An Exploratory Study of Young Children's Thinking on the *Test of Early Language and Literacy (TELL)* using Verbal Report Data and a Cognitive and Normative Model of Test Performance

by

Karen Lori Vavra

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Elementary Education
University of Alberta

ABSTRACT

Early indications of potential school failure include difficulties in early language and literacy development. Thus, prevention and early identification depend on timely and coordinated assessment of language and literacy ability. Despite the critical need, there are no available Canadian assessment instruments to provide the type of diagnostic information needed for early, effective, and appropriate intervention. This exploratory study investigated whether the *Oral Narrative (ON)* and *Oral Reading-Reading Comprehension (OR-RC)* subsections of the newly-developed *Test of Early Language and Literacy (TELL)* measure the skills and processes fundamental to listening and reading comprehension ability and are suitable for their intended purpose and use with children 3- to 8-years old. Children (n = 174) from 3- to 8-years of age completed the literal and inferential comprehension questions for each subsection with the inclusion of think-aloud probes of their reasoning for each item response. Test performance and protocol analyses supported the *TELL* subsections and confirmed the *TELL* as developmentally appropriate across the intended age range and highlighted that comprehension development is not age dependent. Four response patterns underscored the informative portrayal of the information sources used by children to answer and provide reasons for their answers. Collectively, the patterns revealed the diagnostic attributes of the measures for identifying strengths and weaknesses in comprehension. The feasibility of using meta-level questions with young children was established in this first known use of verbal reports with preschoolers and in the first Canadian combined language and literacy comprehension test.

PREFACE

This dissertation is an original work by Karen Lori Vavra. The research project, of which this dissertation is part, received research ethics approval from the University of Alberta Research Ethics Board, The Design and Development of an Early Language and Literacy Screening Test, Pro00017968, January, 1, 2008.

ACKNOWLEDGEMENTS AND DEDICATION

The completion of my doctoral dissertation represents the culmination of many years of required perseverance and focus to complete this challenging scholarly accomplishment. I gratefully acknowledge the significant role of a number of key people whom provided support, guidance, and direction throughout my doctoral studies.

It is with utmost respect and admiration that I thank my supervisor, Dr. Linda Phillips, for her continuous commitment, guidance, and support throughout my doctoral program. It has been a privilege and honour to work with an accomplished and internationally renowned scholar and mentor with unparalleled depth and breadth of knowledge and expertise in cognitive psychology, language and literacy assessment and intervention. At every stage, she inspired and challenged me to think critically, to expand my knowledge and understanding of the many facets of my study, and to focus always on the evidence in the existing research and in my results. She exercised immense patience by allowing me the time and space to work through the mental challenges and provided constructive feedback throughout the research process. Her unwavering support and mentorship motivated me to aspire to the very highest standards of achievement.

It is with profound sadness that I dedicate this dissertation to the memory of Dr. Stephen Norris, a key member of my supervisory committee whose extensive theoretical conceptualizations of measurement and construct validity were fundamental to the design and development of my study. Dr. Norris provided superior guidance and mentorship during my candidacy preparations, data collection, and the early stages of the data analyses for this study. His extensive body of research and scholarship has made a significant contribution to the field of science and literacy education, and assessment. However, it was his stellar character, calm demeanor, and sense of humour which left an indelible imprint on me and anyone who had the privilege of knowing him.

I gratefully acknowledge the members of my supervisory committee, Dr. Denyse Hayward and Dr. Todd Rogers, for their expertise, support, and guidance in critically examining my research and providing constructive feedback in their pursuit of the excellence expected in my work.

I express sincere gratitude to my examining committee, Dr. Peter Afflerbach (University of Maryland), Dr. Jean Clandinin (Examination Chair, University of Alberta), and Dr. George Buck (University of Alberta) for their careful scrutiny of my dissertation and their thought-provoking questions which challenged me to think deeply about my methodology and the implications of my research.

And finally, to my dear family and friends who were steadfast in their support and encouragement through the many varied challenges and successes I experienced as I worked to accomplish my goal. I owe much gratitude and respect to my dear friend and colleague, Diane Gagley, who was always willing to engage in lengthy discussions about my research, serve as my sounding-board throughout my graduate studies, and particularly in the weeks of preparation leading up to my final oral doctoral defense.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Approximately 25% of all children who enter school have a range of different learning, behavioural, and social problems that place them at greater risk for school failure (Lee & Burkham, 2002; Long, 2006; McCain, Mustard, & Shanker, 2007). Many of these children show evidence of early and sustained difficulties in language and literacy development. Specifically, children from low-income families with limited educational background are at higher risk for delayed development and difficulties in a number of areas including reading (Hayward, Das, & Janzen, 2007; Phillips, Hayden, & Norris, 2006); receptive and expressive language (Lonigan et al., 1999); preschool alphabet knowledge and phonological awareness (Lonigan et al., 1998); decoding and comprehension (Levy et al., 2006); and cognitive development (Dickinson & Tabors, 2001; Neuman & Dickinson, 2001). The most prevalent and earliest indicators of school failure include difficulties in early language and literacy development. Most children who experience reading difficulties also exhibit problems with early language development. Thus, early identification of children who are at-risk is dependent upon timely and coordinated assessment of early language and literacy ability. Despite the critical need for accurate and adequate assessments that measure a sufficiently broad range of early language and literacy skills, there are currently no Canadian assessments to provide the kind of diagnostic information needed for early, effective, and appropriate intervention.

The purpose of this research was to investigate the construct validity of a newly-developed test of early language and literacy using verbal report data to establish the underlying cognitive and normative model that explains test performance. Specifically, the study examined the relationship between children's performance and what they reported thinking and reasoning as they completed items on the *Test of Early Language and Literacy* (*TELL*) (Phillips, Hayward, & Norris, 2016). This research is positioned as part of a national study led by Drs. Phillips, Hayward, and Norris at the University of Alberta involving the design and development of the *TELL.* The *TELL* is an innovative, individually-administered diagnostic assessment of early language and literacy skills for children three to eight years of age. A systematic review of language and literacy tests confirmed the need for new and alternative measures (Hayward et al., 2008). Unlike current tests, the *TELL* offers a more comprehensive and thorough assessment of a child's combined language and literacy performance (typically separate tests) including print understanding, letter knowledge, phonological awareness, oral vocabulary, oral narrative comprehension and production, word reading, oral reading and reading comprehension, written spelling and writing. Additionally, it provides a comprehensive research-based rationale for inclusion of each subsection, each test item, and sequence of skills essential for planning informed intervention. The *TELL* has two main purposes: (1) to assess children's language and literacy development in order to determine whether they are at risk of school failure, and (2) to identify specifically their areas of strengths and weaknesses for intervention.

The *TELL* is comprised of eight main components of early language and literacy development. *Print Understanding* measures children's understanding of print concepts (i.e., book knowledge and conventions of the printed word) and environmental print (i.e., function of print in the environment and the ability to distinguish between print and pictures). *Letter-knowledge and Naming* is designed to measure children's ability to recognize and name letters of the alphabet. *Phonological Awareness* is comprised of six subsections to measure children's ability to blend and segment syllables and sounds, and to identify initial and final sounds. *Oral Vocabulary* includes four subsections designed to assess children's ability to label and define pictured nouns and verbs, and define non-pictured nouns and verbs. *Oral Narratives* is comprised of three subsections to measure listening comprehension, story generation, and story recall. *Word Reading* is designed to assess children's ability to recognize differences between pictures and print, words and nonsense words, and to read words. *Oral Reading and Reading Comprehension* contains two main components for assessing children's reading behaviours and reading accuracy, as well as their reading comprehension. The final component, *Written Spelling and Writing*, includes two separate subsections to measure children's ability to spell their names and commonly occurring vocabulary words, and to write a story.

As part of the test validation phase, the current exploratory study focused on the underlying cognitive model that explains performance (Norris, Leighton, & Phillips, 2004; Norris, Macnab, & Phillips, 2007) by incorporating the use of verbal reports of cognitive processing (or think alouds) (Ericsson & Simon, 1980, 1993; Phillips, 1988; Pressley & Afflerbach, 1995) in the administration of the two

TELL comprehension subsections that require reasoning over and above recall: (1) Oral Narratives (ON) and (2) Oral Reading and Reading Comprehension (OR-RC).

In this study, verbal reports were the primary data source for learning about the knowledge, strategies, and principles that children use to respond to test items and tasks (Ericsson & Simon, 1980, 1993; Messick, 1989; Pressley & Afflerbach, 1995). In the context of this study, knowledge is based on the accumulated prior experiences and understanding of language and concepts that children bring to the listening and reading comprehension assessments in order to construct new knowledge. Strategies are the procedures and conscious processes used by the children to integrate their relevant background knowledge with the available information in order to answer and justify their answers to complete the listening and reading comprehension test items. Principles are the underlying beliefs and standards evident in what children say in their justificatory responses for the answers they provided to the listening and reading comprehension test questions. Verbal report data is one source that can be used to identify whether children understood the test items and tasks and whether they answered test items correctly or incorrectly for the right reasons. Children's verbal reports were analysed in accord with the specified scoring criteria and in relation to their performance on the test items to identify the processes and strategies used to respond to test items, to determine whether these were commensurate with the theoretical conceptualization of the construct, and, to examine whether the test items measured what they were designed to measure. Validity evidence was based on a systematic comparison of the test items and tasks, the expected cognitive processes required for successful completion of the tasks, the actual processes and information that children reported using, and the reasons given for their responses.

Background and Theoretical Framework

This section identifies the components and properties of assessment, examines reliability and validity as the two psychometric properties of assessment and presents recent cognitive and normative models of assessment.

Assessment Components and Properties

A leading publication from the National Research Council (NRC) titled *Knowing What Students Know: The Science and Design of Educational Assessment* (National Research Council, 2001) marked an important transition in educational measurement. The NRC report represented the culmination of an intensive three-year study of the most current perspectives on assessment research and practice in cognitive and measurement sciences with specific recommendations to update and reform educational measurement. Although the committee focused mainly on academic achievement testing in science and mathematics, the fundamental principles of assessment are relevant and pertinent to all types of educational measurement including early language and literacy assessment, the primary focus of this study. Ultimately, the quality of assessment is of utmost importance. “Better assessment, curriculum, and instruction could help educators diagnose the needs of at-risk students and tailor improvements to meet those needs” (National Research Council, 2001, p. 18). The NRC report outlined the guiding principles for test development, and interpretation of test results in educational assessment to reflect advances in cognitive and measurement sciences and understanding of thinking and learning.

Assessment is the process of gathering information and “reasoning from evidence” (Mislevy, 1994; 1996) in order to draw inferences about individuals’ competencies, that is, what individuals know and can do related to a particular domain or construct. Since psychological constructs are comprised of mental representations and processes that cannot be observed or measured directly, assessments consist of indirect measures of attributes or qualities that represent the knowledge and performance related to a domain. Judgments and inferences about what individuals know and can do are based on samples of performance in the specific domain (National Research Council, 2001, p. 36).

According to Pellegrino (2002-2003), assessment requires the integration of three main components: cognition, observation, and interpretation. The cognitive aspect refers to the model of how individuals “represent knowledge and develop competence in the domain” (p. 49). The underlying theory of learning of the domain is the main source of specification for the knowledge and skills that are most important to measure (National Research Council, 2001). The observation component includes the selection and design of particular types of tasks or contexts that will elicit demonstrations of the knowledge and skills specified by the cognitive component. The interpretation component is the method for making sense of or applying meaning to the test performance data by drawing meaningful inferences about individuals’ competencies. According to

Pellegrino (2002-03), the three components of assessment are interconnected, interdependent, and operate simultaneously.

> Assessment comes down to which types of evidence or observations are available to help reason about the examinee's competence. What one believes about the nature of learning will affect the kinds of assessment data sought and the chain of inferences drawn…The chain of reasoning determines what to look for in what students say, do or produce and why it constitutes evidence about what they know and do (National Research Council, 2001, p. 43).

Pellegrino (2002-03) acknowledged that while the three aspects (cognition, observation, and interpretation) underpin current views of educational assessment, this has not always been the case, "Much of what we've been doing in assessment has been based on impoverished models of cognition, which has led us to highly limited modes of observation that can yield only extremely limited interpretations of what students know" (p. 49). Reviews of early language and literacy assessments show that many of the current assessment instruments clearly fall into this category. The constructs are often narrowly conceived and the assessments provide an inadequate picture of children's early language and literacy development. Thus, the key to effective assessment is the integration of the three essential components of cognition, observation, and interpretation. Furthermore, the value and effectiveness of an assessment instrument depends on whether the assessment yields accurate results about the specific domain of interest and the extent to which the assessment is fundamentally adequate. The fundamental adequacy of a test is evaluated based on criteria related to two main psychometric properties: reliability and validity.

Psychometric Evaluations of Assessment Instruments

Evaluations of the psychometric properties of an assessment instrument are an important and essential part of test design, development, and use. The main purpose of psychometric evaluations is to determine whether measures provide "enough evidence of trustworthiness to warrant use" (Kame'enui et al., 2006, p. 9). That is, the extent to which the test results are reliable and support valid interpretations and conclusions about individuals' ability and competencies and decisions in accord with the assessment results (Kane, 2006).

Reliability. Reliability is the consistency and stability of test results. Evaluations of reliability include comparison of test scores. An instrument is reliable to the extent that it produces similar scores over multiple administrations with minimal variation or error in the test scores across administrations (Lonigan, McDowell, & Phillips, 2004). There are several different categories of reliability. *Test-retest reliability* evaluates the relationship between two separate administrations of the same measure to the same sample. *Alternate forms reliability* evaluates the correlation between the test scores of two separate forms of a test. *Internal consistency* examines the degree to which the items in a test measure the same construct. *Inter-rater reliability* is the comparison of test results administered by two independent examiners. Evaluations of reliability are reported as correlation coefficients that range between 0 and 1.0 with 0 indicating no relationship or a lack of reliability between the two or more variables and scores approaching 1.0 indicate a high degree of reliability.

Validity. Validity is whether a test measures what it was designed to measure, that is whether content-related evidence is present and solid. The *Standards for Psychological and Educational Tests* states, “Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (American Educational Research Association (AERA), American Psychological Association (APA), and National Council of Measurement in Education (NCME), 2014, p. 11). According to Messick (1989), validity is a unitary concept comprised of two main aspects: (1) interpretation of test results and what the test scores mean, and (2) consequences of the interpretations. Validity is an inquiry process concerned with establishing the adequacy and appropriateness of the inferences derived from test scores and the actions or decisions made on the basis of these inferences.

Several noteworthy changes characterize the most recent conceptualizations of validity. First, validity is concerned with test responses and the inferences or interpretations of test responses within a specific context of use. Validation includes meaningful interpretation of test scores and justification for decisions and actions based on score interpretations. Moreover, there is no single measure to establish definitively whether the interpretation or use of an assessment is valid. Finally, validity is a process of accumulating evidence to support an argument or justification for the specific interpretations and applications of an assessment (Kane, 2006).

Previous conceptualizations of validity named three basic categories of validity evidence (i.e., content, criterion, and construct). However, the most recent views conceive the scope of different sources of validity evidence under one broad category, namely, construct validity. Currently, construct validity is conceptualized in relation to five sources of evidence based on several distinct aspects: (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences (American Educational Research Association (AERA) et al., 2014). The *test content* aspect examines whether the test items and content are sufficiently relevant and representative of the construct being tested. Evidence based on *response processes* is concerned with the fit between cognitive and thinking processes that individuals use to respond to test items and the intended cognitive and thinking processes needed to obtain a correct or acceptable response. The *internal structur*e of an instrument considers the evidence that the test items in order to estimate its content reliability of a test. The *external structure* refers to evidence for the relations to other variables by examining the relationships between test scores with other criterion measures (i.e., similar tests). Finally, the *consequential* aspect refers to the relevant consequences and uses of the test. Messick (1995) states, "Construct validity comprises the evidence and rationales supporting trustworthiness of score interpretations in terms of explanatory concepts that account for both test performance and score relationships with other variables" (p. 743). The potential sources of construct validation contribute to a body of evidence to substantiate the interpretation of test scores and uses. Each source of evidence offers a different perspective on the validation of the interpretation and application of test performance results. It is not the quantity of evidence that matters most but rather the quality of the evidence that is of utmost importance in making validation claims.

According to Messick (1989), there are two main threats to validity: *construct underrepresentation* and *construct-irrelevant variance*. Construct under-representation occurs when “the test is too narrow and fails to include important dimensions or facets of the construct” (Messick, 1989, p. 34). On the contrary, construct-irrelevant variance occurs when the “test contains excess reliable variance that is irrelevant to the interpreted construct” (Messick, 1989, p. 34). In other words, the assessment is too broad if test performance is influenced by factors extraneous to the construct which inadvertently make the test more difficult or easy for particular individuals or groups. For example, test performance may be affected by factors associated with other constructs or test items and formats may contain inadvertent clues which cause test-takers to respond in certain ways using sources of information that are irrelevant to the construct (Messick, 1989).

Substantive aspect of validity. Over the last two decades, the expanded view of validity maintains that in order to establish the trustworthiness of the interpretation of test scores it is important to gather empirical evidence verifying that the assessment tasks actually elicit and measure the cognitive processes that they were designed to measure. According to the NRC (2001) report, “...the trustworthiness of the interpretation of test scores should rest in part on empirical evidence that the assessment tasks actually tap the intended cognitive processes” (p. 207). This source of validation evidence focuses on the extent to which the actual processes used to respond to the test items and tasks correspond and adequately represent the theoretical processes associated with the construct that is being measured.

Messick (1989) identified *protocol analysis* as one approach for analyzing the processes underlying test-item responses and task performance. This analysis centers on collecting verbal reports of thinking during test administration. One method is to ask test-takers to think-aloud as they complete the test items and tasks or to describe retrospectively what they did or thought as they completed the tasks. Another method, referred to as *analysis of reasons*, is to ask test-takers to give reasons for their responses to test items and tasks (Messick, 1989). These methods provide important insight into the types of processes that test-takers use as they complete test items and tasks. Protocol analysis of the verbal report data examines the degree to which these processes match the theoretical conceptualization of the construct. Norris (1992) summarized a number of different applications and purposes for using verbal reports of thinking during testing,

> Because direct evidence on thinking processes can play such an important role in test interpretation, testing theorists have endorsed verbal reports of thinking for amplifying the meaning of the constructs a test is measuring (Cronbach, 1971), for representing those constructs by the mental processes that underlie performance (Embretson, 1983; Embretson, Schneider, & Roth, 1986; Haney & Scott, 1987), for directly analyzing the processes underlying item performance (Messick, 1989), and for specifying the intellectual processes used to perform test tasks (Anastasia, 1988) (p. 156).

Verbal reports of thinking during testing have become a significant and important source of evidence to substantiate the validation of interpretations of test performance and test scores for specific purposes and applications. The next section examines more closely how verbal reports of thinking during testing are used to confirm a cognitive model of test performance.

Cognitive Models of Assessment

Over the last two decades, there has been increasing interest in the cognitive aspect of assessment design, development, and use. The NRC (2001) report states, "…every assessment is grounded in a conception or theory about how people learn, what they know, and how knowledge and understanding progress over time" (p. 20). In the past, tests have been based on behaviourist principles concerned mainly with the outcomes of assessment. In other words, test performance results were used to show the number of test items answered correctly or incorrectly and whether mastery or competency had been achieved. Very little attention has been given to how examinees formulate their answers or how well they understand the underlying concepts (National Research Council, 2001). The basic problem with traditional tests is that the types of inferences that can be drawn from test results are limited to identifying whether individuals demonstrate competence or not. Leighton and Gierl (2007) stress the apparent need for different types of assessments that clearly define the content domain and specify the content and processes being assessed by each test item and task in order to support more specific diagnostic inferences about examinees' cognitive strengths and weaknesses. Tests developed on the basis of cognitive models integrate the principles of cognitive psychological research with educational measurement to design assessment tasks that measure knowledge and understanding, as well as support inferences about the related cognitive processes underlying test performance (Leighton, 2004).

Cognitive models were originally conceived in the computer sciences to simulate human problem solving and mental task processing (Leighton & Gierl, 2007). According to Leighton and Gierl, cognitive models in educational measurement are derived from cognitive psychology. They are defined as a "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have acquired and to facilitate the explanation and prediction of students' performance" (p. 6). Cognitive models provide an important source of information to expand the meaning and application of assessment results and the types of diagnostic inferences that can be made from test performance.

Types of cognitive models. The development of tests on the basis of a cognitive model requires specification of the knowledge, skills, and cognitive processes required to demonstrate competence or successful performance in a particular domain (Leighton, 2004, p. 7). Leighton identified three main types of cognitive models used in educational measurement to organize and understand test performance according to three different models: (1) *domain mastery*, (2) *test specifications*, and (3) *task performance*. Leighton (2004) cautioned that ideally test development should be based on all three models; however, the current reality is that many assessments focus primarily on the first two and do not typically address the cognitive model of task performance.

The cognitive model of domain mastery is typically conceived by consulting with experts to determine the general areas of knowledge and skills required to demonstrate competence in a particular domain. The cognitive model of test specifications includes the design and selection of specific test items and tasks that are representative of the domain. The cognitive model of test specifications is developed in consultation with content experts in accord with theoretical accounts to determine the specific knowledge and skills that characterize the domain and the precise types of test items and tasks that likely elicit a representative sample of the relevant knowledge and skills. Finally, the cognitive model of task performance is generated to "validate and verify the actual set of interconnected knowledge and skills that (individuals) use to respond correctly to test items within a domain" (Leighton, 2004, pp. 7-8). Cognitive models of task performance are generated to empirically test whether the items and tasks elicit the expected types of knowledge and skills that they were designed to measure. The cognitive model of test performance is the central focus of the current study.

The main method for generating cognitive models of test performance is the collection of verbal reports during test administration by asking examinees to explain their rationale or reasons for their responses. These verbal reports are likely "the most direct evidence possible on the knowledge, strategies, and principles that examinees use to answer items on a test" (Norris, Leighton, & Phillips, 2004, p. 293). They offer insight about how examinees interpret the test items and how their background beliefs and understanding influence their judgment and performance (Norris, Leighton, & Phillips, 2004). Verbal reports can reveal patterns of thought related to underlying beliefs and understanding, misconceptions or gaps in learning and thinking, and any other knowledge, strategies, and interpretations not necessarily targeted by the test items and tasks. They are an important source of information for attaining the best explanation of test performance because they represent firsthand why examinees respond to test items the way they do. Cognitive models of test performance strengthen claims of validity and reliability by investigating the actual cognitive processing that examinees use to respond to test questions.

Rationale for cognitive models of test performance. The underlying assumption of the interpretation of test results is that correct answers to test items are an indication that examinees possess the relevant knowledge and skills required to generate the expected response(s) and on the contrary, incorrect answers reveal a lack of proficiency in the related knowledge and skills that the test is measuring. Nevertheless, without empirical evidence to show precisely what types of knowledge and skills the test items are eliciting and whether the knowledge and skills are in fact associated with the theoretical construct and what the test was intended to measure, the claim is unfounded and remains unsubstantiated.

A primary motivation for developing cognitive models of test performance stems from the realization that examinees answer test items correctly or incorrectly for a variety of reasons; some of which are legitimate, while others are not. Magone et al. (1994) pointed out, “The same answer can reflect vastly differing levels of understanding, depending on the processes which were used" (p. 329). Specifically, researchers have found cases in which examinees arrive at "correct answers to questions using patterns of thought that are unrelated to the knowledge and skills targeted by the test item (e.g., Gierl, 1997; Leighton & Gokiert, 2005; Poggio et al., 2005)" (Leighton & Gierl, 2007, p. 5). Conversely, other studies have found that sometimes examinees answer questions incorrectly for good reasons. That is, further probing showed that their thought processes were in fact based on legitimate and sound reasoning (e.g., Haney & Scott, 1987; McKay, 1974; Phillips, 1989). Thus, cognitive models of test performance are necessary and important for empirical confirmation of the *actual* thinking processes, knowledge, and skills that examinees use to respond to test items and to verify whether these processes align with the expected types of responses (Leighton, 2004). The cognitive model of test performance is used to investigate an explanation for the causes of test performance.

According to Norris, Macnab, and Phillips (2007), while cognitive models of test performance provide information about what examinees think and do during testing, they fail to speak to the quality of their thought processes and performance,

> Cognitive models can capture the features themselves and tell us that they occur. However, it is not enough to know that examinees followed certain principles and strategies when arriving at their answers. An account is needed of the appropriateness of those strategies and principles in context, and cognitive models cannot provide such an account. Cognitive models can represent the use of information and what examinees' thinking covers, but they cannot tell whether appropriate use was made of the information or whether examinees' thinking was complete. Such normative judgments fall outside the model" (p. 81).

Cognitive models are bounded by specific parameters which govern what they can and cannot say about test performance. Norris, Macnab and Phillips (2007) explain that while cognitive models can be used to discover the types of cognitive processes entailed in test performance, they are limited in a number of different ways: (a) they do not provide a complete explanation of performance, (b) they do not specify degrees of understanding, and (c) they do not indicate whether test responses are correct or whether response justifications are appropriate. Moreover, cognitive models cannot pinpoint precisely what a test is testing without normative appraisal of which responses are acceptable.

Cognitive and normative models. To determine exactly what a test is testing and whether it is measuring understanding in a particular domain, a combination of cognitive and normative models is necessary. The integration of cognitive and normative models is used to explain test performance and to distinguish between levels of understanding.

> The explanation would be fleshed out by providing detail on the underlying cause of the performance in terms of the knowledge, cognitive strategies, and principles the examinee employed, and a normative appraisal of the appropriateness of the examinee's thinking. The explanation would provide a cognitive model whereby we could see how the performance arose and why it was or was not successful (Norris, Leighton, & Phillips, 2004, p. 293).

The integration of cognitive and normative models strengthens the types of inferences and interpretations that can be made about test performance. Cognitive models provide an account of the reasoning that examinees use to respond to test items and how their reasoning affects their understanding. Whereas, normative models provide an account of what defines understanding on various test items, what is required for determining correct answers, which types of reasoning are valid, and whether examinees demonstrate understanding and good thinking (Norris, Macnab, & Phillips, 2007, p. 88). Normative models analyse the quality of reasoning and the extent to which the justifications for responses are relevant, consistent, and complete with the available information. Cognitive and normative models are used in combination to distinguish achievement from lack of achievement and to appraise the quality of thought processes and strategies used. They also offer insight into the legitimacy of the explanations and approaches to test items and tasks and whether the cognitive processes are associated with competence in the particular domain. Together, the models are used to explain why examinees answer test questions the way they do; to identify their motivation, intentions, and reasons for their answers; and, to predict future performance on tasks requiring similar kinds of knowledge and understanding.

Statement of the Problem

The main objective of early language and literacy assessment is to gather information about children's competencies and more importantly to identify those children who experience difficulties with particular aspects of early language and literacy acquisition. Typically, when children show signs of learning difficulty, a battery of language and literacy assessments is administered to measure skills such as receptive and expressive language; print concepts, environmental print; alphabetic knowledge; and, phonological and phonemic awareness. Based on the test performance results from multiple measures, professionals draw inferences and make decisions about the nature of children's strengths and weaknesses and the type of interventions that are thought to be appropriate. However, it is widely known that these interventions are often less than effective and sometimes even detrimental (Denton et al., 2006; Phillips, Norris & Steffler, 2007). Furthermore, the decisions are often based on an incomplete and inadequate representation of early language and literacy development depending on the types of measures that were used and when the assessments were administered.

There are a number of issues with the current approach to early language and literacy assessment. At present, there is no single Canadian measure to assess the broad range of skills associated with early language and literacy development (Lonigan, 2006; Lonigan, McDowell, & Phillips, 2004). Moreover, the types of early language and literacy measures that are available tend to be narrowly conceived and inadequate for measuring the vast array of skills that comprise early language and literacy. In addition, few early language and literacy assessments provide useful diagnostic information about children's specific strengths and weaknesses or clear direction for the types of interventions that target specific aspects of early language and literacy ability. Furthermore, conventional measures fail to provide sufficient insight into what children know and think (Pellegrino, Chudowsky & Glaser, 2001). Thus, this is an opportune time, based on advances in the sciences of thinking and learning, to study what young children know, how they know it, and how well they are able to use that knowledge. Given what is known, it is imperative to work towards the identification and prevention of early language and literacy difficulties through valid, timely, and effective assessment which in turn can inform better interventions.

Research Purpose and Questions

The purpose of the present study was to explore the relationship between children's performance (three to eight years of age inclusive) on the *Test of Early Language and Literacy (TELL)* and what they report thinking and reasoning as they complete the *Oral Narrative (ON)* and *Oral Reading and Reading Comprehension (OR-RC)* test items. Specifically, the study was guided by one main research question: What thinking and reasoning do children report as the basis of their responses to the *TELL ON* and *OR-RC* test items? In other words, how sound is their thinking and reasoning? The study also addressed several related questions:

a) What information sources did children use (i.e., text information and background knowledge) and how did they use the information sources to respond to the *TELL ON* and *OR-RC* test items? Did they use the information appropriately?
b) Were children's test performance scores related to their verbal report scores on each test item? Specifically, did children who performed well on the test items also show good thinking and reasoning and conversely, did children who performed poorly on the test items show poor thinking and reasoning?

Verbal reports of thinking were the primary data source for examining the processes underlying item responses and test performance on the *TELL.* In addition, test performance and verbal report data were analysed for evidence of construct validation as to whether the items on the *TELL* comprehension assessments measure the skills and processes they were designed to measure.

Rationale for the Research

The verbal report data provided important insight about what children attend to and what information they used to respond to early language and literacy tasks. This research advanced a more complete and thorough understanding of young children's thinking, reasoning, and cognitive processing in the context of a specific and new early language and literacy assessment. Thus, eliciting children's explanations of why they respond the way they do challenged conventional assessment approaches and provided support for more valid testing methods and interventions.

The data from this study was used to compare children's responses against conventional test theory. Typically, conventional test theory supports a unilateral interpretation of test scores focused mainly on the outcome of test performance (i.e., the number of questions answered correctly or incorrectly). Consequently, interpretations and decisions are made on the basis of test results alone. The underlying presumption is that test questions are answered correctly or incorrectly for the right reasons. That is, correct answers are indicative that the examinee has the required knowledge and understanding of the construct being measured and incorrect answers are indicative that the examinee lacks sufficient knowledge and understanding of the concepts being tested. However, existing research provides contrary evidence to show that these assumptions are false (Leighton & Sternberg, 2003; Norris, Leighton & Phillips, 2004; Norris, Mcnab & Phillips, 2007). Studies using verbal reports during test administration reveal that sometimes examinees answer test questions correctly for the wrong reasons and incorrectly for the right reasons. The problem with conventional test theory is that decisions based on faulty assumptions and test results alone can lead to misidentification and misdiagnosis of the nature of children's difficulties and hence, to inappropriate language and literacy interventions. Consequently, validation studies based on verbal report data and a cognitive and normative model of test performance provide an important source of evidence for determining precisely what test items are actually measuring.

Four chapters remain to this body of work. Chapter 2 is divided into four main parts, namely the nature of early language and literacy development; the assessment of the aspects of language and literacy; the theoretical perspective of early reading development; and finally, an overview of the methodology of think-aloud and protocol analysis. Chapter 3 follows with a discussion of the research methodology including the instruments used; sample; data collection, preparation, analysis, and interpretation; and, the ethical considerations. Chapter 4 presents the results and discussion of the oral narrative listening comprehension test and thinking performance as well as the oral reading and reading comprehension test and thinking performance distribution and closes with a summary. The fifth and final chapter provides a summary of the results; limitations; conclusions and contributions of the study; and recommendations for future research and practice.

CHAPTER 2: LITERATURE REVIEW

Overview

The research reviewed is specifically relevant to the construct validation of an early language and literacy assessment using verbal reports of thinking to establish a cognitive model that explains performance. The review of literature has four main sections. The first section includes the research pertaining to early language and literacy development. The second contains a critical review of early language and literacy assessment practices. The third focuses on the theoretical perspective of early reading development. The final section includes a comprehensive review of the reading research with think-aloud and protocol analysis methodology.

Early Language and Literacy Development

Components of Early Language and Literacy Development

The perspectives of early language and literacy have evolved over time. Earlier perspectives emphasized the continuity between emergent literacy skills (i.e., oral language) and later conventional literacy skills (i.e., reading and writing) in which the earlier skills provided the foundation for the development of later skills (Storch & Whitehurst, 2002; Wilson & Lonigan, 2010). In contrast, contemporary views focus more on the simultaneous development and reciprocal relationship between oral language and code-related skills (i.e., learning about print features). Multiple perspectives of emergent literacy acquisition and early reading development (Adams, 1990; Scarborough, 1998; Storch & Whitehurst, 2002; van Kleeck, 1998; Whitehurst & Lonigan, 1998) are reviewed to highlight differences in the conceptualization of the components of emergent and later literacy.

Emergent Literacy Acquisition

In 1998, Whitehurst and Lonigan proposed a new framework for the components of emergent literacy. From their perspective, emergent literacy relies on the development of precursor skills, knowledge and attitudes, and the environment to support later reading and writing development. Whitehurst and Lonigan (1998) classified emergent literacy skills into two separate domains: *inside-out* and *outside-in*. Each domain was comprised of various information sources including the processes and skills that children must acquire in order to develop early reading and writing abilities. The respective domains and related skills and processes were interdependent and developed concurrently. Early reading and writing acquisition depended on the successful coordination and execution of the combined skills and processes.

According to Whitehurst and Lonigan (1998), the *inside-out* domain focuses specifically on the relationship between oral and written language pertaining to the specific attributes, rules, and conventions that govern the translation of print to sounds and sounds to print in an alphabetic system. The *inside-out* domain includes the code-related components of reading comprised of five main processes and skills: knowledge of graphemes (i.e., recognition and identification of the letters of the alphabet), phonological awareness (i.e., recognition of rhyme, recognition and manipulation of syllables and phonemes), syntactic awareness (i.e., sentence grammar), phoneme-grapheme correspondence (i.e., letter-sound correspondence), and emergent writing (i.e., phonetic spelling).

The *outside-in* domain focuses on the meaning of oral and written language. The specific processes and skills related to general knowledge and language are required for the interpretation and understanding of the printed word. The *outside-in* domain includes four main areas: language (i.e., receptive and expressive vocabulary; semantic, syntactic, and conceptual knowledge), narrative (i.e., understanding and producing narrative; story schemas), conventions of print (i.e., knowledge of the direction and format of print; word spacing; left-to-right and front-to-back orientation), and emergent reading (i.e., attending to environmental print and pretending to read). Children must draw upon several sources of knowledge and information to understand the meaning of oral and written language including background knowledge (i.e., prior experience and world knowledge), semantic and syntactic knowledge (i.e., vocabulary knowledge and the meaning of words and sentences), and knowledge of the written context (i.e., story schemas).

Whitehurst and Lonigan (1998) also identified a set of other factors which influence children's emergent literacy development. These factors include phonological memory (i.e., short term memory of phonological information including numbers, nonwords, and sentences), rapid naming of lists of random letters, numbers, or colours, and print motivation including interest in reading books. Much of the current research on emergent literacy development is based on the conceptualization by Whitehurst and Lonigan (1998). There is general agreement that the skills and processes included in each domain are important aspects of emergent literacy development and the related sets of processes and skills develop simultaneously and at different rates depending on children's individual experiences with oral and written language. Based on this conceptualization, research has begun to focus on studying the relationship between and among the various components of early language and literacy development.

More recently, Storch and Whitehurst (2002) adopted a slightly different perspective of emergent literacy based on Whitehurst and Lonigan's original framework. They grouped emergent literacy skills into two different categories: (1) *code-related* skills and (2) *oral language* skills. Similar to Whitehurst and Lonigan, the two sets of skills were considered to be the precursors for later reading achievement. The *code-related* domain includes five main types of skills: (a) conventions of print (i.e., print directionality and format), (b) emergent writing (i.e., writing letters of the alphabet and own name), (c) knowledge of graphemes (i.e., identifying the letters of the alphabet), (d) grapheme-phoneme correspondence (i.e., recognition of letters and their corresponding sounds), and (e) phonological awareness (i.e., manipulation of the individual sounds in words). The *oral language* component covers four main areas: (a) semantic (i.e., word knowledge, expressive and receptive vocabulary), (b) syntactic (i.e., knowledge of word order and grammar), (c) conceptual knowledge (i.e., strategies and processes), and (d) narrative discourse skills (i.e., generating and retelling narrative stories, comprehension, and memory).

Storch and Whitehurst (2002) conducted a longitudinal study of the developmental progression and relationship between the various emergent literacy skills at different stages of development in low-income children. They identified a strong reciprocal relationship between the oral language and code-related skills during the preschool years. According to Storch and Whitehurst (2002), the acquisition and coordination of code-related and oral language skills during preschool and kindergarten is the foundation for reading and writing development in the early elementary grades. They claim that the relationship between code-related and oral language skills, in conjunction with later reading development, changes over the course of the emergent and early literacy period. For example, whereas code-related skills influence oral language skills primarily during the emerging stages of development, code-related skills have little to no effect on oral language skills during grades one and two. However, code-related skills influence reading development in grades one and two. In addition, Storch and Whitehurst report that oral language skills influence code-related skill development mainly during the preschool period but had little to no direct effect on reading development in grades one and two and only an indirect influence on reading comprehension in grades three and four.

Basically, the two frameworks by Whitehurst and Lonigan (1998) and Storch and Whitehurst (2002) are similar. They both conceptualize emergent literacy using similar components but the categories of skills are classified and defined differently to show the evolution of their developmental model.

Early Reading Development

Another important aspect for understanding early language and literacy development is to examine the conceptualization of early reading acquisition. Scarborough (1998) proposed a model of skilled reading development using an analogy of separate strands of a twisted rope to represent the integration of different subskills and abilities. Skilled reading was characterized by two main aspects: the ability to read words and to derive meaning from printed text. To achieve this goal, readers must develop several underlying component skills (or "strands") related to two specific processes: language comprehension and word recognition. According to Scarborough, learning to read requires the development, coordination, and execution of both word recognition and text comprehension and their related component skills. Children who have difficulty developing or integrating the necessary component skills in one or both processes are likely to experience problems in learning to read.

The first aspect of skilled reading has to do with language comprehension. Language comprehension is comprised of several component skills related to oral language: background knowledge (i.e., facts, concepts), vocabulary (i.e., breadth, precision, links), language structures (i.e., syntax, semantics), verbal reasoning (i.e., inference, metaphor), and literacy knowledge (i.e., print concepts, genres) (Scarborough, 1998). In order to comprehend oral and written texts, readers must understand the words and messages conveyed in the text, the syntactic and semantic relationships between the words, and have the background knowledge and inferential skills to interpret the meaning of texts.

The second aspect of skilled reading includes word recognition. Word recognition focuses on the underlying component skills for translating print into language: phonological awareness (i.e., syllables, phonemes), decoding (i.e., alphabetic principle, spelling-sound correspondences), and sight recognition (i.e., identifying familiar words). In order to learn to read words, children must understand the basic principles of the English orthography or the "alphabetic principle": (a) spoken words are comprised of a series of sound units (phonological awareness), (b) letters or graphemes are used to represent the corresponding sound units in words (the alphabetic principle), (c) systematic letter groupings or spelling patterns and their corresponding sounds are used to pronounce words (decoding), and (d) some words fail to correspond directly to common sound-symbol correspondence and therefore must be memorized (i.e., "know" or "rough"). Children actively compare their pronunciation of a word with words in their mental lexicon to confirm whether the word sounds are correct according to their linguistic knowledge and understanding (Scarborough, 1998). Ultimately, proficient word recognition is characterized by a large sight vocabulary of words recognized automatically.

Based on a comprehensive review of research over 20 years, Scarborough (1998) showed that letter identification and phonological awareness, vocabulary, sentence and story recall, and concepts of print are reliable predictors of later reading development and therefore warrant early assessment. The different aspects of early language and literacy assessment and the types of assessments that are currently available are discussed further in the next section.

Early Language and Literacy Assessment Practices

Aspects of Early Language and Literacy Assessment

An important aim of early language and literacy assessment is to ascertain, as early as possible, which children are at-risk for developing later reading difficulties. The first challenge is to determine precisely what to assess with respect to the specific risk factors or predictors of later reading achievement. Correlational research provides one source of information about which aspects of early language and literacy are associated with later literacy achievement. Specifically, the research has identified certain precursors of conventional literacy and approximately when these different precursors exert their greatest influence on later literacy development. Correlational research typically examines the relationship between measures of various early language and literacy components during the preschool period with measures of conventional literacy aspects (i.e., decoding and/or comprehension) during the early elementary grades.

During the last two decades, several research reviews show converging results about the types of skills that lay the foundation for later literacy development (National Early Literacy Panel, 2008; Scarborough, 1998, 2005; Snow, Burns, & Griffin, 1998; Whitehurst & Lonigan, 1998; Storch & Whitehurst, 2002). The main aspects of early language and literacy that have been associated with later literacy development relate to three broad areas: (1) oral language, (2) phonological processing, and (3) print awareness. Each aspect is comprised of a number of different subskills and components.

Oral language. The specific oral language variables that predict later decoding and reading comprehension include receptive and expressive vocabulary (i.e., labeling and naming, and definitional vocabulary) (Goswami, 2001; Nagy & Herman, 1987; National Early Literacy Panel, 2008), oral narrative discourse (i.e., story generation and retell) (Bishop & Edmundson, 1987; Feagans & Applebaum, 1986; McCabe & Rollins, 1994; Vernon-Feagans et al., 2001), understanding grammar (i.e., syntax and morphology) (Scarborough, 1990, 2001; Tunmer et al., 1988), and listening comprehension (i.e., verbal memory for sentences and stories) (Kendeou, Bohn- Gettler, White, & van den Broek, 2008; Kendeou, van den Broek, White, & Lynch, 2009; Storch & Whitehurst, 2002). According to Paris, Carpenter, Paris, and Hamilton (2005), language skills, vocabulary skills, and narrative reasoning influence the development of reading comprehension and overall reading achievement.

Phonological processing. One of the most consistent findings is that phonological processing is an important determinant of future reading achievement (Lonigan et al., 2000; Wagner, Torgesen, & Rashotte, 1994; Wagner et al., 1997). Phonological processing includes phonological sensitivity (i.e., detect and manipulate sound structure), memory (i.e., recall sound-based information), and naming (i.e., retrieval of phonological information from permanent memory) (Adams, 1990; Snow, Burns, & Griffin, 1998; Stanovich, 2000). In addition, phonemic awareness accounts for the greatest influence on later reading success (NICHD Early Child Care Research Network, 2005).

Phonemic awareness skills encompass the ability to perceive and manipulate the sound structure of language (i.e., words, syllables, onsets-rimes, and phonemes).

Print awareness. Print knowledge, or the *concepts of print*, has also been highly associated with later reading and writing achievement. Print knowledge includes the form and function of print. These skills entail knowledge of the alphabet (i.e., letter names and sounds), awareness of the alphabetic writing system (i.e., translating units of print (graphemes) to units of sound (phonemes)), written language (i.e., translating units of sound into units of print), and the directionality of print. At the beginning of kindergarten, letter knowledge is the single best predictor of eventual reading achievement (Adams, 1990).

The onset and development of these different skills emerges throughout the preschool period. Research clearly shows significant variability in skill development during early childhood (NICHD Early Child Care Research Network, 2005). Although many of these skills are related to subsequent reading achievement, the acquisition of the various components does not necessarily guarantee later literacy development. However, the assumption is that the more proficient children are in these skills, the more likely they are to experience success in learning to read and write, to benefit from early reading and writing instruction, and to excel in reading and writing (Lonigan, McDowell, & Phillips, 2004). In contrast, children who lack ability in one or more of these areas may experience problems learning to read and write and face greater challenges with early reading and writing instruction (Snow, Burns, & Griffin, 1998).

Research shows that the various precursor skills are interrelated and influence different aspects of literacy development at different points in time (NICHD Early Child Care Research Network, 2005; Speece, Roth, Cooper & de la Paz, 1999; Whitehurst & Lonigan, 1998). According to Whitehurst and Lonigan (1998), letter knowledge, phonological sensitivity, and emergent writing are among the strongest predictors of reading at the end of first grade when the focus is mainly on learning how to decode print. The code-related skills maintain a strong and direct influence on reading achievement during the early elementary period. Additionally, vocabulary knowledge in kindergarten is one of the best predictors of reading comprehension in third and fourth grade and beyond (Sénéchal, Ouellette, & Rodney, 2006; Storch & Whitehurst, 2002; Wood et al., 2005). Dickinson et al. (2003) suggested that oral language skills exert their greatest influence on emergent literacy development during the preschool years and on later literacy development in third and fourth grade when the focus shifts to meaning-making and comprehension; with only indirect effects during first and second grades when the emphasis is on learning to decode print (Evans, Shaw, & Bell, 2000; Sénéchal & LeFevre, 2002; Storch & Whitehurst, 2002).

Storch and Whitehurst (2002) reported a strong positive association between code-related skills (i.e., phonological awareness) and oral language skills (i.e., vocabulary) during the preschool period (e.g., Burgess & Lonigan, 1998; Chaney, 1992; Lonigan, Burgess, Anthony, & Barker, 1998). However, the association becomes less significant once children enter school. Measures of phonological awareness and print knowledge (i.e., alphabet knowledge and concepts of print) during preschool and kindergarten are important determinants of early reading acquisition. Similar to Whitehurst and Lonigan (1998), they surmised that code-related skills were foundational to the early stages of learning to decode print, whereas oral language skills were key to reading comprehension.

Several important conclusions are drawn from the research involving early language and literacy assessment (NICHD Early Child Care Research Network, 2005; Scarborough, 2005). First, there is considerable variability in the development of skills during the preschool period. In a systematic review of national data in the United States, the Early Childhood Longitudinal Study (ECLS) showed that 66% of children entering kindergarten had already mastered the most prevalent predictors of later literacy achievement, namely, letter recognition and other print concepts (i.e., that print represents language and is read from left-to-right and top-to bottom, in the case of English) (Zill & West, 2001). In addition, 29% of children associated letters with their sounds at the beginning of words and 17% recognized the sounds at the end of words. In addition, 2% of children beginning kindergarten read simple sight words and 1% read complex words in sentences and knew how to read. Conversely, 18% of the children starting kindergarten failed to demonstrate concepts of print or book knowledge and 34% failed to identify the letters of the alphabet.

Another significant finding is that the strength of correlations between measures at three- and four-years of age is just as strong, or stronger as those at five-years of age which suggests the potential for early identification of children with difficulties from a very early age (NICHD Early Child Care Research

Network, 2005). Third, children who experience difficulty in any aspect of oral language during the preschool period are at greater risk for experiencing difficulties in learning to read and write. Finally, no single measure of any unitary aspect of early language and literacy is sufficient to determine definitively which children will inevitably experience reading and writing difficulties. Persistent calls are made for a more comprehensive approach to early language and literacy assessment. To ensure greater accuracy in identifying those children who may need further intervention, it is important to measure a wide range of skills known to be associated with early literacy development (i.e., oral language, phonological processing, and print knowledge) at different points in time during early childhood. There is still much to be learned about how various aspects of early language and literacy relate to one another and to the eventual acquisition of reading and comprehension (NICHD Early Child Care Research Network, 2005). It is also important to point out that any research involving early language and literacy assessment is only as good as the measures used to assess the different aspects of the construct. Thus, research findings are limited by the types and quality of the available measures and exactly what areas of early language and literacy can be assessed.

In a landmark meta-analysis of research over 20 years, Scarborough (1998) found that there was wide variability in the quality of measures used in correlational research to predict future reading scores. To ensure prediction accuracy and useful application of results, she advised that kindergarten screening measures should center on the strongest language and literacy predictors including letter knowledge, print concepts, phonological awareness, expressive vocabulary, sentence imitation, and story recall. In addition to these areas, Davis et al. (2007) added that screening batteries should also include measures of word reading ability, orthographic and syntactic knowledge, and background knowledge. Ultimately, the prevention of reading problems depends on the ability to identify specific areas of deficiency that most closely relate to conventional reading and writing development. The accuracy and significance of early language and literacy assessment and future research depends on whether the measures that are used are of the highest quality. The next section explains how early language and literacy development is measured and discusses the types of assessment instruments that are currently available.

Current Early Language and Literacy Assessment Instruments

It is clear that the construct of early language and literacy is both complex and multidimensional which makes the matter of measurement complicated. The unfortunate reality is that until recently there has been no single empirically-validated test or screening instrument to measure the many facets of early language and literacy for children between three and eight years of age (Lonigan, McDowell, & Phillips, 2004). Therefore, measurement of early language and literacy has been based mainly on an aspectual approach focused on separate dimensions of the broader construct. However, adequate assessment of early language and literacy development must rely on multiple assessments of different aspects of the construct. Moreover, inferences that can be made about early language and literacy development depend on how the different aspects are operationalized and measured. The challenge is whether the available measures are of sufficient quality to adequately assess the full-range of dimensions associated with the construct.

Over the past twenty years, several reviews of early language and literacy and early reading assessments have been conducted (Hayward et al., 2008; Kame'emui et al., 2006; Lonigan, McDowell, & Phillips, 2004; Meisels & Piker, 2000; Stallman & Pearson, 1990). The reviews illustrate that the quality and types of assessment instruments currently available are dismal at best. The main limitations concern the alignment between current perspectives of early language and literacy and the way that each respective aspect is operationalized and measured. Other issues relate to the now-dated norming samples for standardization and the narrow psychometric properties of available tests.

Teale (1990) and Stallman and Pearson (1990) identified a significant discrepancy between the types of assessments used to measure early language and literacy development and the theoretical changes in the conceptualization of early language and literacy. Despite recent developments in the field over the past two decades, the lack of appropriate assessments remains a pervasive problem.

In the most comprehensive review to date, Hayward et al. (2008) examined 25 standardized early language and literacy assessment instruments developed between 1959 and 2007 to measure skills in the following areas: general language, vocabulary and grammar, narrative, phonological awareness (i.e., segmenting, blending, elision, or rhyming), print knowledge (i.e., environmental print or alphabet knowledge), reading (i.e., single word reading, decoding, or comprehension), writing (i.e., letter formation, capitalization, punctuation, conventional structures, word choice, and spelling) and achievement. Their review revealed many important limitations in the current tests. Most of the early language and literacy tests centered on code-related skills such as phonological awareness, alphabet knowledge, print awareness, letter-sound correspondence, and sight words. Oral language tests tended to be unitary measures of expressive or receptive vocabulary. These aspects of early language and literacy are likely most prevalent because they consist of sets of skills that can be clearly defined and relatively easy to measure; the concept is either known or not. However, inclusion of these aspects alone represents a rather narrow perspective of the construct.

Additionally, Hayward et al. (2008) found significant variability in the composition and size of the norming samples. As well, some of the revised tests relied exclusively on outdated norms generated during the original development of the test. That is, the norms were not properly updated for the newer versions of the test. These norming issues are important for the kinds of interpretations that can be made from test results and the relevance of using particular tests with certain populations.

Another noteworthy limitation was that many of the tests were originally developed more than 30 years ago. Although several editions of the tests have been produced over the years, there was very little change in how the construct was operationalized or the types of test items included in the different iterations of the tests. Thus, there is a lack of congruence between current knowledge of the construct and what the available tests measure. Hence, many early language and literacy assessment instruments perpetuate a narrow and outdated view of the construct (Afflerbach, 2007). In addition, the psychometric analyses of these tests tended to be limited and some tests failed to report any information pertaining to reliability or validity evidence, which makes their use questionable.

In another review of 29 early reading assessments for kindergarten to grade three, Kame'enui et al. (2006) found that many tests lacked sufficient coverage of particular skills related to early reading as well as incomplete sampling of particular domains. Similar to the early language and literacy assessments, many of the early reading tests were originally developed many years ago and do not reflect the most current views of reading development. The skills typically measured failed to align with the five main components of early reading: phonemic awareness, phonics, fluency, vocabulary, and comprehension (National Reading Panel, 2000).

Kame'enui et al. (2006) noted that the psychometric evidence for the early reading tests was minimal and the reliability and validity coefficients were "disappointingly low" across all assessment purposes (i.e., screening, diagnosis, progress monitoring, and outcome evaluation) and essential reading components (i.e., phonemic awareness, phonics, fluency, vocabulary, and comprehension). The reliability evidence was often restricted to evaluations of internal consistency with very little reference to other types of reliability (i.e., test-retest, inter-rater, or alternate forms). Furthermore, validity data included mostly evaluations of concurrent relationships with other measures. The main issue with concurrent validity is that one measure is compared with another that supposedly measures a similar construct. This type of validity has obvious limitations given the fact that many measures are of questionable quality. For example, if one measure is compared with another measure that is fundamentally flawed in its conceptualization of the construct, then concurrent validity between the two measures is of little value. In such instances, the only inference that follows is that both measures are similarly flawed in their conceptualization and measurement of the construct. Consequently, Kame'enui and his colleagues (2006) concluded that "many measures do not provide enough evidence of trustworthiness to warrant use" (p. 9).

The most salient finding from these reviews of assessment instruments is that only certain aspects of early language and literacy development are represented in the assessments available to date while other aspects are lacking or non-existent. Specifically, there are substantially more standardized measures for code-related aspects of early language and literacy associated with the alphabetic system and decoding print than for meaning-based components of language comprehension related to reading comprehension and oral narration (Lonigan, 2006; Paris & Hoffman, 2004). Assessments based solely on code-based measures provide an inadequate and incomplete view of a child's early language and literacy development.

Even though comprehension is clearly important to early language and literacy development, surprisingly little attention has been given to this aspect in the preschool and early grades. Moreover, the few early reading and oral language measures that have been developed vary tremendously. They tend to focus on unitary aspects of the construct such as the conventions of print or environmental symbols; phonemic decoding and sight word efficiency; syntactic structure and grammar in sentences; vocabulary understanding; story retelling; oral reading rate, accuracy, fluency, and comprehension of stories with multiple-choice comprehension questions; listening comprehension by selecting a picture that matches orally-presented sentences or answering questions about a picture stimulus; and cloze passages to identify the missing words in sentences. These types of measures, when used alone, typically operationalize comprehension to be narrow and one-dimensional, which fails to coincide with the most current understanding of comprehension as a complex, multidimensional construct. Recently, Paris (2007) noted,

> Reading comprehension of students in grades K-3 is usually assessed with formative (i.e., informal diagnostic tasks that inform instructional decision-making) rather than summative (i.e., scores that summarize performance and allow comparisons among test-takers) measures because the main purpose of assessment with beginning readers is to identify children who need additional instruction. This 'low-stakes' approach may be partly responsible for the lack of rigorous evidence about the validity, reliability, and utility of early assessments (p. 4).

He explained that there are three primary types of informal comprehension assessments currently in use: oral retellings, answering comprehension questions, and completing cloze tasks. The three types of measures are used to assess listening, viewing, and reading comprehension.

In the past decade, there has been a call for the development of "comprehensive, authentic, and valid" comprehension measures for young preschool and early elementary-aged children (Paris & Paris, 2003; van den Broek et al., 2005). “Given the recent advances in our understanding of the complex nature of the reading process, it is time that tests start focusing on the rich multifaceted aspects of comprehension” (van den Broek et al., 2005, p. 126). Furthermore, the call is based on research evidence which shows that reading comprehension skills develop simultaneously with basic literacy skills (i.e., phonological awareness, letter knowledge, vocabulary, etc.) during the preschool period (van den Broek et al., 2005). Moreover, evidence shows that young children make use of inferential processes to “identify meaningful relations and to establish coherence” between events in their life experience or in a literacy context; similar processes to those found in adult reading (van den Broek et al., 2005, p. 115). Specifically, children as young as two-years of age are capable of establishing meaningful connections and making causal inferences between events after viewing a series of pictures, listening to stories, or watching television episodes (Bauer, 1996, 1997; Trabasso & Nickels, 1992; van den Broek, Lorch, & Thurlow, 1996; Wenner & Bauer, 2001).

Previous research shows that narrative comprehension is the basis of future reading comprehension and it is important to examine how these skills manifest in the early years. Two promising new informal measures of children’s narrative comprehension have been developed to assess children’s comprehension of non-print materials using televised narratives (van den Broek, Kendeou, Kremer, Lynch, Butler, White, & Lorch, 2005) and wordless picture books (Paris & Paris, 2003; van Kraayenoord & Paris, 1996). These studies found that preschool and early elementary children’s comprehension of televised narratives and wordless picture books correlates significantly with their reading comprehension measured one to two years later (Paris, 2007). These types of comprehension measures have shown that it is important and possible to measure the comprehension skills of children in preschool and early elementary.

The matter of assessment is important because the measures that are available inadvertently shape how the construct is operationalized in practice. Accordingly, if the only early language and literacy assessments in use focus on measuring only code-related skills, these types of skills will likely take precedence and become the primary targets of instruction. Consequently, the emphasis of assessment should focus on developing new, empirically-validated measures that span the range of the different types of skills associated with current theoretical views of early language and literacy development. The next section examines the theoretical perspectives of early reading development.

Theoretical Perspectives of Early Reading Development

Several theories account for different aspects of the two main components of early reading development: decoding print and understanding the meaning of print. Current understanding of early reading development is informed by theories of word recognition development (Adams, 1990; Chall, 1996; Ehri, 1995, 1998) and theories of comprehension (Anderson & Pearson, 1984; Kintsch, 1994; Norris & Phillips, 1987; van Dijk & Kintsch, 1983). The theories highlight the complexity of early reading development.

Theoretical Perspectives of Word Recognition

Reading process model. The first models of word recognition were developed over two decades ago. Specifically, Seidenberg and McClelland (1989) introduced a theoretical model of visual word recognition that was later reconceptualized by Adams (1990) as a connectionist model to include cognitive and perceptual connections necessary for word recognition and comprehension. Marilyn Adams' book (1990), *Beginning to Reading: Thinking and Learning about Print*, was described as a "near exhaustive compendium of the best research on early reading acquisition" (Stanovich, 1992). The models provide important insights into the roles and functions of four separate processors and the relationship between and among them in the reading process.

The theoretical model of visual word recognition and pronunciation proposed by Seidenberg and McClelland (1989) was derived mainly from examining and modeling the processes of skilled reading. Briefly, the model conceptualizes the reading process as a complex system involving the simultaneous development, coordination, and integration of four complex processors: (1) orthographic, (2) phonological, (3) meaning, and (4) context. The four processors focus on two main aspects of reading: the translation of print into sound and the interpretation of meaning from print. Adams (1990) explained that while each processor serves a separate function in the reading process, skilled reading requires the execution of all four processors. Although the processors are not well-integrated in the beginning stages of learning to read, successful reading achievement relies on the development and coordination of all four processors. Thus, the processors function in an interdependent, complementary, and compensatory manner during the acquisition and mastery of reading ability. All four processors are equally important and necessary for skilled reading.

The orthographic processor focuses on visual perception and interpretation of the printed word including recognition of individual letters and ordered letter sequences. According to Adams (1990), as readers attend to printed words through repeated exposure, they begin to recognize and associate common sets of ordered letter sequences. These "associative connections" allow readers to learn and recognize common spelling patterns (such as: st-, th-, pr-, spl-, -ack, -ell, -it, -op, -unk) used in English orthography and through repeated exposure, these letter-pattern associations are consolidated and eventually recognized automatically. The development of the orthographic processor demands efficient individual letter recognition and identification of spelling patterns in words.

The phonological processor focuses on auditory perception of the sound units associated with words, syllables, and phonemes. Simply put, it focuses on the pronunciation of words and the individual phonemes or sound units of words. Adams (1990) explained that visual processing of a sequence of letters activates the corresponding sound units in the phonological processor. Similarly, associative links between the sound and visual components are solidified and automaticity is achieved through repeated activation of the individual phonemes and their related visual representations.

The meaning processor focuses on the reader's understanding of word meanings, vocabulary, and concepts. According to Adams (1990), word knowledge and understanding depends on previous encounters with the word. The meaning of a word is derived from the collective experiences and exposure to the specific word in different contexts. Novel encounters with the word in different contexts produce changes in the original conceptualization of the word. That is, understanding of the word evolves based on whether the subsequent encounter with the word confirms, adds to, or revises previous understanding. Encounters that evoke similar understandings of the word activate overlapping sets of meaning units while encounters that perpetuate new and different understandings of the word activate different meaning units which broaden and expand the original conceptualization of the word. Similar to the other processors, associative links between words and their meaning are achieved through repeated exposure which ultimately promotes a more complete and comprehensive understanding of the word over time. The orthographic, phonological, and meaning processors are integrally- and reciprocally-related because each processor supports and stimulates activation of the others.

The context processor manages the "ongoing, coherent interpretation of text" (Adams, 1990, p. 138). During activation of this processor, readers monitor and integrate the meanings of individual words and phrases into a "composite interpretation" of the overall text meaning. Readers revise and update their original understanding of the text based on the new information that they encounter as they read the text. Meaningful interpretation of the text may require the integration of information from different parts of the text, as well as information extraneous to the text from the readers' background knowledge or experiences (Adams, 1990, p. 142).

In 1998, van Kleeck used Adam’s (1990) connectionist model to show how the four processors emerge during the pre-literacy stage of early childhood development. She identified a set of related subskills for each processor. According to van Kleeck (1998), the related sets of subskills in each domain emerge prior to learning how to read print. Thus, children acquire many of these subskills simultaneously in the context of oral language. The subskills, in turn, become the foundation for later literacy development. She explained that the orthographic and phonological processors are important for learning about print form, whereas the context and meaning processors provide the foundation for children's early understanding of print meaning.

According to van Kleeck (1998), the context processor is activated from the earliest stages of cognitive development through early experiences with and exposure to oral language. The subskills of the context processor include: world knowledge, syntactic knowledge, narrative development, book conventions, and reasoning (i.e., functions of print). Initially, these subskills begin to develop in response to early stimulation and experiences with language and books. Children’s world knowledge and understanding about events and objects is based on their early experiences in their environment. During the pre-literacy stage, children draw upon their world knowledge and prior experience to make sense of oral texts and in turn, what they learn from oral texts adds to their world knowledge and understanding. In the early stages of development, children’s syntactic knowledge is based on their initial exposure to grammatical structure in the English language through personal interactions and by listening to stories.

This syntactic knowledge is the basis for later word recognition and comprehension of written language. Children begin to acquire narrative skills by listening to stories. They learn about the composition and function of different aspects of story grammar (i.e., setting, characters, initiating events, internal responses, attempts, consequences, and reactions) and how the aspects and story episodes are related. They also learn much about the basic attributes of books and print from their early experiences of listening to books read aloud. For instance, they learn that books are read and that illustrations and print represent important aspects of the story and carry meaning. During early interactions with stories, children begin to engage cognitively and learn to reason about information presented in books. In their early experiences with stories, children ask questions, reason about the events, predict what will happen next, rationalize and make inferences about why events occurred, and relate information from past experiences. The context processor should be well-advanced by the time children enter school.

The meaning processor is comprised of subskills related to vocabulary development and word awareness. Initially, children develop oral vocabulary from personal experiences and listening to stories. During their early interactions with books, they learn metalinguistic terms related to books and print (i.e., page, story, read, letters, words). Word awareness includes word segmentation and word consciousness. According to van Kleeck (1998), word segmentation is the awareness that words are units of language and sentences are comprised of words. Word consciousness is demonstrated when children recognize that the print form of a word is separate from its meaning, identify whether a sound sequence is a word, and that words comprised of the same sounds can mean different things (e.g., bark).

The orthographic processor uses letter knowledge and print conventions. In the emergent literacy phase, children distinguish the names, shapes, and sounds of letters and the difference between print and pictures. The phonological processor includes subskills related to phonological awareness such as syllable segmentation, rhyming (onset/rime), and phoneme segmentation. Children develop conscious awareness of the sound components of words. The research by van Kleeck (1998) provided interesting insight into how the four processors are implicated in the earliest stages of emergent literacy development.

Stages of early reading development. Jeanne Chall (1996) proposed another theory to trace the early stages of reading development. She proposed a continuum of six stages of reading. The early reading or emergent literacy stage includes the development of foundational literacy skills including concepts of print, phonemic awareness, book-handling skills, and recognition that print carries meaning. The second stage marks the beginning of conventional literacy development including sound-symbol correspondences and decoding accuracy with deliberate and effortful decoding. The third stage is confirmation and fluency which Chall (1996) referred to as the "ungluing from print" (p. 8). During this stage, reading fluency develops as readers consolidate decoding skills and increase automaticity with print. It is a time when reading sounds increasingly more natural and conversational as the reader focuses on the prosodic features of text through phrasing, stress, and intonation. The development of fluency and automaticity are necessary for the reader to focus on understanding and constructing meaning from text. The fourth stage is “reading for learning the new” which allows for expanding knowledge and understanding from text. The fifth stage engages readers in considering and critically analyzing “multiple viewpoints” in texts on a particular topic. The final stage of Chall’s theory is “construction and reconstruction” which necessitates synthesizing multiple perspectives from texts to arrive at a unique and personal perspective.

Theory of word recognition development. Finally, Linnea Ehri (1995, 1998) proposed a continuum of four developmental phases aimed at automatic sight word recognition. Sight words are defined as “all words that have been recognized accurately on several occasions” (Kuhn & Stahl, 2003, p. 4). Ehri suggested that in order for words to be recognized instantly and automatically as sight words, the reader must establish a mental representation of the orthographical structure of the word. With increased exposure to the word, readers gradually expand their conceptualization of the word to include its spelling, pronunciation, and meaning.

The four phases of sight word development include the pre-alphabetic phase; the partial alphabetic phase; the alphabetic phase; and the consolidated alphabetic phase. The pre-alphabetic phase maps onto Chall’s (1996) early reading stage which suggests that the reader relies on a visual cue in which sight word recognition is contingent upon memory recall of the visual representation of the word and how the word is pronounced and/or what it means. At this beginning point, letter-sound recognition is not yet developed. During the partial alphabetic phase, readers apply basic sound-letter correspondence to identify words. The full alphabetic phase draws upon the reader's increased phonological awareness and decoding ability to generalize familiar spellings to identify new and unfamiliar words. The reader makes connections between graphemes and phonemes in conventional spellings and begins to establish a core group of sight words that are recognized automatically including words with irregular phonetic spellings. The consolidated alphabetic phase corresponds to Chall's (1996) confirmation and fluency stage of reading. During this phase, the reader develops increased understanding of the orthographic system and begins to recognize familiar letter patterns as holistic units within words. The reader recognizes many words accurately and automatically. All three perspectives of word recognition add to our understanding of how children develop automatic word recognition in the early language and literacy phase. The main focus here has been on the skills essential for the translation of print into language. The theoretical perspectives associated with comprehension are reviewed next.

Theoretical Perspectives of Comprehension

The meaning-based component of early language and literacy development is also explained by theories of language and reading comprehension. Whether children are listening to stories or reading on their own, they engage in similar processes to interpret meaning from text. Of the many well-established theories of reading comprehension, a combination of constructivist and cognitive processing theoretical perspectives ground the data collection and analysis of the current validation study with verbal reports of thinking and reasoning on a reading and listening comprehension assessment. These perspectives include (a) discourse comprehension (van Dijk & Kintsch, 1983) and construction-integration model (Kintsch, 1994), (b) schema theory (Anderson & Pearson, 1984), and (c) perspectival view of reading (Norris & Phillips, 1994). These theoretical perspectives conceptualize comprehension as a dynamic, constructive, and interactive process. The primary focus is on the construction of meaning from text, that is, on how individuals account for the ideas presented in the text and the different information sources used to interpret the text. From these perspectives, comprehension involves the interaction between the reader and the text and the integration of information from different sources such as the text and prior knowledge. Each theory acknowledges the importance of prior knowledge, experiences, and perspectives that individuals bring to the text and how these factors influence textual interpretations (or representations). The selected theories focus on the unobservable mental processes underlying comprehension and the active construction of meaning from text. Each perspective is detailed next.

Theory of discourse comprehension (van Dijk & Kintsch, 1983) and construction-integration model (Kintsch, 1994). Discourse comprehension theory is a bottom-up framework of text processing and meaning-making that deconstructs comprehension into various components. A bottom-up perspective conceptualizes comprehension as a progression of different stages of information processing starting with print and text features (i.e., letter and word recognition), followed by the text content, and then the construction of meaning based on the text. The model applies to both reading and listening comprehension.

Comprehension entails several sets of processes that operate in cycles. First, the process begins with understanding the individual words and propositions in the text (i.e., semantic structures comprise the most basic relational meaning units in text); then it moves to understanding the relationships among the different propositions in the sentences, and finally, the understanding of the overall text meaning (van Dijk & Kintsch, 1983). The semantic structure of discourse is conceived on two levels: the microstructure and the macrostructure. The microstructure of discourse operates at the local level, that is, a mental representation of the meaning of individual propositions and their relations. The macrostructure operates from a more global level, in which the discourse is considered as a whole and as a set of hierarchical propositions (Kintsch & van Dijk, 1978, p. 365). A set of semantic rules governs the relations between the different levels of processing. The discourse is said to be coherent to the extent that the microstructure and macrostructure are congruent. According to Kintsch and van Dijk (1978), "comprehension always involves knowledge use and inference processes" (p. 364). Macro processes entail distinguishing between relevant and irrelevant information in the text base and generating inferences to construct the gist or main idea of the text. The new information is processed in relation to other information sources that are available to the individual (i.e., the text, the context, and their background knowledge) (Kintsch & van Dijk, 1978).

The theory of discourse comprehension was updated and expanded into the *construction-integration model* by Kintsch (1988, 1994), which suggests three levels of mental representations based on various types of information sources: linguistic (e.g., word meaning), conceptual (e.g., sentence meaning), and situational (e.g., integrating text with background knowledge to construct meaning) (Tracey & Mandel Morrow, 2006). Comprehension is the process of constructing a mental representation of the meaning of a text from the text base and situation model. The text base is more or less a literal representation of meaning primarily based on information directly in the text derived from relevant linguistic and world knowledge. The situation model includes an inferential and interpretive representation of meaning comprised of the integration of the text's main ideas in relation to the reader's relevant prior knowledge and expectations. According to Kintsch (1988), the construction-integration model "combines a construction process in which a text base is constructed from the linguistic input as well as from the comprehender's knowledge base" (p. 164) and the integration of the two (i.e., linguistic input and knowledge base) combined into a coherent whole. This integration theory highlights the importance of readers' prior knowledge, text syntactic structure, and meaningful relations between the two for the processing and comprehension of the text content.

Recently, Kintsch (2010) admitted that the one main limitation of his model is that the notion of *propositions* (a fundamental aspect of the theory) has never been clearly defined or operationalized. He states, "I spent most of my career arguing that propositions are the units of meaning, not words" (p. 198). He explains that there has been no direct way to distinguish these units clearly.

Schema theory. Schema theory (Anderson, 1977; Anderson & Pearson, 1984) is a top-down perspective focused on the important influence of background knowledge for monitoring, predicting, inferring, and evaluating text. A top-down perspective of comprehension focuses on what readers bring to the text and how they use information from many different sources (including their knowledge of the text topic, text structure, sentence structure, and vocabulary) to interpret the meaning of the text by making and revising predictions and hypotheses in light of new text information with the aim of constructing meaning from the text. According to this theory, the activation of schemata occurs automatically in response to the concepts encountered in the text. An individual's cognitive structure is comprised of schemata, which are an "organized set of concepts, possibly hierarchically related" (Norris & Phillips, 1987, p. 289). Individuals learn about these complex events or concepts through their previous knowledge and experiences. These schemata are the basis for text interpretation. Through repeated exposure and experience with certain concepts or events, individuals learn about the attributes associated with the concept and organize their knowledge into "orderly systems of procedures and expectations" (Pressley & Afflerbach, 1995).

The information presented in the text provides the stimulus to activate whatever schema the individual has about a particular concept or event. The schema elicits information about the various attributes of the concept, its purpose, the context to which it applies, what it is comprised of, who it is associated with and so on. Individuals have schemas for text content (e.g., people, places, things), reading processes (e.g., decoding, skimming, inferring, summarizing), and different text genres (e.g., narrative, expository) (Tracey & Mandel Morrow, 2006). The schema include the specification of relations between its component parts (i.e., temporal order of events, causal interaction among its components, and spatial relations among events) (Norris & Phillips, 1987, p. 290). The various components of the schema are referred to as nodes, variables, or slots. In the interpretation of text, the individual tries to account for the text information by fitting the information into the various slots within the schema. The activated schemata elicit certain expectations of what the text will be about and where to allocate attention. These schemata also allow the individual to anticipate the content of the text and to process and make inferences about the meaning of the text. Comparisons between prior knowledge and what is presented in the text either confirm the individual's expectations or cause the individual to re-evaluate and possibly revise their interpretation in the face of incongruent information.

Comprehension is achieved when the individual "is able to activate or construct a schema that gives a good account of the objects and events described" (Wilson & Anderson, 1986, p. 33). In other words, comprehension occurs when the textual information is fitted into most of the slots in the schema and the individual is able to provide a coherent account of the text (Norris & Phillips, 1987). However, it is also important to highlight the shortcomings when a reader does not activate the correct schema for the particular context and in turn, how the alternative schema affects the meaning inferred from a text. For example, the sentence, “The hikers came upon a fork in the road” can be interpreted differently depending on the schema that is used. If a reader’s schema for ‘fork’ is limited to a kitchen utensil, the understanding of the text will differ from the context in which the term ‘fork’ means a division in the road in two different directions. Thus, the reading context and the schema evoked by the text plays a key role in comprehension. In other words, an interpretation based on the evidence in the text which fills most slots of the schema results in a coherent account of the text. Whereas, an interpretation based on a specific word meaning that is incongruous with the available text in which no slots in the schema are filled produces an incoherent account of the text.

Perspectival view of reading. In 1994, Norris and Phillips proposed a perspectival view of reading which maintained that readers infer meaning from text by integrating textual information with prior knowledge to generate an interpretation of the text. The perspectival view of reading acknowledges that there is not merely one, singular interpretation of a text. Rather, a number of different interpretations of a single text are possible depending on the individual’s knowledge and perspective brought to the text within a particular reading context, as well as the clarity of the written text.

Norris and Phillips (1994) distinguished between the more traditional ‘god’s eye view’ of reading and their perspectival view of reading. The ‘god’s eye view of reading’ focused on a third-person point of view and decontextualized perspective in which an ‘outsider’ determines the meaning of the text and how it is to be interpreted and which information sources (i.e., knowledge) are relevant to the text. The perspectival view of reading places greater emphasis on ‘first-person intentionality’ (i.e., the reader’s point of view) within a contextualized perspective which may on the surface appear that the meaning of text is ‘relative to the interpreter.’ However, individual interpretations of text depend on the specific information sources (i.e., personal knowledge and experiences) that individuals attend to and how they effectively use their background knowledge to account for the information presented in the text. Readers must be guided by the text as a fundamental source of information with which they must integrate their relevant knowledge sources in order to make acceptable inferences. The perspectival view of reading focuses on the knowledge that readers themselves deem to be important and relevant to particular texts and their ability to make “inferential links between their knowledge and the text information” (Norris & Phillips, 1994, p. 408). Thus, readers are actively engaged in the process of “creating their own relevance” and determining which knowledge they will draw on to generate a coherent and meaningful interpretation of the text.

The perspectival view of reading offers an important distinction between literal and inferential interpretations of text. According to Norris and Phillips (1994), the difference between literal and inferential interpretations of text can be conceptualized on a continuum of meaning which ranges between more or less obvious. On the one end of the continuum, the meaning of a text is more obvious for a reader who is familiar with the topic in a particular reading context and the inferences are considered reliable and based mainly on the explicit text information. On the other end of the continuum, when a reader is not familiar with the topic and the meaning of the text is less obvious, then inferences are less reliable. The degrees of obviousness and reliability of the inferences in textual interpretations are relative to the reader's knowledge of, the purposes for, and context of the reading task.

The perspectival view of reading recognizes that there are qualitative differences in interpretations of a text and not all interpretations are of equal quality. High quality interpretations of text depend on individuals' ability to integrate and account for the information presented in the text using relevant prior knowledge to construct inferences. In contrast, low quality text interpretations do not account for or integrate the available relevant information from the text or the reader's background knowledge. The adequacy of textual interpretations are judged according to certain criteria: (1) completeness and comprehensiveness, and (2) consistency and coherence (Norris & Phillips, 1994, p. 395). The first aspect determines whether the interpretation is *complete* and provides a comprehensive explanation that accounts for the textual information and relevant knowledge applicable to the textual content. The second criterion is whether the interpretation is "consistent with the known facts" which means that it is both accurate and "plausible given what is known" (Norris & Phillips, 1987, p. 302). Both criteria are necessary and must be considered simultaneously in judgments about textual interpretations. Comparative appraisals of interpretations are undertaken using both criteria to identify the extent to which an interpretation of text is justified and sound. Thus, interpretations will have varying degrees of completeness and consistency. Some interpretations will be *more* or *less* complete and *more* or *less* consistent. The aim of inferring meaning from text is to generate an interpretation that is as complete and consistent as possible given the available information. In the context of reading, the interpretation per se is not the most significant aspect, rather the process that is undertaken to arrive at the interpretation and the evidence that the individual uses to substantiate their interpretation is what counts. A key difference between schema theory and the perspectival view of reading is that relevance in text is determined by teachers (schema theory) and in the latter it is determined by the students (perspectival view) in the process of reading. The schema view may lead to dependence whereas the perspectival view aims to create independence on the part of students. The next section examines how think-aloud and protocol analysis methodology has been used to study the reading process.

Think-aloud and Protocol Analysis Research

Think-aloud research using verbal report data and protocol analysis is based on the concept of introspection or the examination of thoughts and mental processes involved in cognitive tasks. The verbal report includes the audio and/or written verbatim record or account of the subject's thinking and behaviours (think-aloud) in relation to the performance of the task. The verbal protocol represents the primary data source for further analysis. Verbal reports provide important information about the thought sequences and underlying cognitive processes for carrying out a task. Protocol analysis involves encoding, analyzing, and categorizing the thought sequences of verbal accounts in order to infer and interpret the underlying processes inherent in the thinking (Newell & Simon, 1972).

Over a century ago, James (1890) used introspective reports of thinking as the primary data source for his theories of psychology. Subsequently, Huey (2009/1908) analysed his own and others' introspective verbal reports of reading as the basis for his conceptualization of reading as a cognitive and meaning-making process. Verbal reports and protocol analysis provide a rich and significant source of data about thinking and problem solving in the study of mental and cognitive processes. The conceptual framework underlying the use of verbal reports as data and protocol analysis is based on information processing theory involving "the study of cognitive activity to understand the structures and processes underlying knowledge acquisition and use" (Taylor & Dionne, 2000). Since the 1980s, extensive research in psychology, education, and cognitive science has expanded think-aloud methodologies to study the underlying cognitive processes of thinking, learning, problem solving, and reading.

Overview of Think-Aloud Methodology

In a seminal publication on the topic titled, *Protocol Analysis: Verbal Reports as Data*, Ericsson and Simon (1984, 1993) provided a critical examination of the collection and analysis of verbal report data. They explained that verbal reports and protocol analysis have been used to study a wide range of topics including second language learning, text comprehension, human factors, cognitive writing processes, IQ test-taking, memory, accounting, learning disabilities, survey question development, test question validation, and computer product testing.

Think-aloud methodology includes three main aspects: (1) performance or execution of a task, (2) verbal report or account of mental processing or "thinking aloud", and (3) protocol analysis. Think-alouds and verbal report data are typically used to investigate cognitive tasks or problem solving such as reading a text or solving a math problem. The think-aloud component centers on the collection of verbal reports. Ericsson and Simon (1984, 1993) distinguished between two main types of verbal reports; *concurrent* and *retrospective* reports. *Concurrent* reports include verbal accounts of actions and thoughts *during* task execution. *Retrospective* reports involve verbal accounts of actions and thoughts *upon completion* of the task. Both types of verbal reports are primarily drawn from information stored in working or short term memory (retrospective reports, however, may in fact draw from long term memory depending on the time lapse between the performance of the task and the verbal reporting). Concurrent and retrospective verbal reports provide a comprehensive source of information about the thinking and cognitive processes involved in task execution and completion.

Ericsson and Simon (1984, 1993) claimed that verbal reports produce a reliable and valid source of data provided that reports are collected under specific conditions. The most important condition is that subjects should be asked to report only on what they are thinking or attending to rather than speculating on and inferring about their thinking or problem solving processes. There is some evidence to show that the process of thinking aloud may slow the speed of task performance (Norris, 1990). Nonetheless, verbal reporting does not appear to alter, interfere with, or compromise the execution of the task or the cognitive processes involved in self-reporting the contents of working or short term memory. Contrary to some claims that verbal reports and think aloud protocols are epiphenomenal, Ericsson and Simon (1993) contended that verbal reports are "highly pertinent to and informative about subjects' cognitive processes and memory structures" (p. 220). They stated,

> Human subjects are not schizophrenic creatures who produce a stream of words, parallel but irrelevant to the cognitive task they are performing. On the contrary, their thinking aloud protocols and retrospective reports can reveal in remarkable detail what information they are attending to while performing their tasks, and by revealing this information, can provide an orderly picture of the exact way in which the tasks are being performed (p. 220).

Thus, verbal reports are recognized as a significant source of information for interpreting and drawing inferences about the implicit processes and strategies underlying the execution of cognitive task performance. The focus of the current review pertains to how think-aloud methodology has been used to study the reading processes of readers ranging in age and ability.

Investigations of Reading Using Think-Aloud Methodology

Fundamentally, reading is conceived as a cognitive and perceptual process for constructing meaning from text which makes think-aloud methodology and verbal reports of reading a viable approach for studying the reading process. Olshavsky's (1976-1977) characterization of reading as a "strategic problem solving process" provided validation for using verbal reports of reading to examine the processes and strategies that readers use while reading. Afflerbach and Johnston (1984) suggested that verbal reporting makes it possible to gain access to a unique source of information about the reader, "…under certain circumstances, (verbal reports) provide veridical descriptions of cognitive processes which otherwise could only be investigated indirectly (p. 308) …Cognitive processes are not directly observable and they must be inferred from available data" (p. 319). Think-aloud methodology provides one means for collecting specific data about what readers think and do while reading in order to infer the types of processes and strategies involved in reading. Pressley and Afflerbach (1995) added that verbal reports offer additional insight about the developmental, cognitive, and affective influences on reading which could not be attained by other means (i.e., product measures).

During the past three decades, a growing body of research has focused on gathering and analyzing verbal reports of reading to investigate different aspects of the reading process. Think-aloud and protocol analysis methodology has been used for many different purposes in reading research: exploratory and descriptive purposes (i.e., inductively); testing hypotheses about reading (i.e., deductively); and developing new theories of reading. Research using verbal report data highlights several different types of reading and text-processing strategies including: inference-making (Phillips, 1988; Trabasso & Magliano, 1996a); main idea (Afflerbach, 1990); summarization (Brown & Day, 1983); and general application of cognitive strategies while reading (Garner, 1982). In addition, think-aloud methodology has been used with many different ages and ability levels and different types of text including narrative and expository.

According to Schellings and his colleagues (2006), think-aloud protocols of reading focus attention on specific components of reading including the cognitive and metacognitive strategies involved in text processing, the motivation and affect of reader responses, and the contextual or situational variables that influence comprehension and understanding. They report, “The data interpreted from the think aloud protocols gave way to new insights into reading comprehension strategies and their interplay with reader characteristics and situational variables” (p. 551). Afflerbach (2000) maintains that verbal protocol analysis can be used as a method of inquiry into single aspects of reading (e.g., a process or strategy), as well as the broader conceptualization of the reading process including the contextual variables which influence the application of reading strategies and processes.

Pressley and Afflerbach (1995) concluded, “Protocol analysis provides compelling evidence that constructive cognition is central to reading, it also proves that reading is more than cognition” (p. 165). In sum, verbal reports and protocol analysis of reading contribute important and significant information about the reading process, how reading acquisition develops, and how readers infer and interpret meaning from text. Think-aloud methodology in the context of reading has shown great potential for providing important insight into the cognitive, affective, and social influences on reading acquisition and development.

Think-aloud Reading Research Review

A comprehensive research review shows that think-aloud methodology has been used extensively as a method of inquiry for investigating the mental processes involved in reading. Specifically, the studies of verbal report protocols of reading have identified a range of different strategies, knowledge sources, and representations that readers construct while reading particular types of texts (Olson, Duffy, & Mack, 1984). The majority of think-aloud studies have been carried out with adult subjects to study the reading processes and strategies of skilled and accomplished readers. Nevertheless, there are a number of studies with adolescents (Campbell, 1999; Cromley & Azevedo, 2004a, 2004b, 2005; Lau, 2006; Janssen, Braaksma, & Rijlaarsdam, 2006) and children as young as four-years of age (Paris & Paris, 2003; Paris & van Kraayenoord, 1996; Tompkins, Guo, & Justice, 2013). In addition, several studies have used think-aloud methodology in the context of reading comprehension assessment.

The extant think-aloud reading research between 1930 and 2015 includes a corpus of more than 200 descriptive and empirical studies from various disciplines including cognitive psychology, speech and language pathology, second language learning, and reading education. The body of research was divided into two main categories: (1) descriptive research on using think-aloud methodology to study reading, and (2) empirical research using verbal report data and protocol analysis of reading. The research review included 56 descriptive research articles and book chapters describing the aspects of think-aloud methodology, the collection of verbal report data, and the different approaches for using protocol analysis to study the reading process. The empirical think-aloud reading research was grouped into four separate think-aloud categories: (a) with children, (b) with adolescents, (c) with adults, and (d) research on reading assessment. In total, the research review included 42 studies of children, 33 studies with adolescents, 57 studies with adults, and 28 studies using think-aloud during reading assessment. Some overlap exists between and among the categories as several studies include participants across various age groups. The number of subjects included in the empirical think-aloud reading research range from as few as two to more than 180 subjects between four years of age to adults (primarily university students).

The empirical think-aloud studies of reading provide detailed accounts about how think-aloud, verbal reports, and protocol analysis have been applied in the context of reading. The studies include a range of ages and ability levels with varied types of reading tasks, contexts, purposes, and materials. There is also significant variability in how the methodology is carried out. Specifically, the cues used to elicit think-aloud; the goals and purposes for collecting verbal reports; the degree of familiarity and practice that subjects are given with think-aloud procedures and processes; and the types of verbal reports collected (i.e., concurrent or retrospective reports). Nevertheless, the main aim of think-aloud reading research is to examine the cognitive processes and strategies involved in the reading process.

The research focuses mainly on the text processing and reading comprehension strategies, reader response, word identification strategies, vocabulary, and text interpretation. Some think-aloud research compares the reading strategies of different types of readers including average and above-average readers, novice and expert readers, younger and older readers, and the reading strategies of deaf readers (Schirmer, 2003). Several studies examined the different types of reading strategies and processes used for comprehension of narrative texts (Laing & Kamhi, 2002; Trabasso & Magliano, 1996a), expository texts (Cote, Goldman, & Saul, 1998; Gillam, Fargo, & Robertson, 2009; Kucan & Beck, 2003; Norman, 2012), as well as both narrative and expository text (Lau, 2006). Moreover, think-aloud reading research features the presence and frequency of use of different kinds of strategies and processes. In addition, the studies compare the types of strategies used in particular types of reading contexts (i.e., reading in areas of expertise, study strategies, problem solving) and for dealing with text comprehension difficulties.

Comprehension assessment think-aloud research. The studies which focused on the use of think-aloud as a method of inquiry to investigate the processes and strategies involved in the assessment of listening and reading comprehension are of particular significance to the current study. The comprehension assessment think-aloud research examines the underlying reasoning inherent in the cognitive response, and decision-making processes during listening and reading comprehension assessment. As mentioned earlier, think-aloud has become an important technique for obtaining relevant evidence to investigate the validity of test performance and to understand what tests actually measure. The primary approach for investigating the cognitive processes used during test performance is to ask examinees why they answered test items the way they did.

Over 30 comprehension assessment think-aloud studies are grouped into four main categories. The first category includes six studies with think-aloud methodology to analyse the psychometric properties of reading comprehension test measures (Alavi, 2005; Anderson, Bachman, Perkins, & Cohen, 1991; Jacobson, 1974; Kavale & Schreiner, 1979; Langer, 1987; Phillips, 1989). In these studies, subjects verbalized their thinking and reasoning as they respond to reading comprehension test questions, mainly multiple-choice test items, in which they explain their rationale for selecting particular answers and their reasons for not selecting other alternatives. The protocol analysis compared subjects' reported reading processes and strategies to their performance in order to validate the test items, scores, and responses. Four other studies (Allen, 1998; Brandao & Oakhill, 2005; Haney & Scott, 1987; Wilson, 1979) focused on analyzing children's reasons for their responses to comprehension questions. In these studies, children between grades 2 and 7 were prompted to explain and justify their answers to reading comprehension questions. This procedure was used to study the relationship between comprehension performance and the quality of reasoning used to answer questions and to examine which information sources (i.e., text or prior knowledge) that the children relied on to answer comprehension questions.

The second cluster of 11 reading assessment think-aloud studies compared reading and test-taking strategies on different types of reading comprehension assessment measures using varied testing methods and item formats (Alvermann

& Ratekin, 1982; Anderson, 1991; Baldo, 2008; Campbell, 1999; Cordon & Day, 1996; Farr, Pritchard, & Smitten, 1990; Nevo, 1989; Pearson & Garavaglia, 1999; Powell, 1988; Sepassi, 2003; Werner & Kaplan, 1950). In these studies, subjects were asked to describe what they were thinking and doing as they answered different types of reading comprehension test questions including multiple-choice, constructed-response, cloze, written retelling, vocabulary, or language comprehension questions. The studies examined the types of thinking processes elicited by various reading comprehension test item formats and assessment measures and compare how the different item formats and assessments influence examinees' thinking processes.

The third category of reading assessment research includes ten studies that use think-aloud to assess reading comprehension (Cromley, 2005; Cromley & Azevedo, 2004a, 2004b, 2005; Lau, 2006; Schellings, Aarnouste, & van Leeuwe, 2006; Scott, 2008; Stahl, Garcia, Bauer, Pearson, & Taylor, 2006; Wade, 1990; Wingenbach, 1982). In these studies, subjects are asked to think-aloud as they read one or more test passages. The think-aloud is one of several different assessment tasks administered to investigate the cognitive processes and strategies that readers use to monitor reading, extend interpretations, and construct meaning from text. The verbal report data is analysed and combined with the results from the other outcome measures of reading comprehension to provide a comprehensive account of how readers process text. The reading assessment think-aloud research includes primarily children and adolescents (with the exception of Baldo in 2008 who studied university students) using a variety of informal and standardized reading comprehension assessment instruments.

Finally, four studies used think-aloud procedures with very young children during story comprehension assessment tasks: (1) van Kraayenoord and Paris (1996) included children between 5- and 6-years of age (n = 46), (2) Paris and Paris (2003) worked with 5- to 8-year olds (n = 158), (3) Tompkins, Guo, and Justice (2013) studied 4- and 5-year olds (n = 42), and (4) Brandao and Oakhill (2005) had a small sample (n = 25) of 7- and 8-year olds. The first three studies used think-aloud to examine children's ability to make inferences and their narrative story comprehension by having them think-aloud during a picture-walk task with a wordless storybook and to explain and justify their thinking in response to comprehension questions. The think-aloud data in these studies showed that children from 4- to 8-years of age recalled specific narrative elements, identified explicit visual story cues, made connections between different text cues, and integrated background knowledge to construct meaning from text. Brandao and Oakhill (2005) used comprehension questions and meta-level probes to examine which information sources that 7- and 8-year olds relied on to answer open-ended literal and inferential comprehension questions after reading several different stories. Children's explanations for their answers to the comprehension questions were evaluated according to how well they utilized relevant text cues and prior knowledge to justify their responses and to demonstrate text understanding. Brandao and Oakhill found that children's justifications were primarily based on either text information or background knowledge.

Thus, it is expected in this study, based on the findings of previous research and allowing for the inclusion of three-year olds, that the primary source of information for children's justifications for correct and incorrect comprehension question responses would likely be either solely information explicitly stated in the text including visual cues or solely children's background knowledge. Thus, there would likely be fewer integrated responses from the three-year olds than for the four to eight year olds because of the level of their emergent language and literacy development. It was also expected that they would likely just repeat their answers to the comprehension questions as their justifications.

Conclusions from the Think-Aloud Reading Research Review

The extensive review of think-aloud reading research has uncovered several important findings pertinent to the current study. The most significant finding is that readers of all ages engage in similar types of processes and strategies during reading. Findings show that novice and expert readers use a variety of literal and inferential strategies to comprehend text including repeating parts of the text, paraphrasing or summarizing, identifying main ideas, explaining, predicting, hypothesizing or speculating, making in-text and beyond text associations, questioning, elaborating, monitoring, and activating prior knowledge, to name a few. The think-aloud reading research has also revealed weaker reading comprehension such as a lack of self-regulation and monitoring during reading, difficulties constructing meaning from text, and problems with integrating text information and prior knowledge to make meaningful connections and to generate inferences (Purcell-Gates, 1991; Scardamalia & Bereiter, 1984; Zabrucky & Ratner, 1992)

Interestingly, the findings show that even very young children are capable of making different types of inferences to comprehend stories that they either listen to or read on their own (Gillam, Fargo, & Robertson, 2009; Laing & Kamhi, 2002; Lepola et al., 2012; Lynch & van den Broek, 2007; Trabasso & Magliano, 1996b). The main difference between younger and older readers has to do with the quantity and quality of the strategies and processes used. In other words, adult readers generally tend to use fewer strategies with a greater degree of sophistication. For example, older, more accomplished readers make a variety of different types of inferences during comprehension of narrative and expository texts including explanatory, causal, predictive, and associative inferences whereas, children focus mainly on goal-oriented inferences.

Another important finding from this research review is that recent studies have shown that children as young as four- to six-years of age are capable of thinking aloud and providing verbal reports of reasoning during different types of reading tasks and contexts (e.g., Farrington-Flint & Wood, 2007; Lynch & van den Broek, 2007; Tompkins, Guo, & Justice, 2013). In fact, the results from these studies provided valuable and encouraging insight about very young children's early reading and listening comprehension processes that there was no reason to believe that using think-aloud and verbal reports with children at even younger ages would be any less effective and informative.

To date, reading research with think-alouds, verbal report data, and protocol analysis has made a significant and important contribution to understanding the dynamic and complex nature of reading (Israel, 2015). Verbal protocols of reading provide detailed and explicit accounts of the critical strategies, skills, processes, contexts, and other knowledge that readers of different ages and ability levels use to process and comprehend text. Think-aloud methodology with verbal report data and protocol analysis is an important means for studying the breadth and depth of reading as a strategic and cognitive process. Ultimately, think-aloud serves an important diagnostic function for providing information about specific strengths and weaknesses in language and reading comprehension.

CHAPTER 3: RESEARCH METHODOLOGY

This chapter describes the instruments used; sample selected; data collection procedures; data scoring and preparation; data analyses and interpretation; and, ethical considerations of the study.

Instruments

Of the eight main sections of the *TELL*, two sections of the test were used to study the relationship between children's test performance and their verbal reports of thinking and reasoning. The *Oral Narratives (ON)* and *Oral Reading and Reading Comprehension (OR-RC)* sections were selected for three reasons. First, relevant subsections from the *ON* and *OR-RC* assessments measure two complementary and related component skills of comprehension namely, listening and reading. Both subsections focus on the interpretation and understanding of oral and print text. Second, the test items in each subsection require reasoning over and above recall (e.g., word reading). Verbal reports of thinking in an assessment context are used to study reasoning. Third, subsections from the *ON* and *OR-RC* assessments focus on the full range of ages for which the test was designed. These subsections provide important insight into oral (listening) and written language (reading comprehension) test performance for children from 3 to 8 years of age.

The main focus of the *ON* and *OR-RC* assessments is on meaning-making. Specifically, the comprehension tasks require children to interpret visual and linguistic textual cues, to consider all available information sources (i.e., oral narratives, visual images, and print text), to make intra- and extra-textual connections in order to understand the relationship between episodes and text content, and to integrate textual information with their background knowledge in order to interpret and understand the overall meaning of text. The specifications for each subsection are detailed next.

***TELL Oral Narrative (ON)* Measures**

The *TELL ON* section is focused on one aspect of oral language: narrative discourse. In the context of the *TELL,* narrative discourse is assessed with three types of measures: (1) story comprehension, (2) story generation, and (3) story recall. Specifically, the *ON* assessments measure children's knowledge of stories and their ability to comprehend, interpret, and construct oral narrations. The *ON* section has a combined total of 96 items for the three different subsections.

The current study focused on only the 14 listening comprehension questions included in the story comprehension subsection to explore the relationship between children's listening comprehension performance and their thinking. The narrative story comprehension tasks require children to answer a set of comprehension questions after listening to a story presented with a set of visual images (i.e., picture series and then a wordless picture book). The other aspects of the *ON* section were not relevant to the focus of this research.

The *ON* listening comprehension subsection is comprised of two separate assessments: (1) *The Unusual Present (UP)* Listening Comprehension Questions (picture series) and (2) *Ice Cream at the Zoo (ICATZ)* Listening Comprehension Questions (wordless picture book). The two *ON* comprehension subsections were intended for children from 3- to 8-years.

Unusual Present (UP). The first oral narrative subsection, *The Unusual Present (UP)*, serves two main purposes: (1) to provide a model for generating a story from a picture series, and (2) to assess story comprehension. The *UP* assessment consists of a five-picture series accompanied by a scripted oral narrative and six listening comprehension questions. The coloured picture series includes a sequence of five pictures depicting a young boy's birthday party and an unusual present. The oral narrative told by the examiner consists of 17 sentences and 191 words. The *UP* story is about a boy and his friends and their growing anticipation and reaction to an unusual shaped and sounding present.

During the *UP* assessment, children are shown the five-picture series while the basic components of story structure are pointed out by the examiner (i.e., *stories have a beginning, things that happen, and an ending*). After listening to the story read by the examiner while viewing the set of corresponding pictures, the pictures are removed and children are asked two literal and four inferential listening comprehension questions.

Ice Cream at the Zoo (ICATZ). The second oral narrative subsection, *Ice Cream at the Zoo (ICATZ),* was designed to measure comprehension of an illustrated oral narrative. The main difference between this assessment and the *Unusual Present* is that the story and wordless picture book used in this subsection are much longer and thus require children to hold more information in memory to complete the tasks. The inclusion of both short and long stories in the oral narrative assessments provide several different measures of narrative discourse comprehension and accommodate for the age range on the *TELL* (3-8 years).

The *ICATZ* subsection assesses children's understanding of oral narrative stories using literal and inferential comprehension questions. The assessment materials include a scripted oral narrative and a wordless picture book. The wordless picture book contains 12-pages of coloured illustrations and the accompanying oral narrative has a repetitive story pattern with 42 sentences and 513 words in total. The title on the front cover is the only print in the book. The oral narrative and wordless picture book include a story about a boy and his mom who take a trip to the zoo and the various mishaps with the boy's ice cream cone. During the *ICATZ* assessment, a wordless picture book and accompanying oral narrative are presented. After viewing the picture book while listening to the story read by the examiner, the book is removed and children are asked three literal and five inferential listening comprehension questions.

TELL Oral Reading and Reading Comprehension (OR-RC) **Measures**

The *TELL Oral Reading-Reading Comprehension (OR-RC)* section is comprised of two subsections: (a) *Read-Talk-Reread-Read (RTRR) Books,* and (b) *A Teddy Bear's Birthday Wish (TBBW).* The *OR-RC* assessments assess children's reading comprehension. During these assessments, children read one or more illustrated storybooks and answer a series of comprehension questions about each respective story.

Read-Talk-Reread-Read (RTRR) **books.** The *RTRR Books* include a set of four illustrated storybooks: *Dogs, Teddy Bear, Teddy Bears*, and *Dinosaur*. All four *RTRR* books are used with children between 3 and 5 years of age and the last two *RTRR* stories, *Teddy Bears* and *Dinosaurs,* are also used with 6 year old children. Each storybook has six pages with one line of print per page accompanied by a complementary illustration. The illustrations are mainly line-drawn black and white pictures with the exception of *Teddy Bears* which includes two coloured pictures (i.e., green and red bears). The texts contain a simple sentence structure with basic vocabulary and concepts such as animals; objects; fruit; numerals and number words; colours; size; emotions; action words (i.e., verbs); and familiar contexts (i.e., in a book, in a car). *Dogs* has a total of 17 words featuring various actions of different sets of dogs. *Teddy Bear* includes six lines of text with a total of 17 words depicting the interaction between a teddy bear, a bee, a bird, and an apple. *Teddy Bears* has 26 total words showing different sets of teddy bears distinguished by their colour (i.e., red and green) or emotions (i.e., happy and sad). *Dinosaur* contains a total of 35 words featuring different sets of dinosaurs or dinosaur body parts distinguished by number (one and two), size (i.e., big and small) and context (i.e., in a book, in a car) with an unexpected story ending.

The *RTRR* procedure is comprised of four phases. During the Read Phase, children are asked what they know about the topic of the story (i.e., dogs, teddy bears, or dinosaurs) and are encouraged to discuss relevant personal experiences related to the story topic. The examiner reads the story aloud while pointing to the printed text. During the Talk Phase, children are asked whether they liked the story and why and whether they had any questions about the story. During the

Reread Phase, children are invited to read along with the examiner as the story was read aloud a second time. The examiner points to the printed text while reading the story. In the final Read Phase, children are asked to orally read the story independently. Upon completion of the *RTRR* procedures, the storybook is removed and children are asked a combination of five literal and inferential comprehension questions after each story.

A Teddy Bear's Birthday Wish (TBBW). The second part of the *OR-RC* subsection, *A Teddy Bear's Birthday Wish (TBBW),* is also designed to measure children's reading comprehension. The assessment requires children between 6 and 8 years of age to read orally a picture book independently and to answer a set of 10 comprehension questions. The picture book, *A Teddy Bear's Birthday Wish*, is comprised of 20 pages with 10 full-page colored illustrations and 20 sentences in total for a total count of 202 words. The storybook is about a young teddy bear who wishes for a teddy rocking chair. After reading the *TBBW* illustrated storybook, the story is removed and children are asked five literal and five inferential comprehension questions.

Data Sources

The current study has two main data sources: (1) children's answers to each comprehension test-item for the *TELL ON* and *OR-RC* assessments, and (2) children's verbal reports of thinking and reasoning for each test-item response.

Comprehension Question Test-Item Responses

The first type of data is comprised of children's answers to the listening and reading comprehension questions on the *TELL ON* and *OR-RC* subsections.

Each subsection measures children’s ability to interpret and construct meaning from text by asking a series of literal and inferential comprehension questions after listening to or reading a story. In order to do well, children must be able to interpret the intent of the comprehension questions and select the relevant information sources pertinent to answering the test items.

Literal comprehension questions require children to recall specific information that is directly and explicitly contained in the title, story, or pictures. Literal comprehension questions focus on concrete aspects of the text content such as the quantity, name, or type of objects or characters; the setting; specific story episodes; and outcomes of story events. These types of questions assess children’s knowledge of stories and understanding of the relevant vocabulary and concepts presented in the text and/or pictures. Thus, one aspect of children’s listening and reading comprehension performance focuses on their ability to recall relevant and pertinent information explicit in the text and/or pictures.

Inferential comprehension questions measure children’s ability to understand, interpret, and recall information implicit in the text. Inferential test items require children to consider the information implied in the text and to select and integrate it with their relevant background knowledge to construct a response to the questions. On the *TELL*, inferential comprehension questions ask children to provide reasons for outcomes of specific events or actions, to predict what might happen next or after the story ends, to compare whether an aspect of the story is similar to or different from another aspect in a different situation or context, or to evaluate whether a story is real or make believe.

Each literal and inferential reading and listening comprehension question response was scored separately according to the specified test-item criteria in the *TELL* test administration manual and test performance booklet. The test-item scoring procedures are described in more detail later.

Verbal Report Explanations for Test-Item Responses

The second type of data for the study included children's verbal report explanations for their test-item responses. Verbal report explanations were obtained in response to specific probe questions asked after each literal and inferential comprehension test item during the alternative test administration in order to determine why children answered as they did. Verbal report data provided evidence for the actual thinking and reasoning processes that children used to justify their responses to the *TELL* test items. A thinking rating rubric (described later) was used to appraise how well children used relevant and available textual evidence and background knowledge to comprehend oral and print texts and to respond to specific test items.

Sample

Preschool and daycare program supervisors and elementary school administrators in Calgary, Alberta, Canada were contacted by telephone to introduce the study. Next, a follow-up letter was delivered to each site to explain the research procedure and to request permission to carry out the research with children. Once permission to conduct the study was granted by a site, a meeting was scheduled with the program supervisors, school administrators, childcare staff, and/or teachers to review the purpose of the research, to specify the participant criteria, and to take questions. Specifically, sample selection was based on the following inclusion criteria:

(a) children from 3 to 8 years of age at the time of the study,

(b) children who were reported to be typically developing,

(c) children who spoke English as their dominant language, and

(d) children for whom parental consent was obtained.

The main exclusionary criteria for this study included:

(a) English as a Second Language (ESL) children, and

(b) children who were known to exhibit language, speech, hearing, visual, behavioural, emotional, cognitive, or neurological difficulties based on previous specialist testing.

Program supervisors, school administrators, childcare staff, and teachers were asked to distribute and collect parent information letters and consent forms for children from 3 to 8 years of age who met the sampling criteria. Children's birthdates reported on the parental consent forms were used to determine chronological age by year.

Ten daycares, two preschools, and eight public elementary schools were contacted initially and children were recruited only from the sites which agreed to participate in the research including seven daycares, one preschool, and seven public elementary schools. In total, 200 parent information letters and consent forms were prepared and distributed and signed consent was obtained for 183 children from 3- to 8-years. Nine children (including 6-three year olds; 1-six year old; 1-seven year old; and 1-eight year old) were excluded from the data set for the current study because they were not able to focus and complete the assessments. The remaining sample consisted of 174 children (83 boys, 91 girls) between three- and eight-years of age. Table 3.1 provides a description of the sample in terms of age, gender, and socioeconomic background. Program supervisors and teachers reported on children's socio-economic background based on school records.

Table 3.1

Sample Demographics

Age in Years	*n*	*Gender*		*SES*		
		M	*F*	*Low*	*Mid*	*High*
3	26	10	16	5	11	10
4	28	11	17	8	11	9
5	29	19	10	5	14	10
6	28	15	13	8	13	7
7	33	14	19	5	19	9
8	30	14	16	10	10	10

Data Collection Procedures

Prior to data collection for the current research, preliminary work was done with a small sample of children from 3- to 8-years of age (n = 12; 2 children per age) using the same inclusion and exclusion criteria for the main research study. The main purpose of the pilot study was to become familiar with the *TELL* comprehension test components and administration procedures and to determine the feasibility of using think-aloud with very young children. Specifically, different types of probe questions were piloted to elicit children's reasons for their answers to the comprehension test items (e.g., Why do you think so? What made you think that? How did you know that? and Tell me more about that.). Children were consistently more responsive and provided better quality explanations to some probe questions than others. The probe that elicited the best quality explanations across the different ages on the *TELL* comprehension assessments was asking children to explain how they knew the answers to the comprehension test items (i.e., How did you know that?) and, consequently, this probe was used during the main data collection.

For the main study, children were assessed individually in a quiet room at their respective daycare, preschool, primary or elementary school. The test components were administered in two sessions of approximately 30-45 minutes each per child within a one week period. In compliance with the developmental order of the *TELL* sections, the *ON* assessments were administered during the first session followed by *OR-RC* assessments in the second session across two days. The *ON* subsection was administered to all of the children in the sample (3-8 years). In the case of the *OR-RC* subsection, children from 3- to 5-years of age completed all four *RTRR* stories, 6-year old children completed only the last two *RTRR* books and the *TBBW* story (depending on whether they could read the latter story more or less independently), and 7- and 8-year old children completed the *TBBW* story. Testing sessions were audio-recorded and transcribed verbatim for scoring purposes and subsequent data analyses.

Each test component was administered in accordance with the standard sequence and procedures outlined in the *TELL* test administration manual with one main variation: children were asked to provide verbal reports of their thinking and reasoning in order to explain their responses to each test item. Within the standard *TELL* test administration protocol for each separate subsection, children were given precise explanations of the assessment tasks and engaged in a discussion with the examiner prior to commencement of the assessment in order to establish rapport and to ensure that children understood the testing procedures and nature and expectations of each task. For example, when children are first introduced to the *RTRR-Dogs* story, they are asked questions such as whether they have a dog, like dogs, the name of their dog or a dog they like, and what they know about dogs. The purpose of the discussion is to activate their thinking about the topic, to establish rapport with the examiner, and to elicit the child's focus and engagement.

The verbal report elicitation procedures for the listening and reading comprehension test items incorporated one main question probe (i.e., How did you know that?). This probe immediately followed children's responses to each comprehension test question in order to find out why they responded to each test item the way they did. Children were asked to give reasons for their answers only after they responded to each test item. The think-aloud probe was used to elicit more detail from children about their thinking and reasoning as they responded to the items on each comprehension measure. The examiner also asked children to clarify any vague or unclear responses (i.e., "I'm not sure I understand what you mean. Tell me more about that.").

Data Scoring and Preparation

The two sets of raw data were scored separately: (1) children's answers to each comprehension test-item for the *TELL ON* and *OR-RC* assessments, and (2) children's verbal reports of thinking and reasoning for each test-item response. The scoring system for each data set is described in turn.

Test Performance Data

Children's responses for the listening and reading comprehension questions were recorded and scored on the test performance record. Test-item responses were scored dichotomously (0, 1) according to the criteria explicit in the *TELL* test administration manual and test performance booklet for each individual test item. Responses that matched the criteria are given 1 point and answers that failed to meet the test item criteria are given 0 points. Some test items include only one acceptable response within the context of the story, whereas others provide several possible responses.

Each *ON* and *OR-RC* subsection yields three scores: total, literal, and inferential comprehension. The total listening and reading comprehension performance scores are equal to the number of correct items out of a total possible score. The total possible *UP* listening comprehension score was out of 6 (total score range from 0 to 6; literal comprehension subscore range from 0 to 2; and inferential comprehension subscore range from 0 to 4). The total possible ICATZ listening comprehension score was out of 8 (total score range from 0 to 8; literal comprehension subscore range from 0 to 3; and inferential comprehension subscore range from 0 to 5).

The total reading comprehension performance score for each *RTRR* book is equal to the total number of questions answered correctly out of a maximum of 5. The range of total reading comprehension scores for each *RTRR* book is between 0 and 5. The literal and inferential reading comprehension subscores for the first three stories range between 0 and 3, or 0 and 2, respectively. For the last Dinosaur story, the literal and inferential comprehension subscores range from 0 to 2 and 0 to 3, respectively. Finally, the total possible reading comprehension scores for the *TBBW* subsection is out of 10 (total score range from 0 to 10; literal comprehension subscore range from 0 to 5; and inferential comprehension subscore range from 0 to 5). Higher scores on each scale represent higher levels of listening and reading comprehension and conversely, lower scores signify lower levels of listening and reading comprehension.

Verbal Report Data

The second source of raw data for this study included audio-recordings and transcripts of children's test responses and verbal reports of thinking and reasoning for each *TELL ON* and *OR-RC* test item and task. To prepare the verbal report data for further analysis, a rating rubric was used to evaluate the quality of children's verbal reports of thinking and reasoning for the comprehension test items. The scale was modeled after the reading and thinking rating scales developed by Phillips (1989) for inferential test items on the *Test of Inference Ability in Reading Comprehension*.

The development of the thinking rubric for this study required analysis of each *TELL* test item for task expectations and scoring criteria. Specifically, each test item was studied to identify the type of question and the information included in the item and corresponding text. Test items were categorized as either literal or inferential comprehension based on the type of information sources required for a correct response according to the scoring criteria.

Scoring verbal protocols required careful inspection of children's test item responses and verbal report explanations to ascertain the information sources and strategies children used to construct their responses (e.g., information from background knowledge; oral or written text; and illustrations). The quality of children's thinking and reasoning for each test item was scored with the four-point scale. The thinking scores ranged from 0 to 3 with 0 indicating low quality thinking and reasoning and 3 representing high quality thinking and reasoning. Verbal report ratings of thinking and reasoning for each test item were recorded on the test performance record. Thus, the second data set for this study was comprised of verbal report ratings for each test item and a combined total for each subsection.

A thinking rating rubric was used to evaluate the quality of children's responses and explanations for the comprehension test items on the *TELL ON* and *OR-RC* components. For each comprehension item, verbal report explanations were assigned a score of 0, 1, 2, or 3.

A thinking score of 0 indicated that the child reasoned on the basis of irrelevant, erroneous, or repeated text information, background knowledge, or both but failed to provide an explanation for the answer given in response to the comprehension test question and as a result provided an inconsistent and incomplete explanation. That is, the child either misunderstood the test question, misconstrued the story information, repeated the test item response or textual information without interpretation, or did not respond to the probe question.

A thinking score of 1 indicated that the child reasoned on the basis of insufficient relevant text information and background knowledge to substantiate the answer given in response to the comprehension test question and as a result provided a partially inconsistent and incomplete explanation. That is, the child mentioned the particular information source used as the basis for the response such as from the story, the pictures, or prior knowledge but did not provide specific and relevant text information and background knowledge to confirm the response.

A thinking score of 2 indicated that the child reasoned on the basis of some of the relevant text information and background knowledge to substantiate the answer given in response to the comprehension test question and as a result provided a consistent but incomplete explanation. That is, the child considered only part of the question or the relevant information sources.

A thinking score of 3 indicated that the child reasoned on the basis of all of the relevant text information and background knowledge to substantiate the answer given in response to the comprehension test question and as a result provided a consistent and complete explanation.

The children's verbal report protocols were scored independently from their test-item performance. It was possible for children to employ good quality thinking and reasoning but still answer a test item incorrectly and in contrast, children's thinking and reasoning for a test item could be of poor quality but the correct answer still achieved. Thus, it was important to evaluate each verbal report of thinking and reasoning for each test item on its own terms.

Children's explanations of thinking and reasoning for each listening and reading comprehension test-item response were scored using the thinking rating rubric by the examiner (i.e., the first coder). For the interrater-reliability check, approximately 25% of the responses for each test item were selected at random and scored by a second coder who had extensive experience using the thinking rubric to score children's explanations for comprehension question responses. The scores from the two coders were checked for agreement using match-mismatch inter-rater reliability and a 0.95 percentage was achieved. Discrepancies in scoring were resolved and a rule-base was followed for scoring the remaining item response explanations.

Data Analysis and Interpretation

The data analysis for this study addressed the main research question and several subsidiary questions: (a) What thinking and reasoning did children report as the basis of their responses to the *TELL ON* and *OR-RC* test items? How sound was their thinking and reasoning? (b) What information sources did children use (i.e., text information and background knowledge) and how did they use the information sources to respond to the *TELL ON* and *OR-RC* test items? Did they use the information appropriately? and (c) Is there a relationship between children's test and verbal report performance on each item? Specifically, did children who performed well on the test items also show good thinking and reasoning and conversely, did children who performed poorly on the test items show poor thinking and reasoning? The data analyses also provided additional insight specifically related to the types of validation questions such as: (a) Did children understand the *TELL ON* and *OR-RC* test items and tasks? (b) Did the *TELL ON* and *OR-RC* test items measure what they were intended to measure? (c) Did the evidence from the performance measures and verbal report protocols support inferences about children's early language and literacy competencies related to oral language and reading?

Protocol Analysis

Protocol analysis of verbal report data was undertaken in four phases.

1. A qualitative "analysis of reasons" (Messick, 1989). That is, a systematic inspection of children's verbal report protocols item-by-item to identify the specific information sources and strategies used by the children to substantiate their test responses.

2. Verbal report protocol analysis required consideration of many factors: the story stimulus and comprehension questions, the child's responses, the scoring criteria, and expected responses for each test item. Moreover, the process involved an appraisal of what constituted a correct answer and what constituted good thinking and reasoning in the context of each test item. Specifically, the analyses required systematic comparison of several aspects in accord with the cognitive model of reading: (a) the evidence available to support the response to the test item, (b) the information and strategies used in the child's actual response and explanation, (c) the specified criteria for evaluating test item responses, and

(d) the information required to generate a complete and consistent response. This part of the analysis provided specific and illustrative examples of children's productive and non-productive strategies and the information sources used in their thinking and reasoning.

3. Analysis of the relationship between test performance and thinking and reasoning. The main goal of this part of the analysis was to explore whether children's reasoning matched their responses and to determine whether children who correctly answered test items also reasoned well and those who answered test items incorrectly did not (Leighton & Gierl, 2007; Norris, 1991; Phillips, 1989). A systematic comparison of children's test responses and explanations was conducted to assess whether children's explanations and justifications for their answers and the information and strategies used were complete and consistent with the relevant available evidence and the question(s) asked. If children got an item incorrect but their thinking and reasoning was of good quality, then the the test item would be called into question.

In the context of oral language and reading assessment, there were several possible manifestations in the quality of thinking and reasoning according to the cognitive and normative model and the children's ability to use the available information sources. One variation includes children with low proficiency who fail to use the textual information or their background knowledge to respond to the test items, or they may rely only on one primary source of information, either textual or background knowledge. Children's responses demonstrate inconsistent and incomplete use of available information sources. Another variation includes children with mid-level proficiency who draw upon some relevant textual information and background knowledge in the context of the test item but fail to draw completely the relevant connections between the two. In these instances, children's responses are considered partially consistent and complete. The third variation includes highly proficient children who use all available information sources, monitor their understanding, and integrate the relevant sources effectively. Their responses are complete and consistent and show evidence of effective integration of all available and relevant information sources.

4. The final phase of protocol analysis focused on the quality of the test items and tasks. This phase involved an examination of the quality of children's thinking and reasoning with their test question accuracy as one way to judge the quality of test items. Several questions were considered: Did the child think poorly but answer the question correctly? Did the child think well but answer the question incorrectly? Did the child think poorly and answer the question incorrectly? Did the child think well and answer the question correctly? On the basis of these questions, inferences were drawn about the quality of test items and what the test was actually measuring (Norris, Leighton, & Phillips, 2004). The first two questions spoke against the quality of test items and the last two questions endorsed the quality of test items.

Cognitive and Normative Model of *TELL* Test-Item Performance

A cognitive and normative model represents the best explanation of children's test performance and the underlying causes of their performance on the *ON* and *OR-RC* test items. This explanation is based on evidence of the knowledge, cognitive strategies, and principles that influence children's test-item responses and the "normative appraisal of the appropriateness" of children's thinking (Norris, Leighton, & Phillips, 2004). The explanation provides an account of "how the performance arose and why it was or was not successful" (Norris, Leighton, & Phillips, 2004, p. 293).

Evidence from children's verbal reports of thinking and reasoning on the *TELL* test items was used to develop a cognitive and normative model of test performance (Norris, Leighton, & Phillips, 2004; Norris, Macnab, & Phillips, 2007). Norris, Leighton, and Phillips (2004) analysed different sets of verbal report data from their collective research studies of test performance and found several emerging patterns of specific features of thought that can be used to explain the causes of test performance and to support the development of a cognitive and normative model of test performance. These features provided the initial framework for identifying emerging patterns in the verbal report data obtained in this study. The primary goal was to ascertain the causes of test performance and the underlying cognitive and normative model that explains performance in the context of the *TELL* assessment. This aspect of the data analysis involved making inferences and drawing conclusions about the processes and factors that influence test performance and the underlying reasons to explain variations in test performance (Kane, 2006; Norris, Leighton, & Phillips, 2004).

Children's test performance and verbal reports of thinking and reasoning on the *TELL* test items were examined with the following aspects in mind: (a) patterns of attention; (b) patterns of dependence, overdependence, underdependence on various sources of information; (c) completeness or incompleteness of thinking; (d) reference to and reliance upon norms and principles of thinking; (e) strategy use and meta-cognition; (f) the generality of thinking; and (g) knowledge sources.

(a) Patterns of attention. Verbal reports of thinking and reasoning were examined to find out what children were attending to as they completed the *TELL* test items. The analysis identified specific aspects of test items or textual information that children focused on and how these related to their test performance and verbal report explanations. For example, some children explain that they knew the answer to the comprehension question because they looked at the pictures. In these instances, their attention is focused primarily on the visual cues in the text. This information is important to determine whether or not children are attending to the pertinent information sources related to the comprehension questions.

(b) Patterns of dependence, overdependence, underdependence on various sources of information. The analysis identified what information sources children used to generate their answers to test items. Of particular interest was whether there was a pattern of dependency and the nature of the dependency on particular types of information sources, and what effect that dependency had on test performance. For example, children might depend too much on the text information or their prior knowledge to respond to particular test items and fail to make meaningful connections and to integrate the relevant and available information sources for inferences required by the comprehension questions.

(c) Completeness or incompleteness of thinking. The quality of children's test performance and verbal reports of thinking and reasoning was reviewed for completeness. That is, an assessment of whether children considered all relevant and pertinent information in the construction of their responses. For example, incomplete thinking was attributed to many possible causes: over-reliance or under-reliance on particular information sources; failure to attend to specific details in the text, test item, or pictures; or lack of recognition that a response is inaccurate, or contains erroneous information irrelevant to the task. Incomplete thinking is an indication that children did not take into account all relevant and available information sources related to the comprehension question. On the other hand, complete thinking would indicate that children considered all relevant text cues and integrated their background knowledge to construct meaning and demonstrate understanding of the text.

(d) Reference to and reliance upon norms and principles of thinking. Verbal reports were examined in order to understand what principles children employed as the basis of their thinking and reasoning. This aspect focused on the norms or criteria that children used to support their thinking and reasoning and how successful or effective these norms were in the particular context. For example, when children ignored the text cues related to the comprehension question and relied solely only on their background knowledge to answer the test item, they based their judgements on merely one information source and failed to consider other relevant story information pertaining to the question.

(e) Strategy use and meta-cognition. The next component for analysis looked for evidence of strategy use and meta-cognition. Information from verbal reports offered important insights into which approaches or strategies children used to answer test items or tasks and how successfully these strategies were applied. The meta-cognitive aspect was concerned with whether children were aware of the strategies they used or their failure to apply appropriate strategies for the task. For example, when children failed to identify the explicit details from the story to answer a literal comprehension question, they apparently did not use effective strategies for monitoring their understanding of the story.

(f) Generality of thinking. Children's test performance and verbal reports were inspected to determine whether their approach and thinking were based on sound principles in the context of the test. On the *TELL* comprehension assessments, this analysis involved an adjudication of whether children's thinking was complete and consistent with the task expectations and the relevant and available information needed to respond to the particular test items and tasks. For example, when children gave vague and generic-type responses to comprehension questions (i.e., 'because it said in the story'), they did not specify the relevant text information pertaining to the comprehension question and their understanding of the text and thinking were lacking.

(g) Knowledge sources. The final aspect of analysis included the knowledge sources children used and how successfully they used them to construct responses to test items. Also of importance here was to determine the knowledge sources applicable to the various test items or tasks. For example,

some children struggled to know when and how to use their background knowledge in order to construct consistent and complete inferences from the story. In some instances, children's answers to the comprehension questions indicated that they did not use relevant background knowledge to make the necessary inferences and to answer the comprehension questions.

The interpretive analysis of the verbal report data was undertaken in consultation with the researcher's supervisor to review test performance records, protocol transcripts, verbal report explanations, and ratings to ensure that all possible interpretations were taken into account. The basic aim of this validation research was to investigate whether there was evidence to show that the test measured what it claimed to measure (Messick, 1989; Shepard, 1993) and whether the evidence justified particular test interpretations and uses pertaining to the relevant aspects of the test and the population for which the test was designed. Verbal report data contributed valuable evidence to show whether the specific *TELL* test items actually measured the cognitive processes that they were expected to measure and, ultimately, provided evidence for the trustworthiness or validity for the interpretive inferences and conclusions that can be drawn from test performance (Messick, 1989, 1995; Norris, 1992). That is, whether the *TELL* comprehension assessments measured the skills and processes fundamental to children's listening and reading comprehension ability.

Ethical Considerations

In accordance with the ethical policies of the *University of Alberta Standards for Protection of Human Research Participants*, ethics approval was granted for the research by the *University of Alberta Faculty of Education and Extension Research Ethics Board* as part of the national research study involving the design and development of the *TELL.* Parents signed consent forms indicating that their child would participate in the interview subject to the right to withdraw at any time during the interview. Strict measures were taken to ensure confidentiality and the protection of the welfare and identities of all the children who participate.

CHAPTER 4: RESULTS AND DISCUSSION

This chapter presents and discusses the test-item frequency distributions for test and thinking performance across the six different age groups on the *TELL Oral Narrative (ON)* listening comprehension and the *Oral Reading-Reading Comprehension (OR-RC)* subsections. In addition, the protocol analyses of the test-item responses and explanations focus primarily on the types of answers given and the quality of thinking and reasoning that children used as the basis of their test-item responses.

The central purpose of this study was to explore the relationship between children's test performance and what they report thinking and reasoning as they completed the listening and reading comprehension subsections of the *TELL ON* and *OR-RC* test items. A comprehensive descriptive analysis of the data was undertaken to address the primary research question: What thinking and reasoning do children report as the basis of their responses to the *TELL ON* and *OR-RC* subsection test items and how sound is their thinking and reasoning? In addition, two other key aspects were considered: (a) the specific information sources that children used; how they used the information sources to respond to the test items; and whether they used the information appropriately, and (b) the relationship between children's test-item performance and verbal report (thinking) scores; whether children who performed well on the test items showed good thinking and reasoning and whether children who performed poorly on the test items showed poor thinking and reasoning.

TELL Oral Narrative (ON) **Listening Comprehension**

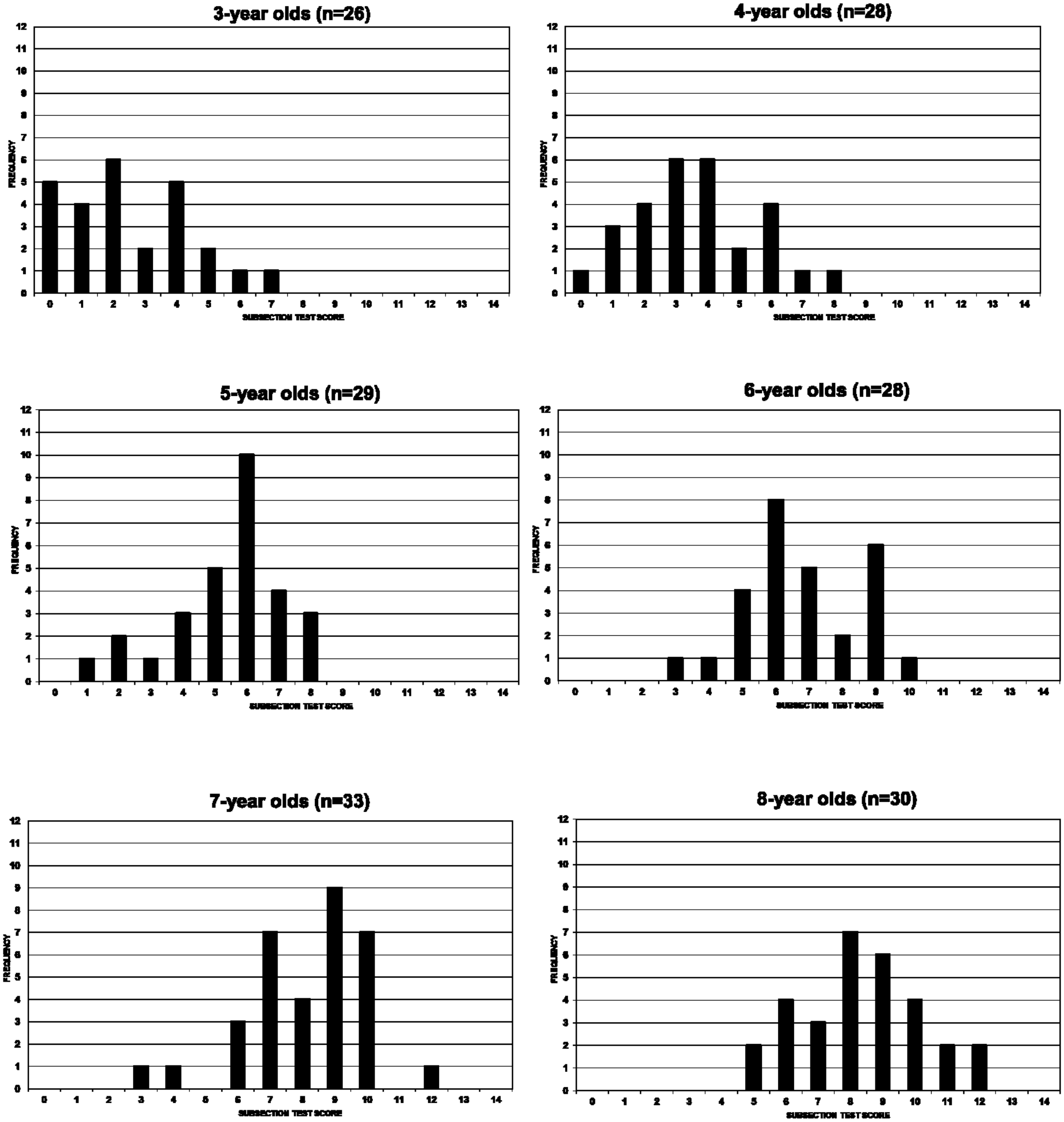
Frequency distributions were calculated for the total listening comprehension subsection test performance according to age and the combined test and thinking scores for each *ON* test item by age.

ON Total Listening Comprehension Test Score Frequency Distribution by Age

Figure 4.1 shows the frequency distribution for the total subsection test scores on the combined *ON* listening comprehension items according to age. These results were derived from the total number of correct responses that children at different ages obtained across two subsections of listening comprehension questions with a maximum possible test score of 14. The *ON* total subsection test scores ranged between 0 and 12 for the 3- to 8-year olds who participated in the study. The frequency distributions revealed that the scores increased with increasing age. Despite comparatively equivalent score ranges for each age, 3- and 4-year olds scored mainly in the lower range and 5- to 8-year olds scored in the mid-to-upper range. The frequency distribution also revealed that some children at the younger ages performed better on the *ON* subsection than children at the older ages. Specifically, two 3-year olds and six 4-year olds had higher total scores on the *ON* subsection than two of the 8-year olds. Overall, the pattern of age-related test performance increases on the listening comprehension question subsections provides supporting evidence of the developmental sensitivity of the measures to discriminate incremental differences in test performance especially between the preschool and school-aged groups in the sample.

Figure 4.1

Oral Narrative (UP & ICATZ) Total Listening Comprehension Test Score Frequency Distribution by Age



3-year olds (n=26)

4-year olds (n=28)

5-year olds (n=29)

6-year olds (n=28)

7-year olds (n=33)

8-year olds (n=30)

ON Test-item Frequency Distribution by Age

Table 4.1 contains item level data. Specifically, the number of children who correctly answered each item with and without justification and the number of children who incorrectly answered each item with and without justification. Keyed answers (test score 1) matched the specified criteria established for a correct response to the test question and unkeyed answers (test score 0) did not match the specified criteria for the respective test items.

The main difference between keyed and unkeyed answers with and without justification is the level of specificity and the degree to which children successfully use the relevant and available information sources to address the particular listening comprehension questions asked and subsequently, their ability to justify their answers. In the first instance, justified keyed and unkeyed test-item answers (thinking score 2 or 3) indicate that children consider the relevant and available explicit and implicit oral and visual text cues in conjunction with relevant background knowledge in either partial (2) or complete (3) justifications for their test-item responses. Justified answers represent good quality thinking and reasoning. Alternatively, keyed and unkeyed test-item answers that are not justified (thinking scores 0 and 1) represent poor quality thinking and reasoning. Low-level justifications for test-item responses are inconsistent and incomplete with the available evidence and children overlook the relevant oral and visual text cues and background knowledge related to the question.

In general, the frequency distributions showed that the proportion of keyed and unkeyed responses varied for each of the fourteen *ON* test items across all age groups. At least half (52%) and up to four-fifths of the sample (82%) scored the keyed answer for six of the fourteen listening comprehension items (literal comprehension *ICATZ* Items 2 and 5; and inferential comprehension *UP* Item 6 and *ICATZ* Items 3, 6, and 8). For the remaining eight *ON* literal and inferential listening comprehension items (*UP* Items 1-5 and *ICATZ* Items 1, 4, and 7), the majority of children (between 52% and 94%) provided unkeyed answers. Five of these items were answered correctly by fewer than 20% of the total sample (literal comprehension *UP* Items 1-2; and inferential comprehension *UP* Items 3-4 and *ICATZ* Item 4) suggesting that these listening comprehension questions posed the greatest challenge overall within this sample.

TABLE 4.1

Frequency and Percentage of Keyed and Unkeyed Answers (Test Scores 1-0) - Justified and Not Justified (Thinking Scores of 2-3 or 0-1) for each TELL Oral Narrative Listening Comprehension Test Item by Age and Total Sample

Test-item/ Age in years (n)	**Keyed Answer- Justified** Test Score 1 Thinking Score 2-3		**Keyed Answer- Not Justified** Test Score 1 Thinking Score 0-1		**Unkeyed Answer- Justified** Test Score 0 Thinking Score 2-3		**Unkeyed Answer- Not Justified** Test Score 0 Thinking Score 0-1	
	f	%	*f*	%	*f*	%	*f*	%
***UP* Item 1-L**								
3 (26)	-	-	1	3.8	-	-	25	96.2
4 (28)	-	-	-	-	-	-	28	100.0
5 (29)	-	-	1	3.4	1	3.4	27	93.1
6 (28)	-	-	3	10.7	-	-	25	89.3
7 (33)	1	3.0	9	27.3	1	3.0	22	66.6
8 (30)	4	13.3	11	36.7	2	6.7	13	43.4
Total (174)	**5**	**2.9**	**25**	**14.4**	**4**	**2.3**	**140**	**80.5**
***UP* Item 2-L**								
3 (26)	-	-	-	-	-	-	26	100.0
4 (28)	-	-	-	-	-	-	28	100.0
5 (29)	-	-	2	6.8	-	-	27	93.1
6 (28)	2	7.1	5	17.8	-	-	21	75.0
7 (33)	2	6.1	9	27.3	-	-	22	66.6
8 (30)	4	13.3	7	23.3	-	-	19	63.3
Total (174)	**8**	**4.6**	**23**	**13.2**	**-**	**-**	**143**	**82.2**
***UP* Item 3-I**								
3 (26)	-	-	-	-	-	-	26	100.0
4 (28)	-	-	-	-	-	-	28	100.0
5 (29)	-	-	1	3.4	-	-	28	96.5
6 (28)	-	-	1	3.6	-	-	27	96.4
7 (33)	-	-	3	9.1	-	-	30	90.9
8 (30)	-	-	8	26.7	-	-	22	73.4
Total (174)	**-**	**-**	**13**	**7.5**	**-**	**-**	**161**	**92.5**

Note. (*f*) = frequency; (*UP*) = Unusual Present; (L) = Literal Comprehension Test Item; (I) = Inferential Comprehension Test Item.

(continued)

TABLE 4.1 (continued)

Test-item/ Age in years (n)	**Keyed Answer-Justified** Test Score 1 Thinking Score 2-3		**Keyed Answer-Not Justified** Test Score 1 Thinking Score 0-1		**Unkeyed Answer-Justified** Test Score 0 Thinking Score 2-3		**Unkeyed Answer-Not Justified** Test Score 0 Thinking Score 0-1	
	f	%	*f*	%	*f*	%	*f*	%
***UP* Item 4-I**								
3 (26)	-	-	-	-	-	-	26	100.0
4 (28)	-	-	-	-	-	-	28	100.0
5 (29)	-	-	3	10.3	-	-	26	89.6
6 (28)	1	3.6	3	10.7	-	-	24	85.7
7 (33)	1	3.0	9	27.3	-	-	23	69.7
8 (30)	3	10.0	4	13.3	-	-	23	76.6
Total (174)	**5**	**2.9**	**19**	**10.9**	**-**	**-**	**150**	**86.2**
***UP* Item 5-I**								
3 (26)	-	-	1	3.8	-	-	25	96.2
4 (28)	-	-	4	14.3	-	-	24	85.7
5 (29)	-	-	8	27.5	-	-	21	72.4
6 (28)	-	-	11	39.2	-	-	17	60.7
7 (33)	-	-	18	54.5	-	-	15	45.5
8 (30)	1	3.3	13	43.4	-	-	16	53.3
Total (174)	**1**	**0.6**	**55**	**31.6**	**-**	**-**	**118**	**67.8**
***UP* Item 6-I**								
3 (26)	-	-	10	28.5	-	-	16	61.5
4 (28)	-	-	13	46.4	-	-	15	53.6
5 (29)	3	10.3	14	48.2	-	-	12	41.2
6 (28)	8	28.6	16	57.1	-	-	4	14.2
7 (33)	6	18.2	21	63.6	-	-	6	18.1
8 (30)	8	26.6	20	66.7	-	-	2	6.7
Total (174)	**25**	**14.4**	**94**	**54.0**	**-**	**-**	**55**	**31.6**

Note. (*f*) = frequency; (*UP*) = Unusual Present; (L) = Literal Comprehension Test Item; (I) = Inferential Comprehension Test Item.

(continued)

TABLE 4.1 (continued)

Test-item/ Age in years (n)	**Keyed Answer- Justified** Test Score 1 Thinking Score 2-3		**Keyed Answer- Not Justified** Test Score 1 Thinking Score 0-1		**Unkeyed Answer- Justified** Test Score 0 Thinking Score 2-3		**Unkeyed Answer- Not Justified** Test Score 0 Thinking Score 0-1	
	f	%	*f*	%	*f*	%	*f*	%
***ICATZ* Item 1-L**								
3 (26)	-	-	1	3.8	-	-	25	96.2
4 (28)	-	-	5	17.9	-	-	23	82.1
5 (29)	2	6.9	8	27.5	-	-	19	65.5
6 (28)	3	10.7	8	28.6	-	-	17	60.7
7 (33)	4	12.2	16	48.5	-	-	13	39.4
8 (30)	6	20.0	12	40.0	-	-	12	40.0
Total (174)	**15**	**8.6**	**50**	**28.7**	**-**	**-**	**109**	**62.6**
***ICATZ* Item 2-L**								
3 (26)	-	-	11	42.3	-	-	15	57.6
4 (28)	-	-	16	57.1	-	-	12	42.9
5 (29)	-	-	21	72.4	-	-	8	27.6
6 (28)	1	3.6	22	78.6	-	-	5	17.9
7 (33)	6	18.2	23	69.7	-	-	4	12.1
8 (30)	4	13.3	23	76.6	1	3.3	2	6.7
Total (174)	**11**	**6.3**	**116**	**66.7**	**1**	**0.6**	**46**	**26.4**
***ICATZ* Item 3-I**								
3 (26)	-	-	5	19.2	-	-	21	80.8
4 (28)	-	-	11	39.2	-	-	17	60.7
5 (29)	2	6.9	16	55.2	-	-	11	37.9
6 (28)	-	-	18	64.3	-	-	10	35.7
7 (33)	3	9.1	18	56.5	-	-	12	36.4
8 (30)	4	13.3	14	46.7	-	-	12	40.0
Total (174)	**9**	**5.2**	**82**	**47.1**	**-**	**-**	**83**	**47.7**
***ICATZ* Item 4-I**								
3 (26)	-	-	-	-	-	-	26	100.0
4 (28)	-	-	-	-	-	-	28	100.0
5 (29)	-	-	-	-	-	-	29	100.0
6 (28)	-	-	1	3.6	-	-	27	96.4
7 (33)	-	-	6	18.2	-	-	27	81.8
8 (30)	-	-	3	10.0	-	-	27	90.0
Total (174)	**-**	**-**	**10**	**5.7**	**-**	**-**	**164**	**94.3**

Note. (*f*) = frequency; (*ICATZ*) = Ice Cream at the Zoo; (L) = Literal Comprehension Test Item; (I) = Inferential Comprehension Test Item.

(continued)

TABLE 4.1 (continued)

Test-item/ Age in years (n)	**Keyed Answer-** **Justified** Test Score 1 Thinking Score 2-3		**Keyed Answer-** **Not Justified** Test Score 1 Thinking Score 0-1		**Unkeyed Answer-** **Justified** Test Score 0 Thinking Score 2-3		**Unkeyed Answer-** **Not Justified** Test Score 0 Thinking Score 0-1	
	f	%	*f*	%	*f*	%	*f*	%
***ICATZ* Item 5-L**								
3 (26)	-	-	12	46.2	-	-	14	53.8
4 (28)	-	-	17	60.7	-	-	11	39.3
5 (29)	-	-	27	93.1	-	-	2	6.9
6 (28)	-	-	26	92.9	-	-	2	7.1
7 (33)	2	6.1	30	90.9	-	-	1	3.0
8 (30)	-	-	30	100.0	-	-	-	-
Total (174)	**2**	**1.1**	**142**	**81.6**	**-**	**-**	**30**	**17.2**
***ICATZ* Item 6-I**								
3 (26)	-	-	12	46.2	-	-	14	53.8
4 (28)	-	-	16	57.1	-	-	12	42.9
5 (29)	-	-	16	55.2	-	-	13	44.8
6 (28)	1	3.6	16	57.1	-	-	11	39.3
7 (33)	-	-	28	85.0	-	-	5	15.1
8 (30)	3	10.0	23	76.6	-	-	4	13.3
Total (174)	**4**	**2.3**	**111**	**63.4**	**-**	**-**	**59**	**33.9**
***ICATZ* Item 7-I**								
3 (26)	-	-	2	7.7	-	-	24	92.3
4 (28)	-	-	5	17.9	-	-	23	82.1
5 (29)	1	3.4	12	41.3	-	-	16	55.2
6 (28)	1	3.6	15	53.5	-	-	12	42.9
7 (33)	2	6.1	24	72.7	-	-	7	21.2
8 (30)	1	3.3	20	66.7	-	-	9	30.0
Total (174)	**5**	**2.9**	**78**	**44.8**	**-**	**-**	**91**	**52.3**
***ICATZ* Item 8-I**								
3 (26)	-	-	10	38.5	-	-	16	61.5
4 (28)	-	-	15	53.5	-	-	13	46.4
5 (29)	-	-	20	69.0	-	-	9	31.0
6 (28)	2	7.1	26	92.9	-	-	-	-
7 (33)	8	24.2	20	60.6	-	-	5	15.1
8 (30)	4	13.3	21	70.0	-	-	5	16.7
Total (174)	**14**	**8.0**	**112**	**64.4**	**-**	**-**	**48**	**27.6**

Note. (*f*) = frequency; (*ICATZ*) = Ice Cream at the Zoo; (L) = Literal Comprehension Test Item; (I) = Inferential Comprehension Test Item.

The distributions of keyed and unkeyed responses showed a consistent pattern as a function of age: the keyed responses increased and the unkeyed responses decreased. That is, 3- and 4-year olds provided the lowest number of keyed (correct) answers and 7- and 8-year olds obtained the highest number of correct answers to the listening comprehension questions. Although the frequency distribution results indicated that 3- and 4-year olds did not perform well on a number of the *ON* test items, it is not the case that these younger children did not respond to the test items. Rather, it means that although some answers were on the right track, the responses did not meet the specified test-item criteria because they were too vague and incomplete. In particular, when asked a question that required two aspects for a correct response, the younger children provided only one of the aspects. For example, the first listening comprehension question for the *UP* story (*Who was at the party?*) required children to name the two children in the story who attended the party. Some younger children only recalled one character's name or gave a general description of the gender of the children at the party (i.e., 'two boys and a girl') which was also consistent with the types of unkeyed answers given by the older children.

In addition, the frequency distribution revealed some test performance discrepancies between the two oldest ages. In fact, there were only two *ON* test items in which 8-year olds clearly outperformed 7-year olds (*UP* Items 1 and 3) and on several test items, their performance was either commensurate with or the 7-year olds performed slightly better and obtained the highest percentage of keyed answers overall (*UP* Items 4 and 5; *ICATZ* Items 1, 3, 4, 7, and 8). Furthermore, the children between 3- and 8- years of age had slightly higher percentages of keyed answers to the literal listening comprehension questions (46% for the total sample) than to the inferential listening comprehension questions (41% for the total sample), even though some literal test items were answered with less accuracy than inferential test items on the *ON* measure.

ON Keyed Answers

According to Table 4.1, less than half of the children in the sample (3-8 years) obtained keyed answers to the listening comprehension test items (42% of the total *ON* test-item responses). Scoring the *ON* test-item responses was straightforward. The test-item response analysis showed that keyed answers (test score 1) were comprised of the essential criteria established for a correct response to the test question indicating that children effectively identified, discriminated, and integrated the relevant story information and/or background knowledge pertaining specifically to the test item. The item response analysis also showed that children who gave keyed responses addressed the test questions directly with clear and specific language use. The protocol analyses indicated that children's justifications for their keyed answers ranged in quality and specificity depending on how well they were able to use the available and relevant information sources to support their answers.

ON keyed answers-justified. According to Table 4.1, there was a relatively small number of 5- to 8-year olds who provided keyed answers with clear justifications for twelve of the fourteen listening comprehension test questions (between 1% and 14% for each test item or 4% of the total test-item responses overall). Moreover, the percentage of keyed answers-justified for literal and inferential listening comprehension test items was similar (5% and 4%, respectively). On the contrary, none of the 3- and 4-year olds in the sample provided sufficient justification for keyed test-item responses. Based on these results, it is fair to assume that most young children may not have developed the concept of justification for an answer (i.e., 'How did you know that?') because they merely repeated the answer. The task was likely unfamiliar and thus appropriate for only the most precocious children for these ages. In fact, there were some instances in which young children responded to the probes in relation to the task at hand but their responses were not of sufficient quality or sophistication to be rated high on the thinking rubric. This finding signals the need for further study of young children's thinking and reasoning by using an alternative method for classifying their response justifications.

Of the keyed responses with proper justification, there were variations in the quality of explanations provided by the children. Approximately 1% of the total *ON* keyed responses were substantiated with consistent and complete justification. In other words, a very low percentage of children explained their answers by inferring and integrating *all* relevant oral and visual cues and background knowledge related to the specific test item. These complete and consistent explanations received the highest thinking score (3) rating and showed sophistication and specificity. The justifications for the remaining 3% of the total *ON* keyed answers in this category were consistent but incomplete (thinking score 2). That is, children's explanations were based on some, but not all pertinent oral and visual cues or background knowledge and were not as comprehensive as the types of responses that met the criteria for the highest thinking score (3). It seemed that the children did not monitor for consistency and completeness. In any case, all of the keyed answers with justification showed that children both understood and interpreted the test items and story well. Furthermore, they were able to synthesize and infer from the relevant oral and visual information and relevant background knowledge to answer both literal and inferential listening comprehension questions.

***ON* keyed answers-not justified.** According to Table 4.1, the percentage of keyed answers to the listening comprehension test questions that were not justified ranged from 6% to 82% across the 14 test items or 38% of the total test-item responses overall. The percentage of keyed answers-not justified for literal and inferential listening comprehension test items was similar (40% and 37%, respectively). When children listen to stories on a day-to-day basis, they are typically not asked about what they have heard and certainly not queried to provide justification for their responses. Thus, it was not surprising that a high proportion of keyed responses for the *ON* test items were not justified on the basis of good thinking and reasoning.

There were several *ON* test items in which all or most of the keyed responses were not justified. In particular, two inferential listening comprehension items, *UP* Item 3 (*Why did Jamal and Ling bring presents?*) and *ICATZ* Item 4 (*Why was the ice cream man surprised to see Ben three times?*), had the fewest keyed answers overall (7% and 6% keyed responses, respectively) and all correct responses to these test items were accompanied by poor reasoning (thinking score 0 or 1). In addition, inferential listening comprehension *UP* Item 5 (*Why is Tommy excited about his unusual present?*) and literal listening comprehension *ICATZ* Item 5 (*Name two animals that ate Ben's ice cream.*) had disproportionately more keyed answers that were not justified (approximately 98% of the keyed answers) than were justified.

A review of the keyed responses for these four test questions confirmed that although some children made the necessary inferences and used the relevant explicit and implicit text cues to provide a correct response to the test items, their justifications fell short because they were often too vague and incomplete and children clearly did not move beyond their original response to provide a consistent and complete justification for their test-item answers. At times their explanations completely overlooked the relevant text cues and were based primarily on mere speculation and background knowledge. Although many children obtained keyed answers to the *ON* test items and demonstrated that they successfully interpreted the questions and comprehended the pertinent information sources, the task of justifying their responses was an entirely different matter and evidently very challenging for most of the sample.

ON **Unkeyed Answers**

Table 4.1 showed that more than half of the children from 3 to 8 years gave unkeyed answers to the listening comprehension test items (58% of the total *ON* test-item responses). When children's answers to the test-items did not correspond to the specified criteria, they were given a test score of 0 for an unkeyed answer. Unkeyed test-item responses signified that children did not attend either to the explicit, implicit oral and visual text cues, or make the necessary inferences related to the particular comprehension test items. Children who provided unkeyed answers to the listening comprehension questions failed to draw upon and integrate the necessary information sources in order to answer the test questions directly and completely.

Although unkeyed test-item responses were assigned the same score (0) and grouped into a single category, the item response analysis revealed differences in the quality and types of incorrect answers. Consequently, different response patterns provided valuable insight into children's listening comprehension skills. Specifically, the item response analysis revealed four main patterns of unkeyed responses with distinct characteristics:

(a) The first and most common pattern included unkeyed answers that were partially-correct. The children understood and addressed the particular question asked but their verbal protocol lacked the specificity required for a complete response including some, but not all test-item criteria. These responses indicated that children experienced difficulty discriminating and inferring from the pertinent story cues to answer correctly the specific test item and failed to monitor their listening. For example, their answers often focused on only one relevant aspect of the story while ignoring other pertinent cues required for a complete keyed response.

(b) The second group of unkeyed answers included some vaguely-related story information which did not match any of the test-item criteria. The children appeared to understand the test question but their responses were so general and vague that they did not address the test question directly. Some children either relied mainly on the picture cues or they conjured an unjustified interpretation of the oral and visual text cues and disregarded the relevant story information altogether.

(c) The third cluster of unkeyed answers consisted of erroneous and irrelevant information unrelated to the question asked. These types of responses suggested that children either misinterpreted or misunderstood the test question or the explicit and implicit oral and visual text cues. The children inadvertently either focused on the incorrect aspect of the story in their response to the test item or they answered a different question from the actual test question asked. Sometimes, children merely repeated random oral story cues verbatim, made false claims, or jumped to conclusions not verified in the story. Moreover, these children were overly-dependent on their background knowledge and often over-generalized or personalized their response to the question while completely ignoring the explicit and implicit oral and visual text cues. Overall, these particular unkeyed responses demonstrated a general lack of specificity, clarity, and precision in expressive vocabulary.

(d) The final category of unkeyed answers included several different variations including ambiguous responses (i.e., vague and general answers that neither showed interpretation of nor addressed the test item directly in relation to the specific story information), ‘I don’t know’ responses, or no response at all. On the basis of these particular unkeyed responses, it remained unclear why children responded the way they did and ultimately, whether they understood the test item or not.

The patterns of unkeyed responses in the current research closely resembled the characteristics of poor comprehension found in previous empirical studies with older children in various reading contexts (e.g., Brandao & Oakhill, 2005; Cain et al., 2001; Carlson, Seipel, & McMaster, 2014; Carlson et al., 2014; Lipson, 1982; McCormick, 1992; McMaster et al., 2012, McMaster et al., 2014; Nicholson & Imlach, 1981; Rapp et al., 2007). Collectively, the response patterns that have emerged across many different studies of comprehension help to inform why children's comprehension breaks down. Not all comprehension difficulties are one and the same. It has become clear that the more that is known about comprehension difficulties, the greater the chance of addressing the source of the problem.

The corpus of unkeyed responses for each *ON* listening comprehension test item was scrutinized to determine the reasons children responded the way they did; whether particular test items or stories were faulty or misleading; and ultimately, whether the test items measured the respective construct intended (i.e., listening comprehension). Particular attention was focused on the *ON* test items (*UP* Items 1-4 and *ICATZ* Item 4) with significantly more unkeyed than keyed responses (between 80% - 94%).

Specifically, the two listening comprehension questions (*UP* Item 3 and *ICATZ* Item 4) with the highest frequency of unkeyed responses were likely the most challenging for children overall due to the more complex reasoning and specificity required. Both test items were ‘why’ questions which required children to integrate their relevant prior knowledge with the pertinent explicit and implicit oral and visual text cues in order to make inferences about the characters’ actions or emotions. In both cases, the unkeyed responses were generally fragmented, incomplete, and lacked reasoning. For example, a keyed response to *UP* Item 3 (*Why did Jamal and Ling bring presents?*) required children to use their relevant prior knowledge and the implicit text cues to make an evidence-based inference about the story context (i.e., Jamal and Ling brought presents for *Tommy’s birthday*). Many unkeyed responses revealed a partial inference (i.e., ‘because it was someone’s birthday’) and the pronoun references used were left unspecified. That is, children did not mention whose birthday or for whom the presents were intended. In general, these vague and incomplete unkeyed test-item responses could be indicative of inferior listening comprehension or a lack of linguistic or cognitive ability with anaphoric referencing.

In addition, *ICATZ* Item 4 (*Why was the ice cream man surprised to see Ben three times?*) had the highest number of unkeyed responses overall on the *ON* measure (94% of the total sample). It is unclear why so many children had difficulty with this inferential listening comprehension test item. Typically, there was a lack of coherence between the test item and the answers given. The most common response pattern showed poor interpretation of the test question and the relevant information sources. For example, some children explained how they knew that the ice cream man was surprised to see Ben, others focused on what happened to the ice cream, and still others gave reasons why Ben returned to the ice cream stand repeatedly (e.g., ‘because he wanted more ice cream’ or ‘because the ice cream fell on the ground’). The remaining unkeyed responses to *ICATZ* Item 4 included ambiguous and erroneous story information or prior knowledge, poor language use, or no response. Regardless, most children did not make meaningful connections between the story information provided (i.e., explicit and implicit text cues) and their relevant prior knowledge to infer that the ice cream man was surprised to see Ben three times because he had already given him ice cream and he did not expect Ben to repeatedly drop his ice cream. Although the inference seemed relatively clear, the question was particularly challenging for most children in the sample nonetheless.

Other test items with relatively low test performance included *UP* literal comprehension Item 2 (*Where was the party?*) and inferential comprehension Item 4 (*What do you think is inside the unusual present and how do you know that?*). The unkeyed responses to these test items were distributed among the four primary response patterns outlined earlier. In particular, some responses were partially-correct (i.e., matching only some of the keyed test-item criteria); others were basically too general and vague; still others used erroneous story information and overgeneralized from prior knowledge; and a small quantity were ‘I don’t know’ or no response.

In any case, while there was no indication that the *ON* listening comprehension test items with the highest frequency of unkeyed responses were particularly difficult or confusing, most children in the sample provided responses which lacked sufficient precision and specificity required for a keyed answer. The high rates of unkeyed responses signified that children generally did not interpret or answer the listening comprehension test items well. They did not attend to the explicit and implicit oral and visual text cues or integrate relevant prior knowledge to make evidence-based inferences related to the individual test items and the story information provided. Thus, unkeyed test-item responses essentially represented gaps and deficiencies in listening comprehension which is precisely what they were designed to measure. Furthermore, the test-item response analysis offered no reason to think that the *ON* questions or stories were problematic or necessitated any further amendments.

ON unkeyed answers-justified. Table 4.1 revealed that a very small number of unkeyed answers to the listening comprehension questions (less than 1%) were justified to some extent on the basis of partially good thinking and reasoning. In these few instances (approximately 5 out of a total of 2,436 test-item responses calculated from 174 responses x 14 test-items), responses to the specific test questions met some, but not all of the keyed item criteria. In other words, the answers were partially-correct but incomplete. Moreover, the explanations for these unkeyed responses obtained a higher thinking score (2) because children provided additional related story information to substantiate their responses. Despite that, four out of five of these explanations did not mention the specific keyed criteria previously omitted from the original response to the test item indicating that children's understanding of the pertinent story information remained incomplete nonetheless. For example, children were required to name two of the three story characters (i.e., *Tommy, Jamal,* or *Ling*) in response to *UP*

Item 1 (*Who was at the party?*). The names of all three story characters were mentioned repeatedly throughout the story. The following transcript illustrates an unkeyed response substantiated with additional relevant story information:

Karen: *Who was at the party?*

T8-14: *Jamal and (pause) um, there was a girl and two boys.*

Karen: *How did you know that Jamal and a girl and two boys were at the party?*

T8-14: *Because, I, I noticed the girl and there were two other guys and one was the birthday boy so he invited two other friends.*

Karen: *How did you know about that? What made you think that?*

T8-14: *Because they all brought presents so I was guessing that they were the only ones that were coming really, and yeah that's really all.*

In this transcript, the 8-year old named only one of the story characters and after some contemplation, he mentioned the gender of the three characters in the story, presumably because he could not recall the names of the other two story characters. When asked to explain his answer to the test item, he elaborated with some accurate and relevant story details regarding the story characters (i.e., "one was the birthday boy…he invited two other friends…they all brought presents") but his explanation did not include the missing keyed criteria (i.e., the names of the other story characters) or an acceptable alternative response specifically related to the question asked. Although the child understood and attempted to address the test item, his response demonstrated apparent gaps in listening comprehension related to the particular characters featured in the story.

Diagnostically, the child's response showed attention to some story detail but a lack of attention to character names. Overall, the data did not reveal any other alternative unkeyed answers to the different listening comprehension test items with consistent and complete justifications to warrant further revision of the questions or the corresponding keyed criteria, or either of the two *ON* stories.

ON unkeyed answers-not justified. Table 4.1 showed that one of the largest pools of responses for many of the test items for each age from 3- to 8-years were unkeyed answers-not justified (between 18% and 94% for each test item or 57% of the total test-item responses overall). More than half of the responses for the literal and inferential listening comprehension test items were unkeyed answers-not justified (54% and 59%, respectively). When children failed to justify their responses, the types of explanations provided were basically further confirmation of the four unkeyed response patterns delineated previously. Unkeyed responses that were not justified, for example, frequently demonstrated poor expressive language with explanations that were fragmented, vague, and unclear (i.e., pronouns and referents were inconsistent and not used appropriately). In many instances, it was difficult to know precisely what children meant. Children who used poor thinking and reasoning as the basis of their test-item responses showed a general lack of understanding of the story content and experienced difficulty integrating the relevant story information into their explanations. In other words, children did not use the available information sources well to explain their thinking. Occasionally, they focused too much on the picture cues reporting what they recalled 'seeing' in the story illustrations which may not have been relevant to the question asked, while overlooking important written story information.

In contrast, some children were overly-dependent on their background knowledge often drawing from their own experiences to explain why they answered the test questions the way they did. Other flawed types of responses by the children consisted of merely reporting from the story verbatim without interpretation or inference; overstating the evidence and making false claims; or reiterating the answer given without offering additional relevant information or further insight to explain their thinking and what information exactly was used as the basis of their answer to the test question. Other faulty explanations included random erroneous story details and ambiguous responses (i.e., "I just knew", or "because it said in the story", or "I saw it in the pictures). Finally, no credit was afforded when no response was offered.

Overall, the rationale that children provided for most unkeyed test-item responses was poorly conceived. Children's explanations did not include the relevant keyed criteria omitted from their original test-item answers nor did the justifications confirm that the children had indeed comprehended the test items, the relevant text cues, or made the appropriate and complete inferences in accordance with the keyed test-item criteria. All such indicators confirmed that the *ON* test items performed well as measures of listening comprehension.

TELL Oral Reading-Reading Comprehension (OR-RC)

The *TELL Oral Reading-Reading Comprehension (OR-RC)* measure included two separate subsections: 7a - Read-Talk-Reread-Read (*RTRR*) and 7b -

A Teddy Bear's Birthday Wish (*TBBW*). The *RTRR* measure has four different stories with a combination of five literal and inferential reading comprehension test items per story (20 test items in total). The *TBBW* measure has one longer story with five literal and five inferential reading comprehension questions (10 test items in total). The *OR-RC* stories and test items range in difficulty from simple to complex and were administered with particular age groups specified in the *TELL* test administration manual. Three- to five-year olds completed all four *RTRR* stories (i.e., *Dogs*, *Teddy Bear, Teddies,* and *Dinosaur*) with a combination of twenty reading comprehension test items (five questions per story). Six-year olds completed the second pair of *RTRR* stories (i.e., *Teddies* and *Dinosaur*) with ten reading comprehension test items in total (five questions per story). And six- to eight-year olds completed the *TBBW* story with ten reading comprehension test items.

According to the *TELL* test administration manual, 6-year olds were included in the administration of the final two *RTRR* stories (i.e., *Teddies* and *Dinosaur*) and *TBBW* reading comprehension measures. In the present study, the total sample of 6-year olds (n = 28) completed the *RTRR-Teddies* and *Dinosaur* measures, whereas only a subsample of the age group (n = 11) also completed the *TBBW* measure due to the difficulty of the latter text and children's inability to read the more complex story independently (i.e., less than half of the 6-year olds in the sample completed both *RTRR- Teddies* and *Dinosaur* and *TBBW* reading comprehension measures). A comparison of test performance for the subsample of 6-year olds (n = 11) who completed both test components is discussed later.

Frequency distributions were calculated for the total *RTRR* and *TBBW* subsection test performance according to age and the combined test and thinking scores for each *RTRR* and *TBBW* test item by age. The reading comprehension test performance results are presented next.

OR-RC Total Subsection Test Score Frequency Distribution by Age

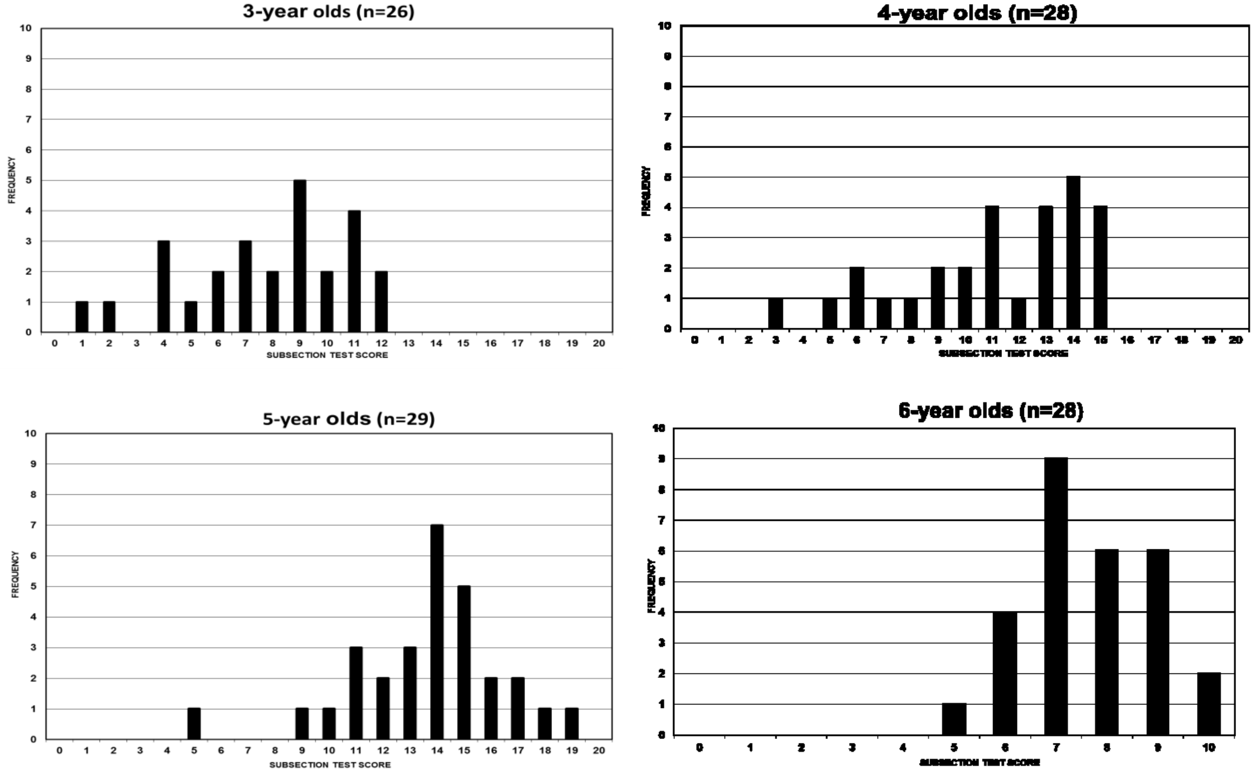
Figures 4.2 and 4.3 present the frequency distribution for total test score performance on the *RTRR* and *TBBW* measures for the different age groups.

Figure 4.2 showed an increasing pattern of total *RTRR* test scores for the children from 3- to 5-years. The *RTRR* total test scores ranged between 1 and 19 (out of a total possible test score of 20) for the 3- to 5-year olds and between 5 and 10 (out of a total possible test score of 10) for the 6-year olds on the *RTRR-Teddies* and *Dinosaur* test items. Overall, children in the younger age groups had total test scores in the low-mid range and children in the older age groups obtained total test scores mainly at the mid-high range on the *RTRR* subsection. However, the frequency distribution also showed that a number of 3-year olds (n = 6) obtained higher total *RTRR* test scores than a few 5-year olds (n = 3).

Figure 4.3 presents the frequency distribution for total test score performance on the *TBBW* subsection for the different age groups. The *TBBW* total test scores ranged between 2 and 9 (out of a total possible test score of 10) for children between 6- and 8-years old. There were consistently small increases in the *TBBW* total test score performance across the ages. The majority of *TBBW* total test scores for each respective age group fell within the mid-range of the scale. Eight year olds showed the greatest test performance variability with total test scores ranging from low-to-high. The *TBBW* total test score frequency distribution also indicated that a few 6-year olds (n = 3) obtained higher total test scores on the *TBBW* subsection than many 8-year olds (n = 13).

Figure 4.2

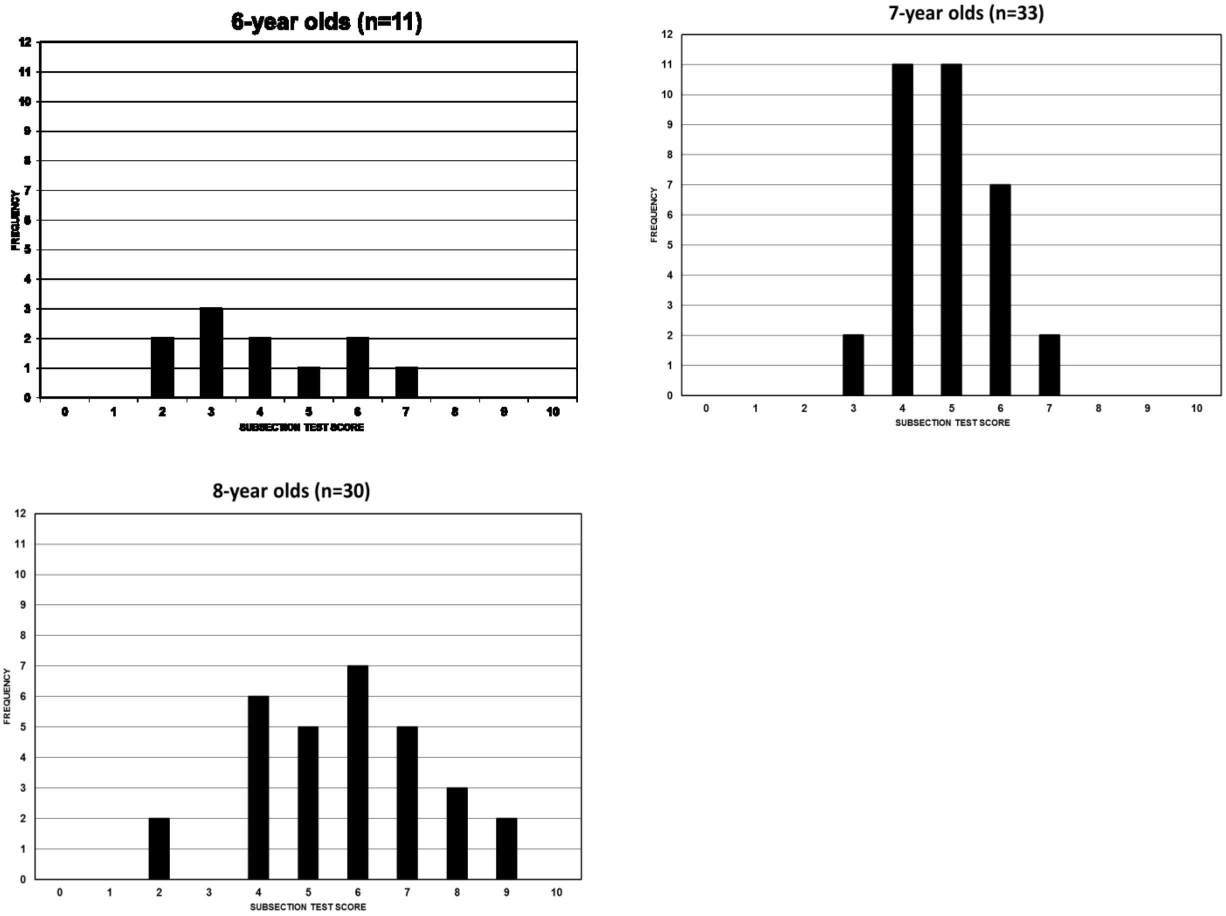
Oral Reading-Reading Comprehension RTRR Total Test Score Frequency Distribution by Age



Note. Total subsection test scores for 3-5 year olds included the test items for all four RTRR books (*Dogs, Teddy Bear, Teddies,* and *Dinosaur)* and the total subsection test scores for 6 year olds included the test items for two RTRR books (*Teddies* and *Dinosaur)*.

Figure 4.3

Oral Reading-Reading Comprehension TBBW Total Test Score Frequency Distribution by Age

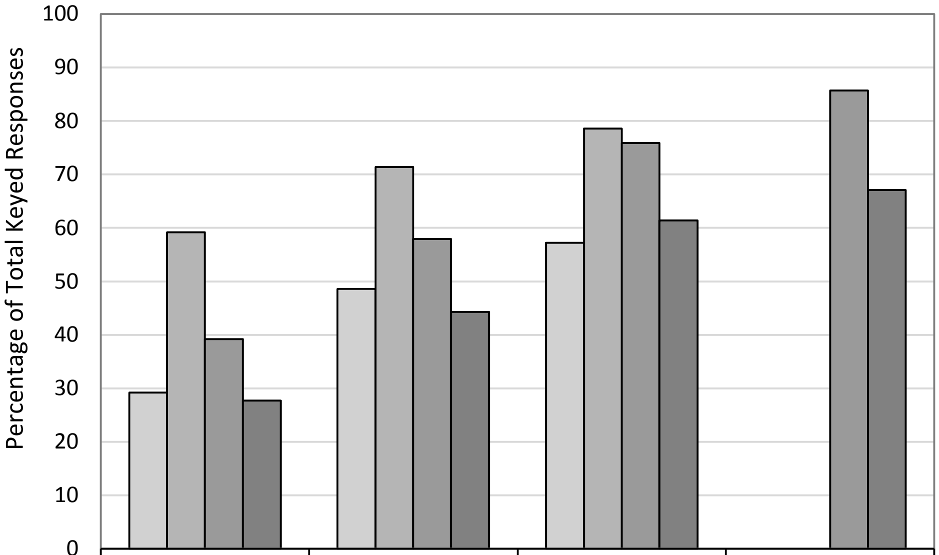


The frequency distribution of total test performance on the *RTRR* and *TBBW* measures support the developmental sensitivity of the respective reading comprehension measures. Notwithstanding a few exceptions in which younger children outperformed older children, there were consistent increases in the total number of keyed responses obtained by the older age groups on each reading comprehension subsection, presumably because, as expected, older children have developed better reading comprehension skills than younger children.

In addition, the subsample of 6-year olds (n = 11) who completed *RTRR- Teddies* and *Dinosaur*, and *TBBW* measures showed a considerable decrease in the total test scores achieved on the respective reading comprehension measures. Specifically, the 6-year old subsample performed significantly better overall on the *RTRR- Teddies* and *Dinosaur* test items (mid-high total test scores) than the *TBBW* measure (low-mid total test scores) and their test performance decreased incrementally across the individual texts and test items (85% keyed responses on *RTRR Teddies*, 69% keyed responses on *RTRR Dinosaur*, and 41% keyed responses on *TBBW*). Granted, the sample was small but the findings were important nonetheless because they confirmed that the reading comprehension measures assessed different levels of skill as intended.

Overall, the *OR-RC* test performance results showed clear differences between the different ages on the *RTRR* and *TBBW* measures. In general, the *RTRR* results indicated that more of the older children (5- and 6-year olds) typically obtained higher total test scores than the younger children (3- and 4-year olds) on the *RTRR* subsection. However, a comparative analysis of the four separate *RTRR* measures revealed an unexpected performance pattern. Figure 4.4 shows the distribution of total keyed responses on each respective *RTRR* story and corresponding reading comprehension test items across the four different age groups. In accordance with the test administration guidelines, 6-year olds completed the latter two *RTRR* measures only (*Teddies* and *Dinosaur*).

Figure 4.4 Percentage of Total Keyed Responses on the RTRR Stories and Reading Comprehension Test Items by Age



	3-year olds	4-year olds	5-year olds	6-year olds
□ Dogs	29	49	57	-
■ Teddy Bear	59	71	79	-
■ Teddies	39	58	76	86
■ Dinosaur	28	44	61	67

Note. Blank spaces (-) indicate that the measures were not administered to this age group as per the test administration manual guidelines.

The comparison of test performance across the four different *RTRR* measures (i.e., *Dogs, Teddy Bear, Teddies,* and *Dinosaur*) revealed surprising variability on the first *RTRR* story and test items in particular. Since the *RTRR* stories and reading comprehension test items were designed to range in difficulty from simple to complex in order to discriminate developmental differences in test performance, it was anticipated that older children would generally perform better than younger children on all four *RTRR* measures. And, accordingly, children at each age would obtain the highest percentage of keyed answers on the first (and presumably, the easiest) *RTRR Dogs* comprehension test items with diminishing performance on the subsequent *RTRR* stories and test items (i.e., *Teddy Bear,* *Teddies, and Dinosaur*). Results only partially supported this expectation with remarkably similar and consistent performance across the different age groups on the four *RTRR* stories. As expected, there were steady age-related test score increases with the youngest age group (3-year olds) obtaining the fewest number of keyed answers and the oldest age group (either 5- or 6-year olds) achieving the highest number of correct responses on each respective *RTRR* story and test items. In spite of that, performance on *RTRR-Dogs* was consistently lower than the subsequent *RTRR Teddy Bear* test items for all three ages (3- to 5-year olds). Moreover, *RTRR Teddy Bear* had the highest percentage of keyed responses overall for all three age groups, followed by the anticipated and progressive decline in test performance on the last two, more difficult *RTRR Teddies* and *Dinosaur* measures.

Several reasons are possible for lower test performance on the *RTRR Dogs* measure. One explanation could be attributed to performance anxiety and lack of familiarity with the task, especially when the testing procedures were initially introduced. That is, reading the first of four stories and then answering several comprehension questions in an assessment context was likely a novel experience for younger children between 3- and 5-years of age. In addition, *Dogs* Item 1 (*What are two things that dogs do?*) produced an unanticipated pattern of unkeyed responses which lowered overall performance on the measure. Specifically, many children across the different ages used prior knowledge as the main frame of reference to respond to the literal comprehension question, as opposed to using the explicit text cues as the primary information source. It was clear that these children interpreted the question differently than intended and as a result there were considerably more unkeyed responses to the test item. Test performance on *Dogs* Item 1 is discussed in more detail later, but suffice it to say that the lower test-item scores affected overall performance on the *RTRR Dogs* reading comprehension measure. And finally, it is possible that the *RTRR Dogs* story and comprehension questions were generally more challenging for this sample of children for some reason. Although the text and test items were not characteristically more difficult than the other *RTRR* measures, further investigation may be warranted nonetheless.

OR-RC **Test-item Frequency Distribution by Age**

A frequency distribution of the *OR-RC* test and thinking scores was compiled to compare performance on the reading comprehension test items within and across the different ages and to examine the relationship between test and thinking performance. Since the listening and reading comprehension scoring procedures for evaluating test-item responses and thinking were identical, the criteria described in the *Oral Narrative* section to differentiate between keyed and unkeyed test-item responses which were either justified or not was also relevant to the *OR-RC* measures and are used here.

Table 4.2 presents the number and percentage of keyed and unkeyed answers either justified or not justified for the individual *RTRR* and *TBBW* test items according to age distributed across the same four categories as the *Oral Narrative* results: (a) Keyed Answers-Justified (test score 1 and thinking score 2 or 3); (b) Keyed Answers-Not Justified (test score 1 and thinking score 0 or 1); (c)

Unkeyed Answers-Justified (test score 0 and thinking score 2 or 3); and (d) Unkeyed Answers-Not Justified (test score 0 and thinking score 0 or 1).

The *OR-RC* test-item performance results in Table 4.2 showed that the proportion of keyed and unkeyed responses varied by test item and age. The *RTRR* test-item frequency distribution showed that children between 3- and 6-years of age obtained more keyed than unkeyed responses (ranging between 51% and 94%) on thirteen of twenty *RTRR* test items (literal comprehension *Dogs* Items 1 and 3, *Teddy Bear* Items 1-3, *Teddies* Items 1-3, and *Dinosaur* Items 1 and 2; and inferential comprehension *Dogs* Item 5, *Teddy Bear* Item 5, and *Teddies* Item 5) and alternatively, they had higher percentages of unkeyed than keyed responses (ranging between 55% and 91%) on the remaining seven *RTRR* reading comprehension test items (literal comprehension *Dogs* Item 2; and inferential comprehension *Dogs* Item 4, *Teddy Bear* Item 4, *Teddies* Item 4, and *Dinosaur* Items 3-5). In other words, greater numbers of children in the sample got these test items incorrect.

TABLE 4.2

Frequency and Percentage of Keyed and Unkeyed Answers (Test Scores 1-0) - Justified and Not Justified (Thinking Scores of 2-3 or 0-1) for each TELL Oral Reading-Reading Comprehension Test Item by Age and Total Sample

Test-item/ Age in years (n)	**Keyed Answer-** **Justified** Test Score 1 Thinking Score 2-3		**Keyed Answer-** **Not Justified** Test Score 1 Thinking Score 0-1		**Unkeyed Answer-** **Justified** Test Score 0 Thinking Score 2-3		**Unkeyed Answer-** **Not Justified** Test Score 0 Thinking Score 0-1	
	f	%	*f*	%	*f*	%	*f*	%
DOGS-Item 1-L								
3 (26)	-	-	13	50.0	-	-	13	50.0
4 (28)	-	-	14	50.0	-	-	14	50.0
5 (29)	-	-	15	51.7	-	-	14	48.3
Total (83)	**-**	**-**	**42**	**50.6**	**-**	**-**	**41**	**49.4**
DOGS-Item 2-L								
3 (26)	-	-	12	46.2	-	-	14	53.8
4 (28)	-	-	11	39.3	-	-	17	60.7
5 (29)	2	6.9	12	41.4	-	-	15	51.7
Total (83)	**2**	**2.4**	**35**	**42.2**	**-**	**-**	**46**	**55.4**
DOGS-Item 3-L								
3 (26)	-	-	5	19.2	-	-	21	80.8
4 (28)	-	-	16	57.1	-	-	12	42.9
5 (29)	3	10.3	20	69.0	-	-	6	20.7
Total (83)	**3**	**3.6**	**41**	**49.4**	**-**	**-**	**39**	**47.0**
DOGS-Item 4-I								
3 (26)	-	-	4	15.4	-	-	22	84.6
4 (28)	-	-	10	35.7	-	-	18	64.3
5 (29)	-	-	9	31.0	-	-	20	69.0
Total (83)	**-**	**-**	**23**	**27.7**	**-**	**-**	**60**	**72.3**
DOGS-Item 5-I								
3 (26)	-	-	4	15.4	-	-	22	84.6
4 (28)	1	3.6	16	57.1	-	-	11	39.3
5 (29)	-	-	22	75.9	-	-	7	24.1
Total (83)	**1**	**1.2**	**42**	**50.6**	**-**	**-**	**40**	**48.2**

Note. (f) = frequency; *(*L) = Literal Comprehension Test Item; (I) = Inferential Comprehension Test Item.

(continued)

TABLE 4.2 (continued)

Test-item/ Age in years (n)	**Keyed Answer- Justified** Test Score 1 Thinking Score 2-3		**Keyed Answer- Not Justified** Test Score 1 Thinking Score 0-1		**Unkeyed Answer- Justified** Test Score 0 Thinking Score 2-3		**Unkeyed Answer- Not Justified** Test Score 0 Thinking Score 0-1	
	f	%	*f*	%	*f*	%	*f*	%
TEDDY BEAR - Item 1-L								
3 (26)	-	-	21	80.8	-	-	5	19.2
4 (28)	-	-	25	89.3	-	-	3	10.7
5 (29)	4	13.8	24	82.8	-	-	1	3.4
Total (83)	**4**	**4.8**	**70**	**84.3**	**-**	**-**	**9**	**10.8**
TEDDY BEAR - Item 2-L								
3 (26)	-	-	19	73.1	-	-	7	26.9
4 (28)	-	-	22	78.6			6	21.4
5 (29)	2	6.9	25	86.2			2	6.9
Total (83)	**2**	**2.4**	**66**	**79.5**			**15**	**18.1**
TEDDY BEAR - Item 3-L								
3 (26)	-	-	23	88.5	-	-	3	11.5
4 (28)	1	3.6	26	92.9	-	-	1	3.6
5 (29)	1	3.4	27	93.1			1	3.4
Total (83)	**2**	**2.4**	**76**	**91.6**			**5**	**6.0**
TEDDY BEAR - Item 4-I								
3 (26)	-	-	4	15.4	-	-	22	84.6
4 (28)	1	3.6	6	21.4	-	-	21	75.0
5 (29)	4	13.8	5	17.2	-	-	20	69.0
Total (83)	**5**	**6.0**	**15**	**18.1**	**-**	**-**	**63**	**75.9**
TEDDY BEAR - Item 5-I								
3 (26)	1	3.8	9	34.6	-	-	16	61.5
4 (28)	-	-	19	67.9	-	-	9	32.1
5 (29)	-	-	22	75.9	-	-	7	24.1
Total (83)	**1**	**1.2**	**50**	**60.2**	**-**	**-**	**32**	**38.6**

Note. (f) = frequency; *(*L) = Literal Comprehension Test Item; (I) = Inferential Comprehension Test Item. *(continued)*

TABLE 4.2 (continued)

Test-item/ Age in years (n)	**Keyed Answer- Justified** Test Score 1 Thinking Score 2-3		**Keyed Answer- Not Justified** Test Score 1 Thinking Score 0-1		**Unkeyed Answer- Justified** Test Score 0 Thinking Score 2-3		**Unkeyed Answer- Not Justified** Test Score 0 Thinking Score 0-1	
	f	%	*f*	%	*f*	%	*f*	%
TEDDIES- Item 1-L								
3 (26)	-		8	30.8	-	-	18	69.2
4 (28)	1	3.6	16	57.1	-	-	11	39.3
5 (29)	3	10.3	22	75.9	-	-	4	13.8
6 (28)	7	25.0	16	57.1	-	-	5	17.9
Total (111)	**11**	**9.9**	**62**	**55.9**	**-**	**-**	**38**	**34.2**
TEDDIES- Item 2-L								
3 (26)	-	-	10	38.5	-	-	16	61.5
4 (28)	-	-	18	64.3	-	-	10	35.7
5 (29)	1	3.4	23	79.3	-	-	5	17.2
6 (28)	5	17.9	21	75.0	-	-	2	7.1
Total (111)	**6**	**5.4**	**72**	**64.9**	**-**	**-**	**33**	**29.7**
TEDDIES- Item 3-L								
3 (26)	1	3.8	17	65.4	-	-	8	30.8
4 (28)	-	-	24	85.7	-	-	4	14.3
5 (29)	1	3.4	26	89.7	-	-	2	6.9
6 (28)	11	39.3	17	60.7	-	-	-	-
Total (111)	**13**	**11.7**	**84**	**75.7**	**-**	**-**	**14**	**12.6**
TEDDIES- Item 4-I								
3 (26)	-	-	3	11.5	-	-	23	88.5
4 (28)	-	-	2	7.1	-	-	26	92.9
5 (29)	1	3.4	7	24.1	-	-	21	72.4
6 (28)	8	28.6	8	28.6	-	-	12	42.9
Total (111)	**9**	**8.1**	**20**	**18.0**	**-**	**-**	**82**	**73.9**
TEDDIES- Item 5-I								
3 (26)	-	-	12	46.2	-	-	14	53.8
4 (28)	-	-	20	71.4	-	-	8	28.6
5 (29)	5	17.2	21	72.4	-	-	3	10.3
6 (28)	9	32.1	18	64.3	-	-	1	3.6
Total (111)	**14**	**12.6**	**71**	**64.0**	**-**	**-**	**26**	**23.4**

Note. (f) = frequency; (TEDDIES) = *RTRR*-Teddy Bears; *(*L) = Literal Comprehension Test Item; (I) = Inferential Comprehension Test Item.

(continued)

TABLE 4.2 (continued)

Test-item/ Age in years (n)	**Keyed Answer- Justified** Test Score 1 Thinking Score 2-3		**Keyed Answer- Not Justified** Test Score 1 Thinking Score 0-1		**Unkeyed Answer- Justified** Test Score 0 Thinking Score 2-3		**Unkeyed Answer- Not Justified** Test Score 0 Thinking Score 0-1	
	f	%	*f*	%	*f*	%	*f*	%
DINOSAUR - Item 1-L								
3 (26)	-	-	15	57.7	-	-	11	42.3
4 (28)	-	-	22	78.6	-	-	6	21.4
5 (29)	2	6.9	24	82.8	-	-	3	10.3
6 (28)	6	21.4	22	78.6	-	-	-	-
Total (111)	**8**	**7.2**	**83**	**74.8**	**-**	**-**	**20**	**18.0**
DINOSAUR - Item 2-L								
3 (26)	-	-	18	69.2	-	-	8	30.8
4 (28)	1	3.6	26	92.9	-	-	1	3.6
5 (29)	-	-	29	100.0	-	-	-	-
6 (28)	7	25.0	21	75.0	-	-	-	-
Total (111)	**8**	**7.2**	**94**	**84.7**	**-**	**-**	**9**	**8.1**
DINOSAUR - Item 3-I								
3 (26)	-	-	3	11.5	-	-	23	88.5
4 (28)	-	-	9	32.1	-	-	19	67.9
5 (29)	-	-	13	44.8	-	-	16	55.2
6 (28)	-	-	16	57.1	-	-	12	42.9
Total (111)	**-**	**-**	**41**	**36.9**	**-**	**-**	**70**	**63.1**
DINOSAUR - Item 4-I								
3 (26)	-	-	-	-	-	-	26	100.0
4 (28)	-	-	-	-	-	-	28	100.0
5 (29)	-	-	6	20.7	-	-	23	79.3
6 (28)	-	-	4	14.3	-	-	24	85.7
Total (111)	**-**	**-**	**10**	**9.0**	**-**	**-**	**101**	**91.0**
DINOSAUR- Item 5-I								
3 (26)	-	-	-	-	-	-	26	100.0
4 (28)	-	-	4	14.3	-	-	24	85.7
5 (29)	2	6.9	13	44.8	-	-	14	48.3
6 (28)	7	25.0	11	39.3	-	-	10	35.7
Total (111)	**9**	**8.1**	**28**	**25.2**	**-**	**-**	**74**	**66.7**

Note. (f) = frequency; *(*L) = Literal Comprehension Test Item; (I) = Inferential Comprehension Test Item.

(continued)

TABLE 4.2 (continued)

Test-item/ Age in years (n)	**Keyed Answer-Justified** Test Score 1 Thinking Score 2-3		**Keyed Answer-Not Justified** Test Score 1 Thinking Score 0-1		**Unkeyed Answer-Justified** Test Score 0 Thinking Score 2-3		**Unkeyed Answer-Not Justified** Test Score 0 Thinking Score 0-1	
	f	%	*f*	%	*f*	%	*f*	%
***TBBW*- Item 1-L**								
6 (11)	1	9.1	10	90.9	-	-	-	-
7 (33)	12	36.4	21	63.6	-	-	-	-
8 (30)	3	10.0	27	90.0	-	-	-	-
Total (74)	**16**	**21.6**	**58**	**78.4**	**-**	**-**	**-**	**-**
***TBBW*- Item 2-L**								
6 (11)	1	9.1	7	63.6	-	-	3	27.3
7 (33)	2	6.1	23	69.7	-	-	8	24.2
8 (30)	2	6.7	24	80.0	-	-	4	13.3
Total (74)	**5**	**6.8**	**54**	**73.0**	**-**	**-**	**15**	**20.3**
***TBBW*-Item 3-L**								
6 (11)	1	9.1	9	81.8	-	-	1	9.1
7 (33)	12	36.4	21	63.6	-	-	-	-
8 (30)	13	43.3	17	56.7	-	-	-	-
Total (74)	**26**	**35.1**	**47**	**63.5**	**-**	**-**	**1**	**1.4**
***TBBW*- Item 4-L**								
6 (11)	-	-	5	45.5	-	-	6	54.5
7 (33)	7	21.2	15	45.5	-	-	11	33.3
8 (30)	5	16.7	20	66.7	-	-	5	16.7
Total (74)	**12**	**16.2**	**40**	**54.1**	**-**	**-**	**22**	**29.7**
***TBBW*- Item 5-L**								
6 (11)	-	-	-	-	-	-	11	100.0
7 (33)	-	-	-	-	-	-	33	100.0
8 (30)	2	6.7	1	3.3	-	-	27	90.0
Total (74)	**2**	**2.7**	**1**	**1.4**	**-**	**-**	**71**	**95.9**

Note. (f) = frequency; *(TBBW)* = Teddy Bear's Birthday Wish; *(*L) = Literal Comprehension Test Item; (I) = Inferential Comprehension Test Item.

(continued)

TABLE 4.2 (continued)

Test-item/ Age in years (n)	**Keyed Answer-** **Justified** Test Score 1 Thinking Score 2-3		**Keyed Answer-** **Not Justified** Test Score 1 Thinking Score 0-1		**Unkeyed Answer-** **Justified** Test Score 0 Thinking Score 2-3		**Unkeyed Answer-** **Not Justified** Test Score 0 Thinking Score 0-1	
	f	%	*f*	%	*f*	%	*f*	%
***TBBW*- Item 6-I**								
6 (11)	-	-	3	27.3	-	-	8	72.7
7 (33)	1	3.0	8	24.2	-	-	24	72.7
8 (30)	3	10.0	12	40.0	-	-	15	50.0
Total (74)	**4**	**5.4**	**23**	**31.1**	**-**	**-**	**47**	**63.5**
***TBBW*- Item 7-I**								
6 (11)	-	-	2	18.2	-	-	9	81.8
7 (33)	-	-	3	9.1	-	-	30	90.9
8 (30)	2	6.7	4	13.3	-	-	24	80.0
Total (74)	**2**	**2.7**	**9**	**12.2**	**-**	**-**	**63**	**85.1**
***TBBW*-Item 8-I**								
6 (11)	-	-	-	-	-	-	11	100.0
7 (33)	-	-	8	24.2	-	-	25	75.8
8 (30)	-	-	5	16.7	-	-	25	83.3
Total (74)	**-**	**-**	**13**	**17.6**	**-**	**-**	**61**	**82.4**
***TBBW*- Item 9-I**								
6 (11)	-	-	-	-	-	-	11	100.0
7 (33)	1	3.0	3	9.1	-	-	29	87.9
8 (30)	5	16.7	5	16.7	-	-	20	66.7
Total (74)	**6**	**8.1**	**8**	**10.8**	**-**	**-**	**60**	**81.1**
***TBBW*- Item 10-I**								
6 (11)	1	9.1	5	45.5	-	-	5	45.5
7 (33)	8	24.2	16	48.5	-	-	9	27.3
8 (30)	6	20.0	16	53.3	-	-	8	26.7
Total (74)	**15**	**20.3**	**37**	**50.0**	**-**	**-**	**22**	**29.7**

Note. (f) = frequency; *(TBBW*) = Teddy Bear's Birthday Wish; *(*L) = Literal Comprehension Test Item; (I) = Inferential Comprehension Test Item.

The *TBBW* test-item frequency distribution for 6- to 8-year olds showed that the majority of children (ranging between 70% and 100%) provided keyed answers to five of the ten *TBBW* reading comprehension test items (literal comprehension Items 1-4; and inferential comprehension Item 10), two of which were answered correctly by most of the children in the sample (*TBBW* Items 1 and 3 had 100% and 99% keyed answers, respectively). The other half of the *TBBW* reading comprehension test items (literal comprehension Item 5; and inferential comprehension Items 6-9) had a higher proportion of unkeyed responses (ranging between 64% and 96%). Three of these test items (*TBBW* Items 7-9) were answered correctly by fewer than 20% of the sample and one test item (*TBBW* Item 5) was answered correctly by only three 8-year olds (4% of the total sample).

For the most part, test-item performance showed a consistent progression of more keyed answers obtained by each older age group on the various *RTRR* and *TBBW* test items (with some minor exceptions) such as *Dinosaur*-Item 4 and *TBBW*-Item 8. Moreover, 3-year olds had the fewest number of keyed responses and 5- to 6-year olds had the most keyed responses on the *RTRR* test items. There was little test performance variability between two or more age groups on several of the *RTRR* test items (i.e., *Dogs* Items 1, 2, and 4; *Teddy Bear* Item 3; *Dinosaur* Items 2 and 4) and on two items in particular, 6-year olds obtained fewer keyed responses than 5-year olds (i.e., *Teddies* Item 1 and *Dinosaur* Item 4). Similar to the *RTRR* results, the number of keyed answers on most *TBBW* test items increased with age (literal comprehension Items 2 and 4; and inferential comprehension Items 6, 7, 9, and 10). In one instance, younger children outperformed older children (i.e., 7-year olds obtained slightly more keyed responses than 8-year olds on *TBBW* inferential comprehension Item 8).

Overall, the consistent increases in the number of keyed responses by age confirmed that the *OR-RC* test items performed as expected by showing the developmental differences in reading comprehension performance. Similar to the *ON* results, the sample of 3- to 6-year olds obtained significantly more keyed responses for the *RTRR* literal comprehension test items as compared to the inferential comprehension test items (74% and 38%, respectively for the total sample). In addition, the 6- to 8-year olds performed considerably better on the *TBBW* literal comprehension questions (71% keyed responses for the total sample) than on the inferential comprehension questions (32% keyed responses for the total sample). These results are consistent with evidence from previous reading assessment research which has shown that literal comprehension questions are often easier and answered more accurately than inferential comprehension questions (Allen, 1998; Bowyer-Crane & Snowling, 2005; Brandao & Oakhill, 2005; Eason et al., 2012; Hua & Keenan, 2014; Keenan, 2014; McCormick, 1992; Paris & Paris, 2003; Pearson, Hansen, & Gordon, 1979).

A comparison of the *ON* and *OR-RC* test performance results also showed that children performed significantly better on the literal reading comprehension test items (*RTRR*-74% literal item keyed responses; *TBBW*-71% literal item keyed responses) than the literal listening comprehension test items (*ON*-46% literal item keyed responses). In contrast, the *ON* and *OR-RC* test performance results for the inferential listening and reading comprehension test items was relatively similar with the former slightly higher (*ON*-41%, *RTRR*-38%, and *TBBW*-32% inferential item keyed responses).

The protocol analyses of children's responses to the *OR-RC* test items revealed similar patterns in the quality of justifications (i.e., thinking score ratings) for keyed and unkeyed answers to those reported on the *ON* listening comprehension responses discussed previously. According to Table 4.2, significant discrepancy was found between the number of keyed and unkeyed reading comprehension test-item responses either justified or not for the different ages on the *RTRR* and *TBBW* measures.

OR-RC **Keyed Answers**

Table 4.2 revealed that more than half of the children in the respective age groups (3-6 years and 6-8 years) obtained keyed answers to the *OR-RC* test item responses (58% of the total *RTRR* test-item responses and 52% of the total *TBBW* test-item responses).

OR-RC keyed answers-justified. Table 4.2 showed that children between 3- and 6-years of age had a considerably lower frequency of keyed answers-justified (i.e., correct test-item answers with test score 1 and justifications with thinking scores 2 or 3) on sixteen of twenty *RTRR* test items (approximately 5% of the total possible *RTRR* test-item responses and 9% of the keyed *RTRR* test-item responses). Of these sixteen *RTRR* test items, *Teddies* Item 5 had the most keyed answers-justified overall (approximately 13% of the total responses for this test item) and *Dogs* Item 5 and *Teddy Bear* Item 5 each had only one keyed answer with justification. Of the keyed-justified *RTRR* responses (n = 98), only a small number of justifications (n = 10) had the highest thinking score (3) in which all relevant information sources and text cues were provided to support the keyed answers. The remaining keyed answers-justified (thinking score 2) were less well-developed and only partially accounted for the relevant information sources or text cues (print or visual) used to justify the response.

As children advanced in age, they produced more keyed responses with justification on the different *RTRR* test items. Specifically, 5-year olds generated more keyed answers-justified on the first set of *RTRR* test items (i.e., *Dogs* and *Teddy Bear*) and 6-year olds had the most keyed answers-justified on the second set of *RTRR* test items (i.e., *Teddies* and *Dinosaur*). There was minimal difference between the percentage of keyed answers with higher-level thinking scores (2 or 3) on the literal and inferential *RTRR* test items (6% and 4%, respectively). Nevertheless, the keyed-response pools for the four remaining *RTRR* test items (*Dogs* Items 1 and 4; *Dinosaur* Items 3 and 4) had thinking scores at the lower end of the scale (0 or 1). These test items and responses are discussed later.

In contrast to the *Oral Narrative* listening comprehension results, there was some evidence that even the youngest children in the sample demonstrated better quality thinking and reasoning in response to a select number of literal and inferential *RTRR* reading comprehension test items. Specifically, there were seven cases in which 3- and 4-year olds obtained keyed answers with thinking scores at the higher end of the scale (thinking score 2) and on two inferential comprehension test items (*Dogs* Item 5 and *Teddy Bear* Item 5), they were the only children in the sample to provide a keyed-justified answer and outperformed older children. These particular test-item responses showed that the young children clearly understood the reading comprehension questions and the main information sources. They discriminated and inferred from the most pertinent print and visual text cues and integrated their relevant background knowledge to construct keyed answers with a clear rationale. These children were precocious.

In fact, one 4-year old even mentioned the precise information source and text cues used to answer *RTRR* literal comprehension *Dinosaur* Item 2 (*How many baby dinosaurs are with the big dinosaur in the story?*). After providing the keyed answer (*two*), the child was asked, "How did you know that?" to which he responded, "Look at the picture…they were with the mommy dinosaur…and she had two babies." The relevant print cues pertaining to the test item (i.e., *One big dinosaur and two small dinosaurs*) did not specify that the three dinosaurs in the illustration were a mother dinosaur with her two babies. Consequently, the justification revealed that the child had in fact used the available relevant information sources. In particular, it showed that the child relied on the visual text cues as the primary information source to justify the test-item response and then used background knowledge to generate an inference about the relationship between the 'big' and 'small' dinosaurs.

To further illustrate, *RTRR Dogs* Item 5 is an inferential reading comprehension question which asked, *What is something that dogs do that is not in the story?* A keyed answer to this test item required children to consider the five activities that the dogs were doing in the story (i.e., run, sit, jump, play, and sleep) and then to use their background knowledge to provide a valid response to the question. Of the keyed-response pool (52% of the total responses to this test item), one 4-year old provided the only correct answer with proper justification. Specifically, the child indicated that the dogs were not doing "tricks" in the story. And when asked how he knew that, the 4-year old explained his thought-process, "Because jumping is not a trick…and all the other ones are not tricks." Presumably, the child considered the relevant text cues in order to verify his answer (i.e., "jumping…and all the other ones are not tricks") and consequently obtained the only keyed-justified response to the test item.

These findings indicate that it is possible for children as young as 3- and 4-years of age to demonstrate good reading comprehension, as well as complex thinking and reasoning skills. The keyed responses to the reading comprehension questions signified that these children integrated the story information well and their justifications revealed the underlying thought processes and information sources used to interpret the comprehension questions in relation to the explicit and implicit text cues. Although these higher-level justifications were rare, particularly at the youngest ages, the finding is important nonetheless because it illustrates the potential for very young children to use sophisticated thinking and reasoning in the reading context.

Children between 6- and 8-years old (n = 74) obtained slightly more keyed-justified answers with higher-level thinking scores (2 or 3) on the *TBBW* reading comprehension measure (12% of the total possible *TBBW* test-item responses and 23% of the keyed *TBBW* test-item responses) compared to the younger children on the *RTRR* reading comprehension measure (5% of total possible *RTRR* test-item responses). The number of keyed-justified answers on the *TBBW* measure also varied by test item (ranging between 3% and 35% for nine of ten *TBBW* questions) with greater frequency on the literal comprehension test items (16%) than the inferential comprehension test items (7%). Although there was a higher percentage of keyed-justified responses with thinking scores at the highest end of the scale (3) on the *TBBW* measure (5% of the keyed *TBBW* test-item responses), the older children were clearly still challenged by the task of justifying their reading comprehension test-item responses. In fact, when one child was asked how he knew the keyed answer to a *TBBW* question, the 7-year old commented, “That’s a stumper.”

The test-item responses and justifications varied between the *RTRR* and *TBBW* reading comprehension measures. The older children not only obtained more test-item response justifications with higher-level thinking scores, but their thought processes and ability to communicate was inherently more advanced and of higher quality. The children who performed well on the *TBBW* questions and provided appropriate justifications for their answers, responded from a more global, macro-level perspective by taking into account the complete story information and utilizing their relevant prior knowledge effectively.

Despite the obvious developmental age differences across the sample on the *TELL* reading comprehension measures, the testing materials (i.e., stories and test items) may also account for some of the performance discrepancy. The *TBBW* story was written for older children and was thus significantly longer and more complex and the corresponding comprehension questions tapped multiple information sources and text cues which required the older children to take more information into account (i.e., integrating a combination of print and visual text cues). Whereas, the *RTRR* stories were much more simplistic with limited text cues and the comprehension questions focused more narrowly on either the print or visual cues and did not require the integration of as much information. Overall, the quality of responses and justifications on the *TBBW* measure were more detailed and sophisticated likely because older children have developed better reasoning and communication skills and the longer story and comprehension questions, albeit more demanding, also offered more text cues for children to consider and reference in their response justifications.

Although there was a much greater proportion of keyed and unkeyed responses that were not justified, the fact that even a small number of keyed answers to most *RTRR* and *TBBW* test items were well-justified meant that at least some children in the sample were able to generate answers to the test questions that matched the keyed criteria and to demonstrate good thinking and reasoning in their test-item response justifications. In other words, the test questions and stories were interpreted with well-made connections among the different information sources, text cues, and prior knowledge. Ultimately, children who achieved keyed test-item responses and used good thinking and reasoning to justify their answers demonstrated greater reading and thinking ability and depth of understanding by accounting for the key story ideas and integrating relevant background knowledge in relation to the questions asked. However, knowing precisely which information sources children did and did not use to answer the comprehension questions offered important insight into why children's comprehension may have been lacking and in turn provided valuable diagnostic information for instruction.

OR-RC keyed answers-not justified. Table 4.2 showed that the category of keyed answers-not justified ranged between 1% and 92% for the *RTRR* and *TBBW* test items. Approximately 53% of the total *RTRR* test-item responses and 39% of the total *TBBW* test-item responses were keyed answers-not justified. Although performance varied by test item, the majority of keyed answers to most *RTRR* and *TBBW* test items were associated with lower-level thinking scores (0 or 1) (ranging between 9% - 92% on the *RTRR* measure and 1% - 78% on the *TBBW* measure). In some cases, all or most of the keyed responses lacked sufficient justification (e.g., *Dogs* Items 1, 4, and 5; *Teddy Bear* Item 5; *Dinosaur* Items 3 and 4; *TBBW* Item 8). Particular attention was given to these test-item response pools to investigate why so many keyed answers were not well-justified.

In order to achieve keyed answers to the various reading comprehension test items, children had to interpret the questions, attend to the pertinent text cues, integrate relevant prior knowledge, and make clear connections between the relevant information sources. Nevertheless, vast numbers of keyed responses with lower-level thinking scores on the reading comprehension test items (91% of the keyed *RTRR* test-item responses and 67% of the keyed *TBBW* test-item responses) revealed that most justifications were generally lacking because the reasoning used to justify the correct answers was not well-executed. Lower-level justifications were mainly inconsistent and incomplete with a number of different variations.

Children often provided rather cursory explanations which lacked sufficient detail to justify their reading comprehension test-item responses. Some justifications vaguely mentioned an information source or text cue but did not make the necessary connections between the test item and the response given. Others showed faulty reasoning by using prior knowledge to speculate and jump to conclusions and did not justify the answer. For instance, one 3-year old gave the keyed answer (*bird*) to *Teddy Bear* Item 3 (*What landed on teddy's head?*). When asked how she knew the answer, the child responded, 'He's trying to lay an egg on him'. Presumably, the child used conjecture to speculate about what the bird was doing on the teddy's head, as opposed to using the explicit visual and print text cues to confirm that a bird landed on the teddy's head in the story.

Similarly, other children ignored the relevant text cues and basically used their prior knowledge to affirm the test-item response. For example, one 4-year old responded to *Dogs* Item 3 (*How many dogs were sleeping?*) with the correct answer (*six*). And when asked how she knew that, the 4-year old responded, 'because I saw six dogs sleeping before' and did not mention the visual or print text cues pertaining to the question asked. In general, the lower-level justifications lacked appropriate and meaningful connections between the various information sources in order to verify the keyed answers.

In addition, several other characteristics of poor quality justifications for keyed test-item responses were evident. Children either just repeated the original answer to the test item without adding any new information to support how they knew the correct answer, or their oral communication lacked sufficient clarity and specificity to understand precisely what they meant, or they could not account for how they knew the keyed answer (i.e., 'I don't know' responses or no responses given). The weakest justifications were so general and vague that they offered little insight regarding the basis of children's test-item answers.

Another pattern of low-level justifications revealed an unwarranted sense of certainty and over-confidence. When some children were asked how they knew the correct answer to the reading comprehension question, their responses included statements such as the following:

- 'Because I know everything!'
- 'Because I'm the smartest kid in my class!'
- 'I don't know, I'm pretty smart.'
- 'cause it's so easy to remember.'
- 'Because I just do.'
- 'I don't know, I just know.'
- 'I was guessing.'
- 'Because I'm good at questions a lot, 'cause I ask my mom good questions.'

Such assertions begged the question about the kinds of messages that children have internalized and whether such self-praise might be detrimental to their ability to think critically. In any case, younger children (3-5 years) were more inclined to give these types of low-level justifications for keyed *RTRR* test-item responses than older children on the *TBBW* measure. Older children (6-8 years) tended to at least reference an information source for their keyed answers to the test questions (e.g., ‘it showed in the pictures’; ‘it said in the story’; or ‘the words told me’). Thus, low-level justifications are likely attributable to children’s developmental and maturity level; low level of reading proficiency; and, their responses seemed to serve as a coping mechanism to overcompensate for the difficulty of the task.

Regardless, children who provided vague and terse justifications had difficulty explaining their thinking and reasoning in response to test questions. As mentioned previously, it is not a common pedagogical practice to have children answer test questions and then to explain how they knew the answer. Thus, the task of reflecting on and expressing their thinking was likely unfamiliar and challenging for many children as evidenced by their unsolicited reactions to the meta-level questioning (i.e., ‘How do you know that?’) in which they openly admitted to having difficulty. Nevertheless, many children did not seem to fully understand what was being asked or precisely what constituted a consistent and complete test-item response justification which would not have been a reasonable expectation of the young children.

It is not surprising that young children experienced difficulty with the meta-level questioning which required them to simultaneously consider the test question, their response to the test item, as well as the information sources, text cues, and background knowledge used as the basis of their response. The task is cognitively-demanding and involves complex thinking and reasoning and it may, in fact, be developmentally beyond the scope of some young children’s abilities.

***OR-RC* Unkeyed Answers**

Table 4.2 showed that 42% of the total *RTRR* test-item responses and 48% of the total *TBBW* test-item responses were unkeyed answers. Although the keyed test-item criteria specified exactly what constituted acceptable answers for the respective reading comprehension questions, unkeyed responses were not one and the same. That is, the quality and nature of unkeyed responses for each test item varied significantly. The unkeyed *OR-RC* test-item responses were primarily distributed amongst the four salient response patterns described previously in the *Oral Narrative* test-item performance analysis with one main difference; namely, the information sources that children used for the respective test items. For the listening comprehension measures, oral and visual story cues were the primary information sources, compared to the oral, print, and visual story cues for the *RTRR* reading comprehension measures and print and visual story cues for the *TBBW* reading comprehension measures. Close examination of the unkeyed reading comprehension test-item responses offered many important insights including: how test items were interpreted, response and error patterns, information sources used, main sources of difficulty and inherent gaps in understanding, quality of reasoning, and finally, whether the items performed well as measures of reading comprehension. Response patterns for the *RTRR* test items with the most unkeyed answers are highlighted next.

An unanticipated pattern of responses to *RTRR Dogs* Item 1 (*What are two things that dogs do?*) emerged in the unkeyed answers (49% of the total sample). The literal reading comprehension question required children to use the explicit print and/or visual cues to name two of the five activities that the dogs were doing in the story (i.e., running, sitting, jumping, playing, or sleeping). Approximately one-half of the 3- to 5-year olds gave an unkeyed answer and two-thirds of these children ignored the explicit story cues altogether and invoked their background knowledge to answer the question (e.g., 'dogs pee and poo'). These types of unkeyed responses were mainly attributed to the generic nature of the question. Consequently, a minor revision of the question would help to ensure the clarity and integrity of the test item to measure literal reading comprehension as intended (i.e., What are two things *in the story* that the dogs do?) (revised on published test).

In addition, *RTRR Teddies* Item 4 (*How did you know from the pictures that some teddies were sad?*) produced low test-item performance (74% total unkeyed responses) because many children were remiss in their interpretation of the question in spite of the obvious test-item cues (i.e., the answer was depicted in the pictures). Specifically, some children ignored both the item and explicit picture cues and were evidently too print-focused (e.g., 'it said one happy and two sad bears' or 'cause it told us that from the words'), while others misconstrued the premise of the question and used their prior knowledge to speculate on the reasons the teddies were sad (e.g., 'because the teddy bears missed their mommy' or 'cause nobody would play with them'). These examples underscored that at times the source of difficulty emanated from the interpretation or comprehension of the question itself, or from an overdependence on background knowledge.

The vast number of unkeyed responses (76% of the total sample) to *RTRR Teddy Bear* Item 4 (*Why did teddy look at the bird and the bee?*) showed that some children jumped to conclusions and made unsubstantiated claims despite the visual and print cues provided in the story. For example, one 5-year old responded, ‘cause he was worried…the bird was going to peck on him or the bee was going to sting him’. There was no evidence in the story to suggest that the teddy was concerned about potential encounters with the bird or the bee. In this instance, the child conjured a more elaborate story interpretation and did not construct an evidence-based inference by integrating and making meaningful connections between the relevant implicit and explicit text cues and the question asked (i.e., the teddy looked at the bird and the bee because they landed on him or because he was waving good-bye to them as they flew away). It is not uncommon for young children to use their imagination in order to fill in gaps and to make up what is not there, particularly during story reading. This tendency is driven mainly by children’s desire to tell a story. However, as shown in the example, the inclination to use irrelevant prior experiential knowledge to elaborate beyond the text can be misleading and distort children’s comprehension of the actual story. Cain and Oakhill (1999) suggested that one source of difficulty that children encounter in making inferences is not having a clear understanding of “when it is permissible to bring in general knowledge from outside of the text in order to make sense of a passage” (p. 491). Ultimately, reading comprehension requires children to distinguish between relevant and irrelevant information sources and to know when and how to use the pertinent cues.

Finally, *RTRR Dinosaur* Item 4 had the most significant disparity between keyed and unkeyed responses (9% and 91% of the total sample, respectively) on the *RTRR* measures. Of the 111 children between 3- and 6-years of age who were asked the test question, only a small number of 5- and 6-year olds (n = 10) achieved a keyed answer. The item is an inferential comprehension question asking children to speculate about what might happen next in the story and how they knew that. Almost half of the children (48%) provided an 'I don't know' response. The remaining unkeyed answers were either implausible, decontextualized responses which were inconsistent with the story content, they were overly print-dependent (i.e., citing something that had already occurred in the text), or no response was given. Since all 3- and 4-year olds, and most 5- and 6-year olds, did not perform well on this inferential reading comprehension test item, it seemed apparent that the skill of predicting was particularly challenging for younger children in general. In any case, the range of unkeyed responses to the different *RTRR* reading comprehension test items highlighted that some test items were inherently more difficult than others for children between 3- and 6-years of age and ultimately, unkeyed answers revealed deficiencies and gaps in their reading comprehension.

The *TBBW* response pools (Items 5-9) with disproportionately greater frequency of unkeyed responses (more than 60%) were examined to determine why the items were difficult for most children in the sample and whether the questions performed well as measures of reading comprehension. In the first instance, test performance (96% total unkeyed responses) on *TBBW* literal reading comprehension Item 5 (*What was the teddy's name?*) was an anomaly since the test question itself appeared simple and straight-forward and the only acceptable answer (*Theodore*) was explicitly-stated in the print. In spite of that, two-thirds of the unkeyed responses were identical ('Teddy') referring to the generic reference for the main character which was used repeatedly in the story, as opposed to the proper name (*Theodore*) which was mentioned only once. To illustrate children's underlying thought process, an 8-year old was asked how he knew that the teddy's name was 'Teddy' to which he replied, 'cause it came up mostly in the story.' Ultimately, the test item effectively discriminated between those children who demonstrated more discerning reading comprehension and were able to identify the explicit print cue referring to the teddy's proper name and those who did not. Teddy is the generic name for any and all teddy bears and thus was not sufficient in this case.

In contrast, *TBBW* Items 6-9 were inferential reading comprehension questions which demanded more extensive reasoning since the keyed answers were not explicit in either the print or visual cues. In general, these inferential reading comprehension test items required children to consider the story content in conjunction with their prior knowledge to generate appropriate inferences for a keyed response. The most common unkeyed response patterns showed that children either relied on only their prior experiential knowledge to respond to the test items and did not take the relevant story information into account (i.e., overly-focused on their general knowledge), or alternatively they were unable to move beyond the text information to activate relevant prior knowledge pertaining to the test items (i.e., too text-dependent). Unkeyed responses were often discrete statements and generalizations not directly linked to the relevant textual information. For example, when asked *TBBW* Item 6 (*Why do you think it is hard to find a rocking chair for a teddy bear?*), some children made fairly overstated claims such as, ‘usually rocking chairs are for older people’ or ‘most places in the world don’t have rocking chairs’. Similarly, several children provided vague answers to *TBBW* Item 7 (*Where might teddy have looked first for a teddy rocking chair?*) ‘in his house’ or ‘in a store’ which did not take the complete story information into account. And on *TBBW* Item 8 (*Why did teddy think that Papa Bear could help?*), children used their prior knowledge for general responses such as: ‘because he’s smart’, or ‘because he’s a grown-up’, or ‘because dads are helpful’. These unkeyed test-item responses indicated that children did not construct the intended inferences and their answers were often too general and disassociated from the text. In this respect, children provided a more immediate, local or micro-level response to the questions relying on their prior knowledge and failing to consider the questions in reference to the story and to take into account the relevant text cues.

Finally, *TBBW* inferential comprehension Item 9 (*Why did the teddy bear despair?*) was answered correctly by less than one-fifth of the children between 6- and 8-years of age (19% total keyed answers). Of the unkeyed responses, most children (approximately 60%) either gave an ‘I don’t know’ response or they mentioned that they did not know the meaning of the term ‘despair’ and offered no further response. Thus, the main source of difficulty was attributed to the unfamiliar vocabulary in the test item. Suffice it to say that only those children who understood the concept were able to address the test question directly.

Ultimately, the *RTRR* and *TBBW* test-item response analyses revealed that unkeyed responses to the various test items were indicative of weaknesses in reading comprehension rather than problems or issues related to the test items or stories on the different measures. Overall, the test items effectively discriminated between those children who were able to interpret what the question was asking, discriminate and integrate the relevant implicit and explicit visual and print cues with their prior knowledge, and make the necessary inferences from those who did not. The extensive review and examination of test-item performance and responses showed strong support that the *TELL OR-RC* stories and test items generally performed well as measures of reading comprehension. The next section examines children's thinking performance and justifications for their reading comprehension test-item responses.

OR-RC unkeyed answers-justified. As shown in Table 4.2, there were no cases where unkeyed answers were justified. The unkeyed responses for all thirty *RTRR* and *TBBW* test-items were exclusively associated with low-level thinking scores (0 or 1). In other words, the justifications for unkeyed responses were essentially inconsistent and incomplete with the available relevant information sources

OR-RC unkeyed answers-not justified. Table 4.2 showed that the frequency of unkeyed *RTRR* and *TBBW* test-item responses varied by test item and were exclusively associated with low-level thinking scores (0 or 1). Overall, approximately 42% of the total *RTRR* test-item responses (between 6% and 91% on the individual test items) and 49% of the total *TBBW* test-item responses (between 1% and 96% on the individual test items) were unkeyed answers that were not justified.

In general, unkeyed test-item responses were comprised of either alternative answers which did not correspond to the keyed test-item criteria or 'I don't know' responses. Although some younger children provided alternative, unkeyed answers to the different *RTRR* reading comprehension test items, they gave 'I don't know' responses more frequently. Whereas, most of the older children gave alternative, unkeyed answers to the *TBBW* test items as opposed to 'I don't know' responses. Children were asked to explain and justify their alternative, unkeyed test-item responses.

Most unkeyed reading comprehension test-item response justifications were not well supported by relevant evidence from the story or prior knowledge. Similar to the keyed test-item responses, younger children also gave more 'I don't know' responses when asked to explain their answers to the *RTRR* test items than older children on the *TBBW* test items. Many justifications for unkeyed test-item responses were based on faulty reasoning. For example, when an 8-year old was asked *TBBW* Item 6 (*Why do you think it is hard to find a rocking chair for a teddy bear?*), the child gave an unkeyed answer, 'because they don't make them.' When asked how he knew that, the 8-year old replied, 'Cause I've never seen one before…if they don't make them…you can't find them so easily.' Although the child explained his thought-process as if it were obvious, the reasoning was spurious nonetheless. Other unkeyed responses with low-level thinking scores merely mentioned the information source (i.e., ‘because it told in the picture’) but did not specify the relevant text cues pertaining to the test item. Still others did not offer any reasoning at all (i.e., ‘I don’t know’ or ‘I just know’ responses).

Overall, children did not generate any new, alternative test-item answers with sufficient evidence to warrant reconsideration of the reading comprehension test items or the keyed response criteria. All incorrect reading comprehension test-item responses had low-level thinking scores. Consequently, poor reading comprehension test performance was associated with poor thinking and reasoning. Accordingly, these findings supplement the validation evidence of the *TELL* reading comprehension measures for assessing and identifying potential difficulties and challenges in reading comprehension.

Summary of *TELL Oral Narrative* (ON) and *Oral Reading-Reading Comprehension* (OR-RC) Assessment Results

The *ON* and *OR-RC* results showed that relatively few children in the sample performed well on the comprehension test items and those that did not perform well showed poor listening and/or reading comprehension. Many children in this study had difficulty constructing meaning from the explicit and implicit text cues and integrating relevant story information with their relevant background knowledge to answer the listening and reading comprehension questions about the stories.

Overall, children performed better on the two reading comprehension subsections (56% total keyed responses) than the *ON* listening comprehension subsection (42% total keyed responses). However, age comparisons showed that 3-, 4-, and 5-year olds performed significantly better on the *RTRR* reading comprehension (39%, 56%, and 68% total keyed responses, respectively) than the *ON* listening comprehension test items (18%, 26%, and 39% total keyed responses, respectively). In contrast, 6-, 7-, and 8-year olds had slightly more keyed answers on the *ON* listening comprehension test items (49%, 58%, and 60% total keyed responses, respectively) than the *TBBW* reading comprehension test items (41%, 49%, and 57% total keyed responses, respectively). These results seem to suggest that the children at the younger ages found the *RTRR* reading comprehension task and test items to be easier than the *ON* listening comprehension test items. Whereas, there was little difference between listening and reading comprehension test performance for the older ages.

Perhaps the best way to put children's test and thinking performance into perspective is by taking into account the collective totals from the frequency distribution of all keyed and unkeyed answers which were either justified or not on the combined *ON* and *OR-RC* measures. Collectively, the majority of keyed answers and almost all unkeyed answers to the *TELL* listening and reading comprehension test items were not justified with good thinking and reasoning according to the rating scale used in this study (thinking scores 0 or 1). Approximately only one-tenth of children's keyed responses to the listening and reading comprehension questions were well justified with good thinking and reasoning and assigned higher-level thinking scores (2 or 3). Although these results showed that some children in the sample were capable of giving better quality justifications for their answers, they seldom did so in the context of these assessments. Moreover, age comparisons revealed sizeable differences in the frequency of well-justified keyed answers between the younger and older age groups.

Of the keyed answers for the 3-, 4-, and 5-year olds on the combined listening and reading comprehension questions (30%, 43%, and 56%, respectively), only a minimal number had higher-level thinking scores (between 1% and 7%). Although the children in these younger age groups knew which information to use to answer certain comprehension questions correctly, they did not grasp what was required for quality response justifications. In other words, they did not know how to use the appropriate information sources to support their answers. Nonetheless, the most compelling finding was the comparable test and thinking performance results among the three older age groups on the combined *ON* and *OR-RC* test items. Of the keyed responses obtained by the 6-, 7-, and 8-year olds on the combined listening and reading comprehension test items (57%, 54%, and 59%, respectively), nearly one-fifth were justified and assigned higher-level thinking scores (18%, 17%, and 19%, respectively). Although these patterns of performance show that the older children in this sample were generally more competent at justifying their answers to questions than younger children, there was little variability or evidence of change among the three older age groups when the results from both comprehension question assessments were taken into account. Altogether, the listening and reading comprehension results showed that children in the sample not only performed better overall on the *OR-RC* subsections (56% total keyed responses) than the *ON* subsections (42% total keyed responses) but they also provided slightly more justifications for their keyed answers on the *OR-RC* test items (7% keyed-justified answers) than the *ON* test items (4% keyed-justified answers).

Furthermore, the evidence from this exploratory study showed that many children in the sample found it difficult to answer the meta-level questions requiring them to think-aloud and to explain the thinking and reasoning underlying their comprehension test-item responses. In some instances, they were visibly perplexed by the think-aloud probes as was the case of one 5-year old when asked to explain how he knew the answers to the comprehension test items, he repeatedly insisted, ‘because I’m the smartest kid in my class!’ Other children merely rolled their eyes or sighed aloud; while others tried to redirect attention elsewhere by commenting on or asking a question about something observed in the environment or simply requesting to finish the assessment and return to their classroom. And, on a few occasions, older children commented on the difficulty of the task. For instance, one 7-year old placed his finger on his temple and remarked, “That’s a *hard* question!” when asked the think-aloud probe. Most children, however, showed no concern with either the comprehension questions or the think-aloud probes used in this study.

The challenges that children encountered with thinking aloud and justifying their answers to questions on the *TELL* comprehension assessments were not entirely unexpected because very few think-aloud studies have been undertaken with children younger than 7- or 8-years of age in reading and assessment contexts (Allen, 1998; Brandao & Oakhill, 2005; Laing & Kamhi, 2002; Lynch & van den Broek, 2007; Paris & Paris, 2003; Paris & van Kraayenoord, 1996; Tompkins, Guo, & Justice, 2013) due to their perceived limited emergent cognitive, language, and reading abilities. In fact, the literature review for the current research identified only four studies which used think-aloud procedures during comprehension assessment with children between age 4- and 8-years inclusive (e.g., Brandao and Oakhill, 2005; Paris and Paris, 2003; Paris & van Kraayenoord, 1996; Tompkins, Guo, and Justice, 2013). Moreover, the current research is one of only a few studies in the early language and literacy literature to collect verbal reports from very young children (e.g., Farrington-Flint & Wood, 2007; Lynch & van den Broek, 2007; Tompkins, Guo, & Justice, 2013). Furthermore, it is the first known study to ask children as young as three years of age to explain their answers to questions in a testing context. Nevertheless, the results from this study's detailed response analyses offered some informative diagnostic possibilities for analyzing children's test-item responses and justifications for those responses. Future research and pedagogical possibilities from these analyses are detailed in the final chapter.

CHAPTER 5: SUMMARY AND CONCLUSIONS

This chapter concludes the study reported herein starting with the summary of results, followed by the limitations; conclusions and implications; and contributions of the study. It closes with recommendations for future research and practice.

Summary of Results

This exploratory validation study examined whether the *Test of Early Language and Literacy (TELL)* subsections on listening and reading comprehension from *Oral Narrative (ON)* and *Oral Reading-Reading Comprehension (OR-RC)* sections measure the skills and processes fundamental to listening and reading comprehension ability and are suitable for their intended purpose and use. The *TELL ON* and *OR-RC* assessments focus on children's story comprehension, specifically, their ability to answer open-ended literal and inferential comprehension questions after listening to and reading stories.

In the current research, children (n = 174) from 3- to 8-years of age completed the *TELL ON* and *OR-RC* questions with the inclusion of think-aloud probes asking them to explain their reasoning for each test-item response. The relationship between children's test responses to the comprehension questions and what they reported thinking and reasoning as they responded to each test item on the listening and reading comprehension assessments was examined thoroughly.

Specifically, the main results from the study revealed variability in children's test performance on the relevant subsections from the *TELL ON* and *OR-RC* comprehension assessments. As expected, the older children in the sample generally performed better overall on the comprehension test items than did the younger children. Furthermore, the *ON* and *OR-RC* test-item performance distribution also revealed that children across the age range generally performed better on the literal comprehension questions than on the inferential comprehension questions for both the listening and reading comprehension assessments.

The protocol analyses of children's test and thinking performance on the *TELL ON* and *OR-RC* test items identified four main response patterns: (1) keyed answers-justified, (2) keyed answers-not justified, (3) unkeyed answers-justified, and (4) unkeyed answers-not justified. Each response category highlighted key differences in children's comprehension skills and their ability to think and reason. The patterns confirmed that the assessments differentiated proficiency in story comprehension (listening, reading) with a clear distinction between good and poor levels of comprehension. In particular, children who obtained keyed answers and used good thinking and reasoning to justify their answers to the listening and reading comprehension test items demonstrated higher performance on story comprehension than did children who gave unkeyed answers with poor thinking and reasoning on the different measures. Moreover, the category of keyed responses that were not justified provided greater insight about the quality of children's thinking and reasoning and the difficulties they encountered on the probes. The few instances of incorrect answers with better quality justifications found in the protocol analysis confirmed that it was rarely ever the case that children substantiated alternative responses to the questions with relevant information. The analyses of the four response categories were the basis for important conclusions about the quality of the *TELL* comprehension assessments and about children’s justifications for their responses to the comprehension questions.

Limitations of the Study

The use of protocol analysis in this validation study was deemed the best approach for analyzing the cognitive processes underlying children’s responses to test items based on the assumption that verbal reports of thinking provide the most direct evidence possible for why children answered the test items the way they did. However, the results of this study must be considered in light of several limitations.

First, the results and generalizations of this study are limited to the specifications of the particular testing instrument used here and the testing context. That is, the results were interpreted within the context of a single measure of listening and reading comprehension and do not necessarily generalize to other comprehension assessment contexts. Thus, it is important to keep in mind that the inferences drawn based on children’s responses to the comprehension questions and think-aloud probes about their comprehension performance, underlying cognitive processes, and the information sources they used to answer the comprehension test items were by no means exhaustive.

Second, the methodology of the study also posed some challenges which had the potential to influence children’s performance on the assessments. One challenge was striking a balance between administering the *TELL* comprehension assessments according to the standard test protocol while at the same time introducing the think-aloud probes as seamlessly as possible after children answered each test item. For verbal report data to be relevant and useful for validation purposes, normal test administration procedures had to be strictly followed and the verbal reporting procedures needed to be the least intrusive and disruptive to test performance as possible (Norris, 1991). Although every effort was made to ensure the integrity of the test administration and to minimize the impact of think-alouds on children's test performance, the inclusion of the probe questions may have altered children's comprehension processes and how they performed on the comprehension test items.

Although there was consistency in the think-aloud probes used during the administration of the *TELL* comprehension assessments in the current study, the procedure was not as fluid and responsive as it could have been. In other words, the results were constrained by the types of think-aloud probes that were asked and the examiner's ability to adapt and be responsive to how children answered each comprehension test item and corresponding think-aloud probe. The extensive review of the assessment transcripts revealed a few instances of oversights and missed opportunities in which questions of clarification would have been prudent and more informative about children's comprehension processes.

Verbal report studies with very young children are rare. In fact, this study was one of the first to use verbal reports in a testing context with children as young as three years of age. Certainly, the personality traits of young children presented some interesting challenges for data collection particularly when children were easily distracted, highly active, or impulsive. Although the young children's self-expression appeared to be visceral, their spontaneous responses were not always relevant to the testing context. Furthermore, the quality of children's verbal reports often depended on their comfort level, willingness to respond to questions, and ability to communicate their thoughts. Thus, the conclusions from verbal reports about children's abilities are constrained by the information that they were willing and able to report which may not represent all that they know. The current study was exploratory and children were engaged to discuss the topic to hand prior to starting the test items. In retrospect, it may have been beneficial to introduce the think-aloud procedures and probes to children prior to the assessment in order for them to become familiar with the questioning and to practice responding to these types of meta-level questions.

The fact that many children in the study had difficulty justifying their answers to the comprehension questions may be due, in part, to a lack of: understanding of what was being asked or how to respond to the probes; communication skills or metacognitive ability to defend their answers; experience with the type of meta-level questioning; and/or instruction on how to explain their thinking. It is reasonable to assume that the preschool children were most likely unfamiliar and lacked experience with the task cognitive demands, and verbal think-aloud with follow-up probe questions. The ability to explain and justify responses to questions necessitates a certain level of cognitive and communicative development coupled with proficiency and the ability and language to express what they are thinking. Since most listening and reading comprehension tasks do not typically ask children to explain their thinking, the inclusion of probes during the assessments in this study were likely a novel experience for many children in the sample and may have introduced a skill that has yet to be developed.

Finally, most verbal report studies are limited by sample size which further constrains their findings and generalizations. However, considering how time- and labour-intensive it was to collect the assessment and verbal report data, this study included a larger sample size than any other documented verbal report study of these age groups. Nevertheless, the sample size of 174 participants with 26 to 33 children in each age grouping imposed certain limitations on the amount of data that was gathered and used as the basis for analyses and interpretations about children's test performance and ultimately, as validation evidence for test score interpretations. The final section describes the most significant contributions of the study.

Conclusions and Implications

The analyses of children's performance on the assessments provided important insights into children's comprehension ability and the quality of the *TELL* comprehension measures. It is widely known that comprehension develops from an early age. The consistent increases in the number of keyed answers across the age groups on each *TELL* comprehension assessment examined herein aligned with the expectation that comprehension ability typically improves with age, particularly in the early stages of development. However, the fact that some younger children in the sample outperformed older children on the test items suggests that the development of listening and reading comprehension ability is not a direct consequence of age. Since comprehension is fundamentally an unconstrained skill that develops progressively and continuously over time and does not follow a predictable developmental trajectory (Paris, 2005), some children will presumably demonstrate greater proficiency sooner than others. Consequently, one of the most significant conclusions from the test performance results in this study is the importance of neither underestimating nor overestimating what children of different ages can and cannot do. Given that there is uneven development in children's comprehension skills is all the more reason to ensure that the assessment tools accurately identify developmental differences in comprehension ability. The evidence from the test performance results in this study showed that the subsections from the *TELL ON* and *OR-RC* section assessments met that expectation.

The analysis of the four response patterns highlighted the differences in children's performance on the test items, the information sources they used to answer the comprehension questions, the quality of their thinking and reasoning, and whether the test items measured the construct of comprehension. The first response pattern (i.e., keyed answers-justified) showed that when children performed well on the individual test items *and* provided quality justifications for their answers based on sound reasoning (i.e., primarily children in the older age groups), they demonstrated superior listening and/or reading comprehension, as well as good thinking and reasoning. The children knew which information to use to answer the comprehension question(s) and how to use the available information sources to justify their test-item responses (i.e., clear explanations for their answers and specification of the information sources they relied on). Moreover, these children demonstrated competence in the particular skills and processes associated with literal and inferential comprehension including the ability: to discern and interpret explicit and implied text cues (oral, print, and visual); to activate prior knowledge related to the text information; to understand intratextual relations between and among ideas presented in the text; and to draw evidence-based inferences and conclusions from all relevant information sources. The keyed and justified test-item responses found in this study exemplified the qualities of good comprehension and confirmed, in these instances, children who performed well on the test items also showed good thinking and reasoning.

The second response pattern included the relatively large number of keyed answers to each test item which were not well justified. Although the children used accurate information sources to answer the comprehension questions, they did not have a clear sense of how to explain their responses nor have awareness of which information sources they used to answer the test items. In these instances, children's answers to the test items were not supported by good thinking and reasoning. The results showed that children from all age groups did not consistently provide complete and consistent justification for their answers to the comprehension questions. Furthermore, three- and four-year olds, in particular, seemed to have the most difficulty and did not know how to respond to the probes.

In contrast, the third response pattern (i.e., unkeyed answers-justified) revealed that children rarely provided a justification for an unkeyed answer. This pattern spoke to the quality of the test items and confirmed that when children did not perform well on the test items, they demonstrated apparent weaknesses in story comprehension because they could not support their alternative answers with sufficient justification.

Finally, the fourth response pattern (i.e., unkeyed answers-not justified) showed that children who performed poorly on the test items typically did not use good thinking and reasoning. These children did not use the proper information sources to answer the comprehension questions nor to explain their answers. In addition, they did not demonstrate understanding of the explicit and implicit text cues in their responses to the test items and did not integrate the text information with their prior knowledge to construct meaning in order to make sense of the text.

The analysis of the unkeyed test-item responses in this study made it possible to ascertain reasonable inferences about the breakdown in comprehension and to identify potential sources of difficulty, namely, which skills were lacking and contributed to poor performance. Children's weaknesses in comprehension often stemmed from their inability to discern which text cues were pertinent; how to make meaningful intratextual connections; and subsequently, how and when to use their relevant prior knowledge to interpret the information in the text. Unkeyed test-item responses also revealed that children often misinterpreted the test item or the story information; had difficulty identifying the explicit text cues and understanding the implied meaning from the information in the text; perseverated on insignificant parts of the text or irrelevant story information; and over-relied upon either the text information or their prior knowledge. Overall, these characteristics were indicative of children's lack of comprehension and understanding of the stories. Thus, the fact that nearly all unkeyed responses to the test items were not justified confirmed that children who performed poorly on the *TELL ON* and *OR-RC* assessments also showed poor thinking and reasoning; yet another source of evidence for the validity of the test measures.

Contributions

The current study makes three important contributions: (1) highlights the attributes of the *TELL ON* and *OR-RC* as valid measures of comprehension ability, (2) confirms that think-aloud is an important source of validation evidence for test score interpretations and that it is feasible to use think-aloud procedures with very young children, and (3) draws attention to the diagnostic potential of the children's think-aloud responses for instruction.

Collectively, the analyses of children's test and thinking performance provided compelling evidence for the validation of the *TELL ON* and *OR-RC* test score interpretations for distinguishing between good and poor comprehension in young children and that the two subsections are developmentally appropriate for assessing young children's understanding of explicit and implicit story information in both oral narrative and reading contexts. Moreover, the thorough and systematic examination of children's responses to the *TELL* comprehension test questions and think-aloud probes revealed the diagnostic attributes of the measures for identifying strengths and weaknesses in comprehension. In particular, the fact that children in this study generally performed better on the literal than the inferential comprehension questions suggested that children found it easier to construct meaning from the explicit text and picture cues as opposed to integrating information from different sources (i.e., explicit and implicit text and picture cues, and prior knowledge). Not only are these results consistent with extant research discussed in the previous chapter, but they also provide evidence that the combination of different types of questions included on the *TELL* comprehension assessments delineated competence and weakness in the skills and processes associated with literal and inferential comprehension.

In the current study, the test measures identified a number of different characteristics of poor comprehension that manifested at the early stages of comprehension development. The primary goal of this type of diagnostic assessment is to be able to identify and address specific areas of difficulty early on in order to avoid compounding issues in later language and literacy development. The key to detecting specific comprehension difficulties relies on having adequate assessments that can properly diagnose which component skills and processes are lacking in order to determine precisely what should be the primary focus of instruction. From a practical perspective, the children who demonstrated weaknesses in comprehension in this study would likely benefit from early comprehension instruction which explicitly teaches about the information sources that are available and necessary for understanding story information with exemplars demonstrating how to reason, make connections, draw inferences, and integrate the relevant information sources (i.e., relevant text cues and prior knowledge). Ultimately, the different analyses from the current investigation found that the design features of the *TELL ON* and *OR-RC* question assessments (including texts, test items, and scoring criteria) are well-developed for the purposes of discriminating between good and poor listening and reading comprehension ability and for identifying specific areas of strength and difficulty.

Another important contribution of the present study is that it provided a new perspective on using think-aloud for test validation purposes. Not only was this research one of the first studies to use think-aloud to evaluate the quality of a listening and reading comprehension assessment and the validity of test score interpretations, but is one of the only known studies to ask children as young as three-years of age to explain their answers to questions. In regards to assessment, think-aloud is indeed one of the only ways to determine whether test items are actually tapping the skills and processes that they were designed to measure and to find out why and how children answered the comprehension questions the way they did. The think-aloud procedures used in this study proved to be complementary to the standard comprehension assessment for investigating children's comprehension skills and, specifically, the source of their weaknesses in comprehension.

The study also provided an informed perspective on the feasibility of using meta-level questions with young children which ask them to explain their answers to questions. The results of the study showed that some children as young as five demonstrated basic metacognitive skills required for this type of questioning but the most informative responses came from children between 6-

and 8-years of age. Ultimately, the study affirmed that think-aloud procedures can be used with young children in different contexts to provide valuable insights into the cognitive processes underlying their task performance.

In conclusion, validation studies such as this are extremely important for the development of quality assessments. Think-aloud provides an effective means for analyzing how children respond to test items and the cognitive processes underlying their task performance in order to determine whether the assessments are measuring the intended construct. The current research has provided compelling supportive evidence that the *TELL ON* and *OR-RC* assessments measure the construct of listening and reading comprehension and can be used to identify children's strengths and weaknesses in literal and inferential comprehension. Ultimately, the *TELL ON* and *OR-RC* measures are an important part of a comprehensive, systematic, diagnostic assessment of early language and literacy skills aimed at screening and identifying comprehension difficulties in young children.

Recommendations for Future Research and Practice

The results from this study's response analyses offered some intriguing diagnostic possibilities for analyzing children's comprehension test-item responses and explanations and warrant further investigation and application in the development and understanding of listening and reading comprehension at these young ages. More research is needed to examine why children have difficulty providing justifications for comprehension question responses. It is important to investigate further the patterns and types of information sources that children at different ages rely upon to justify their responses to comprehension questions and how this reliance influences their comprehension performance. It would be informative to examine young children's justifications for comprehension questions beyond the confines of the standardized testing protocol to probe further what young children think and how they reason in a less restricted context. Future research on alternative methods for tapping children's thinking and reasoning and for categorizing the patterns of their response explanations to comprehension questions is a logical next step in understanding children's emergent comprehension development.

The present study has practical implications for using think-aloud with young children. Given that many children in the study had difficulty responding to the think-aloud probes, it seems that young children need to be taught how to use relevant information sources to explain their answers to questions in many different contexts in order to acquire the kinds of skills that promote deeper and more critical reading. In order for children to improve their ability to think metacognitively, it would be beneficial for them to have opportunities to discuss meta-level questions such as, 'how do you know that?' or 'what made you think that?' or 'why do you think so?' in a variety of shared reading contexts to encourage them to examine and explain their thinking and reasoning. These types of questions may serve as the catalyst for developing critical thinking skills and the ability to judge when and how to use particular information sources for constructing meaning from text. As is the case with reading comprehension, the skill of justification will likely take considerable time to develop and will need to be fostered from an early age in order for children to expand their metacognitive ability. Research on the merits of think-alouds on children's language and literacy proficiency may well point to a significant innovation in how we teach comprehension development.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Afflerbach. P. (1990). The influence of prior knowledge on expert readers' main idea construction strategies. *Reading Research Quarterly, 25,* 31-46.

Afflerbach, P. (2000). Verbal reports and protocol analysis. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 163-178), Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Afflerbach, P., & Johnston, P. (1984). Research methodology: On the use of verbal response in reading research. *Journal of Reading Behavior, 16,* 307-322.

Alavi, S. (2005). On the adequacy of verbal protocols in examining an underlying construct of a test. *Studies in Educational Evaluation, 31*, 1-26.

Allen, C. S. (1998). *Metacognition, reading, and test taking of third graders.* Available from ProQuest Dissertations and Theses database. (UMI Number 9907834)

Alvermann, D., & Ratekin, N. (1982). Metacognitive knowledge about reading proficiency: Its relation to study strategies and task demands. *Journal of Literacy Research, 14*(3), 231-241.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anastasia, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan Publishing.

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal, 75*(4), 460-472.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*(1), 41-66.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 255-292). White Plains, NY: Longman.

Baldo, A. (2008, March). *Comprehension processes of proficient portugese readers*. Paper presented at the Thirtheenth Annual Graduate Students' Conference. New York, NY.

Bauer, P. J. (1996). What do infants recall of their lives? Memory for specific events by one to two-year-olds. *American Psychologist, 51*(1), 29-41.

Bauer, P. J. (1997). Development of memory in early childhood. In N. Cowan & C. Hulme (Eds.), *The development of memory in childhood* (pp. 83-111). East Sussex, UK: Psychology Press.

Bishop, D. V. M., & Edmundson, A. (1987). Language-impaired 4-year-olds: Transient from persistent impairment. *Journal of Speech and Hearing Disorders, 52*, 156- 73.

Bowyer-Crane, C. & Snowling, M. J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *British Journal of Educational Psychology*, *75*, 189–201.

Brandao, A. C. P. & Oakhill, J. (2005). How do you know this answer? Children's use of text data and general knowledge in story comprehension. *Reading and Writing*, *18*, 687–713.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior, 22*(1), 1-14.

Burgess, S. R., & Lonigan, C. J. (1998). Bidirectional relations of phonological sensitivity and prereading abilities: Evidence from a preschool sample. *Journal of Experimental Child Psychology, 70,* 117-141.

Cain, K., & Oakhill, J. V. (1999). Inference making and its relation to comprehension failure. *Reading and Writing, 11*, 489–503.

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference making ability, and the relation to knowledge. *Memory and Cognition, 29*, 850–859.

Campbell, J. R. (1999). *Cognitive processes elicited by multiple-choice and constructed response questions on an assessment of reading comprehension.* Available from ProQuest Dissertations and Theses database. (UMI Number 9938651)

Carlson, S. E., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences, 32*, 40-63.

Carlson, S. E., van den Broek, P., McMaster, K., Rapp, D. N., Bohn-Gettler, C. M., Kendeou, P., & White, M. J. (2014). Effects of comprehension skill on inference generation during reading. *International Journal of Disability, Development, and Education, 61*(3), 258-274.

Chall, J. S. (1996). *Stages of reading development* (2nd ed.). Fort Worth, TX: Harcourt-Brace.

Chaney, C. (1992). Language development, metalinguistic skills, and print awareness in 3-year-old children. *Applied Psycholinguistics, 13,* 485-514.

Cordon, L. A., & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology, 88*(2), 288-295.

Cote, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes, 25*(1), 1-53.

Cromley, J. (2005). *Reading comprehension component processes in early adolescence.* Available from ProQuest Dissertations and Theses database. (UMI Number 3178579)

Cromley, J. G., & Azevedo, R. (2004a). *Testing the fit of three models of reading comprehension with 9th grade students.* Poster presented at the annual meeting of the American Educational Research Association. San Diego, CA.

Cromley, J. G., & Azevedo, R. (2004b). *Using think-aloud data to illuminate a model of high school reading comprehension*. Poster presented at the annual meeting of the American Educational Research Association. San Diego, CA.

Cromley, J. G., & Azevedo, R. (2005). *Testing the fit of four variations of the Inferential Mediation model of reading comprehension.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada. Retrieved on July 10, 2009 from http://azevedolab.autotutor.org/files/publications/Vick,Azevedo,Hofman%20AERA%202005%20FINAL.pdf

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education

Davis, G. N., Lindo, E. J., & Compton, D. L. (2007) Children at risk for early failure. Constructing an early screening measure. *Teaching Exceptional Children*, *39*(5), 32-37.

Denton, C. A., Fletcher, J. M., Anthony, J. L., & Frances, D. J. (2006). An evaluation of intensive intervention for students with persistent reading difficulties. *Journal of Learning Disabilities. 39(5),* 447-466.

Dickinson, D. K., McCabe, A., Anastasopoulos, L., Peisner-Feinberg, E. S., Poe, M. D. (2003). The comprehensive language approach to early literacy: The interrelationships among vocabulary, phonological sensitivity, and print knowledge among pre-school aged children. *Journal of Educational Psychology, 95*(3), 465-481.

Dickinson, D. K., & Tabors, P. O. (Eds.). (2001). *Beginning language with literacy: Young children learning at home and school.* Baltimore: Brookes Publishing.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., Cutting, L. E. (2012). Reader-test interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology*, *3*, 515–528.

Ehri, L. C. (1995). Phases of development in learning to read words by sight. *Journal of Research in Reading*, *18*, 116-125.

Ehri, L. C. (1998). Grapheme-phoneme knowledge is essential for learning to read words in English. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 3-40). Mahwah, NJ: Erlbaum.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197.

Embretson, S., Schneider, L. M., & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement*, *23,* 13–32.

Ennis, R. H. (1962). A concept of critical thinking. *Harvard Educational Review*, *32*, 81-111.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-250.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (2nd Ed.)*. Cambridge, MA: The MIT Press.

Evans, M. A., Shaw, D., & Bell, M. (2000). Home literacy activities and their influence on early literacy skills. *Canadian Journal of Experimental Psychology, 54*, 65-75.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*(3), 209-226.

Farrington-Flint, L., & Wood, C. (2007). The role of lexical analogies in beginning reading: Insights from children's self-reports. *Journal of Educational Psychology , 99*(2), 326-338.

Feagans, L., & Appelbaum, M. I. (1986). Validation of language subtypes in learning disabled children. *Journal of Educational Psychology, 78,* 358-364.

Garner, R. (1982). Verbal report data on reading strategies. *Journal of Reading Behavior, 14*(2), 159-167.

Gierl, M. J. (1997). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 91*, 26-32.

Gillam, S. L, Fargo, J. D., & Robertson, K. S. C. (2009). Comprehension of expository text: Insights gained from think-aloud data. *American Journal of Speech-Language Pathology, 18*, 82-94.

Goswami, U. (2001). Early phonological development and the acquisition of literacy. In S. Neuman & D. Dickinson (Eds.), *Handbook of early literacy research.* New York: Guilford Press.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test ambiguity. In R. D. Freedle & R. P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 298-368). Norwood, NJ: Ablex.

Hayward, D., Das, J. P., & Janzen, T. (2007). Innovative programs for improvement in reading through cognitive enhancement: A remediation study of Canadian First Nations children. *Journal of Learning Disabilities, 40*(5), 443-457.

Hayward, D. V., Stewart, G. E., Phillips, L. M., Norris, S. P., & Lovell, M. A. (2008). *Language, phonological awareness, and reading test directory*. London, ON: Canadian Centre for Research on Literacy and Canadian Language and Literacy Research Network.

Hua, A. N., & Keenan, J. M. (2014). The role of text memory in inferencing and in comprehension deficits. *Scientific Studies of Reading, 18*(6), 415-431.

Huey, E. (2009/1908). *The psychology and pedagogy of reading (Reprint)*. Newark, DE: International Reading Association. (Originally published 1908).

Israel, S. E. (2015). *Verbal protocols in literacy research: Nature of global reading development*. New York: Routledge.

Janssen, T., Braaksma, M., & Rijlaarsdam, G. (2006). Literary reading acitivites of good and weak studnets: A think aloud study. *European Journal of Psychology of Education, XXI*(1), 35-52.

Jacobson, V. (1974). *A linguistic feature analysis of verbal protocols associated with pupil responses to standardized measures of reading comprehension.* Paper presented at the annual meeting of the International Reading Association. New Orleans, Louisiana.

James, W. (1890). *The principles of psychology*. New York: Holt.

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

Kame'enui, E. J., Fuchs, L., Francis, D. J., Good, R., O'Connor, R. E., Simmons, D. C., Tindal, G., & Torgesen, J. K. (2006). The adequacy of tools for assessing reading competence: A framework and review. *Educational Researcher, 35*(4), 3-11.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: ACE/Praeger.

Kavale, K., & Schreiner,R. (1979). The reading process of above average and average readers: A comparison of the use of reasoning strategies in responding to standardized comprehension measure. *Reading Research Quarterly, 15*, 102-128.

Keenan, J. M. (2014). Assessment of reading comprehension. In C. A. Stone, E. R. Silliman, B. J. Ehren, & G. P. Wallach (Eds.), *Handbook of language and literacy: Development and disorders* (2nd ed., pp. 469-484). New York: Guilford Press.

Kendeou, P., Bohn-Gettler, C., White, M. J., & van den Broek, P. (2008). Children's inference generation across different media. *Journal of Research in Reading*, *31*(3), 259-272.

Kendeou, P., van den Broek, P., White, M., & Lynch, J. (2007). Preschool and early elementary comprehension: Skill development and strategy interventions. In D. S. McNamara (Ed.) *Reading comprehension strategies: Theories, interventions, and technologies,* (pp. 27–45). Mahwah, NJ: Erlbaum.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 163-182.

Kintsch, W. (1994). The role of knowledge in discourse comprehension: A construction integration model. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (4th ed., pp. 951-995). Newark, DE: International Reading Association.

Kintsch, W. (2010). Comprehension processes and classroom contexts. In M. G. McKeown & L. Kucan (Eds.) *Bringing reading research to life (*pp. 194-207). New York, NY: The Guilford Press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

Kucan, L., & Beck, I. L. (2003). Inviting students to talk about expository texts: A comparison of two discourse environments and their effects on comprehension. *Reading Research and Instruction, 42,* 1-29.

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, *95*(1), 3-21.

Laing, S. P., & Kamhi, A. G. (2002). The use of think-aloud protocols to compare inferencing abilities in average and below-average readers. *Journal of Learning Disabilities, 35*(5), 436-447.

Langer, J. A. (1987). The construction of meaning and the assessment of comprehension: An analysis of reader performance on standardized test items. In R. Freedle & R. Duran (Eds.), *Cognitive and linguistic analyses of standardized test performance* (pp. 225-244). Norwood, NJ: Ablex.

Lau, K. (2006). Reading strategy use between Chinese good and poor readers: A think-aloud study. *Journal of Reading Research in Reading, 29*(4), 383-399.

Lee, V. E., & Burkham, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, D.C.: Economic Policy Institute.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, Winter*, 1–10.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, Summer*, 3-16.

Leighton, J. P. & Gokiert, R. J. (2005, April). *The cognitive effects of test item features: informing item generation by identifying construct irrelevant variance*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Montreal, Quebec, Canada.

Leighton, J. P., & Sternberg, R. J. (2003). Reasoning and problem solving. In A. F. Healy & R. W. Proctor (Eds.), *Experimental psychology* (pp. 623-48). New York, NY: Wiley.

Lepola, J., Lynch, J., Laakkonen, E., Silven, M., & Niemi, P. (2012). The role of inference making and other language skills in the development of narrative listening comprehension in 4-6-year-old children. *Reading Research Quarterly, 47*, 259-282.

Levy, B. A., Gong, Z., Hessels, S., Evans, M. A., & Jared, D. (2006). Understanding print: Early reading development and the contributions of home literacy experiences. *Journal of Experimental Child Psychology, 93*, 63-93.

Linn, R. L. (1989). *Educational measurement (3rd ed.)*. New York, NY: Macmillan.

Lipson, M. Y. (1982). Learning new information from text: The role of prior knowledge and reading ability. *Journal of Reading Behavior, 14,* 243-261.

Lonigan, C. J. (2006) Development, assessment, and promotion of preliteracy skills. *Early Education and Development*, *17*(1), 91-114.

Lonigan, C. J., Bloomfield, B. G., Anthony, J. L., Bacon, K. D., Phillips, B. M., & Samwel, C. S. (1999). Relations among emergent literacy skills, behavior problems, and social competence in preschool children from low- and middle-income backgrounds. *Topics in Early Childhood Special Education, 19*(1), 40–53.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology,36*(5), 590-613.

Lonigan, C. J., Burgess, S. R., Anthony, J. L., & Barker, T. A. (1998). Development of phonological sensitivity in two- to five-year-old children. *Journal of Educational Psychology, 90*(2)*,* 294–311.

Lonigan, C. J., McDowell, K. D., & Phillips, B. M. (2004). Standardized assessments of children's emergent literacy skill. In B. H. Wasik (Ed.), *Handbook of family literacy* (pp. 525-550). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Lynch, J., & van den Broek, P. (2007). Understanding the glue of narrative structure: Children's on-and off-line inferences about characters' goals. *Cognitive Development, 22*, 323-340.

MacKay, R. (1974). Standardized tests: Objective/objectified measures of 'competence'. In A. V. Cicourel, *K.* Jennings S. Jennings, K. Leiter, R. MacKay, H. Mehan, & D. Roth (Eds.), *Language use and school performance* (pp. 218-247). New York: Academic Press.

Magone, M., Cai, J., Silver, E. A., & Wang, N. (1994). Validating the cognitive complexity and content quality of mathematics performance assessment. *International Journal of Educational Research, 21*(3), 317-340.

McCabe, A., & Rollins, P. R. (1994). Assessment of preschool narrative skills. *American Journal of Speech Language Pathology, 3*(1), 45-56.

McCain, M. N., Mustard, J. F., & Shanker, S. (2007). *Early years study 2: Putting science into action*. Toronto, Ontario, Canada: Council for Early Child Development. Retrieved October 9, 2008 from http://www.councilecd.ca/cecd/home.nsf/7F1BCE63A330D017852572AA00625B79/$file/Early_Years_2_rev.pdf

McCormick, S. (1992). Disabled readers' erroneous responses to inferential comprehension questions: Description and analysis. *Reading Research Quarterly, 27*(1), 54-77.

McMaster K. L., Espin, C. A., & van den Broek, P. (2014). Making connections: Linking cognitive psychology and intervention research to improve comprehension of struggling readers. *Learning Disabilities Research & Practice, 29*(1), 17-24.

McMaster K. L., van den Broek, P., Espin, C. A., White, M. J., Rapp, D. N., Kendeou, P., Bohn-Gettler, C. M., & Carlson, S. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences, 22*, 100-111.

Meisels, S. J., & Piker, R. A. (2000). An analysis of early literacy assessments used for instruction (Tech. Rep. No. 3-002). Ann Arbor: University of Michigan, Center for the Improvement of Early Reading Achievement.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(1), 741-749.

Mislevy, R. J. (1994). Evidence and inference in educational measurement. *Psychometrika, 59*, 439-483.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379-416.

Nagy, W. E., & Herman, P. A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M. McKeown & M. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 19-35). Hillsdale, NJ: Erlbaum Associates.

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy.

National Reading Panel. (2000). *Teaching children to read.* Washington, DC: National Institutes of Health.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* In J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.), Washington, DC: National Academy Press.

Neuman, S. B., & Dickinson, D. K. (Eds.) (2001). *Handbook of early literacy research.* New York: Guilford Press.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*(2), 199-215.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nicholson, T., & Imlach, R. (1981). Where do their answers come from? A study of the inferences which children make when answering questions about narrative stories. *Journal of Reading Behavior, 13*, 111-129.

NICHD Early Child Care Research Network. (2005). Pathways to reading. The role of oral language in learning to reading. *Developmental Psychology*, *41*(2), 428-442.

Norman, R. R., (2012). Reading the graphics: What is the relationship between graphical reading processes and student comprehension? *Reading and Writing, 25,* 739-774.

Norris, S. P. (1990). Effects of eliciting verbal reports of thinking on critical thinking text performance. *Journal of Educational Measurement*, *27*(1), 41-58.

Norris, S. P. (1991). Informal reasoning assessment: Using verbal reports of thinking to improve multiple-choice test validity. In D. N. Perkins, J. Segal, & J. F. Voss (Eds.), *Informal reasoning and education* (pp. 451-472). Hillsdale, NJ: Lawrence Erlbaum Associates.

Norris, S. P. (1992). A demonstration of the use of verbal reports of thinking in multiple-choice critical thinking test design. *Alberta Journal of Educational Research, 38*(3), 155-176.

Norris, S. P., Leighton, J. P., & Phillips, L. M. (2004). What is at stake in knowing the content and capabilities of children's minds? A case for basing high stakes tests on cognitive models. *Theory and Research in Education, 2*(3), 283-308.

Norris, S. P., Macnab, J. S., & Phillips, L. M. (2007). Cognitive modeling of performance on diagnostic achievement tests: A philosophical analysis and justification. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theories and applications* (pp. 61-84). Cambridge: Cambridge University Press.

Norris, S. P., & Phillips, L. M. (1987). Explanations of reading comprehension: Schema theory and critical thinking theory. *Teachers College Record, 89*(2), 281-306.

Olshavsky, J. E. (1976-77). Reading as problem solving: An investigation of strategies. *Reading Research Quarterly, 12*, 654-674.

Olson, G. M., Duffy, S. A., & Mack, R. L. (1984). Thinking out loud as a method for studying real-time comprehension processes. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 253-286). Hillsdale, NJ: Erlbaum.

Paris, S.G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*(2), 184-202

Paris, A. H., & Paris, S. G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly*, *38*(1), 36-76.

Paris, S. G. (2007). Assessment of reading comprehension. *Encyclopedia of language and literacy development* (pp. 1-8). London, ON: Canadian Language and Literacy Research Network. Retrieved July, 1, 2010 from http://www.literacyencyclopedia.ca/pdfs/topic.php?topId=226

Paris, S. G., Carpenter, R. D., Paris, A. H., & Hamilton, E. E. (2005). Spurious and genuine correlates of children's reading comprehension. In S. Stahl & S. Paris (Eds.), *Children's reading comprehension and assessment* (pp. 131-160). Mahwah, NJ: Erlbaum.

Paris, S., & Hoffman, J. (2004). Reading assessments in kindergarten through third grade: Findings from the center for the improvement of early reading achievement. *The Elementary School Journal, 105*(2), 199-217.

Paris, S. G., & van Kraayenoord, C. E. (1996). Story construction from a picture book: An assessment activity for young learners. *Early Childhood Research Quarterly, 11*, 41-61.

Pearson, P. D., & Garavaglia, D. R. (1999). *The impact of item format on the depth of students' cognitive engagement.* Unpublished manuscript, NARP Validity Study Panel.

Pearson, P. D., Hansen, J., & Gordon, C. (1979). The effect of background knowledge on young children's comprehension of explicit and implicit information. *Journal of Reading Behavior, 11*(3), 201-209.

Pellegrino, J. (Winter 2002-03). Knowing what students know. *Issues in Science and Technology*, *19*(2), 48-52.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know.* Washington, DC: National Academy Press.

Phillips, L. (1988). Young readers' inference strategies in reading comprehension. *Cognition and Instruction, 5*(3), 193-222.

Phillips, L. M. (1989). *Developing and validating assessments of inference ability in reading comprehension,* (Technical Report No 452). Champaign, IL: Centre for the Study of Reading, University of Illinois (ERIC Document Reproduction Service Number ED303 767).

Phillips, L. M., Hayden, R., & Norris, S. P. (2006). *Family matters: A longitudinal parent-child literacy intervention study*. Calgary, AB: Temeron Press.

Phillips, L. M., Hayward, D. V., & Norris, S. P. (2016). *Test of early language and literacy*. Scarborough, Ontario: Nelson Publishing.

Phillips, L. M., Norris, S. P., & Steffler, D. J. (2007). Potential risks to reading posed by high-dose phonics. *Journal of Applied Research on Learning, 1*(1), 1-18.

Poggio, A., Clayton, D. B., Glasnapp, D., Poggio, J., Haack, P., & Thomas, J. (2005). *Revisiting the item format question: Can the multiple choice format meet the demand for monitoring higher-order skills?* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Powell, J. L. (1988). An examination of comprehension processes used by readers as they engage in different forms of assessment. (Doctoral dissertation, Indiana University, 1988). *Dissertation Abstracts International*, 50(05), 1287A.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Purcell-Gates, V. (1991). On the outside looking in: A study of remedial readers' meaning-making while reading in literature. *Journal of Literacy Research, 23*(2), 235-253.

RAND Reading Study Group. (2002). *Reading for understanding*: *Toward an R & D program in reading comprehension.* Santa Monica, CA.

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11*(4), 289-312.

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75-119). Timonium, MD: York Press

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. Neuman & D. Dickinson (Eds.), *Handbook for research in early literacy* (pp. 97-110). New York: Guilford Press.

Scarborough, H. S. (2005). Developmental relationships between language and reading: Reconciling a beautiful hypotheses with some ugly facts. In H. W. Catts & A. G. Kamhi (Eds.), *The connections between language and reading disabilities* (pp. 3-24). Mahwah, NJ: Erlbaum.

Scardamalia*,* M., & Bereiter*,* C. *(*1984*).* Development of strategies in text processing. In H. Mandl, N. Stein*,* & T. Trabasso (Eds.), *Learning and comprehension of text (*pp. 379-406*).* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schellings, G. L. M., Aarnoutse, C. A. J., & van Leeuwe, J. F. J. (2006). Third-grader's think-aloud protocols: Types of reading activities in reading an expository text. *Learning and Instruction, 16*(6), 549-568.

Schirmer, B. R. (2003). Using verbal protocols to identify the reading strategies of students who are deaf. *Journal of Deaf Studies & Deaf Education, 8*(2), 157-170.

Scott, D. B. (2008). Assessing text processing: A comparison of four methods. *Journal of Literacy Research, 40*(3), 290-316.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523-568.

Sénéchal, M., & LeFevre, J. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development, 73*, 445-460.

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary in future reading. In D. Dickinson, & S.B. Neuman (Vol. Eds.). *Handbook of early literacy research* (Vol. 2; pp. 173-184). New York, NY: Guilford Press.

Sepassi, F. (2003). How do learners of different language ability perform on the cloze?: A verbal protocol analysis of EFL test takers performance on cloze tests. *Indian Journal of Applied Linguistics, 29*(2), 5-33.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*(1), 405-450.

Snow*,* C. *E. (*2003*).* Assessment of reading comprehension*:* Researchers and practitioners helping themselves and each other. In A. Sweet & C. Snow (Eds.), *Rethinking reading comprehension* (pp. 192-218). New York: Guilford Press.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.) (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Speece, D. L., Roth, F. P., Cooper, D. H., & de la Paz, S. (1999). The relevance of oral language skills to early literacy: A multivariate analysis. *Applied Psycholinguistics, 20*, 167-190.

Stahl, K. A. D., Garcia, G. E., Bauer, E. B., Pearson, P. D., & Taylor, B. M. (2006). Making the invisible visible: The development of a comprehension assessment system. In K. A. D. Stahl & M. C. McKenna (Eds.), *Reading research at work: Foundations of effective practice* (pp. 425-436). New York: Guilford Press.

Stallman, A. C., & Pearson, P. D. (1990). Formal measures of early literacy. In L. Mandel Morrow and J. K. Smith, *Assessment for instruction in early literacy* (pp. 7-44). New Jersey, NY: Prentice-Hall.

Stanovich*,* K. E. *(*1992*).* The psychology of reading*:* Evolutionary and revolutionary developments. *Annual Review of Applied Linguistics, 12*, 3-30.

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers.* New York: Guilford Press.

Storch, A. S., & Whitehurst, G. J. (2002). Oral language and code related precursors to reading: Evidence from a longitudinal structural model. *Early Language & Literacy Development, 38*(6), 934-947.

Taylor, K. L. & Dionne, J. P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*, 413-425.

Teale, W. H. (1990). The promise and challenge of informal assessment in early literacy. In L. M. Morrow, & J. K. Smith (Eds.), *Assessment for instruction in early literacy* (pp. 45-58). Englewood Cliffs, NJ: Prentice Hall.

Teale, W. H., & Sulzby, E. (1986). Emergent literacy as a perspective for examining how young children become writers and readers. In W. H. Teale & E. Sulzby (Eds.), *Emergent literacy: Writing and reading* (pp. vii-xxiv). Norwood, NJ: Ablex.

Tompkins. V., Guo, Y., & Justice, L. M. (2013). Inference generation, story comprehension, and language in the preschool years. *Reading and Writing: An Interdisciplinary Journal, 26,* 403-429.

Trabasso, T., & Magliano, J. (1996a). Conscious understanding during comprehension. *Discourse Processes, 21*(3), 255-287.

Trabasso, T., & Magliano, J. P. (1996b). How do children understand what they read and what can we do to help them? In M. Graves, P. van den Broek, & B. Taylor (Eds.), *The first R: A right of all children* (pp. 160-188). New York: Columbia University Press.

Trabasso, T., & Nickels, M. (1992). The development of goal plans of action in the narration of a picture story. *Discourse Processes, 15*, 249-275.

Tracey, D. H., & Mandel Morrow, L. (2006). *Lenses on reading: An introduction to theories and models.* New York, NY: The Guilford Press.

Tunmer, W. E., Herriman, M. L., & Nesdale, A. R. (1988). Metalinguistic abilities and beginning reading. *Reading Research Quarterly, 23,* 134-158.

van den Broek, P., Kendeou, P., Kremer, K., Lynch, J. S., Butler, J., White, M. J., & Lorch, E. P. (2005). Assessment of comprehension abilities in young children. In S. Stahl & S. Paris (Eds.), *Children's reading comprehension and assessment* (pp. 107-130). Mahwah, NJ: Erlbaum.

van den Broek, P. W., Lorch, E. P. & Thurlow, R. (1996). Children's and adults' memory for television stories: The role of causal factors, story-grammar categories, and hierarchical level. *Child Development, 67*, 3010-3029.

van Dijk, T. A., & Kintsch, W (1983). *Strategies of discourse comprehension.* New York: Academic.

van Kleeck, A. (1998). Preliteracy domains and stages: Laying the foundations for beginning reading. *Journal of Children's Communciation Development, 20(*1*)*, 33-51.

van Kraayenoord*,* C. E., & Paris, S. G. *(*1996*).* Story construction from a picture book: An assessment activity for young learners*. Early Childhood Research Quarterly*, *11*, 41-61

Vernon-Feagans, L., Miccio, A. W., Manlove, E. E., & Hammer, C. J. (2001). Early language and literacy skills in low-income African American and Hispanic children. In S. Neuman & D. Dickinson (Eds.), *Handbook on research in early literacy* (pp. 192-210). New York: Guilford.

Wade, S. E. (1990). Using think-alouds to assess comprehension. *The Reading Teacher*, *43*(7), 442-451.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1994). The development of reading-related phonological processing abilities: New evidence of bi-directional causality from a latent variable longitudinal study. *Developmental Psychology, 30,* 73-87.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., Donahue, J., & Garon, T. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology, 33*, 468-479.

Wenner, J., & Bauer, P. J. (2001). Bringing order to the arbitrary: One to two-year olds' recall of event sequences. *Infant Behavior and Development, 22*, 585-590.

Werner, H., & Kaplan, E. (1950). Development of word meaning through verbal context: An experimental study. *Journal of Psychology*, *29*, 251-257.

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development*, *69*(3), 848-872.

Wilson, M. M. (1979). The processing strategies of average and below average readers answering factual and inferential questions. *Journal of Reading Behavior, 11*, 235–245.

Wilson, P. T., & Anderson, R.C. (1986).What they don't know will hurt them: The role of prior knowledge in comprehension. In J. Orasanu (Ed.), *Reading comprehension: From research to practice* (pp. 31-48). Hillsdale, NJ: Lawrence Erlbaum.

Wilson, S. B., & Lonigan, C. J. (2010). Identifying preschool children at risk of later reading difficulties: Evaluation of two emergent literacy screening tools. *Journal of Learning Disabilities, 43*(1), 62-76.

Wingenbach, N. G. (1982). *Gifted readers: Comprehension strategies and metacognition* (Report No. CS-007-633). Columbus, OH: (ERIC Clearinghouse on Reading & Communication Skills No. ED 182 465).

Wood, F. B., Hill, D. F., Meyer, M. S., & Flowers, D. L. (2005). Predictive assessment of reading. *Annals of Dyslexia, 55*, 193-216.

Zill, N., & West, J. (2001). *Entering kindergarten: A portrait of American children when they begin school*. (Findings from the Condition of Education, 2000; NCES 2001-035). Washington, DC: United States Department of Education, Office of Educational Research and Improvement. http://nces.ed.gov/pubs2001/2001035.pdf

Zabrucky, K., & Ratner, H. H. (1992). Effects of passage type on comprehension monitoring and recall in good and poor readers. *Journal of Reading Behaviour*, *24*(3), 373-391.