be closely matched in content from form to form" (Crocker and Algina, 1986:124). Therefore, the first recommendation is that further research is guided by the conditions of the classroom. Instead of a random sample of judgements, random samples of items from the pool of items of story structure could be used to elicit teacher judgments. Teachers would be assigned randomly to the set of items and a binomial model could be used to address the issues of the dependability of the teacher as a test form. In addition, sample size could be determined by the actual number of judgments rather than the consequence of a random sampling. As suggested by House, Mathison, and McTaggart (1989), external validity could then be the primary validity issue. A related issue is teacher intentional validity, or the validity of cause. It could be determined the extent to which the construct being judged corresponds to an actual student learning objective selected by the teacher. Such efforts would ensure that a teacher's belief system regarding retelling as an assessment procedure is operationalized in terms of a judgment/criterion analysis.

The second recommendation is a study of teachers' judgments with each teacher considered to be a parallel form of the same test and with judgments codified as categories of students. The universe of admissible observations could be defined as broadly as possible, for example: judgments for students who are culturally diverse, judgements for students categorized on a continuum of development, or judgments of students on a retest basis. The magnitude of the variance components would indicate the extent to which each facet contributes to measurement error and would contribute knowledge of the generalizability of teacher judgments.

The third and most important recommendation is a replication of this study. Many small studies like this study could lead to the validation of the construct teacher-as-form-of-a-test. The concerns that frame generalizability are many. First, many replications would provide the framework for a body of knowledge regarding teacher use of an informal measurement instrument. In so doing, the issue of knowledge representation could be explored. Second, replications would increase an awareness among researchers who use teacher judgments that there is available a set of procedures to ensure statistical conclusional validity when teacher judgments are used as data in research. Finally, the important role of teacher judgments in research, in the evaluation of learning, and in teacher decision making would be revealed.

Implications for Teaching

Teacher judgments about student story schema and performance levels were revealed to be valid. However, there was some degree of error generated by individually held belief systems. Therefore, there are several implications for improving the accuracy of judgments.

First, teachers should be led to a high level of awareness of the impact of their personal beliefs in the informal assessment process. Second, teachers must be provided with the theoretical framework from which the items of a test are constructed. Finally, programs for enhancing teacher ability at assessing students' cognitive states and performance levels should be developed.

Instructional theory holds that one assesses growth and then redefines the objective of instruction based on the assessment results. Instructional theory is structure-based and is an outgrowth of the medical model of deficit analysis. Instructional theory does not, however, address the issues of the assessment process in terms of (1) the content objectives of an assessment technique and, (2) the theoretical orientation of the teacher to an assessment procedure. Content objectives are a theory of growth and development, of teaching and learning. Futhermore, when the scores from an assessment procedure are those which are codified by the teacher, the teacher is not only the test, but is also the indicator of three aspects of a test: (1) the domain of knowledge, (2) construct validity, and (3) the only source of reliability of the assessment procedure. There is a point where either the user of the assessment procedure is allowed to match the theory of the assessment to his/her pedagogical belief system, or is allowed the opportunity to know the theory of learning that frames the assessment process.

Assessment is the heart of the instructional process. The alledged purpose of assessment is to provide data for the planning of appropriate learning experiences based upon what the student knows. Teachers, therefore, should have the opportunity to (1) know the knowledge represented in any test, (2) explore the theory of learning that frames a test, and (3) determine the match between their pedagogical beliefs and those represented by the test. This is a process that could occur in pre-service education, should be the focus of teacher inservice when a new test of accountability is imposed on a school system, and must be an ongoing process in the supervision of teachers.

REFERENCES

Anderson, R. C. (1985). Role of the Reader's Schema in Comprehension, Learning, and Memory. In H. Singer & R. Ruddel (Eds.), *Theoretical Models and Processes of Reading*, 3rd ed., (pp. 372-384). Newark, DE: IRA.

Applebee, A. N. (1978). *A Child's Concept of Story*: *Ages 2-17*. Chicago: University of Chicago Press.

Beagles-Roos, J. & Gat, I. (1983). Specific impact of radio and television on children's story comprehension. *Journal of Educational Psychology*, *75*, 128-137.

Bejar, Isaac I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement*,*21*, 175-189.

Black, J. B. & Bower, G. H. (1980). Story understanding as problem solving. *Poetics* *9* , 223-250.

Blank, M. & Sheldon, F. (1971). Story recall in kindergarten children: effect of method of presentation on psycholinguistic performance. *Child Development*, *42*, 199-213.

Borko, H., Cone, R., Russo, N., & Shavelson, R. (1979). Teachers' decision making. In P. L. Peterson & H. Walberg (Eds.), *Research on Teaching* (pp. 136-160). Berkley: McCutchan.

Bower, G. (1976). Experiments on story understanding and recall. *The Quarterly Journal of Experimental* *Psychology* *28*, 511-534.

Bredekamp, S. (1987). *Guidelines for Developmentally Appropriate Practice in Early Childhood Programs Serving Children from Birth through Age 8*, Washington, D. C.:National Association for the Education of Young Children.

Brennan, Robert L. (1983). *Elements of Generalizability Theory*. Iowa City, Iowa: ACT Publications.

Brown, A. (1975). Recognition, reconstruction, and recall of narrative sequences of preoperational children. *Child Development*, *46*, 155-166.

Byers, J. L., & Evans, T. E. (1980). *Using a lens-model analysis to identify the factors in teacher judgment*.

(Research Series No. 73). East Lansing: Institute for Research on Teaching, Michigan State University.

Calfree, R.C. (1983). The design of reading research. Journal of Reading Behavior, 25, 59-80.

Clark, C. M. (1980). Choice of a model for research on teacher's thinking. Journal of Curriculum Studies,12, 41-47.

Clark, C. M. & Peterson, P. L. (1986). Teachers' thought processes. In M.C .Wittrock (Ed.), Third Handbook of Research on Teaching, (pp. 255-296). New York: McMillian.

Clark, C. M. & Yinger, R. J. (1979). Teachers' thinking. In P. L. Peterson, & H. J. Walberg, (Eds.), Research on Teaching (pp. 231-263). Berkeley:McCutchan.

Clark, C.M., Marx, R. W., Stayrook, N.G., Gage, N.L., Peterson, P.L., & Winne, P. H. (1979). A factorial experiment on teacher structuring, soliciting and reacting. Journal of Educational Psychology, 71, 535-552.

Clay, M. (1986) Constructive processes: talking, reading, writing, art, and craft. The Reading Teacher, 39, 764-770.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standarized test items. Journal of Educational Psychology, 78, 141-146.

Coleman, E. B. & Miller, G. R. (1974). The simplest experimental design that permits multiple generalizations. Journal of Reading Behavior, 6, 31-40.

Cook, T. D. & D. T. Campbell (1979). Quasi-Experimentation: Design & Analysis for Field Settings. Boston: Houghton Mifflin.

Crockett, L. & Algina, J. (1986). Introduction to Classical and Modern Test Theory. Holt, Rinehart and Winston.

Deford D.E. (1985). Validating the construct of theoretical orientation in reading instruction. Reading Research Quarterly, XX, 351-367.

Duffy, G. (1982). Response to Borko, Shavelson, and Stern: There's more to instruction decison-making than the "empty classroom." Reading Research Quarterly,XVII, 295-300.

Duffy, G. & Anderson, L. (1984). Teachers' theoretical orientations and the real classroom. Reading Psychology 5, 97-104.

Farstrup, A. E. (December, 1989/January, 1990). Point/counterpoint: state by state comparisions on national assessments. Reading Today, Vol.7, 3, p. 1.

Fine, G. A. & Sandstrom, K. L.(1988). Knowing Children: Participant Observation with Minors. Sage University Paper Series on Qualittive Research Methods, Vol.15. Beverly Hills, CA: Sage.

Frederiksen, N. (1984). The real test bias, Influences of testing on teaching and learning. American Psychologist, 39, 193-202.

Gage, N. L. & Needles, M. C. (1989). Process-product research on teaching: a review of criticism. The Elementary School Journal, 89, 254-300.

Gambrell, L., Pfeiffer, W., & Wilson, R. (1985). The effects of retelling upon reading comprehension and recall of text information. Journal of Educational Research, 78, 216-220.

Griffith, P. L., Ripich D. N., & Dastoli, S. L. (1986). Story structure, cohesion, and propositions in story recalls by learning-disabled and non-disabled children. Journal of Psycholinguistic Research 15, 539-555.

Grinnell, P. C. (1984). Children's conceptions of reading comprehension: A developmental study. In J. A. Niles & L. A. Harris (Eds.), Changing Perspectives on Research in Reading/language Processing and Instruction. Thirty-third Yearbook of the National Reading Conference (pp. 80-84). Rochester, NY: National Reading Conference.

Harste, J. C., & Burke, C. L. (1979). Understanding the hypotheses: Its the teacher who makes the difference. In K. VanderMuelen (Ed.), Reading Horizons: Selected Readings (pp.137-156). Kalamazoo: Western Michigan University Press.

Haney, W. & Madaus, G. (1989). Searching for alternatives to standardized tests: why, whats and whithers. Phi Delta Kappan, 70, 688-697.

Hayes, W. L. (1981). Statistics, 3rd Edition. New York:

CBS College Publishing, Holt, Rinehart and Winston.

Hennerson, M. E., Morris, L. L., Fitz-Gibbon, C. T. (1987). How To Measure Attitudes. Beverly Hills:SAGE Publications.

Hieshima, J. & Sulzby, E. (1985). Product vs. process: The relevance of methodology in the elicitation of story re-enactments for young children. ERIC Document Reproduction Service (No. ED 276986 CS008632)

Hopkins, K. D. (1984). Generalizability theory and experimental design: Incongruity between analysis and inference. American Educational Research Journal, 21, 703-711.

House, E. R., Mathison, S., & McTaggart, R. (1989). Validity and teacher inference. Educational Researcher, 18(7), 11-15, 26.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

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

Irwin, P. A. & Mitchell, J. N. (1988). The reader retelling profile: using retellings to make instructional decisions. In preparation. c.f. L. N. Morrow (1988). Retelling stories as a diagnostic tool. In S. W. Glazer, L. W. Searfoss, & L. M. Gentile (Eds.), Reexamining Reading Diagnosis: New Trends and Procedures (pp. 128-150). Newark: IRA.

Johnston, P. H. (1983). Reading Comprehension Assessment: A Cognitive Basis. Newark: IRA.

Johnston, P. H. (1984). Assessment in Reading. In P. D. Pearson (Ed.), Handbook of Reading Research (pp. 147-182). New York: Longman.

Johnston, P. H. (1987). Assessing the Process and the Process of Assessment in Language Arts. In J. Squire (Ed.), The Dynamics of Language Learning: Research in the Language Arts (pp.335-357). Urbana, IL: National Council of Teachers of English.

Kamil, M. L. (1984). Current traditions of reading research. In P.D. Pearson (Ed.), Handbook of Reading Reserch (pp. 39-62). New York: Longmans.

Kane, M. T. (1986). The role of reliability in criterion referenced tests. Journal of Educational Measurement,23, 221-224.

Kane, M. T. & Brennan, R. L. (1977). Generalizability of class means. Review of Educational Research, 47(2), 267-292.

Kerlinger, F. N. (1986). Foundations of Behavioral Research. New York: CBS College Publishing.

Kindsvatter, R., Wilen, W., & Ishler, M. (1988). Dynamics of Effective Teaching. New York: Longman.

Kinzer, C. D. & Carrick, D. A. (1986). Teacher Beliefs as instructional influences. In J. A. Niles & R. V. Lalik (Eds.), Solving Problems in Literacy: Learners, Teachers, and Researchers. Thirty-fifth Yearbook of the National Reading Conference (pp. 127-134). Rochester: The National Reading Conference.

Kintsch, W., & Kozminsky, E. (1977). Summarizing stories after reading and listening. Journal of Educational Psychology, 69, 491-499.

Kirk, R. E. (1986). Experimental Design, (3rd Ed.) Procedures for the Behavioral Sciences. Belmont, CA: Brooks/Cole Publishing Co.

Kirk, J. & Miller, M. L. (1986). Reliability and Validity in Qualitative Research. SAGE University Paper Series on Qualitative Research Methods, Vol.1. Beverly Hills: Sage Publications.

Madaus, G. F. (1988). The influence of testing on the curriculum. In I. N. Tanner (Ed.), Critical Issues in Curriculum: 87th yearbook of the National Society for the Study of Education (pp. 83-121). Chicago: University of Chicago Press.

Mandler, J. M. (1987). On the psychological reality of story structure. Discourse Processes, 10, 1-29.

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

Mangano, N. & Allen, J. (1986) Teachers' beliefs about language arts and their effect on student beliefs and instruction. In J. A. Niles & R. Lalik (Eds.), Solving Problems in Literacy: Learners, Teachers,and Researchers. Thirty-fifth Yearbook of the National Reading Conference (pp. 135 - 142). Rochester, NY: The National Reading Conference.

Margolis, H., & Nicholas, D. (1986). Teacher perceptions of infuences on choice of reading materials. Ohio Reding Teacher, 20(4), 5-10.

Matthew 7:1-5, The Kings James Version of the Bible.

McGee, L. (1981). The videotape answer to independent reading comprehension activities. The Reading Teacher, 34, 427-433.

McGee,L. M., Ratliff, J., Sinex, A., Head, M., and LaCroix, K. (1984). Influence of story schema and concept of story on children's story compositions. In J. A. Niles & L. A. Harris (Eds.), Changing Perspectives on Research in Reading/language Processing and Instruction. Thirty-third Yearbook of the National Reading Conference (pp. 270-277). Rochester, NY: National Reading Conference.

Messick, S. (1989). Meaning and values of test validation. The science and ethics of assessment. Educational Researcher, 18(2), 5-11.

Meyer, B. J. F. (1975). Text dimensions and cognitive processing. In H. Mandl, N.L. Stein, & T. Trabasso (Eds.) Learning and Comprehension of Text (pp. 3-47). Hills dale, N.J.: Erlbaum.

Moore, S. A. (1986). A Comparison of reading education stu dents' instructional beliefs and instructional practices. In J. A. Niles & R. Lalik (Eds.), Solving Problems in Literacy: Learners, Teachers, and Researchers. Thirty-fifth Yearbook of the National Reading Conference (pp. 143-146). Rochester NY: The National Reading Conference.

Morris. L. L., Fitz-Gibbon, C. T., & Lindheim, E. (1987). How to Measure Performance and Use Tests.Newbury Park: SAGE Pub., Inc.

Morrow, L. M. (1978). Analysis of syntax of six, seven, and eight year old children. Research in the Teaching of English, 12, 143-148.

Morrow, L. M. (1984). Effects of story retelling on young children's comprehension and sense of story structure. In J. A. Niles & L. A. Harris (Eds.), *Changing Perspectives on Research in Reading/language Processing and Instruction*. *Thirty-third Yearbook of the National Reading* Conference (pp. 95-100). Rochester, NY: National Reading Conference.

Morrow, L. M. (1985a). Attitudes of teachers, principals, and parents toward promoting voluntary reading in the elementary school. *Reading Research and Instruction*, *25*(2), 116-130.

Morrow, L. M. (1985b). Retelling stories: A strategy for improving young 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: IRA.

Morrow, L. M., Gambrell, B., Kapinus, B., Koskinan, P., Marshall, N. & Mitchell, J. (1986). Retelling: A strategy for reading instruction and assessment. In J. A. Niles & R. Lalik (Eds.), *Solving Problems in Literacy:Learners, Teachers,and Researchers*. *Thirty-fifth Yearbook of the National Reading Conference* (pp. 73-80). Rochester: The National Reading Conference.

Munby, H. (1982). "The place of teacher's beliefs in research on teacher thinking and decision making." *Instructional Science*, *11*, 201-225.

Neil, D. M., & Medina, N. J. (1989). Standarized testing: harmful to educational health. *Phi Delta Kappan*, *70*, 688-697.

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher*, *18*(9), 3-7.

Nitko, A. (1989, April). "Agreeing to Disagree:Researchers Seeking to Broaden Policy Debate." *Education Week*, 5 Vol. VII, 28, 19.

Nurse, J. R. & McGauvran, M. E. (1987). Metropolitan Readiness Tests. *Metropolitan Readiness Assessment Program*. San Antonio, TX: The Psychological Corporation.

O'Donnell, R., Griffin, J., & Norris, D. (1967). Syntax of Kindergarten and Elementary School Children: A Transformational Analysis. Report No. 3. Urbana IL: National Council of Teachers of English.

Page, J. L. & S. R. Stewart (1985). Story grammar skills in school-age children. Topics in Language Disorder, 16-30.

Peterson, P. L., Marx, P. W., & Clark, C. M. (1978). Teacher planning, teacher behavior, and student achievement. American Educational Research Journal, 15, 417-432.

Peterson, P. L. & Comeaux, M. (1987). Teachers' schemata for classroom events: the mental scaffolding of teachers' thinking during classroom instruction. Teaching and Teacher Education, 3, 319-331.

Pickert, S. M., & Chase, M. L. (1978). Story retelling: an informal technique for evaluating children's language. Reading Teacher, 31, 528-29.

Pfaffenberger, B. (1988). Microcomputer Applications in Qualitative Research. Sage University Paper Series on Qualitative Research Methods, Vol. 14. Beverly Hills, CA: Sage.

Poulsen, D., Kintsch, E., Kintsch, W., & Premack, D. (1979). Children's comprehension and memory for stories. Journal of Experimental Child Psychology, 28, 379 - 403.

Powell, D. A. (1986). Influences on teachers' decision-making in selecting reading textbooks. In J. A. Niles & R. Lalik, (Eds.). Solving Problems in Literacy: Learners, Teachers, and Researchers. Thirty-fifth Yearbook of the National Reading Conference (pp. 147-151). Rochester NY: The National Reading Conference.

Ray, K. J., Lee, S. C., & Stansell, J. C. (1986). New methods, old theories, and teacher education: some observations of writing in a third-grade classroom. In J. A. Niles & R. Lalik (Eds.) Solving Problems in Literacy: Learners, Teachers, and Researchers. Thirty-fifth Yearbook of the National Reading Conference (pp. 152-159). Rochester: The National Reading Conference.

Rose, M. C., Cundick, B. P., & Higbee, K. L. (1984). Verbal rehearsal and visual imagery: Mnenomic aids for learning disabled children. Journal of Learning Disabilities, 16, 352-354

Rowley, G. L. (1976). The reliability of observational measures. *American Educational Research Journal*, *13*(1), 58-60.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. Bobrow (Ed.). *Representation and Understanding: Studies in Cognitive Science*. NY: Academic Press, 213-25.

Rumelhart, D. E. (1977). Understanding and summarizing brief stories. In D. LaBerge & S. J. Samuels (Eds.), *Basic Processes in Reading: Perception and Comprehension* (pp. 265-303). Hillsdale, N.J.:Erlbaum.

Rummelhart, D. E. (1981). Schemata: The building blocks of cognition in comprehension and theory. In J. T. Guthrie (Ed.), *Comprehension and Teaching* (pp. 3-26). Newark:IRA

Rupley, W. H. & Logan, J. W. (1985). Elementary teacher's beliefs about reading and knowledge of reading content: relationship to decisions about reading outcomes. *Reading Psychology*, 6, 145-146.

Rupley, W. H. & Logan, J. W. (1986). Relationship between teacher's beliefs about reading and their reported use of questioning and engagement strategies. In J. A. Niles & R. Lalik (Eds.), *Solving Problems in Literacy: Learners, Teachers, and Researchers*. Thirty-fifth Yearbook of the National Reading Conference (pp. 165-170). Rochester NY: The National Reading Conference.

Sanders, J., Hills, J. R., Merwin, J. C., Trice, C., & Dianda, M. (1989). *Standards for Teacher Competence in Education Assessment of Students*. Unpublished position paper. Western Michigan University, Kalamazoo, Michigan.

Schmitt, M. C., & O'Brien, D. G. (1986). Story grammars: some cautions about the translation of research into practice. *Reading Research and Instruction*, *26*, 1-8.

Sergiovanni, T. (1989). Science and scientism in supervision and teaching. *Journal of Curriculum and Supervision*, 4, 93-105.

Shake, M. C. (1986). Basis for grouping decisions. In J. A. Niles & R. Lalik (Eds.) *Solving Problems in Literacy: Learners, Teachers, and Researchers*. Thirty-fifth Yearbook of the National Reading Conference (pp. 171-177). Rochester NY: The National Reading Conference.

Shavelson, R. J. (1976). Teacher decision-making. In N. L. Gage (Ed.), *The Psychology of Teaching Methods* (pp. 372-

414). 75th yearbook of the National Society for the Study of Education, Part 1. Berkley: McCutchan Publication Corp.

Shavelson, R., Cadwell, J. & Izu, T. (1977). Teachers sensitivity to the reliability of information in making pedagogical decisions. *American Educational Research Journal*, *14*, 83-97.

Shavelson, R. & Russo, N. (1977). Generalizability of measures of teacher effectiveness. *Educational Research*, *19*, 171-183.

Shavelson, R. J. & Stern, P. (1981). Research on teacher's pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research*, *51*, 455-498.

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of Teaching. In M. C. Wittrock, (Ed.), *Handbook of Research on Teaching* (3rd ed.), pp. 50-91. NY:McMillan.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922-932.

Shulman, L. S. (1986). Paradigms and research programs in the study of teaching: A contempory perspective. In M. C. Wittrock (Ed.), *Handbook on Research on Teaching* (3rd. ed.) (pp. 3-49). New York: MacMillan.

Shulman, L. S. and Elstein, A. S. (1975). Studies of problem solving, judgment and decision making: Implications for educational research. In F. D. Kerlinger (Ed.), *Review of Research in Education*, (pp. 3-42). Itasca, IL:F.E. Peacock.

Smith, L. A. and Shephard, M. C. (1988). Kindergarten readiness and retention: a qualitative study of teachers' beliefs and practices. *American Educational Research Journal*, *25*, 307-333.

Smith, P. B. & Peterson, M. F. (1988). *Leadership, Organizations and Culture: An Event Management Model*. Beverly Hills: Sage Publications.

States move to improve assessment picture. (1989, March). *ASCD Update*. p. 7.

Stein, N. (1979). How children understand stories: A devopmental analysis. In L. G. Katz (Ed.), Current Topics in Early Childhood Education (Vol.II), pp. 261-290. Norwood, NJ: Ablex.

Stein, N. L. and Glenn, C. (1979). An analysis of story comprehension in elementary school children. In R.O. Freedle (Ed.), Advances in Discourse Processes: New Directions in Discourse Processing (Vol.2.), pp. 53-120. Norwood, NJ: ABLEX.

Stenner, A. J., Smith III, M., & Burdick, D. S. (1983). Toward a theory of construct validation. Journal of Educational Measurement, 20, 305-311.

Stiggens, R. J. (1988). Revitalizing classroom assessment: The highest instructional priority. Phi Delta Kappan, 69, 363-368.

Stiggens, R. J. & Bridgeford, N. J. (1985). The ecology of classroom assessment. Journal of Educational Measurement 22, 271-86.

Student assessment: Slow growth seen toward improved measures. (1989, March). ASCD Update. pp. 1, 7.

Sulzby, E. (1982). Oral and written mode adaptions in stories by kindergarten children. Journal of Reading Behavior, 14, 51-59.

Taylor, J. & Garcia, C. L. (1987). Three teachers: What they said and what they did. California Reader, 20(2), 17-22.

Tenbrink, T. D. (1974). Evaluation: A Practical Guide For Teachers. New York: McGraw-Hill Book Co.

Thorndyke, P. W. (1977). Cognitive structure in comprehension and memory of narrative discourse. Cognitive Psychol ogy, 9, 77-110.

Tyler, R. W. (1986). Changing concepts of educational evaluation. International Journal of Educational Research, 10, 1-113.

Wiggins, G. (1989). A true test: toward more authentic and equitable assessment. Phi Delta Kappan, 70, 703-713.

Wineburg, S. S. (1989). Rememberance of things past. Educational Researcher, 18(4). 7-10.

APPROVAL SHEET

The dissertation submitted by Josephine C. Logan-Woods has been read and approved by the following Committee:

Dr. Robert C. Cienkus. Chairman
Associate Professor, Curriculum and Instruction,
Loyola, Chicago

Dr. Barney M. Berlin
Associate Professor, Curriculum and Instruction,
Loyola, Chicago

Dr. Jack A. Kavanagh
Professor, Educational Foundations, Loyola, Chicago

The final copies have been examined by the director of the dissertation and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the dissertation is now given final approval by the Committee with reference to content and form.

The dissertation is therefore accepted in partial fulfillment of the requirements for the degree of Doctor of Education.

April 16, 1990
Date

Robert C. Cienkus
Director's Signature