



Theses and Dissertations

2005-12-14

Toward a Domain Theory of Fluent Oral Reading with Expression

Reo H. McBride

Brigham Young University - Provo

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



Part of the [Educational Psychology Commons](#)

BYU ScholarsArchive Citation

McBride, Reo H., "Toward a Domain Theory of Fluent Oral Reading with Expression" (2005). *Theses and Dissertations*. 341.

<https://scholarsarchive.byu.edu/etd/341>

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

TOWARD A DOMAIN THEORY OF FLUENT ORAL
READING WITH EXPRESSION

By

Reo H. McBride

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

Doctor of Philosophy

Department of Instructional Psychology and Technology
Brigham Young University

August 2005

Copyright © 2005 Reo H. McBride

All Rights Reserved

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a dissertation submitted by

Reo H. McBride

This dissertation has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date

C. Victor Bunderson, Chair

Date

Richard R Sudweeks

Date

J. Olin Campbell

Date

David Williams

Date

Stephanie Allen

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the dissertation of Reo H. McBride in its final form and have found that (1) its format, citations and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

C. Victor Bunderson
Chair, Graduate Committee

Accepted for the Department

Richard R Sudweeks
Graduate Coordinator

Accepted for the College

K. Richard Young
Dean, David O. McKay School of Education

ABSTRACT

TOWARD A DOMAIN THEORY OF FLUENT ORAL READING WITH EXPRESSION

Reo H. McBride

Department of Instructional Psychology and Technology

Doctor of Philosophy

Today's educators are in need of tests or rating systems that provide specific and valid feedback to parents, students and programs. This need includes the area of expressive fluent oral reading. One way to address this need is to provide a rating system based on theoretical models that explore how fluency develops. This study explores the dimensions, constructs, or aspects that make up fluency. It also explores whether there is a sequence or order in how fluent oral reading with expression develops and the theoretical reasons for that ordering. This study further addresses whether word recognition or accuracy confounds the ratings of other aspects of fluency.

Such issues may affect the reading community's approach to the teaching of fluency in the schools. For, if there is a developmental ordering of constructs that make up fluency, or if it is found that accuracy (word recognition) is separate from fluency, knowledge of such an ordering and separation can influence paradigms of how we as educators view present approaches to the teaching of reading in the classroom, especially in how we build our students' fluent oral reading skills.

The researcher developed a rating scale to measure fluent oral reading with expression. He found that there are two dimensions providing the most meaningful interpretation to expressive fluent oral reading: accuracy and fluency. The author

provides the rationale and empirical evidence that there is a learning order of subordinate constructs belonging to the fluency dimension. This order, as determined by a many-facets Rasch analysis, is (a) phrasing, (b) smoothness, (c) rate, (d) expression, and (e) confidence. When accuracy is used in the same Rasch analysis, it was found to be easier than phrasing, showing that the method used to select texts easy enough for students was successful. Accuracy was used as a control dimension to assure that fluency constructs could be observed by avoiding confounding the observations of fluent oral reading with word knowledge problems.

Each construct consists of at least two descriptors or indicators, totaling 14 indicators in all. Three indicators load together on accuracy, and ten load together on fluency. An indicator designed for fluency, *Smoothness 2: No Repeats*, also loads on accuracy when included in the factor analysis, but it was found not to be a good indicator of accuracy or fluency. This clarification of number of dimensions and ordering constitutes the beginnings of a domain theory of fluent oral reading with expression (FORE) which provides an empirical description of the developmental sequence of progressive attainments that the average learner achieves on the two primary dimensions.

ACKNOWLEDGEMENTS

First, I thank all the members of my committee who stuck with me throughout: Dr. C. Victor Bunderson – my chair, Dr. Richard Sudweeks, Dr. Olin Campbell, Dr. David Williams and Dr. Stephanie Allen. Many thanks also to the outstanding and extra-mile assistance of fellow graduate student Kairong Wang as the analyst who set up the different Facets models and completed the analysis runs implemented in this study. The analysis could not have been accomplished without her timely efforts. I also thank Dr. Van Newby for his advice, insights and encouragement regarding the domain theory discussed in this research. I thank Michelle Bray, the Instructional Psychology & Technology Departmental Secretary, and Claire DeWitt, Office of Graduate Studies – both of whom kept me in line with deadlines and advice. I also thank my brothers and sisters: Maurice H. McBride, Dane C. McBride, Darla Jean (McBride) Anderson, Tanya Kay (McBride) Skeen, Bonnie Colleen (McBride) Whitehurst, Gina LaRee (McBride) Jones, and their spouses for all their advice and counsel, both emotionally and financially. Speaking of brothers and sisters, Darla Jean (McBride) Anderson served as my main editor. Her dogged determination helped me push myself beyond what I thought I could do. What a jewel she has been! I also thank her daughter, Tamigene Conklin who has been a wonderful advisor as well. Her timely insights and advice are worth more than what money could buy. My thanks also goes to Daniel K. Anderson, husband to Darla and father of Tamigene. His wit and humor provided comic relief to help reduce the stress a dissertation can bring. I also thank my dear friend, Martha Gladden, for her willingness to put her life on hold in order to help me out. I thank my dear, dear,

wonderful parents, Dr. Vearl G. McBride and Mrs. Betty Jean McBride, who continue to inspire and influence me in all aspects of my life: financially, morally, emotionally and in any other way possible that parents aid and assist their children. Finally, I thank my sweet wife, Keri, and my three boys, Christopher, Carson and Alexander, for sacrificing the time I wished I could have spent with them. Without their support, especially that of my dear wife, the gem of my life, this dissertation never would have been accomplished.

TABLE OF CONTENTS

	Page
Acknowledgements.....	vii
List of Tables	xi
List of Figures.....	xii
Chapter	
1. Introduction.....	1
Need	1
Research Questions.....	2
Background.....	3
Development of the FORE Domain Theory Constructs	3
FORE Constructs Supported by Early Reading Theory Development	5
Reasons for a Different Fluency Measurement Instrument and Theory.	6
Purpose.....	8
Definitions of Terms.....	9
2. Literature Review.....	13
Constructs of Fluent Oral Reading with Expression.....	13
Accuracy	14
Smoothness	16
Rate	17
Phrasing.....	19
Expression.....	21
Confidence	22

Issues Pertaining to a Domain Theory	24
Construct Validity.....	27
Validity-Centered Design	27
Category II of Validity-Centered Design.....	29
Design for Inherent Construct Validity.....	29
Content Validity	29
Substantive Process Validity.....	30
Structural Validity.....	30
3. Method	33
Design	33
Sample.....	33
Procedures.....	34
Obtaining Student Participation	34
Text Selection	35
Videotaping Student Reading Performance	36
Selection of Raters	37
Instrument Development.....	37
Training of Raters	38
Collection and Preparation of Data for Analysis	40
Methods for Addressing Research Questions.....	41
Research Question 1	41
Research Question 2	43
Research Question 3	44
4. Results.....	47
Research Question 1	47
Question 1a: Inter-rater Reliability	47

Question 1b: Internal Consistency of Accuracy and of the Five Fluency Constructs	49
Question 1c: Systematic Differences in Rater Means	52
Research Question 2	58
Factor Analysis Results.....	58
Extraction of Factors	59
Factor Rotation.....	62
Research Question 3	70
Setting up the Facets Program	70
Order of Constructs Based on Facets	71
5. Conclusions and Recommendations	77
Discussion	77
Conclusions.....	82
Recommendations for Improving the FORE Measurement Instrument	86
General Recommendations for Future Research	87
References.....	89
Appendices.....	97
A. History of the Development of the FORE Measurement Instrument	99
B. Permissions and Consent Documentation.....	111
C. Reading Texts Used in Study	121
D. Institutional Review Board Permissions.....	129
E. Facets Generated Reports	135
F. Factor Analysis Output with 14 Variables, No Suppressions.....	141
G. Factor Analysis Output with 13 Variables, No Suppressions.....	145

LIST OF TABLES

1	Validity-Centered Design	28
2	Student Participation Numbers and Rates by Grade Enrollment and Gender.....	34
3	Reading Selections and Grade Levels.....	35
4	Alpha Coefficients across Four Raters (Inter-rater Reliabilities)	48
5	Correlations among the Mean Ratings of the 14 Indicators in the FORE Domain	50
6	Internal Consistency Reliability Coefficients for Composite Constructs of Accuracy and Fluency	51
7	Abbreviated Rater Measurement Report from Facets Analysis.....	54
8	Abbreviated Rating Measurement Report from Facets Analysis.....	57
9	Difficulty of FORE Constructs in Logits	72

LIST OF FIGURES

1	FORE Measurement Instrument Version V (FMI-V).....	39
2	Facets Generated Chart Displaying Differences in Leniency and Severity of the Four Raters and also the Difficulty Levels of the 14 Indicators.	53
3	Communalities Derived through the Principal Axis Factor Method of Extraction	60
4	Eigenvalues and Variance Accounted for by Each Factor;	61
5	Three-Factor Structure and Correlation Matrices Compared with Two-Factor Structure and Correlation Matrices	63
6	Three-Factor Pattern Matrix and Two-Factor Pattern Matrix Compared.....	67
7	Factor Plot of FORE Indicators in Rotated Space	69
8	FORE Constructs Ordering Based on Facets Analysis.....	73
9	FORE Indicators Compared with Their Respective Constructs	76
A1	FMI-I, Untested.....	101
A2	FMI-II, Untested	102
A3	FMI-III-A, p.1	104
	FMI-III-A, p.2.....	105
A4	FMI-III-B, Untested, p.1	107
	FMI-III-B, Untested, p.2	108
A5	FMI-IV	110
E1	Facets Generated Table 7.2.1—Rater Measurement Report	136
E2	Facets Generated Table 7.1.1—Examinee Measurement Report	137
E3	Facets Generated Table 7.3.1a—Rating Scale Measurement Report	138
E4	Facets Generated Table 7.3.1b—Rating Construct Measurement Report	139
F1	Three-Factor Structure Matrix	142

F2	Three-Factor Pattern Matrix	143
G1	Principal Axis Factors, Eigenvalues, and Variance Accounted for by Each Factor	146
G2	Scree Plot Showing Eigenvalues	147
G3	Two-Factor Pattern Matrix	148
G4	Two-Factor Structure Matrix	149
G5	Two-Factor Correlation Matrix.....	150

CHAPTER 1

INTRODUCTION

Need

The National Reading Panel (NRP) identified five components of reading instruction: (a) phonemic awareness, (b) phonics, (c) fluency, (d) vocabulary, and (e) comprehension (National Institute of Child Health and Human Development, 2000). Influenced by the National Research Council's (NRC) *Committee on Preventing Reading Difficulties in Young Children* (Committee on the Prevention of Reading Difficulties in Young Children, 1998), the NRP argued that in terms of reading instruction research the education field needs scientifically-based information that is clear and objective. The Panel went on to explain that any conclusions and determinations made should be based on findings obtained from experimental studies. These studies should be characterized by strict methodological rigor, including evidence of "reliability, validity, replicability and applicability" (p. 1-2).

The NRP's recommendations address the fact that today's educators are in need of tests or rating systems that provide specific and valid feedback to parents, students, and programs (Strong-Krause, 2001). This need includes the area of expressive fluent oral reading.

One way to address this need is to construct and provide rating scales (measurement instruments) based on theoretical models that explore how fluency develops. Such models would have to address

1. What comprises fluency in oral reading?
2. Is there a sequence in how fluent oral reading with expression (FORE) develops?
3. If there is such a developmental sequencing or ordering, what are the theoretical reasons for that ordering?

4. How does word recognition or accuracy affect other aspects of fluency?

Answers to these questions may facilitate the reading community's approach to the teaching of fluency in the schools. If there is a developmental sequencing or ordering of constructs or aspects that make up fluency, or if it is found that accuracy (word recognition) is separate from fluency, knowledge of such an ordering and separation can change paradigms of how educators view present approaches to the teaching of reading in the classroom, especially in how to build oral reading skills.

A *domain theory* is a descriptive theory of the development of capabilities along measurable dimensions of learning. While the broader need is for improved approaches to teaching that work better for each student, a *domain theory* is neither a theory of instruction nor of teaching. Although the term *domain theory* is used to honor the initial concept introduced by Messick (1995), another term is *local learning theory* (Bunderson & Newby, 2005, in press). Such a description is local to a narrow domain of learning. This type of theory provides an interpretive framework for the number of dimensions along which the measurement of learning and growth may occur. A domain theory or local learning theory also gives an account and interpretation of the order of development. It is a strong guide for prescriptive instructional approaches but is not itself such a prescriptive theory. The associated measurement instrument is a tool for evaluating alternative instructional approaches in later studies.

Research Questions

In an effort to address the stated issues, needs and concerns, the researcher proposes the following three research questions, which serve as the crux of this study.

1. What is (a) the inter-rater reliability across the four raters for each of the 14 indicators? (b) What is the internal consistency of the measures of fluent oral reading constructs (accuracy, smoothness, phrasing, rate, confidence, and expression) and dimensions in the FORE measurement instrument? (c) Are the

- raters interchangeable in their ratings of students in terms of rater leniency and severity? If not, what are the systematic differences among the raters?
2. How many dimensions of accuracy and fluency are sufficient to describe the domain of fluent oral reading with expression for students in grades 2 through 6?
 3. Using the features of the Facets software, how are the average levels of rating scales that make up each sub-construct located along the dimension(s) of fluent oral reading with expression?

Background

The researcher presents a history of the development of the domain theory of fluent oral reading with expression. He then discusses that well before the NRP was convened, the need for research in reading fluency was made evident.

Development of the FORE domain theory constructs. While working with remedial reading students of all ages, the author observed that the majority of students with whom he worked were those whose local schools had already written off as being hard to work with and less likely to become good readers. In contrast, copies of school records provided by parents showed that most of these students were of average intellectual ability. Most of these readers displayed problems with the continual miscuing of words and skipping lines of text.

Such students often were slow to recognize and pronounce words correctly and lacked smoothness while attempting to read. It was observed that students would take breaths at inappropriate places where there was no punctuation indicating the opportunity to do so. Students' inability to recognize words quickly and accurately, along with their being unable to read smoothly, made phrasing in reading next to impossible. If reading were accomplished at all, there was an inappropriate phrasing of words or groups of words together. As a natural follow-on to these problems, it was observed also that the overall oral reading rate was drastically inhibited. That is, if students were required to

read out loud, following along while the text was read out loud with them, students were unable to keep up with the normal flow or pace (rate) in reading. Then, when asked to read aloud, students demonstrated improper expression. When asked to relate what they had just read many students were unable to tell about what they had read, indicating a lack of comprehension.

Finally, as a result of these reading problems and evidence from school reports, these students showed a lack of confidence in their own reading skills. They had become so discouraged in their reading that they were afraid to try to read, afraid of making mistakes, afraid of being ridiculed. This non-cognitive characteristic, confidence, may be of even greater importance than the cognitive and fluent oral reading skills just emphasized.

The author worked with these students to help remediate their reading using the McBride Reading Program developed initially by Dr. Vearl G. McBride (McBride, 1997, unpublished manuscript), based on his over 40 years of experience in working with individuals who had difficulties in reading. McBride breaks with the traditional or typical route of requiring students to first go through decoding. Instead, his program uses unique methods to promote growth in the recognition of sight words, postponing decoding and phonics, if necessary, until success and confidence have been attained. His methods move students along from the reading of single words, to multiple words, to the oral reading of whole sentences.

McBride's program has not, up to this time, articulated a theoretical basis. Because of its success and the author's intimate familiarity with it, McBride's program provided a base of experience for the theory development activities in this study. Through this study, the author identifies and develops the supporting constructs pertaining to fluent oral reading using aspects of the McBride Reading Program. This study seeks to articulate the constructs which the McBride Reading Program embraces, grounding them

in the reading research literature to show that they all, to greater and lesser degrees, are well known.

FORE constructs supported by early reading theory development. FORE and its theoretical development beginnings stem from the process involved in the development of a domain model of language and communication (reading, writing, listening, replying, public speaking) which took place through extended conversations and iterative attempts between the author and two other investigators, to build both a model and map of the substantive processes involved in early reading. The efforts in developing such a map were validity centered. The component constructs of fluent oral reading are not grounded on just one reading program, but on findings from previous research and on two distinct programs each with a different approach and philosophy.

In addition to the McBride Reading Program, the author was involved in the analysis of constructs found in another widely used reading program: Dr. Grant Von Harrison's *Companion Reading Program*. The Companion Reading Program was a fertile ground for deriving a set of constructs that led to the initial domain map that influenced the development of the fluent oral reading constructs that evolved to their present status in this study. The Companion Reading Program itself, over the course of the last 35 years, has been implemented in 625 schools. It is well-formulated and claims to be a complete instructional system. The program's components are claimed to work together to enable the learner to improve reading achievement significantly while elevating belief in personal ability and enjoyment of reading.

Using the Companion Reading Program as a reference, three investigators, Dr. C. Victor Bunderson, an expert in validity-centered design, Dr. John Wilkinson, an expert in Instructional Technology and in the Companion Reading Program, and the author of this study looked for the constructs of language and communication and how they developed over time. The investigators placed these constructs into a conceptual interpretive framework. Related planning and working meetings took place over several months

where the various constructs unique to the domain of language and communication development were identified and placed into a model of language and communication development. This model presented graphically the investigators' concepts of how language and communication progress through time, with written explanations elaborating how the identified constructs may influence early reading and comprehension.

In addition to the other constructs unique to early reading, the development of the domain map revealed that fluent reading, both silent and oral, is a part of the domain of early reading. However it did not provide details of those constructs inherent in fluent reading. This research, then, is a step toward exploring and investigating the constructs of fluent oral reading with expression and a step in confirming the usability and inter-rater reliability of the related FORE measurement instrument.

Reasons for a different fluency measurement instrument and theory. Before the NRP convened, Lipson and Lang (1991) expressed concern that teachers and parents need good information about reading fluency to distribute time and resources effectively in instructionally appropriate ways. They also stated concerns that there is a lack of agreement and much confusion surrounding suitable approaches to how fluent oral reading should be taught and assessed in the classroom. Reutzel and Hollingsworth (1993) expressed concern that the development of reading fluency is a neglected part of reading instruction despite the fact that many reading authorities consider it to be an important part of the reading curriculum.

Lyon and Moats (1997) summed up the “state of intervention research with a call to refocus attention on fluency” (p. 211). They claimed that it is far easier to gain information about improvements in decoding and word-reading accuracy than it is to obtain information about improvements in reading fluency and automaticity. They stated that such a disparity between these constructs denotes a need to learn more about how the development of componential reading skills contributes to or affects reading rate and

reading comprehension. This need harks back to the research questions, hinting at such issues as the possibility of there being more than one dimension to fluent oral reading. The word *development* suggests that an ordering of componential reading fluency skills may exist that reflects how processes develop and enable other processes. Further, Snow, Burns and Griffin (National Research Council, 1998) assert that there is little effort made in the classroom to develop fluent oral reading skills. Their assertion served as one of the reasons for the NRP report.

Despite the stated concerns for the lack of emphasis on reading fluency in the classroom, there is evidence to suggest that efforts have been made to improve students' reading fluency. For example, the University of Oregon states that it has an assessment designed to test and measure reading fluency. This program of assessment is known as the *Dynamic Indicators of Basic Early Literacy Skills* or *DIBELS* (University of Oregon, 2000). DIBELS measures (a) phonological awareness, (b) alphabetic principle, and (c) fluency with connected text. It is based on a curriculum-based measurement program developed at the University of Oregon.

Another such program of assessment is put forth by Scholastic which claims to generally assess reading and match students to books at appropriate grade levels (Scholastic, 2002). Adding to the list is *The Partnership for Reading* which states that it brings scientific evidence to learning (The Partnership for Reading, 2004). Still one other, *The Reading Genie* (Murray, 2002) gives several suggestions as to how to improve reading in general, providing lesson plans and strategies. It also gives guidance in how to help beginners with oral reading (Murray). The field is also well supplied with informal reading inventories, among which are the *Burns & Roe Informal Inventory* (Burns & Roe, 1993), the *Ekwall & Shanker Reading Inventory* (Ekwall & Shanker, 2000), and the *Leslie & Caldwell Qualitative Reading Inventory – II (QRI-II)* (Leslie & Caldwell, 1995).

The programs and associated assessments mentioned previously claim to assess various dimensions of fluent oral reading. Others only mention how important fluent oral

reading abilities are and recommend strategies in how teachers may improve fluency and other reading skills. These programs state that their assessment instruments have been scientifically tested resulting in high correlations, but do not provide information as to how their identified reading skills are matched up to the measurement scales attached to them. They appear to lack evidence of validity which indicates a scarcity of construct and measurement oriented research in fluent oral reading as defined in this current research. These programs also appear to confound word recognition (accuracy) with fluency, making the observation of such fluency aspects as smoothness and expression unclear.

Purpose

This study sought empirical evidence for key aspects of construct validity in fluent oral reading with expression. In particular the researcher desired to determine the number of constructs needed to span the domain of fluent oral reading with expression and a theorized ordering of learning difficulty for each construct. At the beginning of this study, the hypothesized ordering was accuracy < smoothness < rate < phrasing < expression < confidence, where the less than symbol (<) means that the construct to the left of the symbol is easier than the one to the right. The justification for the use of these terms as constructs is elaborated in the literature review.

The researcher also hypothesized that poor word recognition or lack of accuracy will tend to confound ratings of the other fluent oral reading construct-linked scales. That is, if a student stutters or hesitates when trying to sound out words, that student may give the impression that he or she may have problems with the other fluent oral reading traits, including smoothness, rate, phrasing, expression and confidence. However, if the student were to be given a reading selection in which he or she demonstrated no major problems with word recognition, the other indicators could then be measured more accurately by such an instrument, being less confounded by problems in word recognition. Not only would the indicators be less confounded by problems in word recognition, but an

ordering of which constructs related to those indicators would be easiest or most difficult for students to receive high marks (an ordering of difficulty) could also be established.

The researcher hopes that the domain theory of fluent oral reading with expression and the accompanying measurement instrument will provide a means to more validly assess the fluency constructs involved in identifying words quickly and easily, and in seeing enough of their meaning to speak smoothly and briskly, with good phrasing and expression, and with greater confidence. The intent of this research is to meet the aforementioned needs by (a) identifying the constructs necessary for expressive fluent oral reading and providing the rationale as to why these dimensions were chosen, (b) determining the number of accuracy and fluency constructs sufficient to span the domain of fluent oral reading with expression, (c) proving if an ordering of FORE constructs exists, and (d) determining if a lack of accuracy confounds the other fluent oral reading constructs as hypothesized earlier.

Another underlying reason as to why the domain theory of fluent oral reading with expression and accompanying instrument are both being developed can best be summed up in a simple but profound statement made by Nathan and Stanovich (1991) wherein they suggested that fluency “may be almost a necessary condition for good comprehension and enjoyable reading experiences” (p. 176).

Definitions of Terms

Construct. A *construct* is an unobservable, hypothesized human characteristic that researchers construct in their minds to help them explain or theorize about human performance and conduct (Bunderson, 2005, in press).

Construct-linked-scale-development (CLSD). CLSD is a measurement term that relates to the development of a domain theory. It consists of four stages: (a) construct delineation, (b) construct-linked ordering, (c) invariant scale development, and (d) construct-linked scaling (Strong-Krause, 2001).

Construct validity. This measurement term is used by psychometricians and test developers denoting both the evidential and consequential basis for any score interpretation or use. It is used when addressing the trustworthiness of score interpretations in accounting for and explaining how both test performance and test scores relate to the construct a test developer wishes to measure (Messick, 1995).

Domain. A domain refers to an area of *human activity involving expertise* or area of *academic interest* or specialization (Bunderson & Newby, 2005, in press).

Domain theory. Domain theory is a measurement term used in describing the contents, substantive processes, structures and boundaries of an area of human activity or area of academic interest or specialization i.e. domain. It provides an account of how construct-relevant sources of task difficulty and substantive processes operate at different levels of growth along the scale(s) that cover a particular domain (Bunderson & Newby, 2005, in press; Messick, 1995).

Expected rating. This measurement term refers to the score a many-facet Rasch model predicts a rater will assign to a ratee on a given trait, based on (a) the estimated level of severity the rater exercises in comparison to the severity of ratings other raters give a ratee on that given trait, (b) the estimated difficulty of that trait, and (c) the ratee's estimated level of performance (Myford & Wolfe, 2003).

Facets. This word is used when conducting a Many-Facet Rasch Measurement (MFRM) which refers to group-level main effects for raters, ratees, traits, and any other variables or sources of error (McNamara, 1996).

Fit indices. This term is used to specify the extent to which observed ratings match the expected ratings that are predicted or calculated by the many-facet Rasch model (Myford & Wolfe, 2003).

Fluency. This word is used to describe an individual's level of freedom from word identification problems that could hinder comprehension while reading, and the ability to

read orally with speed, accuracy, and proper expression or proper inflection and phrasing (Harris & Hodges, 1995).

Indicator. The author uses this term to describe an observable, measurable characteristic of a FORE construct.

Local learning theory. This term is used as a synonym for domain theory, but draws attention to the developmental or learning aspects of the explanatory account given by a domain theory (Bunderson & Newby, 2005, in press).

Logit. This term denotes the unit of measurement used in Rasch scaling. It is used to report locations along a logit scale of candidate ability, task difficulty and rater severity, all of which are entered into a facets analysis. The word is pronounced *lo'-git*, with the accent on the first syllable (McNamara, 1996).

Validity-centered design. This term refers to a principled design process for developing, designing and improving (a) learning theories (b) domain theories, and (c) associated construct-linked measurement scales. Validity-centered design also serves as a guide in documenting the evidence for a validity argument (Bunderson & Newby, 2005, in press).

CHAPTER 2

LITERATURE REVIEW

This study focuses on the development of a domain specific theory of fluency and a rating scale designed for use in assessing the constructs in the proposed domain. This literature review combines constructs unique to FORE and a domain theory which explains how those fluency indicators relate to the constructs. Such an explanation is essential if one is to determine the number of dimensions necessary to describe the domain and a possible developmental ordering in terms of leniency for raters and difficulty for students as far as the fluent oral reading with expression constructs are concerned.

Constructs of Fluent Oral Reading with Expression

According to Wolf and Katzir-Cohen (2001), there is not a consensus about what is meant by *fluency* and what its relation might be to the time-related subset of terms most frequently related to it such as (a) automaticity, (b) speed of processing, (c) reading rate or speed, and (d) word recognition rate or proficiency. However, the researcher allowed himself to be guided by the National Reading Panel's (2000) definition which states that fluency means freedom from word identification problems that hinder comprehension and that fluent readers are characterized by the ability to read orally with (a) speed, (b) accuracy, (c) proper expression or inflection, and (d) phrasing (Harris & Hodges, 1995).

The researcher was also influenced by Brenna (1995) and Zutell and Rasinski (1991). They state that the goal of reading instruction is to help children interact meaningfully with a variety of texts. Such interaction requires children to be competent in word recognition and read at a suitable rate. Children also need to understand how to demonstrate the appropriate *pausing* and *intonation*, terms denoting prosody or the phrasing and expression of the orally-delivered words upon written words. Taking

directly from Zutell and Rasinski (1991), the author refers to their *Multidimensional Fluency Scale* wherein they highlighted phrasing, smoothness and *pace* (referred to as *rate* in this study) as necessary components of fluency acquisition.

Based on Zutell and Rasinski (1991), Harris and Hodges (1995), Brenna (1995) and the National Reading Panel (2000) the following constructs are proposed as components of fluent oral reading: (a) accuracy, (b) smoothness, (c) rate, (d) phrasing, (e) expression, and (f) confidence.

Accuracy. According to Farstrup and Samuels (2002) fluency is significant because it places an emphasis on comprehension and that in order to experience good comprehension the reader must be able to identify words quickly and easily. In other words, the reader must have excellent word recognition skills.

The early work of Cattell (1886) and Huey (1968) laid the groundwork for LaBerge and Samuels (1974), Logan (1997), Pikulski (2000) and others in the use of such terms as automaticity or accurate word recognition. Cattell found that his students could name letters and words faster than other symbolic categories such as colors and other more concrete semantic categories such as pictured objects. He was the first researcher to draw attention to the automatic-like rates of recognition achieved in letter naming and word reading, with words read as fast as letters, and the first to point out that reading speeds actually increased when semantic and syntactic information are provided, as in sentences.

Huey (1968), in following up on Cattell's (1886) work, stated that the development of fluent reading involves the steady accumulation and synthesis of increasingly complex essential fluency skills which gradually weld together over time by practice. An integral aspect of this synthesis, according to Huey, was the development of rate of processing, which through repeated practice allows the reader not to have to concentrate on details. Huey's concept of rate of processing and repeated practice

actually causes the entire act of fluent reading to become easier. As fluency skills grow, conscious exertion to read fluently decreases, becoming more automatic.

Building on Huey's (1968) concepts, early research shows that students must become fluent in their ability to identify words. If they do not, they will be less able to respond to the texts they read (Nathan & Stanovich, 1991). When processes of word recognition take little capacity (that is, when readers have achieved automaticity), most of the reader's cognitive capacity can be focused on comprehending the text, criticizing it, elaborating on it and reflecting on it. This may otherwise be known as critical thinking, or at least an early manifestation of it. The opposite is also true when the reader does not possess fluent word recognition skills, therefore leaving much less mental capacity for comprehension.

According to Logan (1997) an individual possessing the properties of fluent word recognition is noted as being fast, effortless and autonomous in his reading. Logan explains that the word recognition process is so fast, that it happens unconsciously. For example, when a fluent reader sees a traffic STOP sign, he automatically reads *stop*. The individual cannot help but read the word. Logan goes on to say that the single most important aspect of fluent reading involves two acts happening simultaneously: the recognizing of words quickly and the comprehending of text at the same time. Supporting Logan's views concerning reading fluency, word accuracy and comprehension, Pikulski (2000) refers to this process as the rapid recognition of words.

Snow, Burns, and Griffin (National Research Council, 1998) further strengthen the conclusions of Logan (1997) by adding that in order to achieve fluency beyond the rudimentary initial levels, one must have sufficient practice in reading different texts. They also go on to say that the ability to acquire meaning from print is so strongly dependent on the acquisition and continuous growth of word recognition accuracy and reading fluency, that both of these should be assessed regularly in the classroom. Such

actions taken permit timely and effectual instructional response when a student displays difficulties or delays in these areas.

It is interesting to note that Snow and her colleagues (National Research Council, 1998) separated the terms word recognition accuracy (henceforth referred to as accuracy) and fluency. Such a separation does not imply that word recognition accuracy is not part of fluency. To the contrary, it is part of fluency, but should be treated as separate from the other constructs. Not to do so allows a student's difficulty in accuracy skills to confound the other indicators of fluency, giving an impression of problems in those areas where none may really exist.

Dwyer (2004) also separated accuracy from fluency in an effort to assist the student in his or her own "management of strategies for accuracy and appropriateness" (p. 2). Such a separation, as highlighted by Snow, Burns, and Griffin (National Research Council, 1998) and Dwyer, imply that reading fluency researchers need to consider accuracy apart from the other fluency dimensions, but not totally separate it from the overall domain of fluent oral reading. Considering reading fluency in this light may avoid the confounding mentioned earlier. Zutell and Rasinski (1991) support this train of thought by stating that they themselves do not include word accuracy errors in their own rating system. However, they do not suggest that such errors are not important. They stress that teachers need to recognize that fluency and accuracy are related but separable dimensions of fluent oral reading. Based on the preceding discussion, accuracy is considered in this study as a key but separable dimension necessary for fluent oral reading.

Smoothness. Although reading smoothness also contributes to fluent oral reading, research regarding the construct of smoothness is very sparse. This could mean that either it is not considered important enough in reading fluency research to warrant such a delineation, or it could mean that those involved in reading fluency research (aside from those cited in this study) have never chosen to consider smoothness as a dimension

worthy of such research. Regardless of the reason, a student is characterized to have good fluency in reading, in part, when his or her reading is generally smooth. Desirable breaks are evident showing phrasing and expression, and if there are other breaks, they are resolved quickly via self-correction. Thus, in smooth reading, any difficulties with words or grammatical structures are rarely encountered and are resolved quickly with the student thus being able to read at a fluid or smooth pace (Zutell & Rasinski, 1991).

Smoothness is also indicated when a student is able to read a given text several times without stopping, stuttering, or manifesting other inhibitors to fluent reading Samuels (1997). Samuels determined that as a student practices reading the same text through various repetitions, the student's ability to read smoothly increases through self-correction of reading problems and growing familiarity with the text.

Brenna (1995) noted in her reading fluency research with children 4 to 6 years of age who were reading fluently prior to the first grade, that fluency should be judged on whether children could conduct meaningful reading with relative smoothness. Relative smoothness, according to Zutell and Rasinski (1991), is also referred to as "automatic word recognition" (p. 215). Based on the preceding discussion, the researcher considers smoothness as a key fluency standard necessary for the acquisition of fluent oral reading.

Rate. The concept of *rate* is closely associated with smoothness. The premise is that the more confident a reader is with reading, the greater the understanding the reader has in knowing when to adjust his reading pace, or rate, depending on the context and content of the given reading selection. Dwyer (2004) states that a reader who has good fluency selects and maintains an appropriate rate and speed during oral reading. He states that the meaning of the passage affects the rate. That is, an individual's reading rate will change, based on the situation or meaning found within the passage. One who possesses good reading fluency skills will speed up or slow down the oral reading rate to create "appropriate emphasis on the meaning of the text" (p. 1). Dwyer continues by stating that

the reader makes the oral reading sound natural, with the rate and speed being well-coordinated.

Looking at rate from another angle, Carver (1992) provides an interesting amalgam of terms. He states that *reading* means to look at words and determine their meaning, and *auding* means to listen to words and determine their meaning. By fusing reading and auding together, Carver coins the term *rauding*, which focuses upon the fact that the comprehension processes underlying typical reading and auding are the same. Rauding, as explained by Carver, refers to the ability of the reader to know when to speed up or when to slow down, when to pause, or when to raise or lower voice intonation, based on what is happening in the passage. Rauding requires the ability to look ahead or to feel what is happening in the story. The rauding, as defined by Carver, suggests that the student comprehends what is happening and knows how to express the emotions appropriate to that passage.

It would seem that the student knowing when and how to raud relates directly to comprehension. According to Shapiro (1989) and Skinner, Cooper, and Cole (1997), students' rates of accurate oral reading have been shown to correlate positively with a number of measures of reading skill such as (a) word identification, (b) word comprehension, (c) inferential comprehension, and (d) literal comprehension. Breznitz (1987) and Skinner et al. (1997) go on to state that increases in reading comprehension may result from increases in rates of reading. Such increases in reading rate and comprehension can only build students' self-confidence in reading.

Although noted earlier when discussing accuracy (word recognition), Logan (1997) stated that a student who possesses the properties of fluent word recognition is fast, effortless and autonomous in his or her reading with that student's fluent reading being so fast and effortless that he or she is unaware reading is actually taking place. Logan also states that a fast reading rate is a needed criterion for automaticity. But what is a fast or appropriate rate of speed supposed to be in order to attain automaticity?

Logan states that there isn't one by explaining that it is very hard to secure a fixed criterion for how fast a reading fluency process must be to be viewed as automatic. Each individual is different and each will vary in speed or rate of reading. Although there is not an absolute criterion for how fast a rate should be in order to obtain automaticity, without automaticity fluency does not occur.

Despite the ambiguity of this last statement, Logan (1997) states unequivocally that there are four properties needed for automaticity to occur: "speed [a proper or fast rate], effortlessness, autonomy, and lack of conscious awareness" (p. 124). Thus, a student possessing these four properties will more likely be able to adjust his or her reading speed according to the text, adhering to the author's syntax (National Assessment of Educational Progress, 2004; White, 2004). The student who possesses automaticity in fluency will know how to adjust the rate of reading speed with little conscious thought appropriate to the situation. For the individual, knowing when to adjust rate and knowing how to read become a developed inherent attribute that grows with time and practice (National Reading Panel, 2000).

Phrasing. The National Assessment of Educational Progress (2004) (NAEP) defines fluency "as the ease or 'naturalness' of reading" (p. 1). The NAEP report goes on to state that the key elements of fluency include the reader's adherence to an author's syntax. Such adherence implies the proper use of intonation, stress, and pauses or proper grouping and phrasing.

While researching fluent oral reading, the NAEP (2004) worked with 1,136 fourth-grade students. The students were given an opportunity to read a selected story silently, then out loud, then answer questions based on that story. They were then instructed to read the story out loud again as if they were reading to someone who had never before heard the story. Their performance was tape-recorded and later rated and scored.

After having established four different rating levels on their own oral reading fluency scale, with 1 being the lowest and 4 being the highest, they found that 55% of their students read at the higher levels of 3 and 4. These students “read in larger phrase groups that consistently preserved the author’s syntax, and they read some or most of the story with expressive interpretation” (p. 2). The study showed further that higher levels of fluency and higher levels of phrasing were associated with higher average reading proficiency scores (NAEP, 2004; White, 2004).

Thirteen years prior to this study Zutell and Rasinski (1991), in addition to noting that the reading performance of those possessing excellent fluent oral reading skills appear effortless or automatic, also noted that those possessing excellent fluent oral reading skills grouped “words into meaningful phrases or clauses” (p. 212). They also stated that good fluency skills included an understanding of how to “project the phrasing and expression of the spoken word upon the written word” (Richards, 2000, p. 534).

Stressing the importance of phrasing when seeking to build fluency skills in students, Johns and Berglund (2002) state that phrasing is an aspect necessary for fluency to occur. They concur with Zutell and Rasinski's (1991) earlier proposition that phrasing involves the chunking or grouping of words into meaningful clusters and appropriate phrases, helping students understand better what they read. Further, Shanahan (2000) states that in order to understand specific passages in reading, the reader must possess appropriate phrasing skills.

Phrasing is necessary for fluent oral reading and fluent oral reading is a contributing factor necessary for comprehension. Cromer (1970), O'Shea and Sindelar (1983) and Rasinski (1990) concluded from their separate research efforts that when text is segmented into appropriate phrasal units by slow but accurate readers, readers experience improved comprehension.

Expression. Reading with expression is derived from the concept of prosody. According to Dowhower (1991), prosody is a general linguistic term to describe rhythmic and tonal features of speech. Because the elements usually cover more than one phoneme segment (e.g., syllables, words, and larger units of speech), they are also called *suprasegmental features*. Prosodic features involve variations in pitch (intonation), stress (loudness), and duration (timing). When these suprasegmental features are present in fluent reading, the term *prosodic reading* is applied (Dowhower, 1987). Prosodic reading, [then], is the ability to read in expressive rhythmic and melodic patterns—educators call it reading with expression. (p. 166)

The National Assessment of Educational Progress (2004) and White (2004) both stress that when an individual reads with good fluency, he or she reads with a natural ease. They also state that one of fluency's key elements is indicated by the expressiveness of one's oral reading, which provides a sense of feeling, anticipation, or characterization.

Four years prior to the NAEP's and White's descriptions of key elements of fluency, the National Reading Panel (2000) stressed in its report the importance of how vital a role proper expression plays in the development of fluency. The NRP indicated that a fluent reader is one who is able to read orally with speed, accuracy, and proper expression, and that those students who develop those skills usually do so by using guided repeated oral reading. The NRP continues by asserting that such practice can substantially improve word recognition, fluency and even reading comprehension.

The NRP's (2000) emphasis on proper expression and fluency appears to be justified when one reads Allington's (1984) earlier assertion that even though most educators have always regarded reading with expression as a necessary and defining feature of skilled fluent reading, defining fluency as being able to read a text quickly, accurately, and with proper expression, fluency remains the most neglected reading skill taught in the classroom. Adding to Allington's stated concerns, Dowhower (1991) asserts

that reading with expression, as a component of fluency “has been a vague instructional phrase, rarely defined explicitly either by teachers or by texts on teaching reading” (p. 165).

Adding strength to the argument that expression is a key element for fluency to occur, Johns and Berglund (2002) use the terms “Radio Reading” (p. 37) and “Say It Like the Character” (p. 43) to describe ways in which to build fluency within a student. The basics for each are very similar: Whether acting as a radio announcer or acting out a character’s part in a play or reader’s theater, a student must read with (a) expression, (b) intonation, (c) proper pitch, and (d) proper stress either to get across a message to a radio audience or the emotions that a character in a play may be experiencing. If read out loud without proper expression, the message may not be understood by the intended audience or the feelings of the character in the play may never be communicated.

Finally, Dwyer (2004) includes expression as part of his rubric for assessing fluent oral reading. He states that an individual who reads with fluency does so by adjusting his tone, inflection, rate, and even speed, which are the prosodic features that Dowhower (1991) spoke of when defining expression. This individual knows how to capture the intended meaning of the passage. Dwyer goes on to say that a fluent reader is confident in his or her choices of “expressive intent, and uses it well to incorporate an oral interpretation of the text that is evident in the reading” (p. 2).

Confidence. Obviously, confidence is not unique to fluent oral reading. It is usually discussed as an outcome resulting from students being positively motivated and experiencing success after practice and work on the other skills that *are* unique to fluent oral reading (LaBerge & Samuels, 1985; Richards, 2000; Samuels, 1997; Zutell & Rasinski, 1991). Although the research literature does not view confidence as an aspect of fluent oral reading per se, the author asserts that it should be included as a necessary component. A student may be a fluent reader and able to practice by him or herself out loud when no one else is present. However, if that student is shy or not sure of him or

herself in front of others, despite having great fluent oral reading skills, attempts made to read out loud in front of others could be discouraging, thus convincing a student not to read out loud at all. If a student were to have poor fluent oral reading skills, the chances are less likely that he or she would be willing to try to read in front of others (Rinehart, 1999).

Confidence in reading builds with repeated practice and improving performance as does any skill. For example, picture an infant learning to walk. He lives in an environment where learning to walk is highly encouraged and expected. Thus, as he practices, his ability to walk increases and so does his confidence. As confidence increases, he tends to pick up the pace, walking greater distances as time goes on. The infant may stumble at times, but with encouragement from parents and family and further practice, stumbles grow fewer and farther between. Then, when a stumble does occur, he learns to self-correct or avoid the problem that may have caused the stumble and grows more confident in increasing his pace and the distances he traverses while walking. The skills necessary for walking gradually move from conscious awareness to unconscious awareness, a state of automaticity.

To a great extent, achieving fluency in reading is much the same. According to Blum and Koskinen (1991), if an instructional setting fosters expertise it is likely that students will understand what they read, learn strategies to improve their reading, feel successful, and be motivated to practice, thus increasing their confidence. Blum and Koskinen go on to stress the importance of regarding oral fluency as a necessary attribute of good reading, emphasizing that a student can be helped to attain fluency through training. Such fluency training has the affect of improving overall reading ability, comprehension, and greater retention of vocabulary. Such an environment, they continue, allows the student to feel more confident about his or her reading and is an activity in which the student will want to participate.

Practice in fluent oral reading motivates students to read and increases their confidence. The theory of automatic information processing as developed by LaBerge and Samuels (1985) emphasized the importance of practice. Practice enables beginning or struggling readers to achieve a level of automaticity in decoding so that they can focus attention on comprehension. LaBerge and Samuels support what Rinehart (1999) and his college and elementary students experienced when using a readers' theater format to improve confidence in fluency, based on successes experienced by the elementary students in that environment.

Two years prior to Rinehart's (1999) action research, Au (1997) stated that when students grow in their self-confidence and command in reading, to include fluent oral reading, they claim greater ownership of their reading skills. As confidence increases, and students experience success after success, their desire and interest in fluent oral reading naturally increases (National Reading Panel, 2000).

The author asserts that confidence can be considered as a key dimension to fluency, as are accuracy, smoothness, rate, phrasing and expression, all of which contribute to the acquisition of fluent oral reading.

Issues Pertaining to a Domain Theory

According to Messick (1995), to understand a domain of learning is to have an understanding of a domain theory. Bunderson and Newby (2005, in press) state that a domain theory is a descriptive theory which provides the contents, substantive processes, and boundaries of a field or area of interest in human learning and growth. A descriptive theory is a middle-range empirical theory that requires a descriptive or *what is* empirical research design. It is the most basic type of middle-range theory (Fawcett, 1999).

According to Polit and Hungler (1995) a descriptive theory addresses three questions: "What are the characteristics of the phenomenon? What is the prevalence of the phenomenon? What is the process [the order, if any] by which the phenomenon is experienced?" (p. 12). To answer these questions, a descriptive theory makes use of

several different kinds of methods, including but not limited to, concept analysis and psychometric (measurement) analyses (Fawcett).

A descriptive theory provides an account of construct-relevant sources of task difficulty and an account of the substantive processes operative at different levels of growth along the scale(s) that span a domain of human learning or growth. Based on measurement instruments (scales) linked to the constructs unique to a domain, “testable predictions can be made about the relationships between tasks, processes and locations along [those] scales” (Bunderson & Newby, 2005, in press, p. 5).

A domain can be thought of as a sphere of activity, concern, or function. A domain infers that it is an area of human activity or an area of academic interest or specialization. A domain theory establishes and identifies the invisible yet important mental processes related to human practices or attributes hypothesized to exist, which researchers desire to measure (Bunderson & Newby, 2005, in press).

The term *construct* is both a noun and a verb. Researchers are not able to examine directly every aspect of a domain. They must at times construct (the verb) conceptual ideas in words and drawings, calling them constructs (the noun) in order to communicate about them in terms of explaining the dimensions or aspects of a domain to others (Bunderson & Newby, 2005, in press).

According to Cronbach (1984), the noun construct comes from the word “construe; a construct is a way of construing—organizing—what has been observed” (p. 133). Despite Cronbach’s use of the word *observed*, what is observed is not always meant to be visible and concrete. For, according to Nunnally and Bernstein (1994), if a variable or aspect is abstract or latent, it is termed a construct. A construct is a concept from scientists’ imaginations used when they attempt to describe a dimension of human behavior only detectable through its effects on observables.

Nunnally and Bernstein (1994) add that scientists populate their theories with constructs. They go on to say that any theory is comprised of two components: “the

measurement component that dictates what constructs are to be measured and the *structural* component that describes the properties of the resulting measures in terms of how constructs interrelate” [italics added] (p. 85). Bunderson and Newby (2005, in press) concur with Nunnally and Bernstein (1994) by saying that when measurement scales are created, they must be linked to the identified constructs of the domain in question so that numbers assigned by the measurement instrument can be interpreted. By creating and documenting the order of difficulty or complexity of these constructs in a theory of the domain, the instrument will have a stronger legitimate claim to meaningful interpretability.

Domain theory is consistent with an alternative to logical empiricism as the usual philosophy of science. Trout (1998) offers *measured realism* as such an alternative. Domain theory also connects with fundamental measurement theory (Krantz, Luce, Suppes, & Tversky, 1971; Luce & Tukey, 1964; Wright, 1999), and a cyclical research methodology similar to Brown's (1992) design experiment framework (Bunderson, 2000, April; Kelly, 2003; Bunderson & Newby, 2005, in press). At its core domain theory is a mathematically robust description of the nature of learner growth in expertise through a domain.

When putting the aforementioned definitions together as a whole, it easy to understand why Bunderson and Newby (2005, in press) used such examples as “calculus, American history, accounting, network engineering, and nursing” (p. 5) when describing examples of a domain. Domain theory implies an area or domain of human learning and growth, or human development. In the area of human learning and growth in fluent oral reading, Kame'enui and Simmons (2001) state “fluent reading is plainly developmental and represents an outcome of well-specified sub-lexical and lexical processes and skills *developed* [italics added] for most children over a bounded period of pedagogical time (e.g. kindergarten to grade six)” (p. 204).

Thus, this study is an investigation of fluent oral reading. The author proposes a domain of human learning and growth complete with the constructs of word accuracy, smoothness, rate (speed), phrasing, expression, and confidence.

With regard to the various methods used to address the questions a descriptive theory presents, Messick (1995) and Bunderson and Newby (2005, in press) provide a means by which to address those questions: the theoretical concept known as construct validity (Messick), and the method for designing, developing, and validating instruments and systems that integrate learning with assessment known as validity-centered design (Bunderson and Newby).

Construct validity. Messick (1989) showed that construct validity was the core of all other aspects of validity and provided the measurement world with six seminal aspects or arguments for construct validity: (a) content, (b) substantive processes, (c) structural, (d) generalizability, (e) external, and (f) consequential. Then he added, “These six serve to function as general validity criteria or standards for all educational and psychological measurement” (pp. 744-745). Messick also talked about the concept of unified validity where he clarified that when an educational measurement instrument is unified, it will be appealing, easy to use, and very user centered. Bunderson placed Messick’s original six aspects of construct validity under two categories: Design for Inherent Construct Validity, comprising of the content, substantive processes and structural validity arguments; and Design for Evidence of Criterion-Related Validity, comprised of the generalizability, external, and consequential validity arguments. He added another category, Design for Usability and Appeal, which encompassed other aspects of the unified validity concept not included in Messick’s six aspects of construct validity.

Validity-centered design. Honoring Messick’s contribution to measurement in education and psychology, Bunderson (2005, in press) proposes an extension of Messick’s unified validity ideas, calling it validity-centered design specifically adding three additional needed aspects of construct validity: overall appeal, usability, and

perceived value, which become the first three aspects, with Messick's original six filling numbers four through nine. According to Bunderson these nine aspects of validity are achieved through a cyclical and continuing design process. The nine goals of validity-centered design are organized into Table 1. The research questions in this study deal strictly with category II. A brief description of category II follows.

Table 1

Validity-Centered Design

Category	Aspects requiring validity argument
I. Design for Usability, Appeal, and Positive Expectations	<ol style="list-style-type: none"> 1. Overall Appeal 2. Usability 3. Perceived Value
II. Design for Inherent Construct Validity	<ol style="list-style-type: none"> 4. Content Aspects 5. Substantive Processes 6. Structural Aspects
III. Design for Evidence of Criterion-Related Validity	<ol style="list-style-type: none"> 7. Generalizability 8. External Aspects 9. Consequential Aspects

Category II of Validity-Centered Design

Design for inherent construct validity. Construct validity is the link between reality and the scores or measures produced by an instrument. This aspect of validity starts with a blueprint. In the blueprint, researchers ask questions dealing with human learning and growth issues specific to the domain. For example, how do we measure the important invisible mental processes related to the valued human practices hypothesized to exist in users when they perform key tasks in the fluent oral reading with expression domain? Do the scales we construct through scoring the questions connect with important aspects of reality? Bunderson (2005, in press). There are three aspects to the blueprint: (a) *content*, which deals with coverage and appropriateness of the tasks and objectives in the domain, (b) *substantive processes*, which address the important but usually invisible mental processes operative while performing tasks in the domain (the unobservable non-concrete constructs discussed earlier), and (c) *structure of the constructs*, which acknowledges that the initial number of construct-linked scales may collapse into a smaller number of separate unidimensional measurement scales.

The structure of the constructs indicates that measurement scales should correspond with a hypothesized, and later validated structure, whose validity argument increases over time. The development of the FORE rating system must be deeply founded in content coverage and appropriateness. It is also founded in substantive processes: the important mental processes used by those whom we would wish to score as more successful on an instrument (Bunderson, 2002). Such a rating system further should be based on structural validity and reliability. It must be easy to use and easy to understand. The reader may wish to refer to Appendix A for a history of the development of the FORE measurement instrument.

Content validity. According to Lennon (1956) and Messick (1989) this aspect of construct validity includes evidence of content relevance, or representativeness and technical quality wherein the key issues are specifications of boundaries of the constructs

or attributes such as (a) knowledge, (b) skills, (c) attitudes, (d) motives, (e) job analysis, (f) task analysis, (g) curriculum analysis, and (h) according to Messick (1995) "... especially domain theory, in other words, scientific inquiry into the nature of the domain processes and the ways in which they combine to produce effects or outcomes".

According to Brunswick (1956), the content validity aspect of construct validity aids educators in assembling tasks that are relevant to the construct domain. It assists them in delineating what people actually do in the performance domain or what characterizes and differentiates expertise in any given domain.

Substantive process validity. The substantive process validity aspect adds to content validity the need for empirical evidence of response consistencies or performance regularities reflective of domain processes (Loevinger, 1957). According to Embretson (1983), substantive process validity provides the theoretical rationale for the observed consistencies in test responses (to include assessment measuring instruments) as well as the process models of task performance. Substantive process validity also provides the empirical evidence that the subjects are actually responding to or participating within the theoretical framework unique to the particular assessment tasks at hand.

Finally, Messick (1995) weighs in on the importance of the substantive aspect by stating that the core concept unique to this particular aspect is *representativeness* with two distinct meanings: the cognitive psychologist's sense of representation (Suppes, Pavel, & Falmagne, 1994), and the Brunswickian sense of ecological sampling or rather, covering all the important parts of the particular domain under study (Brunswick, 1956).

Structural validity. The third aspect of construct validity used in domain theory development and validity-centered design is known as structural validity, wherein the fidelity of the scoring structure to the structure of the construct is appraised (Loevinger, 1957; Messick, 1989). When a researcher develops a measuring instrument of any kind, the instrument should match as close as possible all sub-constructs within a given construct and strictly adhere to the undergirding theory supporting it. In other words,

substantive process validity comes into play when structural validity demands that scoring models be rationally consistent with what is known about the structural relationships inherent in behavioral manifestations of the construct in question (Loevinger, 1957; Peak, 1953).

As Messick (1995) states, “the theory of the construct domain should guide not only the selection or construction of relevant assessment tasks but also the rational development of construct-based scoring criteria and rubrics” (p. 746). Messick’s statement ties in with what he earlier explained, stating that the internal structure of the assessment should be consistent with what is known about the internal structure of the construct domain itself (Messick, 1989).

In conclusion, this research study represents an attempt to define and delineate the characteristics of the domain of fluent oral reading with expression. The researcher demonstrates that by applying category II of validity-centered design both a viable domain theory of learning and growth and an accompanying measurement instrument emerge, assisting the field as it refocuses on the issues pertaining to expressive fluent oral reading.

CHAPTER 3

METHOD

Design

Chapter 3 contains the methods used to conduct this research. It describes the sample of students participating in this study and the measures taken to protect human subjects according to the university's institutional review board policies. Procedures for task administration are herein laid out. These procedures include assessing students' fluent oral reading skills, developing the measurement instrument, selecting and training raters, and collecting and analyzing data. The chapter also includes the methods used to address each research question.

Sample

A total of 200 students in grades 2 through 6 from a student body of 445 participated in this study. The participants were native English speaking and/or second language English acquisition students who were recommended by their individual classroom teachers as being sufficiently fluent in English to be able to read comfortably from English texts out loud. The students were asked to read out loud and be videotaped while doing so. The procedures for obtaining permissions for participation in this study were carefully followed.

The students attended a regular, mid-size central city school located in a city in central Utah. Of the 445 students enrolled, 244 were male and 201 were female. Of those participating in this study, 108 were male and 92 were female. Both Kindergarten and Grade 1 were excluded due to limited development of necessary reading fluency skills. Table 2 provides student participation rates by grade and gender.

Of the 445 students enrolled in the school, 51 students were free-lunch eligible representing 12% of the total student enrollment. Thirty-three students were reduced-price lunch eligible representing 7% of the total student enrollment. Socio-economic status information by grade was not available.

Table 2

Student Participation Numbers and Rates by Grade Enrollment and Gender

Grade	No. of students Enrolled	No. of students participating	% Male participation	% Female participation
2	61	38	33%	30%
3	53	45	51%	34%
4	65	55	52%	32%
5	62	34	19%	36%
6	66	28	23%	20%
Total	307	200	36%	30%

Procedures

Obtaining student participation. In accordance with the policies of the Institutional Review Board of Brigham Young University, permission was obtained from the participating school district, participating school principal, parents, and students (see Appendix B for supporting documentation). Students in second through sixth grades were given two informed-consent forms: one for the parent or legal guardian to sign and one for the student to sign. Every student who returned the two forms was given a candy bar regardless of whether or not he or she was allowed to participate. Once the forms were

collected, they were arranged into a random order. Students were called upon to participate based on that order and whether or not they were in attendance on the day that participation was needed. If an authorized student was absent he or she was scheduled in to participate on a different day.

Text selection. Teachers in the second through sixth grade were asked to supply texts from which they had been or were currently reading with their students. The researcher used these texts and selections from books to obtain excerpts for student use. The reading excerpts chosen for this study came from five sources: (a) Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (University of Oregon, 2000), (b) Highlights for Children (2003), (c) A Wrinkle in Time (L'Engle, 1973), and (d) third-grade reading primers. Reading grade levels and Lexile measures are displayed in Table 3. All reading selections are shown in their entirety in Appendix C.

Table 3

Reading Selections and Grade Levels

Text name	Reading level	Lexile measure
The Ant Hill	Grade 1	640L
Going to the Swimming Pool	Grade 2	750L
The Sun	Grade 3	820L
Brave Irene	Grade 3	1000L
A Brick to Cuddle Up To	Grade 4	910L
A Wrinkle in Time	Grades 5 - 6	910L

Videotaping student reading performance. Each student received the described interventions listed in one session, found below. The interventions varied from student to student based on how fluently the student read out loud. For the purposes of this study and at the request of the elementary grade teachers, a session lasted no more than 15 minutes per student. The procedure used to obtain the samples of oral reading included the following steps:

1. Each student was asked to sit in a chair beside the researcher with a digital video camera four to five feet away in front, placed on a tripod, facing the student and the researcher.
2. Each student was provided with a text at his or her grade level, then filming began.
3. The reading selection was placed on a clipboard with a microphone connected by wire to the digital camera.
4. The researcher pointed to 10 to 15 randomly-chosen words on the page.
 - A. The student read the selected words aloud.
 - B. If the student were able to say the words correctly with minimal to no prompting from the researcher, the researcher moved the student to step 5.
 - C. If the student were not able to say a majority of the words correctly, a lower grade-level text was provided.
 - D. Steps 4a through 4c were repeated until the student could accurately read a majority of the words.
5. The student was then asked to read through the selection silently to become familiar with the material. The filming stopped until the student finished reading the selection.
6. The filming started again and the student was asked to read aloud the first paragraph from the reading selection.

- A. If the student read the passage with smoothness, as determined by the researcher's professional judgment, the researcher allowed the student to continue with the selection.
 - B. If not, the student was given a lower grade selection from which to read until a text was found that presented a comfortable reading level for the student.
7. The student was then thanked for his or her willingness to help and sent back to the classroom.
 8. The cycle was then repeated for the next student.

Upon completion of the videotaping, videotape footage was transferred from the digital video cassettes using a Macintosh PowerBook G-4 and its related iMOVIE video-editing software. Once the video footage was arranged into appropriate movie lengths for DVD burning the author used iDVD software installed on the Macintosh PowerBook G-4 to create the DVDs necessary for the raters to use.

Selection of raters. Four raters were chosen who either had experience with the McBride Reading Program or in tutoring children with reading problems. The constructs used in the instrument, while derived and confirmed in large part through the literature review, are well understood by practitioners of the McBride program. In addition, the tutoring procedures listed in this chapter are consistent with practices in the McBride program. The experience of the selected raters with this program ranged from teaching students privately to home schooling their own children in reading. The raters consisted of three females and one male and ranged in age from 31 to 85. Each rater lives in the southeastern part of the United States.

Instrument development. FORE measurement instrument (FMI) versions I and II were never tested, but were used in the development of version III-A. A pilot study using FMI-III-A led to one other version, III-B, based on some minor changes in the wording of the instrument. Version III-B was to be used in a validation study. However, the raters who were selected for that study suggested further improvements, resulting in version IV.

A more detailed history of the development of these instruments is found in Appendix A. The FMI-V (Figure 1) is a product that reflects the true intent of validity-centered design wherein both the researcher and the subjects conjointly participate in the process of making changes in instrument format and wording to improve usability. Through a series of practice sessions rating student videos, adjudications to compare results from rater to rater, and discussions, the team of researcher and raters sought to improve each rating scale. The key was to clarify each observable aspect of the six FORE constructs. The FMI-V is the result of this endeavor.

The many changes in instrument wording and format as it evolved over this series of studies are documented in another paper (McBride, 2005). This paper describes the development of the FORE instrument through version V and the interplay of this development process with rater training materials and procedures.

Training of raters. The training took place in the individual raters' homes, one-by-one and as a group wherein a detailed discussion about domain theory and constructs took place. The construct terms were then related to fluent oral reading. Each of the constructs and sub-constructs as they appear in version FMI-V was explained to and then reviewed by each rater. They were required to explain back to the researcher what each of the constructs meant. The raters then viewed a video clip of students reading, while the trainer identified the constructs. Raters were shown both good and bad examples of fluent oral reading constructs, as displayed by the students. Discussion took place regarding how each of the rating categories should be interpreted and applied.

Practice was provided through three cycles of rating followed by discussion. Ten students were rated in each cycle. The researcher directed the raters to view ten students that were randomly chosen. Upon completion of the ratings of each of the 10 students by each of the four raters, the researcher met with the raters for adjudication and further training. Following each session the ratings given were discussed and compared.

Rater Number:		FORE MEASUREMENT INSTRUMENT - Ver. V		Student Number:	
Accuracy		Smoothness		Expression	
Automaticity in recognizing words		Looks ahead		Stress and intonation	
5	Automatically recognizes all words	5	Consistently looks ahead	5	appropriate stress & intonation
4	Independently corrects errors	4	Mostly looks ahead	4	mostly appropriate stress & Intonation
3	Some errors with some correction	3	Sometimes looks ahead	3	Inappropriate stress and intonation
2	Lots of errors, some self-correction	2	Choppy	2	Little stress and intonation
1	Lots of errors, no self correction	1	Word - by - word	1	Monotone - same note
Pronounces words correctly		No repetitions		Conversational in manner	
5	Pronounces All words correctly	5	No repetitions	5	Always natural expression
4	Independently corrects errors	4	Repetitions mildly impede smoothness	4	Mostly natural expression
3	Some errors with some correction	3	Moderately impede smoothness	3	Somewhat natural expression
2	Lots of errors, some self-correction	2	Significantly impede smoothness	2	Little natural expression, mostly forced
1	Lots of errors, no self correction	1	Severely impede smoothness	1	Monotone - same note
Reads text as written		No Elongations (words or pauses)			
5	No omissions or inserts	5	No elongations		
4	Independently corrects errors	4	Elongations mildly impede smoothness		
3	Some errors with some correction	3	Moderately impede smoothness		
2	Lots of errors, some self-correction	2	Significantly impede smoothness		
1	Lots of errors, no self correction	1	Severely impede smoothness		
Phrasing		Rate		Confidence	
Punctuation		Pace		Student is sure of him/herself	
5	Observes All punctuation	5	Always appropriate to situation	5	Bold
4	Observes Most punctuation	4	Almost Always appropriate	4	Slightly Hesitant
3	Observes Some punctuation	3	Sometimes appropriate	3	Somewhat Hesitant
2	Omits Most punctuation	2	Almost Never appropriate	2	Mostly Hesitant
1	Omits All punctuation	1	Never appropriate	1	Timid
Phrase boundaries		No inappropriate breaths		Commands positive attention from others	
5	Strong sense of phrase boundaries	5	No inappropriate breaths	5	Riveting
4	Good sense of phrase boundaries	4	Occasional inappropriate breaths	4	Mostly interesting
3	Some sense of phrase boundaries	3	Mid-sentence pauses for breaths	3	Somewhat interesting
2	Weak sense of phrase boundaries	2	Multiple breaths in sentences	2	Barely interesting
1	No sense of phrase boundaries	1	Breaths between most words	1	Boring

Figure 1. FORE measurement instrument version V (FMI-V)

Formative improvements to the instrument were suggested during each session, and minor revisions of the instrument made each time. Each session served to test the revisions, as well as giving the raters further training. The raters requested that the third session consist of students who displayed obvious problems in fluent oral reading, and this was done.

The raters then proceeded to rate the 200 students. Upon completion of most of the second and third grade, the researcher met with the raters and adjudicated the ratings one last time. Ratings that had a spread of two or more points were identified and discussed with the pertinent raters until the disparities were minimized to a spread of no more than one point. At this point, the time consuming and costly adjudication sessions were discontinued for the remainder of the student ratings, relying on the training effect of the four previous adjudication sessions to carry through with common interpretations and with common standards. The inter-rater reliability results in the next section provide evidence of the extent to which this goal was obtained.

Collection and Preparation of Data for Analysis

From the original ratings provided by the raters using the FMI-V, the researcher transcribed the ratings into an Excel spreadsheet. With the help of a fellow researcher, the rating inputs were double-checked for accuracy and input errors were corrected. With the help of two fellow researchers, the original data that was input into the Excel spreadsheet was reformatted for use with the SPSS and Facets software.

Each rater was given a different order in which to view the students in an effort to reduce possible order effects while rating the students. Raters were then asked to rate each of the students using the FMI-V.

As previously described, adjudication was used with three sets of 10 students each in an iterative process of training and formative instrument changes. The results of these practice ratings were not retained for analysis. Once the initial ratings of second and third graders were completed the author met with the raters and adjudicated the scores where

needed. This last adjudication served as a final effort to assure as much reliability and common interpretation of constructs in the rest of the ratings as possible. No further instrument changes were made, and these ratings were retained for analysis.

Myford and Wolfe (2002) support adjudication by emphasizing that “While it is important to review those cases in which there is obvious rater disagreement, in many instances such disagreements can be readily resolved by adjusting scores for differences in rater severity” (p. 319). Thus, adjudication was necessary because the students in this study received scores from raters on several of the various scales that differed in two or more points from the ratings given by other raters on the same scales. It was important that a discussion take place with the goal that raters agree on ratings and adjust disparities so that the adjudicated ratings were either the same or consisted of only one score point difference.

McNamara (1996) supports the views later espoused by Myford and Wolfe (2002) about the need for adjudication stating that even when raters receive proper training, significant differences may still exist in the ratings they assign to a ratee. McNamara goes on to state that adjudication is needed to decrease extraneous differences between raters. In this study the adjudication process was used as a part of the training. It was not possible to carry it through the adjudication of all 200 students and all four raters.

Methods for Addressing Research Questions

Research question 1. What is (a) the inter-rater reliability across the four raters for each of the 14 indicators? (b) What is the internal consistency of the measures of fluent oral reading constructs (accuracy, smoothness, phrasing, rate, confidence, and expression) and dimensions in the FORE measurement instrument? (c) Are the raters interchangeable in their ratings of students in terms of rater leniency and severity? If not, what are the systematic differences among the raters?

This question deals with the extent to which the ratings are reliable and thus permit the use of average ratings in research question 2. Part a of research question 1

addresses the average inter-rater reliability across the four raters for each of the 14 indicators. The author used the reliability program found within the SPSS Graduate Pack 12.0 for Windows to perform an inter-rater reliability analysis on the initial ratings. For each of the fourteen indicator scales, four columns of ratings across all students were entered into the reliability program. This program provided correlations between the different raters, as well as the internal consistency reliability coefficient across the four raters. Part b addresses the internal consistency of the measures of fluent oral reading across constructs and dimensions. Again, the researcher used the reliability program found within the SPSS Graduate Pack 12.0 for Windows. He totaled the scores each rater gave a student for each of the FORE indicators then computed the average of the scores. He used these averages to determine the internal consistency of constructs and scales for both fluency and accuracy.

The Facets software enabled the researcher to determine to what extent the raters differed in terms of relative leniency or severity (research question 1c). This Facets analysis also identified the ordering of the individual indicators. This provides a preliminary answer to research question 3.

In summary, research question 1 deals with category II of validity-centered design which includes content, substantive process, and structural aspects of validity. Although methods for assessing content validity empirically are not a part of this study, content aspects were considered in the literature review. Substantive process validity and structural validity are the key parts of this study. Substantive process issues deal directly with initial scale construction. Are the indicators developed for each scale representing subordinate aspects of the main dimensions, internally consistent, and reliable? What about the two main dimensions of accuracy and fluency?

Research question 2. How many dimensions of accuracy and fluency are sufficient to describe the domain of fluent oral reading with expression for students in grades 2 through 6? This question deals with the structural aspect of validity. The

methods for addressing research question 2 included performing a factor analysis with an oblique rotation to show the number of factors and the correlations between two or more factors. This method necessitated conducting a Kaiser Test to determine the number of eigenvalues greater than 1.0. It also necessitated addressing the meaningfulness of the resulting factors.

Using the SPSS Factor Analysis program, a Principal Axis Factor (PAF) method with Promax rotation was implemented. The researcher used the PAF method because he wanted to begin with communalities in each of the diagonals of the correlation matrix instead of 1's. When 1's are used the unique variance found in each rating scale is incorporated into the factors extracted. This unique variance is not of interest in studies like this one where the variables (individual rating scales) have been carefully designed to link to constructs in a theorized structure.

That which is common to the fluency rating scales as a whole, and to each subordinate construct within fluency, is of primary interest, so communalities rather than ones (1's) were used in the diagonal of the correlation matrix. The researcher used the Promax rotation method because factor extraction methods, including PAF, extract uncorrelated or orthogonal linear combinations of FORE's observed variables (indicators). The Promax rotation provides an oblique rotation (adjusting these linear combinations in a Procrustean manner) that allows factors to be correlated. It provides a confirmatory step toward the hypothesis stated earlier concerning the fluent oral reading dimensions of accuracy and fluency. The hypothesized correlation between the accuracy and any fluency dimensions can be determined directly from the Promax solution. The Promax rotation was selected rather than the direct Oblimin rotation (another oblique rotation program). If a direct Oblimin rotation were used the researcher would have had to guess at providing a delta or k , and the researcher did not wish to do so, thus his choice in using the Promax rotation method.

The literature review showed a number of investigators who argued that the two dimensions of accuracy and fluency, while inter-related, should be considered separately. The hope and expectation in this study is that there are two dimensions. By factor analyzing the three accuracy rating scales in the same analysis with the 11 fluency rating scales, the hypothesis that accuracy and fluency are separate but related dimensions of fluent oral reading may be verified. Evidence from the inter-rater reliability analysis was used to determine if it were appropriate for the researcher to use the average ratings on each of the 14 scales in performing the factor analysis.

Research question 3. Using the features of the Facets software, how are the average levels of rating scales that make up each sub-construct located along the dimension(s) of fluent oral reading with expression?

The researcher applied the Facets program to perform a many-faceted Rasch analysis. A Facets model was constructed which grouped the individual descriptors for the subordinate constructs of accuracy, smoothness, rate, phrasing, confidence and expression together. The model in effect displays the average of each group of descriptors pertaining to a single aspect along a single dimension. If the factor analysis shows only two factors, and if these are highly correlated, only one Facets analysis needs be performed. It will show the ordering of the average location of the several accuracy scales, in comparison with the one to four scales for each of the fluency aspects.

Analysis of the data in this way provides information as to how the average levels of rating scales are located along a composite dimension which goes between the correlated dimensions of accuracy and fluency. This ordering provides also the developmental or learning sequence which is the goal of research question 3. If no ordering is found, but if all or most of the scales were to fall in the same location along the scale, it could provide evidence that no developmental sequence for fluent oral reading exists. If an ordering is found, it could provide evidence that a developmental sequence does exist and that a local learning theory (or domain theory) of fluent oral

reading with expression has empirical support. The theory would also have a measurement instrument with which to conduct future research.

CHAPTER 4

RESULTS

Research Question 1

The following analysis, while addressing research question 1, also addresses the content and substantive processes unique to the domain of fluent oral reading. Research question 1 asks: (a) What is the inter-rater reliability across the four raters for each of the fourteen indicators? (b) What is the internal consistency of the measures of fluent oral reading constructs (accuracy, smoothness, rate, phrasing, expression, confidence) and of the fluency dimension in the FORE measurement instrument? (c) Are the raters interchangeable in their ratings of students in terms of rater leniency and severity? If not, what are the systematic differences among the raters?

Table 4 lists the six fluent oral reading constructs and their related 14 indicators as well as the abbreviations used to denote them. The constructs are listed in the order hypothesized prior to the results of this study. Sequence numbers used are those found in the Facets output.

Question 1a: Inter-rater reliability. The reliability program of SPSS, release 12, was used to conduct the inter-rater reliability analysis. This reliability analysis helps determine how well raters' judgments on particular indicators or scales correlate with each other. It produces a summary reliability statistic known as Cronbach's alpha coefficient (Worthen et al., 1999). Table 4 reports that the alpha reliabilities for the 14 indicator scales range from .73 to .87, indicating that the ratings obtained from the four raters tended to be highly intercorrelated. In other words, the four raters were consistent with each other in a relative sense, meaning that the students who were rated above the mean on a particular indicator by one rater tended also to be rated above the mean of that indicator by other raters. Similarly, students who were rated below the mean by one rater

were generally rated below the mean by other raters. The alpha coefficient does not provide any evidence

Table 4

Alpha Coefficients across Four Raters (Inter-rater Reliabilities)

Constructs	Indicator sequence numbers	Indicator abbreviations	Indicators	Alpha coefficients
Accuracy				
	1	Acc1: AIRW	Acc1: Automaticity-in-recognizing-words	.80
	2	Acc2: PRON	Acc2: Pronunciation	.75
	3	Acc3: RTAW	Acc3: Reads-text-as-written	.78
Smoothness				
	4	Smo1: LA	Smo1: Looks-ahead	.86
	5	Smo2: NR	Smo2: No-repetitions	.79
	6	Smo3: NEWP	Smo3: No-elongated-words-or-pauses	.73
Rate				
	7	Rate1: PACE	Rate1: Pace	.81
	8	Rate2: NIB	Rate2: No-inappropriate-breaths	.76
Phrasing				
	9	Phr1: PUNCT	Phr1: Punctuation	.74
	10	Phr2: PB	Phr2: Phrase-boundaries	.77
Expression				
	11	Exp1: SAI	Exp1: Stress-and-intonation	.86
	12	Exp2: CIM	Exp2: Conversational-in-manner	.84
Confidence				
	13	Con1: SISS	Con1: Student-is-sure-of-self	.87
	14	Con2: CPA	Con2: Commands-positive-attention	.86

about whether some raters were more severe or more lenient than others. Further analysis is needed to determine whether some raters tended to systematically assign lower or higher ratings on the average. That issue is addressed in research question 1c.

Table 4 shows that half of the 14 fluency indicators have an inter-rater reliability coefficient of .80 or higher. Since half of the indicators are less than .80, it would be beneficial to focus to a greater extent on rater training in future related studies.

Under accuracy, raters display more agreement in *Acc1: Automaticity-in-recognizing-words* (Acc1: AIRW) than with its sister indicators. Under smoothness, there is a wide disparity. Rate also reveals a disparity in the reliability of its two indicators. Phrasing shows the least amount of rater consistency with both indicators falling below .80. Both expression and confidence reveal higher reliabilities (hovering around .86) as compared to the other FORE constructs.

Question 1b: Internal consistency of accuracy and of the five fluency constructs.

Using the mean rating for each of the 14 indicators computed by averaging the ratings obtained from the four raters on each indicator, it is useful to examine the correlations among these variables as a first step in considering the internal consistencies of the ratings. The correlations between all possible pairs of the 14 indicators are displayed in Table 5. The correlation matrix shows that most of the fluency indicators show high correlations with their sister fluency indicators, which is evident in the lower portion of the matrix. The accuracy indicators correlate well among themselves, but show less correlation with fluency's indicators. These findings support the author's hypothesis that accuracy and fluency should be considered as separate dimensions.

Table 5 also shows that *Smo2: No-repetitions* (Smo2: NR), has a low correlation with the accuracy indicators. Most significantly, Table 5 reveals that Smo2: NR has the lowest correlation with any of its sister fluency indicators.

Table 5

Correlations among the Mean Ratings of the 14 Indicators in the FORE Domain

FORE indicators	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Acc1: AIRW - 1.	-	.70	.62	.60	.43	.63	.56	.44	.21	.37	.40	.41	.59	.49
Acc2: PRON - 2.		-	.53	.33	.33	.33	.30	.20	.11	.21	.16	.16	.33	.23
Acc3: RTAW - 3.			-	.53	.46	.53	.42	.31	.29	.38	.36	.36	.44	.43
Smo1: LA - 4.				-	.34	.79	.82	.67	.42	.69	.75	.78	.88	.86
Smo2: NR - 5.					-	.32	.31	.32	.27	.28	.18	.23	.31	.29
Smo3: NEWP - 6.						-	.77	.60	.26	.52	.57	.60	.72	.69
Rate1: PACE - 7.							-	.65	.43	.66	.70	.75	.81	.81
Rate2: NIB - 8.								-	.42	.56	.56	.60	.66	.68
Phr1: PUNCT - 9.									-	.74	.55	.56	.38	.74
Phr2: PB - 10.										-	.72	.74	.64	.72
Exp1: SAI - 11.											-	.95	.79	.91
Exp2: CIM - 12.												-	.81	.92
Con1: SISS - 13.													-	.88
Con2: CPA - 14.														-

Using the mean ratings of the four raters on each of the 14 scales, the researcher computed the internal consistency reliability of the six constructs. He then computed the internal consistency reliability of all five fluency subconstructs together as if they could be considered one dimension. Table 6 reports the resulting alpha coefficients

Table 6

Internal Consistency Reliability Coefficients for Composite Constructs of Accuracy and Fluency

Facets variable	Composite construct summed over indicators	Alpha coefficient
1	Accuracy: 3 original indicators	.82
	Accuracy: 2 only (less Acc3: Reads-text-as-written)	.82
	Accuracy: 3 original plus Smo2-No-repetitions	.79
2	Smoothness: 3 original indicators	.74
	Smoothness: 2 only (less Smo2-No-repetitions)	.86
3	Rate: 2 original indicators	.78
4	Phrasing: 2 original indicators	.85
5	Expression: 2 original indicators	.97
6	Confidence: 2 original indicators	.94
	Fluency: 11 original indicators	.95
	Fluency: 10 indicators (less Smo2-No-repetitions)	.96
	Composite of Accuracy (A), Fluency (F), Phrasing (P)	.77
	AFP minus A	.81
	AFP minus F	.50
	AFP minus P	.70

calculated using the SPSS reliability program. Table 6 shows that the three accuracy indicators, when added together, have a reliability coefficient of .82. It also shows that when Smo2: NR is added to accuracy's composite, accuracy's score decreases to a reliability coefficient of .79. On the other hand, smoothness' reliability coefficient of .74 increases to that of .86 when Smo2: NR is deleted from the smoothness composite. The preceding discussion exemplifies how a reliability study using the average ratings taken from the fluent oral reading with expression construct-linked scales sheds light on the substantive processes, coherence, and internal consistencies of the six constructs, both separately and together. Information from the correlation matrix (see Table 5) shows that the Smo2: NR rating scale does not correlate highly with either the accuracy or fluency indicators, and may need to be dropped.

Question 1c: Systematic differences in rater means. Part c of question 1 concerns the rating leniency or severity of each of the raters. The Facets generated chart in Figure 2 displays output derived from analysis using a four-facet model. The common metric for all of the facets is the logit scale found in the first column starting from the left. The second column (examinees) shows the distribution of students (higher ability students at the top of the chart, lower ability students at the bottom of the chart). The third column gives the line up of the four raters, ranking in severity from top (severe) to bottom (lenient). The 14 different indicators as a facet are shown in the fourth column, corresponding to the labels used in Table 4. Indicators that were hardest for students to receive a high rating appear at the top, with the indicators that were the easiest for students to get a high rating at the bottom. The fifth column shows the locations in logits of the boundaries between the five rating scale rubric values, 1 through 5, with 1 and 5 being at the ends, having some indeterminacy.

Acc1: AIRW and Acc2: Pronunciation (Acc2: PRON) were the easiest FORE indicators in which students could obtain a high rating. This was expected inasmuch as

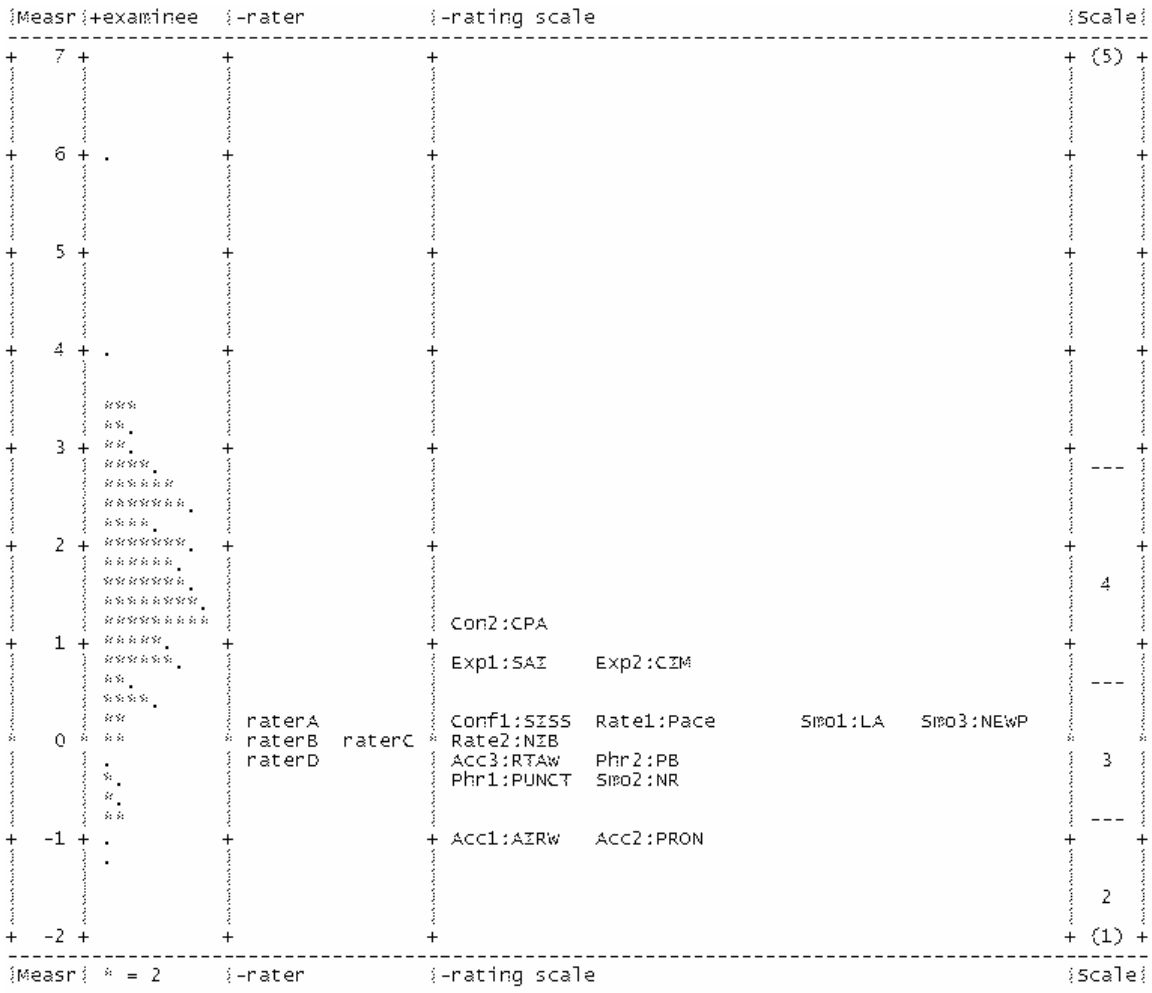


Figure 2. Facets generated chart displaying differences in leniency and severity of the four raters and also the difficulty levels of the 14 indicators

the researcher gave an informal pretest to the students at the beginning of each videotaped session to determine an independent or comfortable word recognition level of text for each. Greenwood, Abbot, and Tapia (2003) advise that fluency should only be assessed when students have texts they can read at an independent level (98% of the words correctly identified). Figure 2 provides evidence that the experimental manipulation to assure this condition was successful. The higher ratings in accuracy are not indications that it is easier than the other constructs, since a selection of harder texts could have made accuracy the most difficult scale to obtain a high rating.

A condensed version of the rater measurement report produced by *Facets* is shown in Table 7. (The complete version of this report is Facets generated Table 7.2.1, displayed in Figure E1). Each row in Table 7 corresponds to one of the four raters. The entries in the observed average column display the mean rating assigned by each rater averaged across all 14 indicators and all 200 students. These averages are reported in the same 1-5 metric as the original ratings. The fact that the mean rating assigned by rater A is lower than the means of the other three raters provides an initial indication of systematic differences between the raters.

Table 7

Abbreviated Rater Measurement Report from Facets Analysis

Rater	Observed average	Severity measure	Standard error
A	3.8	.17	.03
B	3.9	.07	.03
C	4.0	-.08	.03
D	4.0	-.17	.03

Note. Reliability of separation index = 0.96. Fixed chi-square = 89.4,

d.f. = 3, $p < .001$

The entries in the severity measure column are reported in logits. Positive numbers in this column are indicative of raters who assigned ratings that were more severe than the other raters, while negative numbers are indicative of raters who lenient compared to the other raters. The relative toughness of the four raters varies from Rater D who was most lenient with a measure of -0.17 to Rater A who was most severe with a measure of +0.17. Thus, the range in relative toughness was about one-third of a logit ($.17 - (-.17) = .34$).

The standard errors reported in Table 7 describe the precision of the estimated severity/leniency measures for each rater. In this case, the severity/leniency estimates for all four raters have a standard error of .03 and the distance between the most severe and the most lenient rater is equivalent to 11.33 standard errors.

Table 7 includes two other sources of evidence about the severity/leniency effect. The fixed chi-square statistic reported at the bottom of Table 7 provides a test of the null hypothesis that the difference in the rater severity measures is not statistically significant. As reported in Table 7, the chi-square value is 89.4 with 3 degrees of freedom and $p < .001$. Hence, the omnibus null hypothesis can be rejected and one can conclude that at least two of the raters differ in degree of severity.

The reliability of rater separation index reported at the bottom of Table 7 should not be interpreted like a traditional reliability coefficient (McNamara, 1996, p. 140). Instead of summarizing to what degree the raters are *reliably similar*, this statistic summarizes the extent to which they are *reliably different*. Values of this statistic can range from zero to 1.00. Myford and Wolfe (2004) assert that this statistic reflects “potentially unwanted variance between raters in the levels of severity exercised” (p. 196). High values of this statistic indicate that “there are discernible statistically significant differences between the severe and lenient raters” (Myford & Wolfe, 2004, p. 196). The .96 value reported for the reliability of rater separation index at the bottom of Table 7 further substantiates the differences in severity/leniency between the four

raters in this study and supports the conclusion that the four raters are not interchangeable.

The Facets analysis also provides two additional reports: an examinee measurement report (see Facets output Table 7.1.1, Figure E2) and a rating scale measurement report (see Facets output Table 7.3.1a, Figure E3). These reports supply a reliability of separation index, but couched in terms of examinees (ratees) or persons and in terms of rating scales. Although a reliability of separation index nearing 1.0 in the abbreviated rater measurement report is not desirable, a high index approaching 1.0 is very desirable in both the examinee (person) measurement report and rating scale measurement report.

Figure E2 provides a reliability of separation index equaling .95 (see Facets output Table 7.1.1—examinee measurement report). According to Myford and Wolfe (2003), the reliability of separation index for persons is analogous to Cronbach's alpha coefficient. This reliability of separation index for persons represents the ratio of true variance to observed variance in the person ability estimates. In other words, the .95 person separation index shows that the raters, while not being interchangeable in their ratings in terms of severity, were in fact able to differentiate between the students in their fluent oral reading abilities.

Figure E3 provides a reliability of separation index equaling .99 (see Facets output Table 7.3.1a—rating scale measurement report). The high value in this case is a good indicator that central tendency error (Myford & Wolfe, 2003) was not present, revealing that the raters were able to distinguish between the performances of the students on the 14 different FORE rating scales.

Whereas Figure 2 displays an ordering of the 14 FORE indicators, Table 8 provides greater detail of that ordering, ranging from the easiest to the most difficult in terms of receiving high ratings from the raters (see also Facets output Table 7.3.1a, Figure E2). Acc1: AIRW and Acc2: PRON cluster at difficulty measures less than -1.0

Table 8

Abbreviated Rating Measurement Report from Facets Analysis

Original FORE indicator sequence numbers	FORE indicator abbreviations	Difficulty measures
1	Acc1: AIRW	-1.08
2	Acc2: PRON	-1.05
9	Phr1: PUNCT	-0.49
5	Smo2: NR	-0.34
10	Phr2: PB	-0.28
3	Acc3: RTAW	-0.21
8	Rate2: NIB	0.07
13	Con1: SISS	0.14
6	Smo3: NEWP	0.15
4	Smo1: LA	0.18
7	Rate1: PACE	0.26
11	Exp1: SAI	0.71
12	Exp2: CIM	0.74
14	Con2: CPA	1.20

logits. While the low difficulty ranking of the accuracy scales was a deliberately contrived artifact of the text selection method, the ordering of all of the fluency rating scales gives an important result related to research question 3 which will be discussed later. The next group of indicators clusters between -0.50 and -0.20 logits. The third

group of indicators clusters between 0.05 and 0.30 logits. The fourth group of indicators clusters between 0.70 and 0.75 logits.

It is surprising that *Con2: Commands-positive-attention* (Con2: CPA) was the most difficult of all the indicators coming in at 1.20 logits, while it was found to be of lesser difficulty than either of the expression scales in the related measurement instrument study (McBride, 2004). Closer inspection of the changes in this rating scale made during the training / adjudication sessions revealed that the highest rating, 5, was defined as *riveting*. This rating hardly ever was given to any reader, which made Con2: CPA much more difficult relative to expression and other indicators than it had been before the changes were made in this study. It is desirable to have no construct-irrelevant variance in any construct, so the wording of this rating scale, needs to be reconsidered before additional uses.

Research Question 2

Factor analysis results. The method for addressing research question 2 involved performing a factor analysis. A factor analysis examines the structure within either one factor or the structure and correlations between two or more factors. The SPSS Factor Program was used with principal axis factoring and the Promax as the rotation method. The output of this program provided a Kaiser Test to determine the number of eigenvalues greater than 1.0. A more telling method to decide the number of factors is to consider the meaningfulness of the rotated factor solution in determining the best argument for a certain number of factors.

To determine if a factor analysis would be appropriate for this study, a Bartlett's Test of Sphericity was conducted. This test evaluated whether the correlation matrix approximated an identity matrix. The results of the test (2884.372, $df = 91$, $p < .0001$) shows that no off-diagonal elements approached 0.0, as in an identity matrix, so the factor analysis model was appropriate. These findings are confirmed also through the high inter-correlations within the accuracy and fluency constructs, as previously shown in

Table 5. The factor analysis is based on the averages of the ratings of the four raters A, B, C, and D. Using these averages is justified inasmuch as the inter-rater reliability coefficients are sufficiently high, illustrating a sound and appropriate methodology.

The factor analysis shows that the communalities (Figure 3) for the 14 variables range from a high of .93 for Con2: CPA to a low of .29 for Smo2: NR. The reader will recall that Smo2: NR was the indicator already found to be lacking in consistency with the other fluency ratings. Despite the lack of internal consistency of the indicator Smo2: NR with the construct of smoothness for which it was originally designed, it was included in the factor analysis to shed light on what other variables and constructs it might be related to instead. The analysis started with the correlation matrix in Table 5. The factor analysis used a linear model that reproduced the correlation matrix with a much smaller number of latent variables modeled as factors, compared to the 14 variables that created the raw correlations in Table 5.

Extraction of factors. Two standard methods of determining the number of factors to extract were used: the number of eigenvalues greater than 1.0 and meaningfulness. It was hypothesized that there were two correlated factors, accuracy and fluency. This hypothesis was also found in the literature (Dwyer, 2004; Logan, 1997; National Research Council, 1998). But as the reader will see in the following discussion, the factor analysis reveals that a feasible interpretation is that there are three, not two, dimensions. However, the initial hypothesis of meaningfulness, as will be seen, was approximated quite well by the factor analysis results because the third factor was weak, and highly correlated with fluency. Figure 4 shows that the first three eigenvalues are 8.13 for principal axis factor 1, 1.83 for axis factor 2 and 1.02 for axis factor 3. Together these three factors account for 78.4% of the variance among the 14 variables.

FORE Indicators	Initial	Extraction
Acc1: AIRW	.710	.817
Acc2: PRON	.552	.551
Acc3: RTAW	.539	.557
Phr1: PUNCT	.627	.862
Phr2: PB	.746	.768
Smo1: LA	.858	.867
Smo2: NR	.307	.291
Smo3: NEWP	.723	.690
Rate1: PACE	.781	.768
Rate2: NIB	.560	.505
Exp1: SAI	.919	.830
Exp2: CIM	.924	.886
Con1: SISS	.855	.863
Con2: CPA	.927	.934

Figure 3. Communalities derived through the principal axis factor method of extraction

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	8.130	58.068	58.068	7.912	56.517	56.517	7.653
2	1.828	13.057	71.125	1.498	10.700	67.217	4.255
3	1.018	7.269	78.394	.777	5.548	72.765	3.251
4	.712	5.089	83.483				
5	.521	3.721	87.203				
6	.459	3.281	90.485				
7	.347	2.481	92.966				
8	.240	1.711	94.677				
9	.193	1.381	96.059				
10	.193	1.379	97.438				
11	.154	1.097	98.535				
12	.100	.714	99.249				
13	.058	.417	99.666				
14	.047	.334	100.000				

Figure 4. Eigenvalues and variance accounted for by each factor

Factor rotation. With Kappa set at 4.0, the rotation produced two other sources of significant information: the factor structure matrix and the factor pattern matrix, both which are discussed in the following. According to Gorsuch (1983), investigators who use factor analysis results usually interpret the structure matrix because it contains the correlations of the variables with the factors. This is quite understandable because researchers are generally familiar with what a correlation is. Also, the structure matrix is generally more stable across studies than is the pattern matrix. The problem is that when the correlations among the factors are quite high, all the common variance in the variable that is shared across the two or three factors is present, and the separation is not clear. However, in the pattern matrix the separation is clear because the variance attributable to all the other factors is removed. The researcher presents both types of matrices, beginning first with the factor structure matrix, then moving on to the factor pattern matrix.

The researcher had hypothesized earlier that this data set would reveal two primary dimensions, fluency and accuracy. But there were actually three: (a) fluency, (b) accuracy and (c) phrase boundaries (Figure 5). The term *phrase boundaries* was chosen to denote that the weak doublet factor 3 consists of the two rating scales punctuation and phrase boundaries. Punctuation plays a significant role in forming phrases, thus the term phrase boundaries. The researcher was surprised to see the analysis reveal a third factor, which is shown to be a minor doublet factor with the two phrasing indicators on it (Figure 6) with an eigenvalue of only 1.018 (see Figure 4). As Gorsuch (1983) pointed out, researchers can learn about unknown variables from known variables by studying not only the values found in a factor structure matrix, but also the values found in a factor pattern matrix. The researcher now explores the meaningfulness inherent in these two matrices.

	Three Factor		
	1 (Fluency)	2 (Accuracy)	3 (Phrase Boundaries)
Con2: CPA	.963		.530
Con1: SISS	.920	.549	
Smo1: LA	.920	.606	
Exp2: CIM	.913		.609
Exp1: SAI	.882		.594
Rate1: PACE	.871	.539	
Smo3: NEWP	.776	.626	
Rate2: NIB	.706		
Acc1: AIRW	.576	.893	
Acc2: PRON		.731	
Acc3: RTAW		.730	
Smo2: NR		.513	
Phr1: PUNCT	.505		.922
Phr2: PB	.752		.768

	Two Factor	
	1 (Fluency)	2 (Accuracy)
Con2: CPA	.962	.549
Exp2: CIM	.937	
Exp1: SAI	.906	
Smo1: LA	.895	.693
Con1: SISS	.886	.643
Rate1: PACE	.852	.625
Phr2: PB	.799	
Smo3: NEWP	.729	.691
Rate2: NIB	.705	
Phr1: PUNCT	.570	
Acc1: AIRW	.525	.911
Acc3: RTAW		.721
Acc2: PRON		.705
Smo2: NR		.499

Three Factor	1 (Fluency)	2 (Accuracy)	3 (Phrase Boundaries)
1 (Fluency)	1.000	.534	.512
2 (Accuracy)	.534	1.000	.130
3 (Phrase Boundaries)	.512	.130	1.000

Two Factor	1 (Fluency)	2 (Accuracy)
1 (Fluency)	1.000	.584
2 (Accuracy)	.584	1.000

Figure 5. Three-factor structure and correlation matrices compared with two-factor structure and correlation matrices

Appendix F shows the factor analyses with three factors, no absolute values of loadings of any size were suppressed (see Figures F1 and F2). In order to clarify the meaningfulness, the researcher suppressed the absolute values less than .50 in the structure matrices for both a three-factor and two-factor solution. These two structure matrices are shown in Figure 5. Starting with the three-factor matrix (upper left in Figure 5) the reader will notice that it is sorted from the largest to the smallest loading variables on factor 1 (fluency), which has as its highest-loading variables the indicators hypothesized to be the most advanced, those of confidence. The other fluency indicators show high correlations with that factor, as expected. Note that Acc1: AIRW pulls over somewhat to fluency. This seems logical when one considers that in order to read aloud fluently, one must be able to recognize words automatically. However, Acc1: AIRW still correlates more highly with accuracy (factor 2), which was expected.

In contrast, even though *Phr1: Punctuation* (Phr1: PUNCT) loaded on factor 1 (fluency) with a .51 and its partner *Phr2: Phrase-boundaries* (Phr2: PB) loaded on that same factor with a .75, they both are actually more related to factor 3 (phrase boundaries). Phr1: PUNCT loaded on factor 3 with a .92; Phr2: PB loaded on factor 3 with a .77, higher than its loading on factor 1.

The factor analysis also shows, in the three factor correlation matrix (bottom left of Figure 5) that the possible phrase boundaries factor is highly correlated with fluency (.51) but not with accuracy (.13). It further shows that accuracy and fluency are well correlated (.53).

The indicators of confidence, expression and phrasing group with themselves, as compared to the smoothness construct, which has three indicators that vary widely. This could be evidence that smoothness is tied in to all the other aspects of fluency and accuracy. The indicators of accuracy are grouped together as well. However, Smo2: NR did not load at all on the fluency or phrasing factors. That which was thought to be part of

fluency, loaded only on to accuracy. Despite Smo2: NR's loading, the correlation with accuracy is the lowest when compared with the other accuracy correlations.

Turning now to the two factor structure matrix (upper right of Figure 5) the reader will notice that this matrix shows a strong fluency factor with all ten indicators present. Additionally, Acc1: AIRW loaded .53 on the fluency factor, showing how important automaticity is throughout the FORE domain. Smo2: NR still loads with accuracy at .499. As in the three factor structure matrix, the two factor structure matrix also is sorted from the largest to the smallest loading variables on factor 1 (fluency). This time, however, the highest-loading variables are Con2: CPA, Exp2: *Conversational-in-manner* (Exp2: CIM) and Exp1: *Stress-and-intonation* (Exp1: SAI) – very close to that which was hypothesized. Smo1: *Looks-ahead* (Smo1: LA) and Con1: *Student-is-sure-of-self* (Con1: SISS) also correlate highly with fluency. Such correlations are understandable when one considers that a reader must be confident enough to look ahead when reading, allowing for one to see the phrases and wording, thus knowing when to vary pitch, rate and expression, with confidence growing as one's fluent oral reading skills increase.

Factor 2 (accuracy) is defined by high loadings on the three accuracy indicators, but also picks up Smo1: LA and Smo3: *No-elongated-words-or-pauses* (Smo3: NEWP, both having correlations of .69. The accuracy factor is further defined by Con1: SISS at .64, Rate1: *PACE* at .62 and Con2: CPA at .55. These correlations suggest that automaticity leads to smoothness, confidence and rate. The factors correlate but have distinct meanings. The analysis also shows that accuracy has a high correlation with fluency in the two factor solution, that of .58 (see lower right in Figure 5).

While often not interpreted, the pattern matrix is meaningful when the factors are known or suspected in advance, as in this study, and is useful in interpreting the factor pattern. Gorsuch (1983) states that the values in a pattern matrix are considered as reference vector correlations. These correlations reflect the unique relationship of the factor to the variable, which is statistically independent of the other factors.

The three-factor and two-factor pattern matrices are presented in Figure 6 with absolute values less than .2 suppressed. The findings reinforce what was discussed in the structure matrix. Any common variance attributed to other factors was removed, providing a clearer interpretation of the relationships between the factors and the variables (FORE indicators). Although Phr1: PUNCT disappeared from factor 1, its value remained the same in factor 3. Phr2: PB's absolute values in factors 1 and 3 were reduced. Running a two-factor pattern matrix with absolute values less than .2 suppressed shows an excellent fluency factor with ten of the 11 indicators loading on fluency. Smo2: NR loads on accuracy, but is not very strong. Smo3: NEWP also loads on accuracy, somewhat less than on fluency.

In the earlier discussion concerning alpha coefficients, the internal consistency among the three factors (fluency, accuracy and phrasing) was .77. If accuracy were deleted from the group, the internal consistency coefficient would jump to .81 (see Table 6). If fluency were deleted from the group, the correlation would drop to .50. This is another indication that phrasing and accuracy are not highly related to each other. It may also be an indication that even though fluency and accuracy are related, with both contributing to fluent oral reading, they both should be assessed separately.

In summary, the analyses show that the third factor, related to phrasing or detecting phrase boundaries is very weak. Keeping a third factor with a marginal eigenvalue of only 1.018 is questionable, especially in light of the more meaningful interpretation of the two-factor solution. Figure 4 illustrates that there clearly are two dominant factors, which after rotation could be readily interpreted as fluency and accuracy. It also shows a third factor, which is interpreted as phrase boundaries. The inherent meaningfulness is enhanced using the two factor approach, but it is useful to compare the structure and pattern matrices with the three factor approach. For, as Gorsuch (1983) poignantly stated, "Indeed, proper interpretation of a set of factors can

	Three Factor		
	1 (Fluency)	2 (Accuracy)	3 (Phrase Boundaries)
Con2: SISS	.993		
Con1: CPA	.954		
Exp2: CIM	.937		
Exp1: SAI	.901		
Smo1: LA	.869		
Rate1: PACE	.837		
Smo3: NEWP	.732	.260	
Rate2: NIB	.632		
Acc2: PRON		.813	
Acc1: AIRW		.803	
Acc3: RTAW		.696	
Smo2: NR		.516	
Phr1: PUNCT			.922
Phr2: PB	.442		.533

	Two Factor	
	1 (Fluency)	2 (Accuracy)
Exp2: CIM	1.044	
Exp1: SAI	1.012	
con2: CPA	.974	
Phr2: PB	.830	
Con1: SISS	.775	
Smo1: LA	.744	.258
Rate1: PACE	.739	
Phr1: PUNCT	.638	
Rate2: NIB	.637	
Smo3: NEWP	.494	.402
Acc1: AIRW		.917
Acc2: PRON	-.245	.848
Acc3: RTAW		.687
Smo2: NR		.473

Figure 6. Three-factor pattern matrix and two-factor pattern matrix compared

probably only occur if at least S[tructure] and P[attern] [matrices] are both examined” (p. 208).

These findings illustrate that the phrasing construct still relates better with fluency than with accuracy. Such a finding may be an indication that phrasing will load more fully in factor 1 (fluency) in future research with theory-guided changes in the rating scales for phrasing, increased rater training and further refinement of the measurement instrument. As a theorist and instructional designer, the researcher recommends that the most useful interpretation for teachers and tutors who might use the instrument is the two-factor interpretation.

As a further check on this interpretation, the entire factor analysis was rerun with only 13 variables (see Appendix G). Smo2: NR was omitted because of its low communality and low correlations with either accuracy or fluency. Perhaps the existence of this largely unrelated variable adds construct-irrelevant variance into the matrix and contributes to the appearance of the weak 3rd factor. This reanalysis had eigenvalues of 7.96, 1.71, and .94. There is no evidence of the marginal phrasing factor given the Kaiser rule of using only eigenvalues > 1.0. The interpretation of this two-factor solution with 13 variables is the same as that of the two-factor solution presented above. The correlation between the two factors, accuracy and fluency, was slightly higher, at .59.

The two-factor interpretation is seen visually in Figure 7, which plots the two-factor solution with all 14 variables. Note that the factor 2 axis is the rotated factor identified as accuracy. Note that the three accuracy variables cluster together and are joined by Smo2: NR, which seems to relate better with automaticity than with smoothness, but not significantly. This cluster of four points is separated from the fluency variables which cluster around the fluency axis (factor 1) of the plot.

As mentioned in the discussion of the factor structure table, the variables that most clearly define fluency are confidence, expression and smoothness (especially Smo1: Looks-ahead). The factor analysis also reveals that Smo2: NR has the lowest

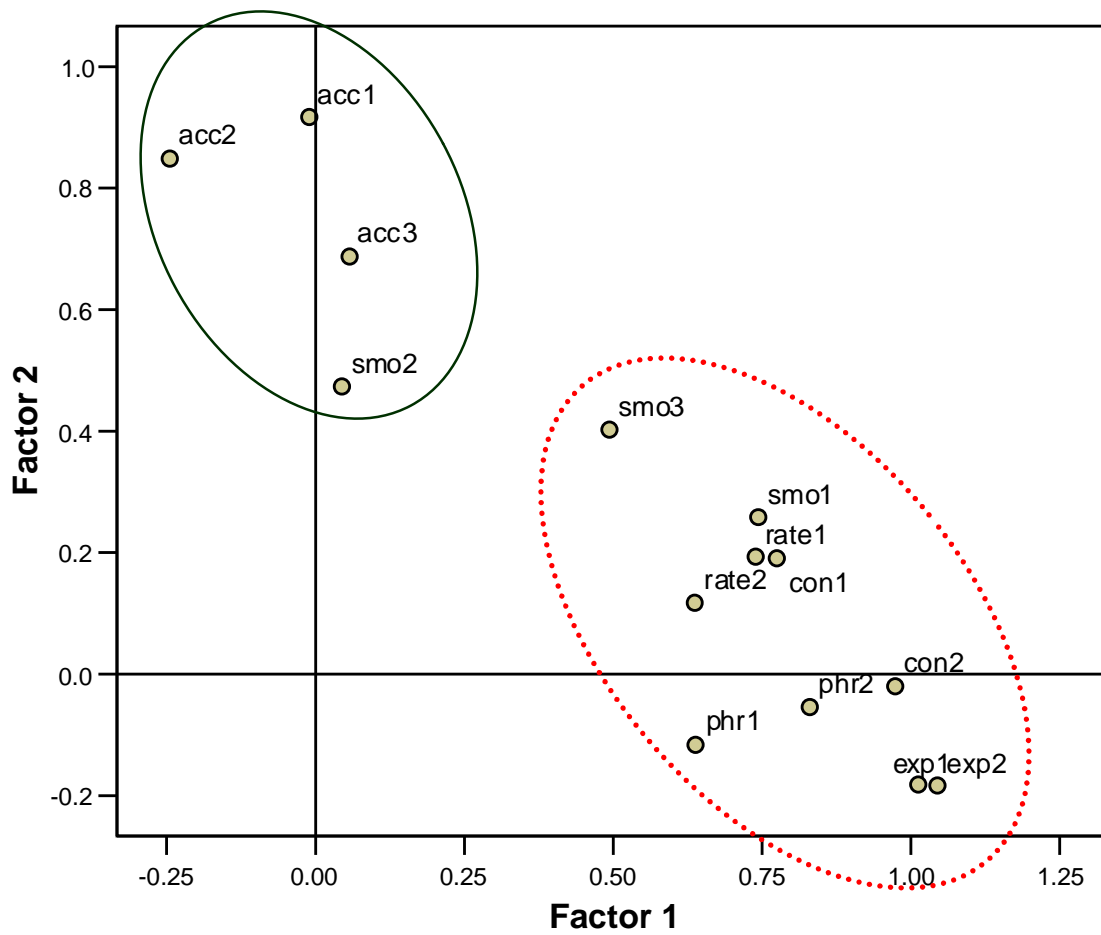


Figure 7. Factor plot of FORE indicators in rotated space

communality and has the least variance in common with the other rating scales.

This concludes the results of the investigation of research question 2. The answer to question 2, at least for this data set and instrument, is that there are two dominant dimensions (fluency and accuracy). Under some circumstances, such as the inclusion of the variable Smo2:NR which has little in common with either of the two factors, a lesser and questionable dimension (phrase boundaries) appears, which correlates strongly with the fluency factor but not with accuracy. All the indicators hypothesized to assess fluency are clearly aligned with that dimension rather than with accuracy, except that the smoothness indicator Smo3 is somewhat related to accuracy. Finally all three of the rating scales for accuracy clearly align with that factor. The high correlation between factors of .58 (.59 in the 13 variable analysis) and the shared positive loadings of the variables on both dominant factors show that, fluency and accuracy are parts of one integrated oral performance called fluent oral reading with expression. It is not only important and useful to consider these two separately in the practice of learning to read and in the assessment of reading skills, but it is also supported by the structural aspect of validity. Two correlated, but separable factors exist in the ratings.

Research Question 3

Setting up the Facets program. In addressing this question, the analyst set up the Facets analysis using three facets: (a) students, (b) raters, and (c) constructs (composed of from two to three construct-linked rating scales). Recall that in this study there were four raters, six constructs, and 14 indicators. Two analyses were run: one with 11 fluency indicators and one with all 14 rating scale indicators grouped into the six construct categories.

After examination of these two outputs, the ordering of the five fluency constructs is the same in each analysis. Rasch analysis assumes unidimensional scales; the evidence of question 2 is that there are two dominant and one minor, unidimensional scales. However, the high correlation of .58 between the accuracy and fluency dimensions make

it possible to include both accuracy and fluency indicators in the same analysis (see Figure 5). The Rasch model found the resultant of the three: a vector closer to fluency than accuracy. Presenting results from the analysis of all 14 variables allowed accuracy to be viewed in relation to the overall difficulty ordering of the fluency dimension. Thus, the output that combined fluency and accuracy was selected for presentation in this section of the analysis.

Order of constructs based on Facets. Chapter 1 hypothesized an ordering of learning with regards to the FORE constructs as identified in this study: accuracy < smoothness < rate < phrasing < expression < confidence. The Facets analysis of the data provided an ordering different in the order hypothesized. Accuracy remained where it was as predicted, the easiest of all to obtain high ratings. Phrasing moved into second place for easiness, and smoothness dropped from second into third place. Rate ended up in fourth place, with expression and confidence also switching places with each other. The actual order of difficulty in learning how to read fluently aloud with expression was accuracy < phrasing < smoothness < rate < confidence < expression.

Figure E4 contains Table 7.3.1b, a Facets generated rating construct measurement report. An abridged version is found in Table 9 which shows the FORE constructs according to how difficult it was for a student to receive a high rating, measured in logits. There is a distinct separation between accuracy and phrasing and also between phrasing and smoothness, both showing a difference of 0.11 logits. However, smoothness and rate are much closer at 0.05 logits. The spread between rate and confidence is much larger, 0.14 logits. Confidence and expression, however, have a difference of only 0.01 logits. The larger table (see Figure E4) shows that the standard error of this location measure is 0.01 in all cases. Setting two standard errors as a comfortable margin of significance between adjacent scales, it is seen that only the difference between Confidence and Expression is non-significant. With refinements in the theory, in the training of raters,

Table 9

Difficulty of FORE Constructs in Logits

Construct	Measure in Logits
Accuracy	-0.24
Phrasing	-0.13
Smoothness	-0.02
Rate	0.03
Confidence	0.17
Expression	0.18

and in the wording of the FORE measurement instrument (especially Con2: CPA), these differences may increase in significance.

As in Figure 2, the Facets generated chart in Figure 8 displays output derived from an analysis using a four-facet model. The first column, *Measure*, is an interval measurement scale in logits from -1 to +2. The second column, *Examinee*, shows a normal curve in the distribution of students. The third column, *Raters*, reveals the relative leniency of the raters. This ordering of rater leniency is exactly the same of that shown under research question 1. The fourth column, *Rating Construct*, shows the ordering of the FORE constructs, which consists of the composites of the rating scales belonging to each construct. The last column, *Scale*, provides the sum of the ratings on all 14 scales—a FORE total rating score.

In examining Figure 8, starting with accuracy and moving up the logit scale, the reader will note that rate and smoothness appear to share the same point on the logit scale and that confidence and expression appear to share a point on the logit scale as well. The

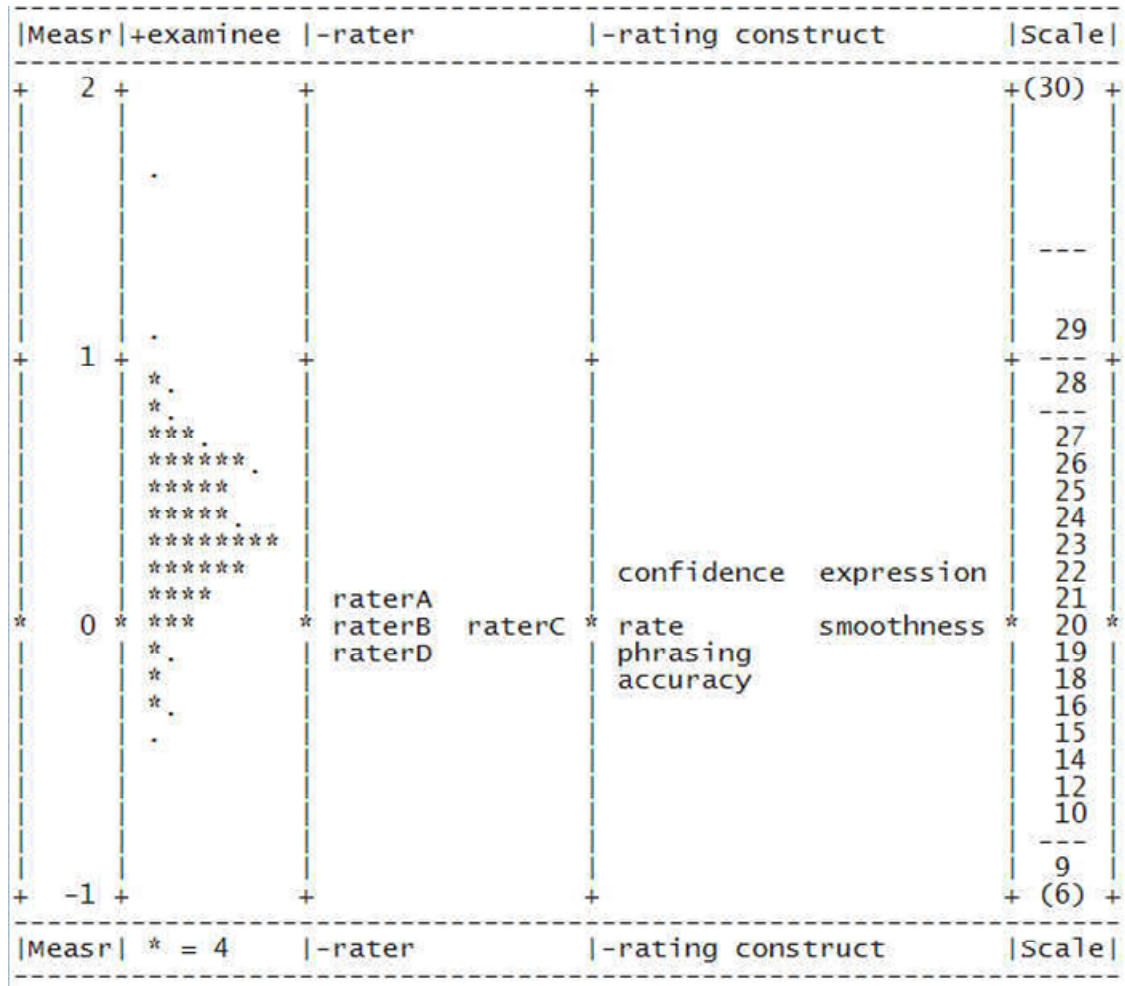


Figure 8. FORE constructs ordering based on Facets analysis

perceived sharing of such levels of difficulty shows that the differences between smoothness and rate are very small, and between expression and confidence, negligible, as shown in Table 9.

In the question 3 analysis, it is the composite of the individual constructs of smoothness, rate, phrasing, expression, and confidence that will be ordered. No single indicator is a perfect indicator of the larger construct, especially where there is wide variability among the indicators.

In summary, after accuracy, phrasing comes in second, with smoothness as third, rate as fourth, confidence as fifth, and expression as sixth. The actual ordering based on the Facets analysis makes sense when one considers all the FORE constructs and their unique aspects (see Figure 3 and Table 8). Simply put, as students gain mastery in the basics of reading, they will more likely have confidence in their ability to read aloud with greater expression.

The Facets analysis showed that the phrasing construct comes next after accuracy in order of difficulty, not smoothness as originally hypothesized. This is understandable when one realizes that the raters considered phrasing not as a student's looking ahead and taking in the meaningful rise and fall of a complete phrase, but as a student's being attentive to punctuation and phrase boundaries—which may be a separate and much easier category of looking ahead. Hence, it is easy for students to glance ahead and see punctuation marks, but harder to look ahead and take in phrase meaning while reading, contributing to the overall smoothness and fluency in their reading skills. Smoothness' influence spreads throughout both accuracy and fluency, contributing to the connectivity of these two highly correlated, yet separable factors (see Figure 7).

Rate was next on the difficulty scale, after smoothness, indicative of students having mastered aspects of the easier, less difficult constructs. Thus, they can control to a greater degree their oral reading rate, appropriately changing it as demanded by the text.

As students improve in these skills, their own recognition of their growing mastery leads to increased confidence. Raters see this and rate that student as more likely to be able to command the positive attention from others that good fluent reading engenders. Increased confidence in turn leads to greater courage to be more expressive in interpreting the text. In other words, the student who has increasingly mastered the skills of accuracy, phrasing, smoothness and rate will have the confidence to use proper stress and intonation that conveys meaning while reading orally, doing so in an expressive, conversational manner. These causal, or at least enabling connections between the different levels of total rating scores in fluent oral reading need further confirmation as hypotheses. The existence of the FORE measurement instrument positions educators to conduct research to further evaluate these developmental hypotheses.

The author combined Figures 2 and 8, creating Figure 9. The comparison revealed an interesting phenomenon. In most instances, the FORE indicators lined up with their corresponding constructs according to difficulty. The three indicators of accuracy (a) Acc1: AIRW, (b) Acc2: PRON, and (c) Acc3: *Reads-text-as-written* (Acc3: RTAW), corresponded with accuracy as being the easiest. At the other end of the logit scale were Exp1: SAI, Exp2: CIM, and Con2: CPA, aligned as being the most difficult for students to receive a high rating. The rest of the fluency indicators are interspersed in between.

In conclusion, concerning questions 2 and 3, according to the pattern matrix in Figure 6, and component plot in Figure 7, and now combined with the Facets analysis output in Figure 8, one can see how the FORE indicators are grouped and ordered in their respective constructs. By examining Figure 7, it is easy to see how the 14 indicators load on the two separate FORE dimensions: fluency and accuracy. More importantly, the findings discussed support the hypothesis stated in previous chapters that there is a developmental order in how fluent oral reading is learned, showing which constructs may be easier to master, and those that may be more difficult.

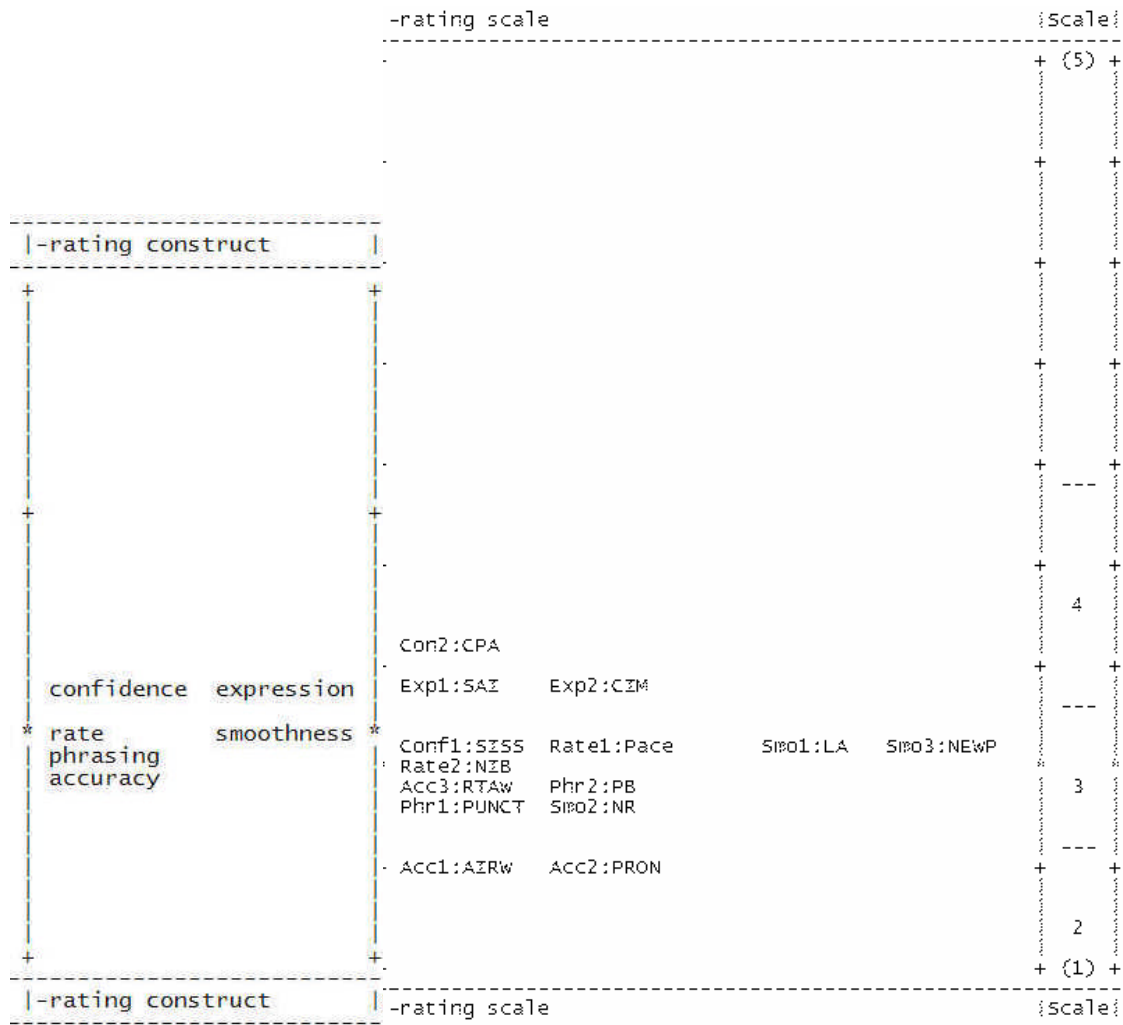


Figure 9. FORE indicators compared with their respective constructs

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

Discussion

This study was designed to aid the reading research community in answering the call to refocus attention on fluency. It was also designed to assist today's educators by offering a rating system that provides specific and valid feedback to parents, students and programs in the area of expressive fluent oral reading.

The researcher constructed a measurement instrument based on theoretical rationales for construct validity in the area of fluent oral reading. He sought and found empirical evidence for key aspects of construct validity in FORE. In particular, the researcher found the number of dimensions needed to span or cover the domain of fluent oral reading, and statistically and analytically determined a possible ordering of the subordinate constructs along the FORE fluency dimension.

The researcher determined empirically that word recognition or accuracy may tend to confound the rating scores of the other FORE construct-linked scales. That is, if a student were to be given reading material that is above his or her instructional or frustrational reading level, that student may display problems that cannot be rated unambiguously. The student may stutter or hesitate when trying to sound out these words. Such a confounding gives the impression that that student may have problems with fluent oral reading constructs rather than with word knowledge (accuracy). However, if the student were to be given a reading selection in which he or she demonstrated no major problems with word recognition, the other constructs could then be measured without being confounded by problems in accuracy.

Based on the findings of this study, it appears from a measurement perspective that the local learning theory (domain theory) of fluent oral reading with expression and accompanying measurement instrument may contribute to the field by providing a means

to assess the fluency constructs described in this study with greater validity. These constructs include identifying words quickly and easily, seeing enough of their meaning to speak with good phrasing, smoothness, at an appropriate rate, doing so with greater confidence, and proper expression.

The guiding principles of this study followed category II of the validity-centered design process: design for inherent construct validity. Category II consists of three separate validity arguments discussed earlier in this study: (a) a content validity argument, (b) a substantive process validity argument, and (c) a structural validity argument. The content validity argument was not addressed empirically or through a separate research question in this study. It was addressed by a literature review which identified the FORE constructs that ultimately found their way into the measurement instrument. The review showed that these constructs had been considered and studied in the literature and no key omissions were identified. The literature also contains discussions of the two main dimensions of fluent oral reading herein named accuracy and fluency.

Although the factor analysis revealed a third, weak and questionable dimension, phrase boundaries, evidence indicates that the phrase boundaries dimension is not a stable dimension distinct from fluency. It did not appear when one of the variables sharing very little common variance with either accuracy or fluency was not included in the analysis (see Appendix G). Therefore improving the instrument further by dropping this rating scale (Smo2: NR) removes construct-irrelevant variance that enables this factor to appear. In addition, by improving what is being rated for *phrasing*, and providing more specific training for the raters, this less meaningful factor will most likely not emerge as separate from the fluency factor in future studies. The two strong dimensions of fluency and accuracy have been reported in several other research reports in the literature as separate but correlated dimensions that are combined in the act of fluent oral reading. This study

replicates these studies, gives the magnitude of the correlation between the factors, and adds detail to the structure of the fluency dimension.

Evidence was provided through the analyses associated with research questions 1 and 2 that the instrument scales have adequate reliability, and that the constructs and subconstructs designed to work together do so. One major exception was found, the smoothness indicator Smo2: NR was shown by internal consistency reliability analysis and by factor analysis to share too little variance with either factor to be useful.

Research question 1 dealt with the inter-rater reliability across four raters. The researcher addressed research question 1 using two different methods: an analysis of the data using the SPSS reliability program, and with a rater analysis augmented by the Facets program. The SPSS program used correlational analyses and Cronbach's alpha reliability coefficients to assess the inter-rater reliability among the four raters for each of the 14 rating scales. The Facets program ordered students, raters, and rating scales onto a common interval measurement scale with the logit as the common interval.

The Facets analysis shows the order of relative leniency or severity among the four raters. It also shows that although the raters were able to avoid central tendency errors, their ratings are not interchangeable in terms of their leniency or severity. The Facets analysis provides further an informative preview of the ordering of the FORE constructs themselves, addressed in research question 3, by showing the relative difficulty of the FORE indicators in terms of which indicators are easier or more difficult for a student to receive a high rating. There is a general ordering from accuracy through phrasing, smoothness, rate, confidence, and expression. However, there is some within-construct spread, especially among the three smoothness indicators and the two confidence indicators. The data analyses further indicated that the tendency for *over-fits* or *mis-fits* is low, the fit indices falling within acceptable levels as discussed in chapter 4.

The data analyses dealing with substantive process validity were divided into methods used in connection with the first two research questions. The analyses revealed a

high correlation between the indicators within the individual FORE constructs, leading to internal consistency alpha coefficients ranging from .74 for smoothness to .97 for expression. However, the Smo2: NR indicator is not consistent with the other smoothness indicators. Nor is it consistent with the accuracy indicators, suggesting that perhaps it should be dropped from future versions of the instrument. Without this indicator the internal consistency of smoothness was .86. The internal consistency reliability of a combined scale using all 10 fluency indicators (dropping Smo2: NR) is high with an alpha coefficient of .96. If Smo2: NR is left as part of the original 11 fluency indicators, the internal consistency reliability of such a combined scale would be reduced to .95.

The structural aspect of inherent construct validity was addressed by research questions 2 and 3 dealing with the number of dimensions, or factors, and the ordering of elements (individual indicators or rating scales, as well as groupings of them into constructs) along those dimensions. The factor analysis of the means across raters of all 14 variables shows that the two hypothesized factors fluency and accuracy account for 71% of the variance. Adding a third factor with an eigenvalue of only 1.018 increases this to 78% of the variance. When this third factor is rotated and interpreted, it has only two main variables defining it, both dealing with the subconstruct of observing phrase boundaries. Interpretation of the three-factor vs the two factor solution led to the conclusion that the better solution would be to use a two-factor solution, where the two phrase boundaries indicators were accounted for fully by their strong relationships to the fluency factor and their weak relationships to the accuracy factor. When the 13 rating scales are analyzed, omitting Smo2: NR, the first two factors account for 74.4% of the variance instead of 71.1% in the 14 variable analysis (see Appendix G). This analysis confirmed that the two-factor solution is the best one.

The two-factor solution, interpreted as a strong fluency factor (the focus of this research) and smaller accuracy factor (used in this study primarily to control for word knowledge) is consistent with other evidence from the literature and from an earlier

construct validation study (McBride, 2004). As anticipated from the literature, the fluency and accuracy factors are highly correlated. The correlation in this study is .58 between the two rotated factor axes in the 14 variable two-factor solution (.59 in the 13 variable solution), and .53 in the three factor solution.

Addressing, in part, the structural validity aspect of validity-centered design's category II, the interpretation of the factor plot gives insight into the clustering of the three accuracy indicators on one factor (accuracy). The two-factor pattern matrix shows that both indicators of expression, along with the Con2: CPA indicator of confidence and the Phr2: PB indicator of phrasing, are the best markers of the fluency factor, while the Acc1: AIRW and Acc2: PRON indicators of accuracy are the best markers of the accuracy factor. The ordering of the loadings on the fluency factor in the two-factor solution gives another preview of the final aspect of the structural validity argument: the ordering of the constructs which was addressed by research question 3.

The final aspect of the structural validity argument developed in this study is addressed through research question 3. This question deals with the ordering of the constructs along what was hoped in advance would be a single fluency scale. The method used was to create a model using the Facets analysis program. This model had three facets: learners, raters and constructs. The order of constructs shown in this analysis was the evidence sought. Although not reproducing the order exactly as hypothesized, the Facets analysis shows a reliable ordering of the fluency constructs, with all but two of them, confidence and expression, being reliably different in difficulty. The ordering is accuracy < phrase boundaries < smoothness < rate < confidence < expression. The method used in this study was designed to make accuracy the easiest in order to reduce the possible confounding of word knowledge with fluency. Thus, word knowledge could be contrived to be the most difficult in another study. It is a correlated, but separate dimension.

Conclusions

The fluent oral reading with expression domain theory now has three of the nine arguments for validity discussed earlier in this study: (a) a content validity argument based on a literature review and influenced by an analysis of two existing reading programs, (b) a substantive process validity argument, and (c) a structural validity argument. Both (b) and (c) are based on the data and analyses in this study and in the previous measurement instrument development study (McBride, 2004). Within the parameters set by this study, the dimensions, constructs, and indicators of the local learning theory (domain theory) of fluent oral reading with expression have high correlations and internal consistencies amongst and between each other. It was confidently expected in advance that the factor structure would yield two dimensions, as in an earlier validity study using version IV of this instrument (McBride, 2004). Although a third, weak phrase boundaries factor was observed, the most useful interpretation is that of a two factor solution—fluency and accuracy. This is most consistent with all of the theoretical rationales and evidence available.

Some literacy experts assert that accuracy and fluency cannot be separated. Some believe that fluency has no developmental aspect to it, although word knowledge or accuracy is believed to be developmental. Finding the dominant two-factor structure and the ordering of the fluency constructs close to what was hypothesized gives credence to the local learning theory of progressive attainments in the domain of fluent oral reading with expression.

There was also a risk that fluency would split up further into more than one factor. In this data set it did, but evidence was given that the third factor, phrasing, would merge back into fluency with more refining of the instrument and training of the raters. There is considerable coherence among all of the fluency indicators, save one, Smo2: NR, which is found to have lower correlations with fluency than it did with accuracy, but the

relations with accuracy were still too low to use it further for rating accuracy in its present form.

There is also new evidence to suggest an ordering by which the FORE constructs may be taught, learned, and assessed. For those involved in improving the fluent oral reading of students in the schools, the implications are worthy of further investigation. First, educators must ensure that fluency is taught when students have texts that they can read at the independent level. Doing so can help prevent the indicators unique to accuracy from confounding the observation of constructs unique to fluency. Second, educators must become aware that there is a developmental sequence in teaching fluent oral reading with expression.

This study provides evidence that there is a theoretically coherent developmental sequence in fluent oral reading in students from the second through the sixth grades. Perhaps this ordering has been disguised in other studies which left it confounded with accuracy. Perhaps the question has not been asked and the data constructed and analyzed in the manner shown in this study.

The findings in this study bear some serious food for thought. Fluent oral reading must be practiced as a unified act that combines all of the constructs considered in this study simultaneously. In the face of this reality, there is still value in understanding the theoretical ordering of the constructs of fluency as developed by a local theory of learning specific to this domain. Further, the domain theory discussed and the analyses conducted herein provide evidence that accuracy or word knowledge needs to be controlled. If accuracy is not controlled, reading educators may be unable to observe and provide instruction to help in the development of the other constructs.

The data analyses reveal that after accuracy, the easiest fluency construct for students to receive a high rating is phrasing. Interpretation of the results of this study show that the rating scales designed to assess phrasing did not capture the whole concept of looking ahead and seeing and planning to speak the whole phrase. Rather, what was

observed was looking ahead to see the punctuation marks and phrase boundaries. Other rating scales could be constructed to assess a more difficult form of phrasing. Using the current rating scales, teachers would teach that looking ahead to see the punctuation boundaries would aid in smoothness and rate. Smoothness comes next as it reveals good habits of looking ahead and poor habits of halting and timidity. Although this study did not address poor eyesight, poor eye convergence or poor health, lack of smoothness, particularly Smo2: NR, could be influenced by these factors. In fact, problems in overall accuracy and fluency could be affected by these factors. Smoothness involves looking ahead to see the next entire phrase, tying in with a more holistic definition of phrasing—looking ahead to see the the meaning of the whole phrase, not just noting its boundaries.

The smoothness construct of looking ahead for phrases and meaning is critical to development of the later constructs of rate, confidence, and expression. Looking ahead and speaking smoothly make possible reading at a fast but appropriate rate. In teaching, after smoothness with looks-ahead-for meaning is accomplished, appropriate rate may be tutored in oral reading sessions. Table 9 shows that the difficulty measure of the smoothness construct is -0.02 logits. Rate is at 0.03 , yielding a 0.05 difference. Since the standard error of both of these constructs is $.01$, this difference of 0.05 is meaningful.

These accomplishments pave the way for the development of confidence, which can be influenced by teacher encouragement, but is probably more deeply motivated by self-observation of smoothness and rate. Hearing oneself speak smoothly and at a brisk rate can build confidence. Confidence in turn can provide a strong foundation for good expression. However, Table 8 shows that Con2: CPA was the most difficult of all the FORE indicators at 1.20 logits. One may conclude that the Con2: CPA indicator resulted in such a high difficulty score because in order for the student to command enough positive attention from the raters to obtain a rating of 5, his oral presentation had to be *riveting*. Only five student readings received this riveting rating. A single word change can sometimes have a large effect on empirical difficulty. In version IV, this word was

not used in the CPA rubric (McBride, 2004). However, it was surprising to observe that there were students who were very sure of themselves, yet whose reading was not smooth. Table 8 also shows that Con1: SISS is at 0.14 logits—obviously much more easily attained than its sister indicator of Con2: CPA. In order for students to receive a high rating in Con2: CPA, students had to show mastery of the other 13 indicators.

The average difficulty levels of the two indicators placed the confidence construct just under expression. Table 9 showed that the expression construct was the most difficult FORE construct at 0.18 logits, with confidence coming in just behind it at 0.17 logits. This is only a 0.01 difference and with standard errors of .01 for both confidence and expression, is not a statistically reliable difference, as is the larger difference found for confidence and expression in the FMI-IV validity study (McBride, 2004). Although the total act of oral reading involves all of the fluency constructs together, the process of getting there can be viewed as developmental. Knowing this developmental sequence can be very helpful for teachers and tutors as they engage students in small oral reading groups and provide feedback.

Entertaining the hypothesis that fluent oral reading is developmental, and trying to account for data found herein, introduces hypothesized causal connections or prerequisite structures of both theoretical and pedagogical interest. Fluent oral reading may be taught and learned using a sequence of attainments, even though all aspects must be practiced, observed, and rated as a totality.

This study answers in part the NRP's call for the reading community to draw its attention back to fluency. It provides a concise, easy-to-use rating system for today's educators in the area of expressive fluent oral reading. Further, this study lays a new foundation toward the building of a domain theory of fluent oral reading with expression. This foundation has as its footings validity arguments based on a comprehensive view of validity and how it may be achieved through an iterative process of design, evaluation, and improvement. So far, for the FMI-V, initial evidence for validity-centered design's

category II has been developed. Evidence for categories I and III have yet to be developed. The guiding framework of validity-centered design suggests steps for future research.

Recommendations for Improving the FORE Measurement Instrument

Many qualitative observations emerge from a study like this, which are not captured by the quantitative methods used to answer the main research questions. The following list provides some of the observations developed by the researcher, in consultation with the raters, which could help future studies and future applications.

Use only second through fourth graders. By the time students reach the fifth grade, poor readers are less likely to be willing to volunteer for a study. The sample discussed in chapter 3 showed considerably fewer fifth and sixth grade students participating (tendency for self-selection) as compared to the other grades.

Use raters of diverse backgrounds. The raters in this study consisted of three females and one male, all Caucasians and all living in the Mid-Atlantic region of the United States. Future studies should include raters from different ethnic backgrounds, as their input may provide greater insights in developing the instrument, especially so it can be used with a broader student population. This is also necessary to develop a validity argument for the generalizability aspect of validity-centered design.

Give more specific, definitive examples of different levels and scorings in the training. This will help clarify the constructs and rating scales for the raters, reducing disparity in overall ratings and also the need for adjudication.

Use texts of same length by reading grade level. Those students reading at a second grade level or below could read a 100 word selection. Students reading at a third grade level or higher could read a 200 word selection. Having specific word counts would allow for a more uniform defining of what constitutes a specific rating in the FORE MI rubrics. It would also make an objective measure of words per minute easier to obtain.

Use texts that have dialogue. Inasmuch as this study deals with fluent oral reading with expression, one of the best ways to observe or hear expression is to have a student read a conversation between two or more characters in a dialogue. Dialogue makes expression more natural. Not only does this make the reading more interesting for the student, it creates a better training situation for the rater and makes it easier for the rater to observe expression.

Provide vision screening for participating students. Raters stated that numerous students in this study were having difficulty in seeing the words on the page. Such vision problems negatively affected their accuracy and smoothness. It may be that the variable Smo2: NR, which did not share enough common variance with accuracy and fluency to be useful on either of those two dimensions, is tapping in part visual problems. For the sake of further development of the instrument, students who are found to have difficulty with their vision should be excluded from the study. For general use, consider creating observational rating procedures to rate visual problems. At least provide a space on the instrument for a rater to note the possibility of vision problems, in order to identify students who may need professional vision correction.

Rename the construct phrasing to phrase boundaries on the measurement instrument. Since the definitions for phrasing and expression overlap so much, it is better to separate the measurable aspects of punctuation and recognition of entire phrases from the intonation and inflection that expression typifies. It is the expression category that indicates the reader's overall comprehension of the passage as well as his/her ability to relay that meaning to a listener.

General Recommendations for Future Research

Conduct repeated trials. Future fluency studies should be conducted to show whether or not the structure formulated in this research remains constant over repeated trials, where the student reads, the teacher gives minimal suggestions, and the student reads again, repeating this one or more times. Studies combining fluent oral reading with

expression and dynamic assessment of repeated readings should also be conducted. This would provide greater insights to the instructor as to how specific interventions could be tailored to the individual needs of the student. The benefit to the student would be improved fluent oral reading skills – perhaps rapid improvement by focusing on just what the student needs to progress. These studies should be conducted where raters see at least three separate readings in a separated and randomized sequence with interventions given to students between each reading. Raters can be aware that various interventions are given, but only see the performance of the student.

Build the remaining six validity arguments. The FORE measurement instrument should be correlated with other instruments of its kind, thus building an external validity argument. It should also be the focus of a design experiment, thus building a consequential validity argument over time. Focus groups should be conducted with educators who use this instrument in the classroom. Doing so would help build the overall appeal, usability and perceived value of the entire rating system.

It is important to investigate the ease of use of the FORE instrument in the regular classroom, and how often the results of ratings, integrated with instruction, are feasible and helpful. Inasmuch as the FORE measurement instrument has been used primarily with native English speaking students and only one school population, it would be beneficial to learn how the instrument fares with those whose native language is not English, thus building a generalizability validity argument. This study should be replicated over several school populations and several school districts with an increased number of raters, to establish consistency regardless of reader or rater diversity.

REFERENCES

- Allington, R. (1984). Content coverage and contextual reading in reading groups. *Journal of Reading Behavior, 16*, 85-96.
- Au, K. H. (1997). Ownership, literacy achievement, and students of diverse cultural backgrounds. In J. T. Guthrie & A. Wigfield (Eds.), *Reading engagement: Motivating readers through integrated instruction* (pp. 168-182). Newark, DE: International Reading Association.
- Blum, I. H., & Koskinen, P. S. (1991). Repeated reading: A strategy to enhancing fluency and fostering expertise. *Theory Into Practice, 30*, 195-200.
- Brenna, B. A. (1995). The metacognitive reading strategies of five early readers. *Journal of Research in Reading, 18*, 53-62.
- Breznitz, Z. (1987). Increasing first graders' reading accuracy and comprehension by accelerating their reading rates. *Journal of Educational Psychology, 79*, 236-242.
- Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*, 141-178.
- Brunswick, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California.
- Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99-136). Washington, DC: American Psychology Association.
- Bunderson, C. V. (2000, April). *Design experiments, design science, and the philosophy of measured realism: philosophical foundations of design experiments*. Paper

- presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Bunderson, C. V. (2002). How to build a domain theory: On the validity-centered design of construct-linked scales of learning and growth. Paper presented at the Institute of Objective Measurement Workshops, April, 2002, New Orleans, LA.
- Bunderson, C. V. (2005, in press). Developing a domain theory, defining and exemplifying a learning theory of progressive attainments. In *Advances in Rasch measurement, volume one (Vol. 1)*. P.O. Box 1283, Maple Grove, MN 55311: JAM Press.
- Bunderson, C. V., & Newby, V. A. (2005, in press). The relationships among design experiments, invariant measurement scales, and domain theories. In *Advances in Rasch measurement, volume one (Vol. 1)*. P.O. Box 1283, Maple Grove, MN 55311: JAM Press.
- Burns, P. C., & Roe, B. D. (1993). *Burns/Roe informal reading inventory: Preprimer to twelfth grade* (4th ed.). Boston, MA: Houghton Mifflin Company.
- Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*, 79-89.
- Cattell, J. M. (1886). The time it takes to see and name objects. *Mind, 41*, 63-65.
- Cromer, W. (1970). The difference model: A new explanation for some reading difficulties. *Journal of Educational Psychology, 61*, 471-483.
- Cronbach, L. J. (1984). *Essentials of psychological testing*. New York: Harper & Row, Publishers, Inc.
- Dowhower, S. L. (1987). Effects of repeated reading on second-grade transitional readers' fluency and comprehension. *Reading Research Quarterly, 22*, 389-406.
- Dowhower, S. L. (1991). Speaking of prosody: Fluency's unattended bedfellow. *Theory Into Practice, 30*, 165-176.

- Dwyer, K. (2004). *Northwest Regional Educational Laboratory K-3 developmental continuum oral reading rubric for: Fluency, rate, expression, self-monitoring*. Retrieved July 13, 2004, from www.nwrel.org/assessment/pdfRubrics/k3devcontoral.PDF
- Ekwall, J. L., & Shanker, E. E. (2000). *Ekwall/Shanker reading inventory* (4th ed.). Boston, MA: Allyn and Bacon.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179-197.
- Estes, E., & Slobodkin, L. (1944). *The hundred dresses*. New York: Harcourt Brace Jovanovich.
- Farstrup, A. E., & Samuels, S. J. (2002). *What research has to say about reading instruction* (3rd. ed.). Newark, DE: International Reading Association.
- Fawcett, J. (1999). *The relationship of theory and research* (3rd ed.). Philadelphia: F. A. Davis Company.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Greenwood, C., Abbot, M., & Tapia, Y. (2003, January 31, 2003). *Reading fluency table*. Retrieved August 14, 2004, from http://www.lsi.ku.edu/jgprojects/cwptlms/html2002/ProjectManagement/reading_fluency_table.htm
- Harris, T. L., & Hodges, R. E. (1995). *The literacy dictionary*. Newark, NJ: International Reading Association.
- Highlights for Children Inc. (2003). *A brick to cuddle up to*. Retrieved April 4, 2003, 2003, from <http://www.sde.state.id.us/naep/info/reading04-sample-1.pdf>
- Huey, S. E. (1968). *The psychology and pedagogy of reading*. Cambridge, MA: MIT Press.

- Johns, J. L., & Berglund, R. L. (2002). *K/H reading resources: Fluency, questions, answers, evidenced-based strategies*. Dubuque, IA: Kendall/Hunt Publishing Company.
- Kame'enui, E. J., & Simmons, D. C. (2001). Introduction to this special issue: The DNA of reading fluency. *Scientific Studies of Reading*, 5, 203-210.
- Kelly, A. E. (2003). Research as design. *Educational Researcher*, 32 (1), 3-4.
- Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations for measurement, vol. 1*. New York: Academic Press.
- LaBerge, D. & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293-323.
- LaBerge, D. & Samuels, S. J. (1985). Toward a theory of automatic information processing in reading. In H. Singer & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (3rd ed., pp. 689-718). Newark, DE: International Reading Association.
- L'Engle, M. (1973). *A wrinkle in time*. New York: Bantam Doubleday Dell Books for Young Readers.
- Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and psychological measurement*, 16, 294-304.
- Leslie, L., & Caldwell, J. (1995). *Qualitative reading inventory II*. New York: Addison Wesley Longman Inc.
- Lipson, M. Y., & Lang, L. B. (1991). Not as easy as it seems: Some unresolved questions about fluency. *Theory into practice*, 30, 218-226.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 45, 517-544.
- Logan, G. D. (1997). Automaticity and reading: perspectives from the instance theory of automatization. *Reading and Writing Quarterly*, 13, 123-146.

- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1-27.
- Lyon, R. G., & Moats, L. C. (1997). Critical conceptual and methodological considerations in reading intervention research. *Journal of Learning Disabilities, 30*, 578-588.
- McBride, R. H. (2004). *Structural and Substantive Process Validity of a Rating Scale Instrument for Fluent Oral Reading*. Provo, UT: Available from author.
- McBride, R. H. (2005). *The interplay of training materials development and rating scale development in the instruction of raters*. Provo, UT: Available from author.
- McBride, V. G. (1997, unpublished manuscript). *Procedures for teaching rapid remedial reading*.
- McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.
- 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. *American Psychologist, 50*, 741-749.
- Murray, B. (2002). *The Reading Genie*, 4 July 2002, from <http://www.auburn.edu/~murraba/>
- Myford, C. M., & Wolfe, E. W. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement, 3*, 300-323.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.
- Nathan, R. G., & Stanovich, K. E. (1991). The causes and consequences of differences in reading fluency. *Theory Into Practice, 30*, 176-184.

National Assessment of Educational Progress. (2004). *Reading rockets: Launching young readers: How fluently do our children read?* Retrieved July 13, 2004, 2004, from <http://www.readingrockets.org/article.php?ID=100>

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4754)*. Washington DC: U.S. Government Printing Office.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4745)*. Washington DC: National Institute of Child Health and Human Development.

National Research Council. (1998). Recommendations for practice and research. In C. Snow, M. S. Burns & P. Griffin (Eds.), *Preventing reading difficulties in young children* (pp. 313-344). Washington DC: National Academy Press.

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

O'Shea, L. J., & Sindelar, P. T. (1983). The effects of segmenting written discourse on the reading comprehension of low- and high-performance readers. *Reading Research Quarterly*, 18, 458-465.

Peak, H. (1953). Problems of observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 243-299). Hindsdale, IL: Dryden Press.

Pikulski, J. J. (2000). *Preventing reading problems: Factors common to successful early intervention programs*. Retrieved 2 October 2000, from <http://www.eduplace.com/rdg/res/prevent.htm>

Polit, D. F., & Hungler, B. P. (1995). *Nursing research: Principles and methods* (5th ed.). Philadelphia: Lippincott.

- Rasinski, T. V. (1990). Effects of repeated reading and listening-while-reading on reading fluency. *Journal of Educational Research*, 83, 147-150.
- Reutzel, D. R., & Hollingsworth, P. M. (1993). Effects of fluency training on second grader's reading comprehension. *Journal of Educational Research*, 86, 325-331.
- Richards, M. (2000). Be a good detective: Solve the case of oral reading fluency. *The Reading Teacher*, 53, 534-539.
- Rinehart, S. D. (1999). "Don't think for a minute that I'm getting up there": Opportunities for readers' theater in a tutorial for children with reading problems. *Journal of Reading Psychology*, 20, 71-89.
- Samuels, S. J. (1997). The method of repeated readings. *The Reading Teacher*, 50, 376-381.
- Scholastic. (2002). *Scholastic Assessment*. Retrieved July 4, 2002, from <http://teacher.scholastic.com/products/assessment.htm>
- Shanahan, T. (2000). *Teaching fluency in the high school*. Unpublished manuscript, University of Illinois at Chicago.
- Shapiro, E. S. (1989). *Academic skills problems: Direct assessment and intervention*. New York: Guilford.
- Skinner, C. H., Cooper, L., & Cole, C. L. (1997). The effects of oral presentation previewing rates on reading performance. *Journal of Applied Behavior Analysis*, 30, 331-333.
- Stenner, A. J. (1996). *The Lexile Framework for Reading*. Retrieved 17 April 2004, from www.lexile.com
- Strong-Krause, D. (2001). English as a second language speaking ability: A study in domain theory development. *DAI-A 61/12*, 4746.
- Suppes, P., Pavel, M., & Falmagne, J. C. (1994). Representations and models in psychology. *Annual Review of Psychology*, 45, 517-544.

- The Partnership for Reading. (2004). *The partnership for reading: Bringing scientific evidence to learning*. Retrieved 6 March 2004, U.S. Department of Education, from <http://www.nifl.gov/partnershipforreading/explore/fluency.html>
- Trout, J. D. (1998). *Measuring the intentional world: Realism naturalism, and quantitative methods in the behavioral sciences*. New York: Oxford University Press.
- University of Oregon. (2000). *Dynamic Indicators of Basic Early Literacy Skills*, Retrieved March 6, 2004, from http://dibels.uoregon.edu/dibels_what.php
- White, S. (2004). *Listening to children read aloud: Oral fluency*. Retrieved July 13, 2004, from <http://nces.ed.gov/pubs95/web/95762.asp>
- Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, 5, 211-239.
- Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and assessment in schools* (2nd ed.). New York: Addison Wesley Longman.
- Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65-101). Mahwah, NJ: Lawrence Erlbaum.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370.
- Zutell, R., & Rasinski, T. V. (1991). Training teachers to attend to their students' oral reading fluency. *Theory Into Practice*, 30, 211-217.

APPENDICES

Appendix A

History of the Development of the FORE Measurement Instrument

The FORE instrument is the product of efforts to develop a local learning theory of progressive attainments in the domain of fluent oral reading with expression. A good domain theory provides a conceptual framework to assist and guide a researcher in designing and critiquing different measurement instruments and teaching methods. A good measurement instrument is tied to the conceptual framework so it also provides the means to measure attainments with high validity, and thus it both guides the design of and assesses empirically the effectiveness of teaching methods.

The FORE instrument is a tool designed to be used diagnostically by educators who are interested in identifying, backed by the evidence of an increasingly strong validity argument, the profile of strengths and weaknesses of the FORE skills found in their students. It is also an instrument that is continually evolving, with its validity argument strengthening through continual use and testing over time.

For example, the FMI-I (see Figure A1) which was a prototype that was never tested, but was developed as an initial starting point, consisted of a single five-point rating scale, with each point connected to a rubric describing each of the levels “1, 2, 3, 4 and 5.” These levels ranged from a high frustrational level in reading to a high level of expertise. A convention in theory-building found useful by domain theory researchers is to define clearly the characteristics of the least and most proficient persons. The hoped-for simplicity of FMI-I was that it was only one scale, which made it very appealing from a practical view (category I of validity-centered design), since rating takes time and effort. The downside, though, was that FMI-I appeared too complicated for others to learn easily and use accurately. Despite the single scale, the rubric descriptors appeared far too detailed and cumbersome for a rater to use effectively and quickly. FMI-I needed to change resulting in FMI-II (see Figure A2).

The FORE Measurement Instrument - Version I

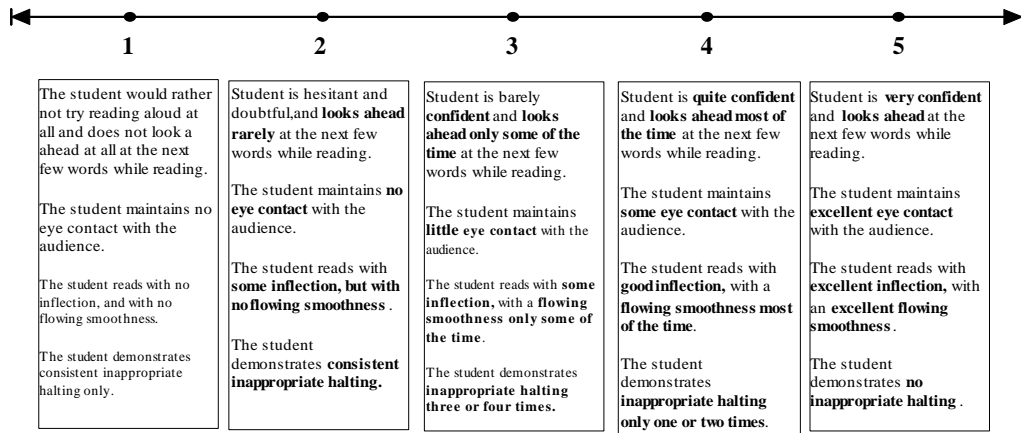
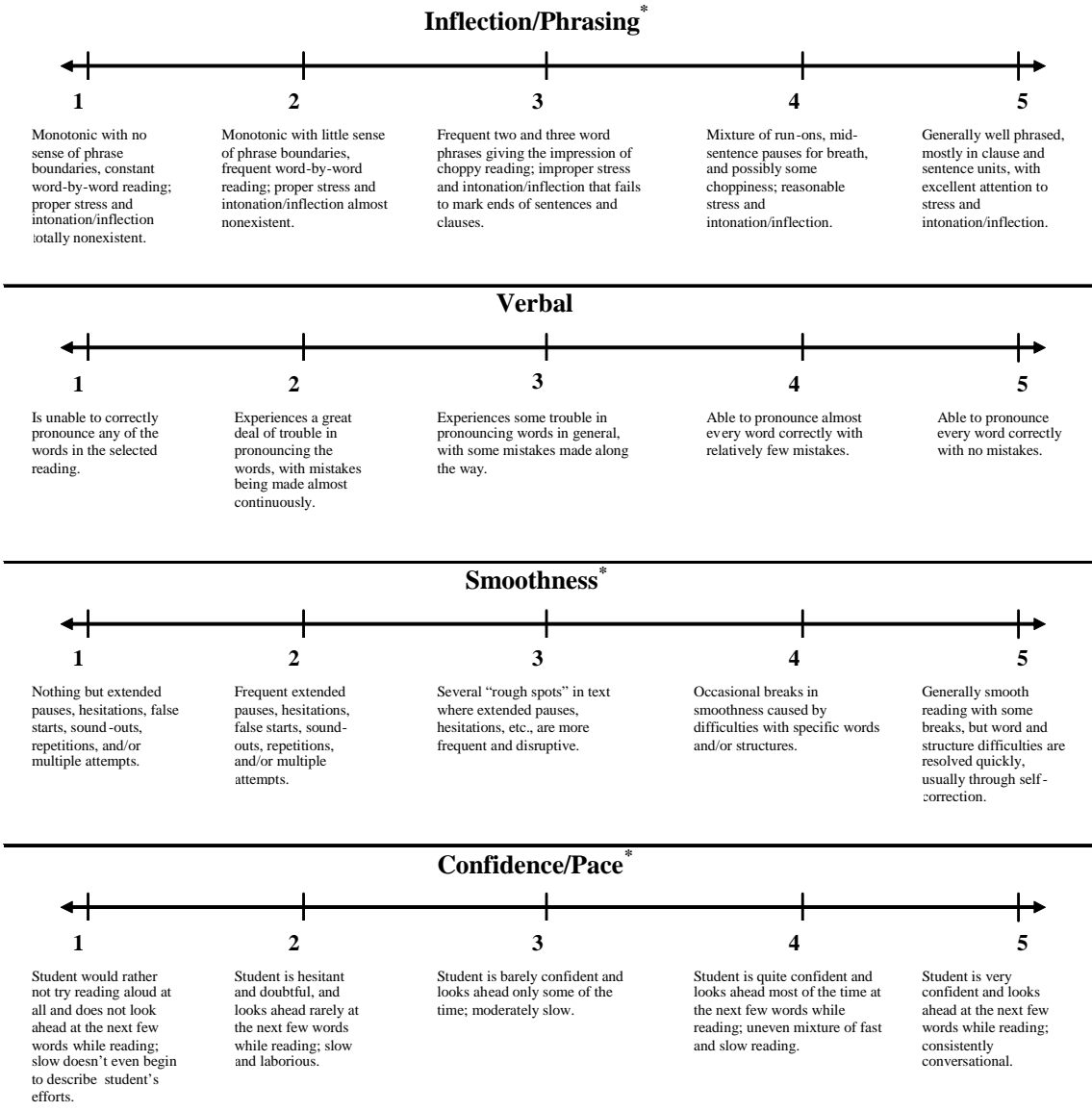


Figure A1. FMI-I, untested

**A Domain Theory of Fluent Reading Fluency:
A Measurement Instrument – Take Two**



 *Part of the rubric descriptions shown above come from scales found in the following reference: Jerry Zutell, "Training Teachers to Attend to Their Students' Oral Reading Fluency" in *Theory Into Practice*, Volume XXX, Number 3, Summer 1991, pp. 211 – 217. I combined my own rubrics with his. This was done to make the scales used and constructs studied/tested more complete. rhm

Figure A2. FMI-II, untested

FMI-II of the FORE instrument consisted of four separate five-point scales (1 through 5), with a separate scale for each construct. The reader will note that the constructs attached to the scales of this version are not the same as those contained in the most current version, evidencing the continual evolution of the instrument. Instead of the word *expression* the term *inflection* was used. The literature, it was found (Dowhower, 1991; National Reading Panel, 2000), did not really use inflection. Rather, it more commonly used the terms *intonation* and *expression*. The author found, however, that the term *intonation* was used mainly with phonological aspects of reading (National Research Council, 1998). The author, therefore, chose the term of *expression* to replace the original term of *inflection*.

The reader will also note how *inflection* (now *expression*) and *phrasing* were originally connected to the same scale. It seemed appropriate at the time, but as the thinking about fluent oral reading progressed, the author came to realize that it would not be wise to try to measure two constructs on one scale. The same rationale applied to the pairing of *confidence* and *pace* and then later to their subsequent separation. The term *rate* was predominantly used more in the literature than *pace*. Each point was supplied with a rubric describing what a “1, 2, 3, 4 and 5” was from high frustrational levels in reading (1) to high levels of expertise (5). The advantage of FMI-II was that the rubrics were short and easy to read quickly and accurately when rating students in their reading fluency skills. Despite this advantage, the FORE measurement instrument was not ready to be used. Even without testing the instrument, it appeared that a few more changes needed to be made.

FMI-II still kept five points to each scale. Upon examining FMI-II, the author, along with his doctoral advisor, determined that the rubrics for each point seemed very similar. For example, what difference was there between *most of the time* or *almost always* or *usually*? To reduce redundancy and make the instrument easier to use, FMI-III-A was created (see Figure A3). FMI-III-A reduced the points or levels

**FORE Measurement Instrument
- Version III-A -
Page 1**

Smoothness

	1	2	3	4
<i>no extended pauses, full starts, no sound-outs, no repetitions</i>	not smooth	somewhat smooth	mostly smooth	always reads smoothly
<i>student appears to look ahead</i>		inconsistently	usually	consistently

Comments:

Word Recognition & Pronunciation

	1	2	3	4
<i>recognition and pronunciation</i>	consistently makes mistakes	many mistakes	some mistakes	essentially no mistakes

Comments:

Phrasing

	1	2	3	4
<i>outstanding sense of phrase boundaries</i>	none, constant word-by-word reading	little, mostly word-by-word reading	usually shows, has mid-sentence pauses for breath, possibly some choppiness.	complete, demonstrates appropriate clause and sentence units

Comments:

Figure A3. FMI-III-A, page 1

**FORE Measurement Instrument
- Version III-A -
Page 2**

Expression

stress and intonation convey plausible meaning; notes the ends of clauses and sentences.

- | | | | |
|---|---|--|--|
| 1 | 2 | 3 | 4 |
| monotone, lacking any stress and intonation | mostly monotone, little stress and intonation | some stress and intonation, not always appropriate | reasonable & appropriate stress and intonation |

Comments:

Rate

student's oral reading speed or rate (pace) is conversational in manner

- | | | | |
|---|---|---|--|
| 1 | 2 | 3 | 4 |
| rate is very inappropriate to the situation | rate is somewhat inappropriate to the situation | rate is mostly appropriate to the situation | rate is totally appropriate to the situation |

Comments:

Confidence

student is very sure of him/herself, has complete confidence

- | | | | |
|---------------|------------------|--------------------|---|
| 1 | 2 | 3 | 4 |
| not confident | barely confident | somewhat confident | very confident, commands attention from others while reading, |

Comments:

Figure A3 continued. FMI-III-A, page 2

of progress from five to four, changing the rubrics accordingly. One further change was made to FMI-II. The reader will recall that the scale for *smoothness* was listed first in the instrument. In keeping with the hypothesized ordering the author felt that the scale for *word recognition & pronunciation* should be placed first, so the two were switched. He now had an instrument, FMI-III-A that could be given a trial run in a pilot study.

The pilot study resulted in a number of formative improvements. The four raters involved (school teachers and teacher education graduate students), who were representative of the population that would be using the FORE instrument, provided feedback as to the wording. Their input was very enlightening, leading to changes in some of the wording of the rubrics, clarifying the explanations attached to each of the points, as well as certain of the descriptors located to the left of each scale. These changes resulted in the creation of FMI-III-B (see Figure A4), and paved the way for the related measurement project (McBride, 2004).

With the creation of FMI-III-B, came also the need for an expanded study with many more students. Four new raters were chosen and training began. These raters (all female, ranging in age from 30 to 75) were either experienced in using the McBride Reading Program or had had experience in tutoring children.

From the outset, FMI-III-B had problems. The raters experienced numerous difficulties as they worked to learn to use it. In keeping with the principles of validity-centered design (user-centered design) the raters provided feedback. Further changes were made based on that feedback. For example, the 1 to 4 rating scale was a problem. FMI-III-B did not allow for any middle ground. Also, the descriptions under the constructs were at times vague and often seemed to overlap.

For example, the description under phrasing “pays attention to punctuation” could not be distinguished from expression’s “notes the ends of clauses and sentences.” Also, rate’s definition of “conversational in manner” seemed more like expression’s

**FORE Measurement Instrument
- Version III-B -
Page 1**

Word Recognition & Pronunciation

<i>automaticity in recognizing & pronouncing words correctly</i>	1	2	3	4
	consistently makes mistakes	many mistakes	some mistakes	essentially no mistakes

Comments:

Smoothness

<i>has full starts, no extended pauses, no sound-outs, no repetitions</i>	1	2	3	4
	not smooth	somewhat smooth	mostly smooth	always reads smoothly
<i>student appears to look ahead</i>		●	●	
		inconsistently	usually	consistently

Comments:

Phrasing

<i>outstanding sense of phrase boundaries, pays attention to punctuation</i>	1	2	3	4
	none, constant word-by-word reading	little, mostly word-by-word reading	usually shows, has mid-sentence pauses for breath, possibly some choppiness.	complete, demonstrates appropriate clause and sentence units

Comments:

Figure A4. FMI-III-B, untested, page 1

**FORE Measurement Instrument
- Version III-B -
Page 2**

Expression

stress and intonation convey plausible meaning; notes the ends of clauses and sentences.

- | | | | |
|---|---|--|--|
| 1 | 2 | 3 | 4 |
| monotone, lacking any stress and intonation | mostly monotone, little stress and intonation | some stress and intonation, not always appropriate | reasonable & appropriate stress and intonation |

Comments:

Rate

student's oral reading speed or rate (pace) is conversational in manner

- | | | | |
|---|---|---|--|
| 1 | 2 | 3 | 4 |
| rate is very inappropriate to the situation | rate is somewhat inappropriate to the situation | rate is mostly appropriate to the situation | rate is totally appropriate to the situation |

Comments:

Confidence

student is very sure of him/herself, has complete confidence

- | | | | |
|---------------|------------------|--------------------|---|
| 1 | 2 | 3 | 4 |
| not confident | barely confident | somewhat confident | very confident, commands attention from others while reading, |

Comments:

Figure A4 continued. FMI-III-B, untested, page 2

“reasonable and appropriate stress and intonation.” “No sound outs” under smoothness seemed to fit better with “automaticity in recognizing and pronouncing words correctly.” For the raters to appropriately determine what a student was doing, a more succinct arrangement of descriptors (indicators) under the constructs was called for.

The researcher, in partnership with the raters, revised the descriptions or definitions into more observable indicators. The construct of “word recognition and pronunciation” was changed to “accuracy.” “No sound outs” from smoothness was combined with “automaticity in recognizing words” and placed under accuracy, Pronunciation was made a separate indicator as “pronounces words correctly.” “Reads text as written” was added under accuracy because so often students insert words that are not in the text. The researcher and raters decided to save space and time by using a numerical score rather than circling a position on a graph. A five-point master key from “Never” to “Always” was created for the majority of the indicators. An exception was made for “Looks ahead” where each number required a detailed description. Hence, working together as a research team, the raters and researcher created a new and improved FMI-IV, (Figure A5) which had six constructs and 14 observable indicators.

This new version was used in a related measurement project (MP), to which the reader may want to refer (McBride, 2004) for details. The analyses of the MP’s data set showed that there are two factors that comprise fluent oral reading with expression: accuracy and fluency. The analyses also showed that there is a developmental ordering of the FORE constructs: (a) accuracy, (b) smoothness, (c) rate, (d) phrasing, (e) expression, and (f) confidence. The McBride (2004) study showed that the order was accuracy, smoothness, phrasing, rate, confidence, expression, This ordering became the hypothesized ordering for the present dissertation study.

Rater Number: _____

Student Number: _____

MASTER KEY: 1 = Never 2 = Almost Never 3 = Sometimes 4 = Almost Always 5 = Always *KEY FOR "Looks ahead" 1 = word-by-word 2 = very choppy 3 = mid-sentence pauses for breath 4 = mildly choppy 5 = consistently looks ahead	Accuracy	Rating	
	Automaticity in recognizing words - (no sound outs)		
	Pronounces words correctly		
	Reads text as written (no extra words inserted)		
	Smoothness		
	COMMENTS:	*Looks ahead	
		Full starts	
		No elongated words or pauses	
	No repetitions		
	Rate (pace) – (appropriate to the situation)		
	Phrasing		
	Strong sense of phrase boundaries		
	Observes punctuation		
	Expression		
	Stress and intonation conveys plausible meaning		
	Conversational in manner		
	Confidence		
	Student is sure of him/herself		
	Commands positive attention from others		

Figure A5. FMI-IV

Appendix B
Permissions and Consent Documentation

Reo, Thank you for the update. I want to be clear that the principal and literacy coordinator must be comfortable with the continuation of this project. We always desire to support such work but it is hard for me to evaluate your research from here. Please work very closely with Dennis. I will ask him in this e-mail to visit with me and Asst. Superintendent Ray Morgan if he has concerns about your next step. As always, parent permission slips are required as well as oversight of your research while being conducted. Thanks, **(Superintendent)**

>>> "Reo H. McBride" <rh2@ > - 11/5/03 3:00 PM >>>
Dear **(Superintendent)**,

This is Reo McBride. You approved my getting to work with (the school) last semester, for which, again, I greatly appreciate and thank you. The measurement project data analysis and write up is nearly complete, and, when it is finished, I'll supply the instrument plus training materials to **(Principal of The School)**.

I know you are busy, so I'll be quick. The message below spoke of "other research" in connection with the measurement project. As part of my doctoral research, I need to go into **(The School)** again to videotape record this year's second graders (this time, my own son will be one of them!). The procedures will be the same as last time, except that **THIS TIME**, I will be giving the students some informal feedback between each of the three readings, encouraging them, pointing out some trouble spots here and there as needed, and modeling for them how the reading should take place. It should take about the same amount of time.

I have already spoken with **(Principal of The School)**, explaining what had taken place earlier when the previous **(Principal of The School)** was there. He and his 2nd grade teachers are ready and willing to assist me. They even already have a room set aside for me to use (their reading resource room)!

Again, this will all be done by me. The teachers will only need to assist me in distributing the parental permission forms to the students, and getting them back from the students. I want to keep any disruptions to a minimum.

I appreciate your past help, and hope for your help once more in my research endeavors to help children in the improvement of their reading skills.

Sincerely,

Reo H. McBride
HP: 801-371-2113 Mobile: 801-360-2658

-----Original Message-----

From: **(Superintendent)**
Sent: Wednesday, April 02, 2003 9:52 AM
To: rh2@

Cc: **(Principal of The School)**

Subject: Re: Reo McBride's Research project

Reo, Your research is approved. Good luck to you. Please work closely with **(Principal of The School)** to minimize disruption to the regular operations of the classrooms in which you are working. Thanks, **(Superintendent)**

>>> "Reo H. McBride" <rh2@ > - 3/31/03 4:18 PM >>>

Dear **(Superintendent)**,

I am Reo McBride, a graduate student working toward my Doctorate in the Instructional Psychology & Technology Department in the McKay School of Education. I have been working with your secretary, , to arrange an appointment with you. I totally can understand how busy you are, with your having to wear two hats, as explained to me. Whew!

I recently dropped off with a copy of my Measurement Project Proposal for you to peruse. It has already gone through the McKay School of Education approval processes. I have planned the project in such a way as to not be very disruptive, just videotape children reading so as to develop and test a measurement instrument in regards to fluent oral reading with expression.

I would greatly appreciate it if you could work me in sometime in your schedule. Maybe a brief discussion with me, even over the phone, will help the process along. **(Principal of The School)**, at **(The School)**, has already expressed an interest and willingness to help, pending your approval of course. If the measurement instrument does in fact, perform as we hope it does, after it has been refined, I am more than willing to let the school district use it, especially **(The School)**!

Other research I am needing to do, based on this Measurement Project, cannot be done until I find out, one way or the other, if I can carry on with this project as I am hoping to do.

Your help in this matter is GREATLY appreciated. I hope to hear from you at your earliest possible convenience.

Sincerely,

Reo H. McBride
Phone: 371-2113

-----Original Message-----

From: **(Superintendent)**

Sent: Tuesday, February 25, 2003 8:02 AM

To: **(Principal of The School)**

Subject: Re: Research project

(Principal of The School), You have no obligation in this matter. Generally, I allow research if we have some interest in it (or if it is useful to us) and it is not too disruptive. This research looks interesting but by agreeing to examine the proposal, I make no commitment to move ahead.

Thanks, **(Superintendent)**

>>> **(Principal of The School)**- 2/24/03 10:13 PM >>>

We discussed that this morning. If he is willing to gear it to 2nd grade, I am, but I will need to check with them. I believe they will be willing.

>>> **(Superintendent)** 02/24/03 17:10 PM >>>

Yes, After **(The University)** Institutional Review Board has approved the research, then he need to send an abstract of the proposal to me and include any instruments he might use. Also, do you agree to use your teachers in his research? **(Superintendent)**

>>> **(Principal of The School)**- 2/24/03 4:13 PM >>>

One of the **(The School)**'s parents is working on a doctorate in Instructional Science at **(The University)** and therefore is involved in a research project. His project is to test the validity of a reading test for K-3. He has asked if we would participate and knows that he must first get approval from the district. How does he proceed? Is there a form?

(Sorry, I am sure I don't have all the details correct about his project, because I wasn't clear on whether he developed the test or is testing someone else's.)

Thanks for your help with this. I know we are inundated with research requests. I would like to help him if we can. **(Principal of The School)**

Reo H. McBride, Doctoral Candidate
(The University)
A Central Utah City, USA
W.P. 1-801-422-3536
H.P. 1-801-371-2113
Email: rhm2@email.byu.edu

26 February 2003

(Principal of The School)

Attention:

Dear

Please find attached a Letter of Confidentiality and Letter of Permission for those children whose parents will allow them to assist me in my research.

As part of my doctoral program in (the university in a central city in Utah), USA, I have the opportunity to complete research related to the area of fluent oral reading with expression (FORE). Specifically, I have created a measurement instrument that measures the qualities or characteristics of FORE. But in order to validate the instrument, or determine if it really measures what I hypothesize it measures, I need to videotape your school's 2nd – 6th grade students three times each. Students will be chosen randomly, and will only be identified as "Student 1 or Student 2" as appropriate. These videotaped recordings will take place on the same day, in the same setting, and last only about 10-15 minutes a piece. The parent or parents of each child may be present if they wish. Reading materials used will be grade-level, age-and-content-level appropriate, and in compliance with (the university)'s standards for the use of human subjects. Recordings will take place in whatever room you designate as appropriate.

Once videotaped, I will seek the help of three to four reading teachers to use the instrument to rate each subject's reading. While this effort is to determine if the FORE measurement instrument validly measures the constructs or qualities unique to FORE, the teachers, if they are associated with the students, may think of some teaching activities to assist the students in their reading skills. At no time will any identifying information be reported in the project. Confidentiality will be protected at all times. Be assured that a child's participation is considered to be voluntary, and that refusal to participate will involve no penalty. The child, parent and school may discontinue participation at any time.

The only individuals who will have access to this videotape and the ensuing results, will be the raters, the five professors who sit on my Doctoral Committee, and myself. All

professors currently reside in the State of Utah, and are part of the **(The University)** faculty, and are bound by the same standards of confidentiality and ethics.

Due to the fact that my research is in the area of English reading by those whose native language is English, my desire is to contact the parents of students whose first language is English. I hope to do this through the help of your well-qualified 2nd - 6th grade teachers, who have so graciously assisted me in the past. I also would like the parents/guardians/child to sign the appropriate forms, as required by **(The University)**'s Institutional Review Board (IRB) for Human Subjects.

The final form of the FORE Measurement Instrument, and its related studies, and the training materials associated with it, will be freely available to you, upon request – a gesture of appreciation for **(The School)**'s help

Thank you ahead of time for your help. Your assistance in this matter is greatly appreciated.

Yours faithfully,

Reo H. McBride

Letter of Confidentiality

To Whom It May Concern:

This letter is to inform you that I, Reo H. McBride, will not share any information about **(The School)** its faculty, staff, or students. The research that I am undertaking is designed to study the qualities and characteristics of fluent oral reading with expression (FORE). Any subsequent studies stemming from this research will not be used to reveal in any way, information regarding the school, its staff, faculty or students.

Any data collected will only be used to determine if the measurement instrument validly measures the characteristics unique to FORE. At no time will any identifying information pertaining to a child be reported in the project, or any subsequent studies. Confidentiality will be protected at all times. Children will only be identified as "Student 1, or Student 2," for administrative purposes only during the study.

Those who will have access to the data collected, to include the videotaping and ensuing results, will only be the raters, and the five professors who sit on my Doctoral Committee, and myself. These professors currently reside in the State of Utah, and are part of the **(The University)** faculty. As scientists in the field, they understand the need for strict confidentiality, as do I.

Signed this date, _____

Reo H. McBride _____

W.P. 1-801-422-3536

H.P. 1-801-371-211

Email: rhm2@email.byu.edu

Dear Parents,

18 March 2004

Attached is information concerning research that will be performed here at **(The School)**. We would appreciate it if you and your child could read and sign the appropriate letters/forms and return them to _____ no later than _____. Your help in this matter is greatly appreciated.

The _____ Grade Team

_____ School

Parent/Guardian Informed Consent Form

I hereby allow my child, _____, to be videotaped by Reo H. McBride, who is a graduate doctoral student with (the university). I understand that this videotaped recording will only be used to assist Mr. McBride in his research of the qualities and characteristics of fluent oral reading with expression (FORE). I understand that there will be three videotaped segments of my child reading, (*read and record, feedback, read and record, feedback, read and record*) and that the videotaped session will take place on the same day, in the same setting, lasting a total of only about 15 minutes, and that I am invited, if I so choose, to be present during the videotaping session. I also understand that my child will be given feedback pertaining to his/her reading between each videotaped segment.

I further understand that the reading materials used will be grade-level, age-and-content-level appropriate. I also understand that only those having access to the videotaped recording will be the raters, who will view the videotaped recordings of the children reading, the professors who sit on Mr. McBride's Measurement and Design Project Committees, the professors who sit on Mr. McBride's Doctoral Committee, and Mr. McBride himself.

In as much as only videotaping is involved, I understand that the risks toward my child will be minimal. If my child does feel some initial nervousness about having to perform in front of a camera, I understand that Mr. McBride, as the principal investigator, will strive to help my child feel at ease. My child will be free to stop his/her participation at anytime. The benefits resulting from my child's participation, however, will be the development of a valid measurement instrument that accurately measures the constructs unique to fluent oral reading with expression, as well its supporting theory.

I further understand that participation is voluntary, and that my child may discontinue participation at any time. I understand also that if I do not return this form, my child **WILL NOT** participate in this study. If I have any questions or concerns, I may contact Mr. McBride at the following numbers: W.P. 1-801-422-3536. H.P. 1-801-371-2113. email: rhm2@email.byu.edu (do not include a period at the end of the email address).

I understand also that if I have any questions regarding the rights of my child as a participant in this research project, I may contact Dr. Shane S. Shulthies, Chair of the Institutional Review Board, 120 B, Richards Building, (**The University**). phone, (801) 422-549.

Parent/Guardian Signature: _____

Date: _____

Child Informed Consent Form

I, _____, agree to let Mr. McBride videotape me three times while reading. I understand that he will use the videotaped recording only to help him in his measurement and design projects and/or dissertation research. I also understand that no identifying information will be reported in this project or any studies resulting from this project. I also understand that he will provide feedback to me after each time I read.

Also, I understand that my participation is voluntary, and that should I feel too nervous about being videotaped, that I may refuse to participate. I also understand that if I do choose to participate, that I will be contributing to the development of a valid measurement instrument that will aid in the improvement of children's reading skills as well as that instrument's supporting theory.

Student Signature: _____

Date: _____

Appendix C
Reading Texts Used in Study

A Brick To Cuddle Up To

Imagine shivering on a cold winter's night. The tip of your nose tingles in the frosty air. Finally, you climb into bed and find the toasty treat you have been waiting for—your very own hot brick.

If you had lived in colonial days, that would not sound as strange as it does today. Winters were hard in this New World, and the colonists had to think of clever ways to fight the cold. At bedtime, they heated soapstones, or bricks, in the fireplace. They wrapped the bricks in cloths and tucked them into their beds. The brick kept them warm at night, at least for as long as its heat lasted.

Before the colonists slipped into bed, they rubbed their icy sheets with a bed warmer. This was a metal pan with a long wooden handle. The pan held hot embers from the fireplace. It warmed the bedding so well that sleepy bodies had to wait until the sheets cooled before climbing in.

Staying warm wasn't just a bedtime problem. On winter rides, colonial travelers covered themselves with animal skins and warm blankets. Tucked under the blankets, near their feet, were small tin boxes called foot stoves. A foot stove held burning coals. Hot smoke puffed from the small holes in the stove's lid, soothing freezing feet and legs. When the colonists went to Sunday services, their foot stoves, furs, and blankets went with them. The meeting houses had no heat of their own until the 1800s.

A Wrinkle In Time

It was a dark and stormy night.

In her attic bedroom Margaret Murry, wrapped in an old patchwork quilt, sat on the foot of her bed and watched the trees tossing in the frenzied lashing of the wind. Behind the trees clouds scudded frantically across the sky. Every few moments the moon ripped through them, creating wraithlike shadows that raced along the ground.

The house shook.

Wrapped in her quilt, Meg shook.

She wasn't usually afraid of weather.---It's the weather on top of everything else. On top of me. On top of Meg Murry doing everything wrong.

School. School was all wrong. She'd been dropped down to the lowest section in her grade. That morning one of her teachers had said crossly, "Really, Meg, I don't understand how a child with parents as brilliant as yours are supposed to be can be such a poor student. If you don't manage to do a little better you'll have to stay back next year."

During lunch she'd rough-housed a little to try to make herself feel better, and one of the girls said scornfully, "After all Meg, we aren't grammar-school kids any more. Why do you always act like such a baby?"

Brave Irene

Mrs. Bobbin, the dressmaker, was tired and had a bad headache, but she still managed to sew the last stitches in the gown she was making.

“It’s the most beautiful dress in the whole world!” said her daughter, Irene. “The duchess will love it.”

“It is nice,” her mother admitted. “But, dumpling, it’s for tonight’s ball, and I don’t have the strength to bring it. I feel sick.”

“Poor Mama,” said Irene. “I can get it there!”

“No, cupcake, I can’t let you,” said Mrs. Bobbin. “Such a huge package, and it’s such a long way to the palace. Besides, it’s starting to snow.”

“But I love the snow,” Irene insisted. She coaxed her mother into bed, covered her with two quilts, and added a blanket for her feet. Then she fixed her some tea with lemon and honey and put more wood in the stove.

With great care, Irene took the splendid gown down from the dummy and packed it in a big box with plenty of tissue paper.

“Dress warmly, pudding,” her mother called in a weak voice, “and don’t forget to button up. It’s cold out there, and windy.”

Going to the Swimming Pool

On a hot summer day there's nothing I like better than going to the pool. Besides cooling off in the water, there are lots of things to do. I can swim laps or have races with my friends. I can do a cannonball when I jump in. I like to jump in with a big splash when my friends are not looking so I get them all wet. Sometimes I pretend I'm a giant whale, and sometimes we play games like water tag.

There are water slides at the pool, too. One slide is very tall and crooked. It tosses you out in the water when you get to the bottom. The other slide is wide and you can go down it with your friends.

I like the wave pool the best. My friends and I watch for the big waves to come our way. We body surf on top of the wave and let it move us across the pool. Sometimes my friend and I get on a raft and wait for the wave to push us. Sometimes we just float along when the waves come.

You can have a great time if you just remember the rules: no running and no pushing anyone into the water. Mom has her own rule. She says we should always remember to wear our sunscreen to protect our skin.

The Ant Hill

Dad and I took a hike in the woods. We walked for a long time and stopped to take a rest. We sat down on a log and had a drink of water. A big hill was nearby.

Dad said, "Look, there's an ant hill."

I walked up to the hill and took a closer peek. At first it looked just like a dirt hill. Then I noticed a few ants running around. I looked closer. I saw little ants carrying pieces of mushroom. The pieces were almost as big as the ants.

"What are they doing, Dad?" I asked.

"They're taking food inside the hill. They probably have thousands of ants to feed inside." Dad said, "Watch this." He gently poked a twig into a small hole on the hill. All of a sudden, many ants came out.

"The ants are on alert, trying to protect their hill," he said. I bent down to look closer. Some ants climbed on my shoes.

"We better leave now," Dad said. Dad and I walked and walked until we were home. Now whenever I see one ant, I stop and think about the city of ants they might be feeding and protecting.

The Sun

Did you know sunshine actually comes from a star? That's because the sun is a star just like the millions of stars in the sky. The sun is so big that more than a million Earths can fit inside it. The sun is not the biggest star, though. Many stars are actually bigger and brighter. The sun looks bigger and brighter because the Earth is closer to the sun than any other star.

The sun is a huge ball of glowing gases. It's so hot you could never touch it. The temperature of the surface is one hundred times hotter than the hottest summer day. Sometimes the sun's surface gets so hot it creates solar flares. Sometimes solar flares cause difficulty on Earth. Solar flares can cause static on radio stations. When solar flares are large, they can even cause electric power failures.

The sun's energy reaches us in the form of heat, light, and radio waves. The sun is millions of miles away from us. It takes the sunlight about eight minutes to travel to the Earth.

The sun gives us light and heat. Without it, no plants, animals, or humans could grow or survive. The sun gives off so much light it can be converted into solar energy. Solar cells convert sunlight into electricity. Solar cells can be used to provide power for cars and lights. Some solar cells are as small as a stick of gum and some are as big as a football field.

Appendix D
Institutional Review Board Permissions

K. RICHARD YOUNG, PH.D.
Associate Dean

March 19, 2003

Reo McBride
138 West 1940 North #160


Dear Reo:

Your research proposal entitled, "Construct-Linked Scales Measuring Fluent Oral Reading with Expression," was determined to be Expedited status. Thank you for submitting your proposal to the Institutional Review Board for Human Subjects Subcommittee. You are approved to begin your research. This approval is good for a maximum of one year, at which time, and sooner as need arises, the study will be reviewed again if the work is still in progress.

The research appears to pose minimal risk to human subjects and meets the Federal guidelines. It is determined that the research as outlined provides adequate protection for human subjects and the consent form will provide the necessary information. Any changes in the consent form, data collection instruments, or procedures will need to be reviewed by the Human Subjects Committee.

Since you are planning to conduct your research in a Public School Partnership district, you will need to contact the research director of the district for official approval to do research in the district.

Sincerely,


K. Richard Young
Associate Dean for Research

seh

INSTITUTIONAL REVIEW BOARD
FOR HUMAN SUBJECTS

November 11, 2003

Reo McBride
138 West 1940 North #160

Dear Reo:

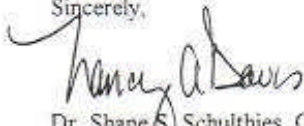
Thank you for your recent correspondence concerning the changes/addendum you have made to your protocol entitled "Construct-Linked Scales Measuring Fluent Oral Reading with Expression." The research appears to pose minimal risk to human subjects and meets the Federal guidelines.

You are approved to continue your research. This protocol will be reviewed on the original approval date, or sooner as need arises, if the work is still in progress.

Please notify Nancy Davis of any changes made in the instruments, consent form, or research process before instigating the alterations, so that they can be approved.

If you have any questions, please let us know. We wish you well with your research!

Sincerely,



Dr. Shane S. Schulthies, Chair /
Nancy A. Davis, CIM, Administrator
Institutional Review Board for Human Subjects
SSS/sgf

Message

Page 1 of 2

Nancy Davis

To: rhm2@

Subject: RE: Reo McBride's Research at ADDENDUM

Reo:

You are approved to increased your subject pool. Good luck with your research.

Nancy

Nancy A. Davis

Research & Creative Activities

-----Original Message-----

From: rhm2@email

Sent: Monday, December 08, 2003 10:25 AM

To: Nancy Davis

Subject: FW: Reo McBride's Research at ADDENDUM

Dear Nancy,

In regards to our recent phone conversation, the Principal at [redacted] has agreed to let me videotape the 1st, 3rd, 4th, 5th, & 6th grade students while reading out loud, in order to obtain more subjects. This is not just a nicety on his part. I asked him this in particular because the measurement instrument data my professor, Dr. Vic Bunderson, and I are collecting requires a statistical procedure, called a factor analysis (FA) to be completed. An FA will require at least 120 subjects (more is better, but that is the minimum). Unfortunately, I was not aware that an FA would need so many subjects (I am new at statistics!). When I received permission from the parents of the 2nd grade children, I was only able to net 40 permission slips allowing me to videotape students. This number of 40 is way too small to be used with an FA, and obtain valid results. Our goal is to increase the validity argument of the measurement instrument. Performing the FA with the greater number of subjects will enable us to do just that.

The same procedures will follow as before, with letters being sent to both parents and students. I have attached those letters to this email.

Please accept this letter as an addendum to the research protocol that I am conducting.

Your help has been and continues to be greatly appreciated! ☺

From: Reo H. McBride
138 W. 1940 N. #160

Email: rfm2@

To: Skylar Rencher
IRB Secretary
A-261

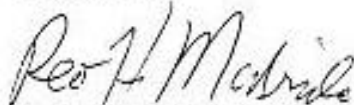
20 February 2004

Re: Annual Review of Approved Research IRB for Human Subjects;
"Construct-Linked Scales Measuring Fluent Oral Reading with Expression

The work continues on as before. I should be finished with videotaping the 3rd graders at next week. Once I have finished with them, I plan to videotape the 4th graders, 5th graders, 6th graders, and, time permitting, the 1st graders. (The videotaping of the other grades was previously approved by the principal of the school, and Nancy Davis of the IRB committee.) The extra grades are needed so as to collect a greater amount of data as I build the validity argument in the area reading fluency research I am conducting. Due to the numbers and extra time needed to videotape these additional grades, I request a month's extension. The same IRB informed consent criteria will be carried out as before.

Administration, teachers and students have been very supportive and willing to help. My mentor and advisor, Dr. Bunderson, and I have been asked to present our research findings to the principal and his faculty upon completion. We will gladly accommodate their request, following all IRB requirements in protecting individuals' confidentiality.

Respectfully,



Reo H. McBride
Graduate Student
Instructional Psychology & Technology Department

INSTITUTIONAL REVIEW BOARD
FOR HUMAN SUBJECTS

March 18, 2004

Reo McBride
138 West 1940 North #160

Dear Reo:

In accordance with policy and Federal Regulations, your research proposal, "Construct-Linked Scales Measuring Fluent Oral Reading with Expression" (Protocol Number: 03-0294) was re-reviewed. All active protocols must be reviewed annually by the Human Subjects Committee. The research continues to pose minimal risk to human subjects and meets the Federal guidelines. It is determined that the research as outlined provides adequate protection for human subjects and the consent form will provide the necessary information.

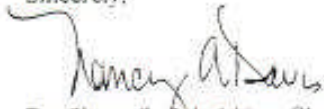
You are approved to continue your research. This approval is good until March 18, 2005 (a year from your last annual review date).

Enclosed is a date stamped consent form. Please use this in obtaining consent. We will be sending a continuing review form before the expiration date. Please fill this form out in a timely manner to insure that there is not a lapse in your approval.

Any changes in the consent form, data collection instruments or procedures will need to be reviewed by the Human Subjects Committee.

If you have any questions, please let us know. We wish you well with your research!

Sincerely,



Dr. Shane S. Schulthies, Chair
Nancy A. Davis, CIM, Administrator
Institutional Review Board for Human Subjects
SSS/sgf

Enclosure

Appendix E
Facets Generated Reports

obsvd Score	obsvd Count	obsvd Average	Fair-M Average	Model Measure	Model S.E.	Infit MnSq	Zstd	Outfit MnSq	Zstd	Estim. Discrm	Exact Obs %	Agree. Exp %	N rater
11140	2783	4.0	4.09	-.17	.03	.91	-3.3	.93	-2.3	1.09	50.7	42.9	4 raterD
11035	2784	4.0	4.05	-.08	.03	1.15	5.1	1.13	4.2	.84	45.9	42.8	3 raterC
10906	2800	3.9	3.98	.07	.03	1.06	2.1	1.06	2.1	.94	47.6	42.6	2 raterB
10729	2790	3.8	3.93	.17	.03	.81	-7.4	.87	-4.9	1.13	45.9	42.2	1 raterA
10952.5	2789.3	3.9	4.01	.00	.03	.98	-.9	1.00	-.2				Mean (Count: 4)
153.4	6.8	.1	.06	.13	.00	.13	4.9	.10	3.6				S.D.

RMSE (Model) .03 Adj S.D. .13 Separation 4.62 Reliability .96
Fixed (all same) chi-square: 89.4 [d.f.: 3 significance: .00
Rater agreement opportunities: 16671 Exact agreements: 7922 = 47.5% Expected: 7104.0 = 42.6%

Figure E1. Facets generated Table 7.2.1—rater measurement report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Num examinee
560.5	24.0	23.4	23.49	.35	.07	.98	-.2	.99	-.1		Mean (Count: 200)
75.9	.4	3.1	3.15	.31	.02	.41	1.4	.41	1.4		S.D.

RMSE (Model) .07 Adj S.D. .30 Separation 4.27 Reliability .95
Fixed (all same) chi-square: 3585.3 d.f.: 199 significance: .00

Figure E2. Facets generated Table 7.1.1—examinee measurement report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Model Measure	Model S.E.	Infit Mnsq	Infit Zstd	Outfit Mnsq	Outfit Zstd	Estim. Discrm	Nu rating scale
3491	796	4.4	4.45	-1.08	.06	1.09	1.6	1.09	1.5	.89	1 accuracy1
3491	798	4.4	4.44	-1.05	.06	1.35	5.9	1.39	5.9	.55	2 accuracy2
3314	796	4.2	4.23	-.49	.06	1.24	4.3	1.19	3.3	.73	4 phrasing1
3259	795	4.1	4.16	-.34	.05	1.23	4.2	1.20	3.5	.70	7 smoothness2
3247	797	4.1	4.14	-.28	.05	1.01	.1	1.03	.5	.98	5 phrasing2
3231	799	4.0	4.11	-.21	.05	1.25	4.5	1.26	4.6	.66	3 accuracy3
3116	796	3.9	3.98	.07	.05	.89	-2.1	.88	-2.3	1.11	10 rate2
3099	798	3.9	3.95	.14	.05	.87	-2.6	.84	-3.3	1.19	13 confidence1
3071	792	3.9	3.94	.15	.05	.99	-.2	.97	-.5	.99	8 smoothness3
3073	795	3.9	3.93	.18	.05	.89	-2.1	.86	-2.9	1.17	6 smoothness1
3053	799	3.8	3.89	.26	.05	.73	-5.7	.76	-5.2	1.28	9 rate1
2867	799	3.6	3.65	.71	.05	.90	-2.0	.90	-2.1	1.14	11 expression1
2856	799	3.6	3.64	.74	.05	.97	-.5	.97	-.5	1.03	12 expression2
2642	798	3.3	3.37	1.20	.05	.62	-8.8	.62	-8.7	1.43	14 confidence2
3129.3	796.9	3.9	3.99	.00	.05	1.00	-.2	1.00	-.4		Mean (Count: 14)
228.5	2.0	.3	.29	.62	.00	.20	4.0	.20	3.9		S.D.

RMSE (Model) .05 Adj S.D. .62 Separation 11.76 Reliability .99
Fixed (all same) chi-square: 1931.4 d.f.: 13 significance: .00

Figure E3. Facets generated Table 7.3.1a—rating scale measurement report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit Mnsq	Infit ZStd	Outfit Mnsq	Outfit ZStd	Estim. Discrm	N rating construct
20464	799	25.6	26.03	-.24	.01	1.19	3.5	1.19	3.4	.41	1 accuracy
19750	799	24.7	25.11	-.13	.01	1.40	7.1	1.40	7.0	1.12	2 phrasing
18923	799	23.7	24.07	-.02	.01	.71	-6.4	.70	-6.6	.46	3 smoothness
18543	799	23.2	23.60	.03	.01	.75	-5.5	.75	-5.4	1.32	4 rate
17247	799	21.6	22.04	.17	.01	.76	-5.4	.76	-5.3	1.30	6 confidence
17169	799	21.5	22.00	.18	.01	1.16	3.0	1.15	2.9	1.29	5 expression
18682.7	799.0	23.4	23.81	.00	.01	1.00	-.6	.99	-.7		Mean (Count: 6)
1207.1	.0	1.5	1.48	.15	.00	.27	5.3	.27	5.3		S.D.

RMSE (Model) .01 Adj S.D. .15 separation 12.86 Reliability .99
Fixed (all same) chi-square: 969.4 d.f.: 5 significance: .00

Figure E4. Facets generated Table 7.3.1b—rating construct measurement report

Appendix F

Factor Analysis Output with 14 Variables, NoSuppressions

	Three Factor Structure Matrix with No Suppressions		
	1 (Fluency)	2 (Accuracy)	3 (Phrase Boundaries)
Con2: CPA	.963	.445	.530
Con1: SISS	.920	.549	.370
Smo1: LA	.920	.606	.407
Exp2: CIM	.913	.328	.609
Exp1: SAI	.882	.314	.594
Rate1: PACE	.871	.539	.406
Smo3: NEWP	.776	.626	.223
Rate2: NIB	.706	.422	.400
Acc1: AIRW	.576	.893	.121
Acc2: PRON	.286	.731	.044
Acc3: RTAW	.474	.730	.246
Smo2: NR	.320	.513	.229
Phr1: PUNCT	.505	.223	.922
Phr2: PB	.752	.374	.768

Figure F1. Three-factor structure matrix

	Three Factor Pattern Matrix with NoSuppressions		
	1(Fluency)	2(Accuracy)	3(Phrase Boundaries)
Con2: CPA	.993	-.089	.034
Con1: SISS	.954	.057	-.126
Exp2: CIM	.937	-.192	.155
Exp1: SAI	.901	-.187	.157
Smo1: LA	.869	.150	-.058
Rate1: PACE	.837	.097	-.035
Smo3: NEWP	.732	.260	-.185
Rate2: NIB	.632	.077	.066
Acc2: PRON	-.158	.813	.019
Acc1: AIRW	.189	.803	-.080
Acc3: RTAW	.031	.696	.139
Smo2: NR	-.051	.516	.189
Phr1: PUNCT	-.030	.119	.922
Phr2: PB	.442	.069	.533

Figure F2. Three-factor pattern matrix

Appendix G

Factor Analysis Output with 13 Variables, NoSuppressions

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	7.964	61.261	61.261	7.716	59.355	59.355	7.471
2	1.708	13.140	74.401	1.369	10.533	69.888	4.834
3	.940	7.229	81.630				
4	.554	4.262	85.892				
5	.489	3.764	89.656				
6	.354	2.723	92.379				
7	.240	1.844	94.224				
8	.197	1.519	95.742				
9	.193	1.487	97.230				
10	.154	1.181	98.411				
11	.100	.769	99.180				
12	.059	.451	99.632				
13	.048	.368	100.000				

Figure G1. Principal axis factors, eigenvalues, and variance accounted for by each factor

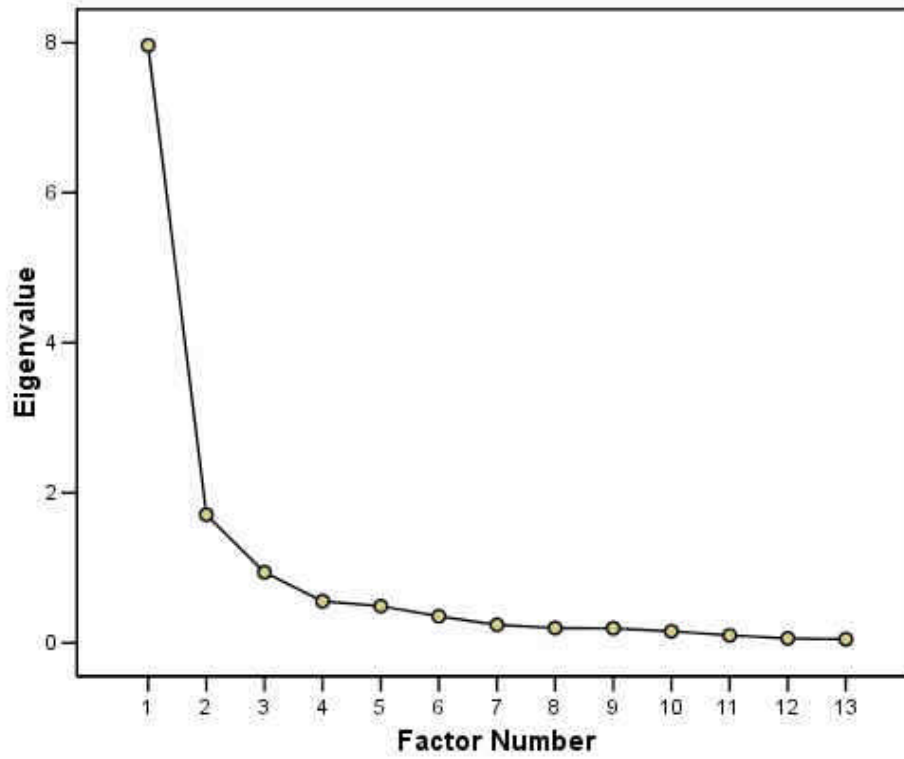


Figure G2. Scree plot showing eigenvalues

	Two Factor Pattern Matrix with NoSuppressions	
	1(Fluency)	2(Accuracy)
Exp2: CIM	1.040	-.174
Exp1: SAI	1.003	-.163
Con2: CPA	.967	-.010
Phr2: PB	.848	-.079
Con1: SISS	.758	.214
Smø1: LA	.732	.273
Rate1: PACE	.728	.207
Phr1: PUNCT	.671	-.165
Rate2: NIB	.643	.104
Smø3: NEWP	.477	.425
Acc1: AIRW	-.040	.950
Acc2: PRON	-.254	.851
Acc3: RTAW	.077	.642

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Figure G3. Two-factor pattern matrix

	Two Factor Structure Matrix with NoSuppressions	
	1(Fluency)	2(Accuracy)
Con2: CPA	.961	.562
Exp2: CIM	.937	.442
Exp1: SAI	.906	.430
Smo1: LA	.894	.707
Con1: SISS	.885	.663
Rate1: PACE	.851	.639
Phr2: PB	.801	.423
Smo3: NEWP	.728	.707
Rate2: NIB	.704	.485
Phr1: PUNCT	.573	.232
Acc1: AIRW	.523	.927
Acc2: PRON	.250	.701
Acc3: RTAW	.457	.688

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.

Figure G4. Two-factor structure matrix

Factor Correlation Matrix

Factor	1(Fluency)	2(Accuracy)
1(Fluency)	1.000	.592
2(Accuracy)	.592	1.000

Extraction Method: Principal Axis Factoring.

Rotation Method: Promax with Kaiser Normalization.

Figure G5. Two-factor correlation matrix