EXPLANATORY ITEM RESPONSE MODELING OF A READING COMPREHENSION ASSESSMENT

by

SUNHI PARK

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2020

DISSERTATION APPROVAL PAGE

Student: Sunhi Park

Title: Explanatory Item Response Modeling of a Reading Comprehension Assessment

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy, and Leadership by:

Kathleen Scalise	Chairperson
Gina Biancarosa	Core Member
Julie Alonzo	Core Member
Sylvia Thompson	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2020

DISSERTATION ABSTRACT

Sunhi Park

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

June 2020

Title: Explanatory Item Response Modeling of a Reading Comprehension Assessment

This dissertation study intended to explore the validity of items by modeling the relationship between the item and person properties and item difficulty a reading comprehension assessment, the Multiple-choice Online Causal Coherence Assessment (MOCCA). This study used explanatory item response modeling (EIRM) which was useful to help explain how the properties were associated with responses to items. Results from the linear logistic test model analysis indicate that the item difficulty was significantly associated with text complexity and story features. An item was more difficult if a text passage in a given item was longer with less familiar and less concrete words in a less familiar topic, and if the passage was not child-centered and/or realistic, having the goal and a main character in the same sentence, implying a goal in the later sentence, and having a goal not met with no positive emotion in the end. Results from the late regression analysis indicate that race/ethnicity represented as white/non-white, socioeconomic status represented as free and reduced meals, special education participation, and EL status were statistically significantly related to the responses on the MOCCA items. English learners among subgroups were unique in that they were assessed on the same reading comprehension assessment as non-ELs. Their different performance was easily ascribed to their limited English language proficiency. To explore whether any items on MOCCA

might involve construct-irrelevant factors beyond the language differences between ELs and non-ELs, twelve items were identified with showing different group responses through IRT-based Differential Item Functioning (DIF) analysis. Results indicate that the twelve DIF items did not show particularly distinctive text features compared to no DIF items when they were reviewed by means of text complexity and story features both of which were used as predictors in the LLTM analysis. The findings from DIF analysis served to detect whether a given assessment measured an intended construct equally for all subgroups, and whether there were any items indicative of unexpected behavior on the assessment. The results provided information of the characteristics of items which are related to test validity as well as fairness.

CURRICULUM VITAE

NAME OF AUTHOR: Sunhi Park

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, U.S.
University of Manchester, Manchester, U.K.
Pusan National University, Busan, South Korea

DEGREES AWARDED:

Doctor of Philosophy, Qualitative Research Methods, 2020, University of Oregon
Master of Arts, Teaching English to Speakers of Other Languages (TESOL), 2009, University of Manchester
Bachelor of Arts, English Language Education, 1993, Pusan National University

AREAS OF SPECIAL INTEREST:

Development of reading assessment
Exploring the way to improve and design valid, reliable, and equitable reading comprehension assessment
Item Response Modeling and Differential Item Functioning
Applying psychometrical methods to understand how an assessment works for the participants
Teaching and assessing ESL/EFL learners
Developing and improving curriculum, instructions, and assessments for ESL/EFL learners
Backwash effect of assessment on professional development and curriculum
Integrating the findings and knowledge of test development into educators' professional development and curriculum for learners

PROFESSIONAL EXPERIENCE:

Graduate Research Assistant for Assessment Innovation Project (AIP),
Center on Teaching and Learning, University of Oregon, June 2017—present

- I have conducted data analysis for descriptive statistics, reliability, validity, and ROC analysis for innovated DIBELS 8th Edition, using mainly R.
- I have taken responsibility of developing Maze, a subtest of DIBELS 8th Edition, from passage development to data analysis.

- I have conducted literature review, write and review technical manuals, and revise administration guides for DIBELS 8th Edition.

Graduate Employee for Administrative Licensure Program, University of Oregon, Fall 2016—Summer 2018

- I assisted the program director and coordinator with revising the exemplar materials, preparing for program materials, and grading papers.
- I monitored teaching and learning during the course for the participants in class and via Zoom.

Graduate Employee for the Center on Diversity and Community (CoDac), Fall 2015—Spring 2016, University of Oregon

- I conducted data analyses for understanding the extent of diversity among faculty and students, and for creating reports.
- I created the surveys on Qualtrics and analyzed the results to report.
- I assisted with the events hosted by CoDac and Devision of Equity and Inclusion.

PUBLICATIONS:

Choo, S., **Park, S.**, & Nelson, N. (2020). Evaluating spatial thinking ability using Item Response Theory: Differential Item Functioning across math learning disabilities and geometry instructions. *Learning Disability Quarterly*, 0731948720912417.

Scalise, K., Irvin, P. S, Alresheed, F., Zvock, K., Yim, H., **Park, S.**, Landis, B., Meng, P., Kleinfelder, B., Halladay, L., & Partsafas, A. (2018). Accommodations in digital interactive STEM assessment tasks: current accommodations and promising practices for enhancing accessibility for students with disabilities. *Journal of Special Education Technology, 33*(4), 219-236.

Scalise, K., **Park, S.**, Mustafic, M., & Greiff, S. (rejected). Regaining lost opportunities: PISA designs for a next generation sequel to spiraling contextual questionnaires. *Assessment in Education: Principles, Policy & Practice.*

Anderson, D., **Park, S.**, Alonzo, J., Tindal, G. (2015). *An Exploration of Differential Item Functioning with the easyCBM Middle School Mathematics Tests: Grades 6-8* (Technical Report No. 1501). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

ACKNOWLEDGMENTS

The journey of dissertation is similar to a quest to find breakthroughs whenever I is stuck with something. And at the same time it is a blessing that I realize there is always someone who is willing to support me, ready to share me with what they have, expertise, guidance, love and care. I would like to give my special thanks to my sister who has always been supporting me in my life. Though she was not with me physically, she has been my friend, my shoulder to lean on whom I can share with all good or bad while I was struggling.

I would like to express deepest appreciation to each of my committee members. I thank Dr. Scalise for her guidance in pursing my academic career as a researcher. She was a wonderful resource as an advisor with her expertise as well as patience and consideration. She always gave me answers and solutions, enabling me to move forward till completing this dissertation research. I am deeply grateful to Dr. Biancarosa, who supported my academic achievement and provided me with research opportunities and financial support. I thanked her for allowing me to use her project MOCCA data for my dissertation study. Dr. Alonzo was a great reviewer who provided me with feedbacks and comments for my dissertation. Dr. Thompson was a wonderful researcher with whom I pleasantly shared perspectives and issues about English learners in the U.S.

I also wish to express my deepest gratitude to my friends who have sent their support and encouragement for my academic success. Especially I thanked Therese for her therapeutic, emotional, and spiritual support. She was an amazing friend who was always concerned about my physical and mental wellbeing and academic progress, and listened to me with affection and consideration. I also thanked HyeonJin for her support

as a researcher and a friend. As she was ahead of me in the academic career path, she was a wonderful information provider about research methods as well as life as an international student. Her encouragement and sharing her experiences from Nebraska relaxed my anxiety and helped me find ways out whenever I struggled with my research. I also thanked Hyunsook and Sangchil who were wonderful reviewers and commenters on my dissertation writing with linguistic expertise and practical experiences.

Lastly, I give my everlasting thanks to my friend, Hyeon. Thanks to her, I was motivated to study in Japan and finally to pursue my doctoral degree in the U.S.. She has been an amazing mentor as a friend and as a researcher. I wish she would overcome her fight against cancer with courage.

For my beloved mom and sister

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION	1
1.1 Overview and Background	1
1.2 Validity and Validation	6
1.3 Fairness in Testing	8
1.4 Research Questions	10
1.5 Organization of the Dissertation	12
II. LITERATURE REVIEW	14
2.1 Reading Comprehension	14
2.1.1 Definition	14
2.1.2 Construction-Integration model (CI model)	15
2.1.3 Situation model	17
2.2 Text Features: Factors That Make a Text Difficult	18
2.2.1 Word difficulty	19
2.2.2 Syntactic complexity	21
2.2.3 Coherence and cohesion	23
2.2.4 Text genre	25
2.3 Person Properties: Factors Depending on Learners	26
2.3.1 English learners in the U.S.	28
2.3.2 Academic achievement of English learners	29
2.3.3 ELs and textual effects on assessment	31

Chapter Page

III. METHODS 37

3.1 Multiple-choice Online Causal Comprehension Assessment (MOCCA) 37

3.1.1 Theoretical framework and construct 37

3.1.2 Item format 40

3.1.3 Administration 42

3.1.4 Scoring and interpretation 43

3.1.5 Psychometric features 44

3.1.6 MOCCA in the study 44

3.2 Text Features as Item Property Predictors 45

3.2.1 Text complexity indicators 45

3.2.2 Story features coding 47

3.3 Participants and the Data Set 52

3.4 Data Analytic Methods 53

3.4.1 Explanatory item response models. 54

3.4.2 Differential item functioning (DIF) 60

IV. RESULTS 62

4.1 Phase I: Understanding of the MOCCA Data 63

4.1.1 Descriptive analyses 63

4.1.2 Item analysis by means of classical test theory 72

4.1.3 Selecting a psychometric model 74

4.1.4 Anchoring 78

4.1.5 IRT estimation results by form 79

Chapter Page

4.2 Phase II: Explanatory Item Response Modeling 81
4.2.1 Linear logistic test model and text features 81
4.2.2 Latent regression analysis and person properties 96
4.3 Phase III: English Learners and Differential Item Functioning 100
4.3.1 IRT-based differential item functioning 101
4.3.2 Reviewing DIF items through text features 103
V. DISCUSSION AND CONCLUSIONS 107
5.1 Summary 108
5.1.1 Phase I: Understanding the MOCCA assessment 108
5.1.2 Phase II: EIRM and external variables 109
5.1.3 Phase III: ELs and DIF items 114
5.2 Limitations 115
5.3 Discussion 118
5.3.1 Phase I: Statistical and psychometric understanding of MOCCA 119
5.3.2 Phase II: EIRM and external variables 120
5.3.3 Phase III: ELs and DIF items 132
5.4 Implications 135
5.5 Conclusions 138
APPENDICES 140
A. CTT ITEM DIFFICULTY BY ITEM BY FORM 140
B. IRT ITEM DIFFICULTY BY ITEM BY FORM 142
REFERENCES CITED 144

LIST OF FIGURES

Figure Page

1. Screenshot of MOCCA Item in Directions…….. 42

2. The Plots of Standard Error of Measurement of Rasch Model..... 76

3. The Plots of Standard Error of Measurement of 2PL Model..... 77

LIST OF TABLES

Table		Page
1.	Models as a Function of the Predictors …….....	56
2.	Summary of the Four Models	60
3.	Demographic Information by Subgroup.....	65
4.	Descriptive Statistics by Subgroup	66
5.	Descriptive Statistics of MOCCA Text Construction	70
6.	Model Fit Comparison Indices of the Relative Fit.....	77
7.	IRT Estimation of the Parameters	79
8.	Estimates of the Effects of Text Complexity Indicators	84
9.	Estimates of the Effects of Story Features.....	88
10.	Estimates of the Effects of Story Features via LLTM plus error.....	92
11.	Estimates of Variance Parameters by Model by Form.....	93
12.	Model Comparisons: Rasch, LLTM, and LLTM plus Error…….....	94
13.	Results of Chi-Square Tests.....	95
14.	Estimates of the Effects via Latent Regression Analysis	97
15.	Estimates of Variance Parameters.....	99
16.	Model Comparisons	99
17.	Results of Chi-Square Tests.....	100
18.	DIF Items Across Forms.....	102
19.	Comparison of Text Complexity for DIF and no-DIF Items.....	104
20.	Comparison of Story Features for DIF and no-DIF Items.....	106
21.	Descriptive Summary of Text Features on Item Difficulty of MOCCA…… ..	114

CHAPTER I

INTRODUCTION

1.1 Overview and Background

Reading comprehension ability assessment has been conducted in various formats with various intentions over time. The scores have been used in the past in some settings mostly to rank students or to distinguish broadly between good readers and poor readers. Currently, scores from reading comprehension assessments are used for many additional purposes such as determining students' proficiency in a variety of reading elements; identifying students' strengths and weaknesses; and evaluating how successful individual students, courses, or reading programs have been in achieving the intended objectives. Administrators may also use reading comprehension assessments as part of teacher and school accountability policies (Afflerbach, 2007; Hughes, 2008). The results of reading assessment may also be used today for identifying the sources of difficulty that limit students' comprehension of text, for instance identifying student weaknesses in cognitive processes such as encoding or decoding (Gorin & Svetina, 2012).

As the purposes and demand for reading comprehension assessments have evolved and expanded, the need to develop appropriate reading comprehension assessments technically adequate for the emerging purposes has also grown. Assessments developed for a certain purpose need to provide reliable scores that are appropriate for the decisions that will be made from them to fit for the purpose. The Common Core State Standards require students to read texts that have appropriate complexity and grade level across their education (Nelson, Perfetti, Liben, D., & Liben, M., 2008). This requirement

affects not only curriculum and instruction, but also test development, interacting with why the assessment is taken and who takes the assessment and emphasizing the significance of valid and fair assessment for the target population.

Measurement modeling has served test development by providing a theoretical basis for measurement and principles for item selection and score equating (Embretson & Reise, 2013). Measurement models are used to describe the location of person ability and item difficulty on an assessment, providing information for psychometric analysis to understand an assessment. An explanatory approach to measurement modeling can explain how responses are generated by means of item properties (e.g., questions, responses) and person properties (e.g., gender, language proficiency), and expand our understanding of whether an assessment measures what we intend to measure with respect to the degree of validity and fairness (De Boeck & Wilson, 2004; Wilson & Moore, 2011).

Validity and fairness are overarching concepts inherent in test development. Correct, consistent and fair interpretations and uses of test scores depend on the extent to which the assessments in use measure the intended construct and lack bias. In other words, an assessment should measure accurately what it intends to measure and incorporate all the participants in testing without favoring one group over another for reasons not relevant to the intended construct being assessed. Thus, it is important to verify whether the developed assessments satisfy the demands and needs for the use and interpretation of the assessment scores. Findings from studies that examine different aspects of technical adequacy in relation to validity and fairness have provided essential

information used to support the intended uses of the assessments and improve test design for future purposes.

A fairly recently developed reading comprehension assessment, the Multiple-choice Online Causal Comprehension Assessment (MOCCA) (University of Oregon, University of Minnesota, California State University, Chico, Georgia State University, and University of North Dakota) was developed with the goal of identifying poor comprehenders and good comprehenders. An examinee on the MOCCA assessment is required to fill in the missing sentence to complete a narrative text coherently. Classification of a reader as a good comprehender on MOCCA implies that they have an ability to comprehend the text by means of appropriate causal inferences made from the cues of the text.

Some cognitive psychologists claim that comprehension takes place when a reader connects words and propositions verbally described in the text, and mentally constructs the situations from the direct or inferential connection (e.g., Gorin, 2005; Graesser, Singer, & Trabasso, 1994; van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). In other words, only understanding the text on the local level does not complete comprehension. To build a mental situation for comprehension, a reader needs to make causal inferences in coherently connecting the information presented in the text. MOCCA is designed to measure readers' ability to comprehend the text passages coherently based on respect to inference making.

In this dissertation study, I explored the relations between item difficulty and text factors as cues in the MOCCA assessment. I analyzed the degree to which item difficulty of MOCCA varied depending on the text factors, specifically examining the factors

related to linguistic complexity and story structures. The factors inherent in text may be related to item difficulty in a reading comprehension assessment. Long text with unfamiliar words is likely to be more difficult for readers than short text with everyday words. MOCCA consists of a series of narrative text passages followed by selected response questions in a multiple-choice format. Given this structure, the characteristics of the text passages used in the assessment might play a significant role in the challenges presented to students by the questions.

I also explored the appropriateness of MOCCA for diverse student populations. I analyzed the relations between the examinees' responses and their individual characteristics such as gender, race, socioeconomic status, special education status and language proficiency status. It may be important to understand whether or to what degree person characteristics affected examinees' test performance. The findings will help understand whether the assessment is psychometrically appropriate for the subgroups within the test population.

I focused on English learners (ELs) especially among the subgroups and explored whether they were affected in an unintended way when they were assessed on the same scale with non-English learners as a result of causes other than their language proficiency. If ELs perform differently than non-ELs in a reading comprehension assessment that heavily depends on text passages, their different performance tends to be ascribed to their limited English proficiency (e.g., Abedi, 2002). However, some studies point to these students' different of background knowledge and limited access to content subject matter such as English Language Arts rather than language proficiency as the cause for their different performance (e.g., Carrell, 1983). Thus, there might be some

features of the assessment itself that affected ELs' performance rather than the construct intended to be measured. ELs' different responses to items on reading assessments might be due to their unfamiliarity with the story structures inherent in the texts rather than linguistic knowledge. Texts as items on reading assessments might be constructed without considering the unique characteristics of the EL population such as their diverse background knowledge or their different native or home language structure, and the interpretations and uses of the scores from such assessments may not be as appropriate to make decisions for them.

I approached these research goals by using psychometric and statistical methods, namely tools from Explanatory Item Response Modeling (EIRM; De Boeck & Wilson, 2004), to understand potential effects of textual and individual factors on the item difficulty and responses in a reading comprehension assessment, MOCCA, and IRT-based Differential Item Functioning (DIF), to detect items working differently between ELs and non-ELs on MOCCA. These specific methods were used to help explain the relations between the factors and test performance that demonstrated sources of item (or text) difficulty in the assessment.

My working premise is that if the sources of item difficulty from text features in a reading comprehension assessment are identified and thus predictable, this information may be useful for deciding whether test items provide evidence in support of construct validity. This information also will shed light on how to design assessment observations that better reflect the intended constructs. My hope is that the findings will provide guidance to those developing or improving reading comprehension assessments. Insights learned might help with designing items and building item pools with appropriate texts

according to target examinees' cognitive characteristics (Embretson & Wetzel, 1987; Gorin, 2005). The findings might also provide some guidance useful for teaching reading comprehension when developing curriculum that presents features and processes when readers have difficulty in fully comprehending the assigned texts.

This introductory chapter includes three main sections. First, I consider what validity and fairness imply in test development. Second, I present the research questions that will be explored in the remaining chapters. Finally, I summarize the remaining chapters.

1.2 Validity and Validation

The *Standards for Educational and Psychology Testing* (AERA, APA & NCME, 2014; hereafter referred to as the *Standards*) define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p.11) and the process of validation as "accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations" (p.11). The definition emphasizes the importance of interpreting test scores adequately and appropriately, and the processes to gather evidence to evaluate the accuracy and appropriateness of the interpretations. To be sure, validity is a summary that integrates the evidence and consequences of score interpretation and use. Thus, validating scores implies the processes of evaluating the meaning and consequences of the assessment to justify score interpretation and use (Messick, 1995).

In the process of validation, the *Standards* (AERA, APA & NCME, 2014) suggest gathering multiple sources of evidence that may explain different aspects of validity.

They classify five sources of validity evidence: test content, response processes, internal structure, relation to other variables, and consequences of testing.

- Evidence based on test content can be acquired from analyzing the relationship between the test construct and the test content. Evidence about test content can partially be associated with the differences in score interpretation across subgroups of test takers through construct underrepresentation (i.e., failure to include important aspects of the construct) or construct-irrelevance (i.e., extraneous influences to the intended construct).
- Evidence based on response processes can be obtained from analyzing cognitive responses regarding whether test takers actually perform or respond to the items in alignment with the construct.
- Evidence based on internal structures can be gathered from investigating the extent to which test items and test components conform to the construct being measured.
- Evidence based on relations to other variables can be obtained by analyzing the relationships of test scores with external variables such as criteria measures and other tests measuring the same or similar constructs.
- Evidence for consequences of testing can be gathered and evaluated regarding the extent to which the proposed interpretations of test scores are used as intended.

A sound validity argument is not established by evidence based on only one source. It needs to include multiple dimensions of validity evidence and integrate them into a coherent relation to the interpretations of test scores intended for particular uses.

Text-based reading passages are a crucial part of reading comprehension assessments, and the way the texts are constructed can affect test takers' responses and, accordingly, test scores. Analyzing the texts in a reading comprehension assessment provides one source of validity evidence that can support the interpretations of the scores for the intended uses.

1.3 Fairness in Testing

The *Standards* (AERA, APA, & NCME, 2014), emphasizing fairness in testing as a validity issue to consider test takers and test users in all stages of testing, define a fair test as reflecting "the same construct(s) for all test takers. And scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct" (p.50). This definition emphasizes that any subgroups among test takers should not be disadvantaged based on individual characteristics such as gender, race, ethnicity, age, disability, socioeconomic status, language outside of the demands of the construct being assessed. It also proposes a fair test should take into consideration the way to address measurement bias that can cause different interpretations of test scores obtained by members of different subgroups.

Minimizing measurement bias and increasing fairness is important in testing so all test takers can have the opportunities to prove what they know and can do on a test without obstructions (AERA, APA, & NCME, 2014). In the processes of test design, test content, test context, test response, and opportunity to learn can be threats to fair and valid interpretations of test scores for the intended purposes (Sireci & Faulkner-Bond.

2014, 2015). Thus, they need to be cautiously evaluated for all test takers across all stages of testing.

English learners are an important subgroup in U.S. education and introduce challenges to developing valid and equitable measurement. Many studies find ELs perform worse than non-ELs in standardized assessments in contents such as mathematics and science. One cause of ELs' poor performances can be linguistic features of the assessment that are not intended to be measured yet introduce construct irrelevant variance (e.g., Abedi, 2010). A finding that supports the argument that construct irrelevant confounds related to EL status is that ELs performed no worse than their non-EL peers when appropriate accommodations were provided (e.g., Abedi, 2002). Different performance by EL students also can be found in reading comprehension assessments in which English proficiency is essential, and by definition, ELs have more limited English proficiency than non-EL students. However, it is not appropriate nor valid if English proficiency impacts test scores on assessments that claim to measure academic constructs other than English language proficiency (Sireci & Faulkner-Bond, 2015), although the use of language for communication and other purposes may well be part of academic constructs such as in STEM assessment. To consider whether a reading comprehension assessment is fair for all the participants, evaluation needs to take place about whether the language proficiency required to respond is appropriate in measuring a test taker's ability, in the given construct, and whether the items measure in the same way for all the participants.

1.4 Research Questions

In the current study, I explored young readers' measured ability and performance on a reading comprehension assessment, the Multiple-choice Online Causal Comprehension Assessment (MOCCA), which is intended to measure the causal inference ability of young readers. In MOCCA assessments, readers are required to select a sentence to fill in a blank in a reading passage with the goal of completing the passage coherently (Carlson, Seipel, & McMaster, 2014). As a reading comprehension measure, MOCCA is constructed in a uniform item format throughout the form: a seven-sentence passage as an item and three responses (i.e., one correct and two incorrect). This uniformity was expected to contribute to concentrating on textual factors by controlling differences in item formats during the processes of analysis.

This dissertation study explored the effects of textual and personal factors to explain how those responses were generated to the items of the MOCCA assessments, based on a measurement model of MOCCA that estimated the person and item parameters for each item on the assessment. From the findings of the effects of person properties on the differences in reading comprehension between persons on MOCCA, I further investigated whether there were items that might measure different constructs for ELs as a group compared to non-ELs on the MOCCA assessment, and reviewed the identified items with respect to the textual features used as the predictors in the explanatory models. The findings provided descriptive and explanatory information related to the use of MOCCA as a fair reading comprehension assessment valid for measuring the causal inference ability of young readers.

My research questions were:

Research Question 1.

Are item response models well-fitting for generating the parameter estimates of young readers' reading comprehension ability proficiency on the MOCCA assessment? If so, what do the results indicate regarding reading ability in the data set used in this study?

Research Question 2.

How are text features such as text complexity indicators and story features associated with explaining the item difficulty of MOCCA?

The text complexity indicators include quantitative indices such as word count, word familiarity, word concreteness, type-token ratio, and narrativity. The story features include qualitatively coded features of story structure such as whether the main character in the story is a child and the story is accordingly realistic, whether there is a second agent, whether the goal and the main character appear in the same sentence, whether the goal is explicit and presented in the first or second sentence of the story, whether the goal is met in the end, and how the ending is emotionally accepted.

Research Question 3.

How are person properties associated with young readers' performance on MOCCA?

The person properties include gender, race/ethnicity categorized as white and non-white, socioeconomic status represented as free and reduced-price meals (FARM), special education status, and English learner status.

Research Question 4-a.

Do MOCCA items exhibit Differential Item Functioning (DIF) between English learners and non-English learners?

Research Question 4-b.

What is the relation between the textual features presented in Research Question 2 and the items functioning differently between ELs and non-ELs on the MOCCA assessment?

1.5 Organization of the Dissertation

This study was designed with the overall goal of understanding how the examinees processed the items in a reading comprehension assessment. Research questions were constructed based on the design, which followed statistical and psychometric modeling to achieve the study goals.

The research questions were divided into three phases according to modeling. Phase I included Research Question 1, which was intended to explore the statistical and psychometric characteristics of the MOCCA reading comprehension assessment. Phase II involved Research Question 2 and Research Question 3, both of which applied explanatory item response modeling to explain the differences in MOCCA item difficulty induced by text features and varying responses of the examinees related to personal characteristics. Phase III covered Research Question 4-a and Research Question 4-b, which were intended to understand whether different responses as a group were detected on MOCCA items between ELs and non-ELs and what features were related to those group differences on the items. Research questions, organized by these three phases, were explored and findings are described in the following chapters.

Chapter 2 synthesizes the empirical literature in relation to understanding the reading comprehension model as the construct in the reading comprehension assessment, followed by describing the factors of text and person related to the extent of reading comprehension. Chapter 3 describes the research methodology including MOCCA as the measure, data analysis procedures, and the statistical modeling used for data analysis. Chapter 4 presents the results and findings of this study, and finally Chapter 5 concludes with a summary of my findings and the limitations of the study, followed by discussion and implications of the dissertations study, connecting it to prior research and suggesting areas for future work.

CHAPTER II

LITERATURE REVIEW

This chapter provides an overview of research related to the current study examining a causal comprehension assessment in reading. The degree of comprehension depends on both the reader and the text. During the process of comprehension, the reader as an agent and the text as an object are primary components, and the reader establishes meaning through interacting with the text. It is important to understand how comprehension is achieved through text, what text factors are related to the degree of comprehension that may lead to different item difficulty in a reading comprehension assessment, and why some test takers perform differently than others for some items when they respond to the items in the same text passages.

2.1 Reading Comprehension

The construct intended to be measured in a reading comprehension assessment is "reading comprehension." This section describes the way that construct is defined, and describes a reading comprehension model dominantly used in U.S. education to explain the way a reader constructs meanings from the text for full comprehension of the text.

2.1.1 Definition

What reading comprehension represents is defined with somewhat overlapping ideas across the field. The National Reading Panel (2000) defined reading comprehension as "an active process that engaged the reader …(by) the construction of the meaning of a written text through a reciprocal interchange of ideas between the reader and the message in a particular text" (p. 4-39). The RAND Reading Study Group was sponsored in 1999

by the Interagency Education Research Institute and charged with proposing strategic guidelines for how to improve reading comprehension outcomes for all students. In the report published by the RAND Reading Study Group (2002), reading comprehension was defined as "the process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (p. 11). This definition involves three elements in reading comprehension: the reader who does reading and comprehending as an agent, the text which is read and comprehended as a target, and the activity of reading and comprehending as a process (ibid.). The *Literacy Dictionary* (Harris & Hodges, 1995) defined reading comprehension as "the construction of the meaning of a written communication through a reciprocal, holistic interchange of ideas between the interpreter (i.e., the reader) and the message in a particular communicative context" (p. 39), in this case the written text.

To sum up the definitions from the multiple sources, reading comprehension can be understood as the processes of constructing meaning through an interaction of ideas between the reader and the message from the text. That means reading comprehension needs three elements: a reader who reads the text, the text that is read and interpreted, and the processes through which the reader constructs meanings from the text.

2.1.2 Construction-Integration model (CI model)

One of the reading comprehension models that is proposed to explain how a reader constructs meanings from texts during the processes of comprehension is the Construction-Integration (CI model) proposed by Kintsch (1998). The CI model is perceived as a dominant reading comprehension theory in current U.S. educational research (Caccamise & Snyder, 2005). The National Reading Panel (2000) and RAND

Reading Study Group (2002) applied the CI model to their theoretical and conceptual framework when they addressed the requirements and improvements of teaching, learning, and assessment of reading comprehension for early and secondary learners.

Under the CI model, reading comprehension is viewed as a cyclical process that consists of two phases: a construction phase and an integration phase (Caccamise & Snyder, 2005; Kintsch, 1998). In the construction phase, meanings are locally constructed in a rough but inaccurate mode from the text base, a reader's background knowledge, and goals. In the integration process, locally constructed meanings are integrated into the global, coherent context of the text. During the processes of comprehension, all possible meanings are activated in the construction phase, inappropriate representations unfit to the context are deactivated in the integration phase, and then a new solution is activated in involving both phases. The processes are cyclically repeated until a reader constructs an appropriate and coherent mental representation of the meanings and comprehends the given text successfully.

The CI model distinguishes multiple levels of mental representations constructed by a reader: a surface level, a global level, and the situation model (Caccamise & Snyder, 2005; Graesser, Mills, & Zwaan, 1997; McNamara, Kintsch, E., Songer, & Kintsch, W., 1996; McNamara, Ozuru, & Floyd, 2011; RAND Study Group, 2002). A local, surface level representation indicates extracting meanings from the exact wording and syntax directly expressed in the text, operating during the construction phase. A global level of representation, operating during the integration phase, consists of a set of propositions such as statements and idea units which are constructed from the meanings of the text base. Meanings constructed at the text base do not ensure a reader's full comprehension

of the text at a deeper level. When a reader establishes an appropriate and coherent situation model from integrating meanings from the text, the reader successfully completes comprehension. Thus, the situation model is considered as the most important level of representation.

2.1.3 Situation model

Under the CI model, successful reading comprehension depends on how to construct a situation model from integrating the information expressed in the text with a reader's background knowledge (Graesser et al., 1994; McNamara et al., 1996). A situation model is a referential microworld constructed from the text, along with a reader's prior knowledge, and the model includes the people, setting, states, actions, and events that are explicitly or inferentially expressed in the text (Caccamise & Snyder, 2005; Graesser et al., 1994; Kintsch, 1998; Zwaan, 1999; Zwaan & Radvansky, 1998). A reader is supposed to achieve full and successful comprehension if the representations from the local level of the text are associated with the global level to form a coherent and appropriate situation model of the mental representations (Kintsch, 2005).

A reader establishes a situation model through inferences made from the meaning representations converted from the explicit words and sentences in the text. The way a situation model is formed varies depending on text genre (Graesser et al., 1994; McNamara et al., 1996). In narrative text, the model is processed on the basis of the characters, objects, settings, actions, event, processes, plans, thoughts, and emotions of characters, and other details of the story. A reader's world knowledge and experiences about these actions, goals, and emotions will play a part in generating inferences from a narrative text to form a situation model in the text (McNamara & Kintsch, 1996). In

informational or expository text, the situation model is operated on the basis of the subject matter content which is mostly new or unfamiliar to readers. Thus, readers who read an expository text generate fewer inferences because they have insufficient background knowledge about the topics in the text. Given the characteristics of text genre, narrative text may be adequate to examine the way of inference generation and situation model formation.

How to construct a situation model also differs by a reader's idiosyncrasy. A reader activates background knowledge and integrates that knowledge to construct a situation model. The extent to which the situation model is appropriately and coherently formed from the text varies depending on the extent to which the reader activates own background knowledge to make causal inferences to integrate the information across the sentences from the text and construct the model (Dowell, Graesser, & Cai, 2016; Graesser, McNamara, & Kulikowich, 2011; Graesser et al., 1994; Kintsch, 1998; McNamara & Kintsch, 1996; McNamara et al., 1996).

2.2 Text Features: Factors That Make a Text Difficult

Reading comprehension begins with the text. The process of comprehension starts when a reader interacts with the text and extracts meanings from the local level of the text (Graesser et al., 1994; RAND Study Group, 2002). A test taker in a reading comprehension assessment is asked to read the text passages and answer the questions related to reading (Alderson, 2005). A test taker as a reader will have a different degree of comprehension depending on the text difficulty and accordingly respond to the items in the assessment differently. Thus, the text factors in association with text difficulty will

have a significant effect on the extent to which a test taking reader comprehends the text and performs on the assessment.

Many researchers have examined the text factors associated with the different experiences that readers have during the processes of reading and comprehending an assigned text. Alderson (2005) enumerated the text variables in relation to text difficulty as text topic and content considering background knowledge, text types and genre including expository and narrative texts, text organization of paragraphs, and linguistic variables such as syntactic complexity and vocabulary difficulty. The RAND Reading Study Group (2002), based on the situation model, enumerated sources of text difficulty as the surface code (vocabulary and syntax), propositional text base, mental model, pragmatic communication, and discourse structure and genre. The commonly proposed textual factors in relation to text difficulty are explored here: language features (i.e., word difficulty and syntactic complexity), cohesion that initiates a reader to construct situational meaning from the text, and text genre.

2.2.1 Word difficulty

Word difficulty is strongly related to text decoding and meaning construction. If a text has difficult words, a reader is likely to struggle with comprehending the text due to lack of word knowledge or difficulty of meaning construction. The characteristics of difficult words are examined with respect to "syllable length, frequency of word usage, and type of word (i.e., content vs. function word)" (Davey, 1988, p.67). That is, a word is more difficult if it has longer syllables, is used less frequently, and/or is more content-loaded. Among the characteristics, word length and word frequency are regarded as more significant indicators in predicting word difficulty, and many text difficulty metrics such

as Degrees of Reading Power: DRP Analyzer (Questar Assessment, Inc.), Lexile (MetaMetrics), and Coh-Metrix (University of Memphis) use both word length and word frequency as their major indicators in quantitatively measuring text complexity (Nelson et al., 2012).

Word length refers to the number of syllables per word, and words having more syllables, i.e., long words, are considered to be more difficult than short words. When a word has more syllables, readers take more time to read and decode its meaning (Just, Carpenter, & Woolley, 1982). If it takes longer to identify words to get meanings, it will take longer to comprehend the sentence or text containing such words.

Word frequency has a relationship with word familiarity. Words that occur more frequently in a corpus of English text are considered to be easier because they are more familiar to the reader and thus are processed more quickly by the reader (Leroy & Kauchak, 2014). Low-frequency words in a sentence tend to make it slower or more difficulty for a reader to extract meanings from the sentence and comprehend the sentence. Sometimes one or a few unfamiliar words will make some readers struggle with comprehending the entire sentence or text (Cain, Oakhill, & Bryant, 2004; Cain, Oakhill, & Lemmon, 2004; Graesser et al., 2011; Perfetti & Stafura, 2014).

In addition to word length and word frequency, word concreteness contained in word meaning is an important feature in assessing word difficulty (Feng, Cai, Crossley, & McNamara, 2011). Word concreteness indicates the extent to which a word is concrete or non-abstract. Words referring to material things that we can touch, hear or see are more easily and quickly perceived and understood than words referring to abstract ideas (Feng et al., 2011; Gilhooly & Logie, 1980; Graesser et al., 2011).

Word difficulty can also be estimated in terms of word syntactic features (e.g., part of speech, tense, nominalization) (Graesser et al., 2011; RAND Reading Study Group, 2002). Parts of speech assigned to each word (i.e., syntactic category of word such as noun, verb, or adjective) are assessed in combination with word frequency to measure word difficulty (Nelson et al., 2012). The syntactic categories are divided into content words (nouns, main verbs, adjectives, adverbs) and function words (e.g., prepositions, determiners, pronouns) (Graesser et al., 2011).

The characteristics of word difficulty such as word length, word familiarity, and word concreteness may be associated with text difficulty and may impact the degree of text comprehension because comprehension begins with constructing meanings locally from the exact wording in the text (Graesser et al., 1997; McNamara et al., 2011; Perfetti & Stafura, 2014; RAND Reading Study Group, 2002).

2.2.2 Syntactic complexity

Another factor having a strong influence on comprehending a text is syntactic complexity of the sentences in the text. Syntactically complex sentences contribute to the load of working memory during a process for comprehension (Alderson, 2005; RAND Reading Study Group, 2002). One indicator estimated for syntactic complexity is sentence length, that is, the number of words per sentence (Davey, 1988). Freedle and Kostin (1991, 1993) and Freedle, Kostin, and ETS (1992) reported that text sentence length (i.e., the average words per sentence), text paragraph length (i.e., number of words in the longest paragraph and number of words in the paragraph containing relevant inference information), and passage length have a significant effect on item difficulty in reading comprehension items. They found that reading comprehension items on the SAT,

Graduate Record Examination (GRE), and Test of English as a Foreign Language (TOEFL) requiring readers to find main ideas or make inferences from the given text passages were more difficult for test takers when the sentence, the paragraph, or the text was longer.

In addition to length, syntactically complex sentences make readers slow or difficult to comprehend. Syntactically complex sentences include those that are embedded, dense, ambiguous, or ungrammatical. According to the RAND Reading Study Group (2002), sentences are perceived as difficult by readers when the sentences have left-embedded syntax before the main verb (e.g., the subject is modified by relative clauses before the main verb), when sentences have a syntactically dense clause (i.e., a high ratio of explicit statements or higher-level syntactic constituents per word), when the head noun is modified by too many adjectives and adverbs (i.e., dense noun-phrase), and when syntactic structures are ambiguous with two or more syntactic structures assigned to a sentence (e.g., The sentence, "The teacher said on Monday she would give a quiz," means either the teacher told the class on Monday about the quiz or the quiz would be taken on Monday).

In contrast, easier sentence syntax refers to shorter sentences, fewer words per noun phrase, fewer words before the main verb of the main clause, and fewer logic-based words. Readers tend to better and more quickly comprehend text with such simple syntax (Graesser et al., 2011). Similar to word difficulty, text difficulty manifested through text length and syntactic complexity can affect the extent to which a reader comprehends a text.

2.2.3 Coherence and cohesion

According to the Construction-Integration model (CI model), multiple levels of mental representations are involved in text comprehension: a local surface code representation from the exact words and syntax of text, a global representation of explicit propositions from the text base, and building up a situation model by integrating the local and global representations (Caccamise & Snyder, 2005; Graesser et al., 1997; McNamara et al., 2011). Readers complete meaningful comprehension when they form a coherently meaningful situation model through inferences generated from the cohesive cues represented at both the local and global levels of text (Graesser et al., 1994).

Cohesion differs from coherence. Cohesion is an explicit characteristic of the text, whereas coherence refers to a characteristic of the reader's mental representation of the text content (Graesser et al., 1994). Text cohesion represents the extent to which elements within the text provide explicit cues to help readers relate information within and across sentences in the text. Cohesive elements in a text are based on explicit linguistic elements such as words and their combinations. When these linguistic elements within the text are explicitly cued in relation to the information across the text, the text is considered highly cohesive. When many inferences are required to construct a coherent representation (such as if explicit elements are insufficiently provided in the text), the text is considered low-cohesive (McNamara, Louwerse, McCarthy, & Graesser, 2010; McNamara et al., 2011).

Cohesion is an objective property of the explicit cues in the text that help a reader to interpret ideas meaningfully in a local level and to connect them with other or higher-level ideas in the global units such as topics (Crossley, Kyle, & McNamara, 2016). These explicit cues of cohesion function as a device to address the relations between sentences

or clauses that form the story and as a tie of the events and propositions by indicating whether more than one event are causally related (Cain, 2003; Graesser, McNamara, & Louwerse, 2003; Graesser, McNamara, Louwerse, & Cai, 2004, p.193). The cohesion devices include connectives (e.g., before, after, because), givenness (i.e., ratio of pronouns to nouns), type-token ratio (i.e., word repetition across a text), lexical overlap (i.e., overlap between nouns, arguments, stems, content and function words), and synonymy overlap (overlap of synonyms across sentences and paragraphs) (Crossley et al., 2016). Readers can construct causally coherent relations between structural elements and propositions depending on the extent of the skills and knowledge that they bring to understanding the explicit cues suggested by cohesion devices. If a text provides more cohesive devices via the explicit cues, readers can construct more coherent inferences by means of the devices to build up a situation model and comprehend the text better (e.g., Zwaan & Radvansky, 1998).

Cain and Oakhill (2006) reported that some poor comprehenders did not exhibit differences in extracting meanings from words and syntactic structures from the text in comparison with good comprehenders, but they showed some lack of comprehension subskills such as inference and integration skills. Thus, poor comprehenders can comprehend a text better if they read a text that contains more explicit cohesion devices or if they can improve the subskills to locate and integrate the cohesion devices to generate coherent representations.

In summary, establishing a coherent situation model depends on the extent to make causal inferences from the cohesive clues from the text, which leads to successful comprehension.

2.2.4 Text genre

Text genre is also considered to influence text difficulty (Alderson, 2005; RAND Reading Study Group, 2002) and is posited to affect the extent to which a reader comprehends the overall meaning of a text at a deep level (McNamara et al., 2011). The *Literacy Dictionary* (Harris & Hodges, 1995) classifies four major text genres— exposition, narrative, description and argumentation— depending on the way ideas are arranged in order to deliver an effective message. The RAND Reading Study Group (2002) also categorizes text into four genres (three of which parallel the classifications of Harris and Hodges (1995) and one of which (persuasion) differs): exposition, narration, description and persuasion.

The two most frequently used text genres in education may be exposition and narrative. Exposition or expository text is intended to propose or explain knowledge or ideas, while narrative refers to a story, actual or fictional (Harris & Hodges, 1995). Expository text is often used to introduce new knowledge or provide technical information that readers do not typically have the background knowledge about. Narrative texts often address topics focused on everyday experiences such as friendship and love in a story that involves characters, goals, settings, events, and times. Readers may generate more inferences during the comprehension of narrative text than expository text because they have background knowledge to apply to the situations addressed in the narrative text (Graesser et al., 1994; Graesser, McNamara, & Louwerse, 2003; McNamara, et al., 2011).

Most children can easily access narrative texts for comprehension because they have experiences and knowledge similar to the events, settings, and actions described in

the narrative text, and because they are familiar with the simple structure of the narrative story (i.e., a sequence of casually or temporally related events) (McNamara et al., 2011; Nelson, 1996; Olson, 1985; Williams et al., 2005). On the other hand, young readers tend to struggle more with comprehending expository texts, particularly those involving abstract and logical relations. Such expository texts have heavier processing demands to comprehend the texts because the syntactic structures are more complex and the information is more domain-specific (Graesser et al., 1994; McNamara et al., 2011). Thus, texts from the narrative genre may be most suitable for assessments designed to assess young readers' ability to comprehend text from inference generation and situation-model constructions.

2.3 Person Properties: Factors Depending on Learners

In reading comprehension assessment, the first task for a test taker is to read the test passages, then answer questions related to what has been read. To comprehend, the test taker needs to have a range of capacities such as working memory, language knowledge such as word and syntactic knowledge, background knowledge, and motivation (RAND Reading Study Group, 2002).

The individual characteristics as well as abilities that a test taker possesses as an active agent of comprehension will determine the extent of comprehension and subsequent responses on a reading comprehension assessment. Collectively, different responses between groups depending on a certain characteristic may be due to cognitive development such as age or to intentional or unintentional factors such as topic familiarity favoring of one group over the other. Scores of the assessment having such

bias as topic cannot be justly interpreted because they are influenced by construct-irrelevant factors (AERA, APA, & NCME, 2014).

For example, age can make a difference in scores when children of different ages take the same reading comprehension assessment. Because older children tend to have better vocabulary knowledge and verbal skills, more developed working memory, and more background knowledge, they are likely to achieve higher scores than younger children on the same assessment (Cain, Oakhill, & Bryant, 2004). Gender is also considered as a factor that may be related to the degree of reading comprehension ability, but the effects are different according to different research studies. The different results may derive not only from the specific sample but from what is being studied. For example, Logan and Johnston (2009) found that girls had better reading comprehension ability and more reading fluency compared to boys in their study. But when some researchers examined why boys achieved more than girls on college admission tests in the U.S. and U.K., they found that the test tasks and text topics favored one gender over the other (Brantmeier, 2003; Hyde & Linn, 1988; Mau & Lynn, 2001). The results indicated the boys were likely to have more cohesive clues to make causal inferences than the girls because the topics and tasks in the assessments were more familiar to boys.

Among the individual characteristics, English proficiency may have a substantial effect on the process of a reader's text comprehension and on the assessment outcomes. Second language learners or English learners who learn to read in English at a later age do tend to develop language proficiency more slowly than non-English learners (Tabors & Snow, 2001). In addition to different language proficiency, they tend to have diverse background knowledge including culture and different accessibility to formal education.

Their unique characteristics above and beyond language proficiency may be associated with the outcomes in a reading comprehension assessment.

2.3.1 English learners in the U.S.

The current trend in the U.S. education involves increasing support for English learners at many schools where English is a main language for teaching and learning. The U.S. Department of Education (2016) defines English learners as individuals "whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual the ability to meet the challenging State academic standards and to successfully achieve in classrooms where the language of instruction is English" (p. 43). According to this definition, English learners can have difficulty with understanding teachers' instructions presented in English because they have English language limitations. Moreover, many of them may fail to meet academic standards due to poor performance on the standardized tests that intend to measure their academic abilities in subject matter areas, when the assessments are constructed in English and without scaffolds or supports to adjust the unintended consequences of the language difficulty.

The enrollment of English learners has been gradually increasing at all K-12 school levels across the nation. The *Condition of Education 2018* (McFarland et al., 2018) reported that the percentage of public school students who were classified as ELs in the U.S. increased to 9.5%, or 4.8 million students in fall 2015 as compared to fall 2000 (8.1%, or 3.8 million students). More ELs were enrolled in lower grades than in upper grades in fall 2015. Among all the students in fall 2015, 16% or more of ELs were enrolled in kindergarten, first and second grades at public schools. In contrast, the

percentage of EL students in upper grades reduced with only 3.9 % of students classified as ELs at twelfth grader. This reduction in the EL population in the upper grades is likely the result of students identified at earlier ages being reclassified and no longer identified as ELs in later grades.

The most frequently spoken home language for ELs in the U.S. was Spanish in fall 2015 (McFarland et al., 2018). At that time, Hispanic EL students consisted of about 78% of EL student enrollment overall, followed by Asian students as the next largest racial/ethnic group among ELs. In summary, the number and proportion of English learners are consistently increasing at U.S. public schools. Their enrollment is concentrated mostly in the lower grades, and they speak a variety of languages at home, although the largest proportion of EL students in the U.S. are Spanish speakers. EL students represent diverse racial/ethnic groups. They are not a singular group with common backgrounds, language, or life experiences.

2.3.2 Academic achievement of English learners

One of the most challenging issues that English learners and educators in U.S. schools are faced with may be related to English learners' academic achievement. EL students often have lower achievement on standardized assessments, and lower skills for performance in the classroom, both of which effect their preparation for higher education (Sireci & Faulkner-Bond. 2015). In the U.S., the National Assessment of Educational Progress, or NAEP, provides assessment results for subject areas such as mathematics, reading, writing, and science. At the national level in 2013 (Kena et al., 2015), the percentage of ELs in the fourth grade who were able to achieve at or above *basic* level was 59% in mathematics and 31 % in reading compared to their non-EL counterparts'

achievement of 85% and 72% respectively. The achievement gap between the two groups was larger in the eighth grade, 31% versus 76% in mathematics and 30% versus 80% in reading. Compared to non-ELs, ELs showed lower achievement with large gaps in mathematics and reading in both grades.

The larger gap in reading than mathematics between the two groups especially at the early grades can be explained by language proficiency which is a part of the abilities assessed in reading (Abedi, Bailey, et al., 2005). That the achievement gap between ELs and non-ELs was wider in upper than lower grades as well as in reading than mathematics in 2013 NEAP might be because the content subjects are increasingly intensified over time, and because a nature of reading assessments requires more English language proficiency than mathematics.

Differences in assessment performance between EL and non-EL students are not surprising. When ELs take the same assessments as native English learners, their academic abilities are measured and compared on the same scale. In addition to the underlying potential confound of English language proficiency, different cultural and educational backgrounds may also impact the comprehension processes used by ELs and their responses to some items compared to their English speaking counterparts (e.g., Abedi, 2002; Abedi, Leon, & Mirocha, 2005; Fitzgerald, 1995).

According to the *Standards* (AERA, APA, & NCME, 2014), construct-irrelevance refers to "the degree to which test scores are affected by processes that are extraneous to the test's intended purpose" (p.12). Test takers cannot successfully perform on the test if the test has construct-irrelevant variance. And the decisions made as a result of test performance will not be fair to test takers if they are affected by irrelevant

variables such as English proficiency and cultural background that are not intended to measure on a test. Thus, construct-irrelevance is regarded as "a prime threat to fair and valid interpretations of test scores" in the *Standards* (AERA, APA, & NCME, 2014, p.54) and in the *Guidelines for the Assessment of English Language Learners* (Pitoniak et al., 2009).

Also, a reading comprehension test may include construct-irrelevant sources as measurement error if the test contains materials that are too easy or too difficult for the level intended to be measured, if the test content prompts extreme emotional reactions, or if the reading passages are based on subject matter unfamiliar to a subgroup of the population (AERA, APA, & NCME, 2014). These factors may be at play as well in assessment for English learners.

2.3.3 ELs and textual effects on assessment

Research has been conducted to identify the factors that influence the often weaker test performance of English learners on standardized assessments. The *Guidelines for the Assessment of English Language Learners* (Pitoniak et al., 2009) listed linguistic, cultural, and educational background as potential confounding factors in relation to English learners' achievement on standardized assessments. Language factors imply that ELs have different language backgrounds and diverse proficiency and literacy levels in their native language as well as in English. Educational background factors affect ELs' performance on the assessment because ELs may have different degrees of knowledge of the language and the content area measured by the assessment depending on the extent of their participation in schooling and exposure to standardized testing. ELs' diverse cultural

backgrounds can also disadvantage their test performance because they are not familiar with the mainstream American culture.

Language factors. Language factors are regarded as the main powerful construct-irrelevant measurement error that interact with ELs' weaker performance in standardized testing compared to their counterparts. Abedi and his colleagues (Abedi, 2002, 2004, 2010; Abedi & Gándara, 2006; Abedi & Lord, 2001; Abedi, Hofstetter, & Lord, 2004; Abedi, Leon, & Mirocha, 2005; Abedi, Lord, & Plummer, 1997; Shaftel et al., 2006) emphasize linguistic features as a major factor in the performance gap between EL and non-EL students in content area assessments and as a source of measurement error that may threaten the validity of such tests.

Abedi (2010) and Abedi et al. (1997) found the ELs' poor performance on the eighth-grade NAEP mathematics test were related to some linguistic features such as unfamiliar or infrequent vocabulary, passive voice constructions, long noun phrases with prenominal modifiers, compound sentences such as conditional clauses (e.g., *if*-clauses), relative clauses, long question phrases, abstract or impersonal presentations, and negation. Although NAEP now includes text-to-speech, thus resolving the direct reading issue, vocabulary and comprehension challenges for spoken English can remain.

In the past, the effects of these linguistic features on ELs' test performance in mathematics tests were different depending on the tests or grades. Among the linguistic features listed, item length had a negative correlation with the achievement of not only ELs but non-ELs (Martiniello, 2009). The total number of words in the item was a significant predictor of the total sample's achievement on the results of Kansas general math assessments (KGMA) when there were no interaction effects between linguistic

features including item length and group membership (i.e., general students, ELs, and students with disabilities) (Shaftel et al., 2006). The effects of mixed linguistic complexity features were consistently significant for ELs' achievement on math tests (Abedi et al., 1997; Martiniello, 2009). These findings suggest that language factors including syntactic and lexical complexity holistically affect the extent to which learners with limited English proficiency comprehend a text in the standardized assessment. More cognitively complex language factors may or may not be a focus of the construct, so should be considered as to their relevance in the measures.

Background knowledge. Cultural factors also affect ELs' achievement in testing because they may be associated with topics unfamiliar to ELs and may complicate ELs' understanding of items. Thus, the impact of culture on test performance is examined as a source of construct-irrelevant variance for ELs (Carrell, 1981, 1987; Johnson, 1981, 1982). Johnson (1981, 1982) investigated the effects of cultural background knowledge on EL university students' reading comprehension and found the EL participants recalled better the English text in relation to their culture regardless of the syntactic and semantic complexity. The findings imply that English learners may have difficulty with comprehension when they are unfamiliar with the text content because of lack of cultural background knowledge.

Some researchers report different results investigating the effects of background knowledge on ELs' text comprehension. Carrell (1983) described background knowledge as "general knowledge of the world and the extent to which that knowledge is activated during the mental process of reading" (p. 183) and specified three components of background: context (i.e., whether a title or picture page precedes the text passage),

transparency (i.e., whether specific, concrete items within the text provide textual cues to the content area of the text), and familiarity (i.e., whether the reader has prior knowledge or experience of the content of the text). Carrell (1983) did not find significant effects of the components of background knowledge on EL readers' recalling the given passages.

The finding was consistent with that of Barnitz and Speaker's (1991) study in which EL readers were more likely to rely on linguistic knowledge to comprehend a text rather than background knowledge. When a reader read a text written in a native language, the reader was more accessible to background knowledge of the content area of the text and had a better chance to comprehend the text (Bransford & Johnson, 1972, 1973). However, a second language learner might not fully make use of the background knowledge in the text because the first task for them was to understand the meaning of words and the structures used in the text (Barnitz & Speaker, 1991; Carrell, 1983).

Some devices were implemented to help second language learners to have more background knowledge from the text, controlling the linguistic complexity of text. For example, analogies were added for EL adults in a science test, but the device did not help them to have more content knowledge because the analogies gave the EL test takers extra linguistic burdens (Brantmeier, 2005). This finding also implied that it was the language factors that impeded ELs' comprehension rather than background or content knowledge.

English learners and reading comprehension assessment. Some researchers (Barnitz & Speaker, 1991; Brantmeier, 2005; Carrell, 1983) countered the assumption that background knowledge would be strongly related to the extent of ELs' reading comprehension. Instead, they identified linguistic complexity as a major source of item difficulty for English learners. This implies ELs need to acquire a certain level of

knowledge of vocabulary and syntactic structures in a second language, here English, and then can demonstrate their genuine ability on the assessment. However, others may treat language complexity as construct-irrelevant variance in ELs' performance on the content assessment, even in language assessments, depending on the intended inferences for the assessment.

Some efforts have been made to reduce ELs' language burden in content assessments, with some success. For example, when the items were linguistically modified by simplifying unnecessarily difficult vocabulary and complex syntactic structures on a math test, the ELs did better than those who took non-modified tests, or the achievement gap between ELs and non-ELs was reduced (Abedi et al., 1997; Abedi & Gándara, 2006; Abedi et al., 2004; Abedi & Lord, 2001; Martiniello, 2009). In another study, Shohamy (1984) found the achievement of EL learners was improved when the questions were presented in their first language and suggested multilingual tests for second language learners.

Language proficiency including linguistic knowledge and skills is not always construct-irrelevant for English learners. A reading comprehension test intends to measure the ability to read and comprehend a text written in a particular language, in the case of the United States, English. Linguistic knowledge of the dominant language will be a part of abilities that test takers are required to have to comprehend the passage.

Individual learners' experiences in language, learning, and the world may be greatly diverse, and the effects of background knowledge, linguistic features, and learning experiences are varied. The findings in relation to English learners' performance on assessment are mostly obtained from adolescent or adult participants (e.g., Anderson,

1991; Brantmeirer, 2003, 2005; Johnson, 1981, 1982). Looking at the demographic characteristics of English learners in the U.S., about 86% of ELs were enrolled in Kindergarten and elementary schools in fall 2015 (McFarland et al., 2018). Thus, it will be worthwhile to explore young ELs' performance on the reading comprehension assessment that also assesses non-ELs at the same scale.

Whether English language proficiency is a source of construct-irrelevant variance for young English learners, or a portion of the construct, i.e., reading comprehension ability, to be acquired by native English speaking, poor reading comprehenders as well as English learners needs to be examined further. Likewise, whether ELs exhibit varying item response patterns to specific features in the texts used in reading comprehension assessment compared to their non-EL peers also needs to be investigated, so I contribute to this work with the five research questions listed in Research Question section. I next explore methods for investigating these questions.

CHAPTER III

METHODS

This chapter describes the measure, the textual features coded, participants and data analytic methods that were used in this dissertation study to examine the effects of item and person properties on different item difficulty and responses of a reading comprehension assessment.

3.1 Multiple-choice Online Causal Comprehension Assessment (MOCCA)

The Multiple-choice Online Causal Comprehension Assessment (MOCCA) measure used in this study was developed to measure reading comprehension ability of young readers who are enrolled in grades 3, 4, and 5. The measure intends to identify the students who struggle with comprehending text and diagnose the nature of their poor reading comprehension (Carlson, Seipel, & McMaster, 2014; Davison, Biancarosa, Seipel et al., 2018). It is an untimed, online multiple-choice assessment that consists of short narrative reading passages and three choices per item (one correct and two incorrect responses). An examinee is required to select a causal inference choice as correct to fill in the missing sentence and complete the story coherently. (Carlson, Seipel et al., 2014; Davison, Biancarosa, Seipel et al., 2018).

3.1.1 Theoretical framework and construct

As described earlier in the literature review chapter, successful reading comprehension involves establishing a coherent mental representation through integrating information from the text and generating a meaningful situation (e.g., Graesser et al., 1994; Kintsch, E., 2005; Kintsch, W., 1998). A situation model is associated with the

way of establishing the mental representations during the comprehension processes. The model illustrates that a reader comprehends the text successfully when the reader constructs all the possible meanings from the explicit words and syntax in the text, makes causal inferences to integrate the meanings, and finally makes an appropriate and coherent situation. If a reader fails to form the coherent and appropriate situations beyond understanding the text at the local and surface level, the reader cannot successfully comprehend the text. Coherent situations required for successful comprehension are constructed by making causal inferences that bridge text elements and related information (e.g., Cain, 2003; Cain & Oakhill, 2006; McMaster et al., 2012; Rapp, Broek, McMaster, Kendeou, & Espin, 2007).

MOCCA was developed to assess the ability of examinees' making causal inferences for successful text comprehension and identify the different types of comprehension processes among the examinees (Carlson, Seipel et al., 2014; Carlson, Van den Broek et al., 2014; Davison, Biancarosa, Carlson et al., 2018; Davison, Biancarosa, Seipel et al., 2018). When they fail to make causal inferences, poor comprehenders tend to resort to specific cognitive processes such as paraphrase or elaboration (Carlson, Seipel et al., 2014; Davison, Biancarosa, Seipel et al., 2018).

Compared to good reading comprehenders, poor reading comprehenders appear to struggle with generating a coherent situation model from the text because they are likely to fail to make causal inferences from the explicit text features even though they sometimes have proper language knowledge to understand the text at the local and surface level (Cain & Oakhill, 2006; Graesser et al., 1994). They tend to employ their own cognitive processes to develop situations during the comprehension processes (e.g.,

Carlson, Van den Broek et al., 2014; Rapp et al., 2007). MOCCA measures the specific processes that poor comprehenders preferably adopt rather than the process of making appropriate inferences and identifies two types of poor comprehenders depending on their preferences:

- *Paraphrasers* who tend to construct meanings from a text by means of more paraphrasing during the comprehension processes rather than generating causal inferences compared to average or good comprehenders (Cain & Oakhill, 2006; Carlson, Seipel et al., 2014; Davison et al., 2018; McMaster et al., 2012).
- *Elaborators* who tend to develop more elaboration in comprehending text by employing inaccurate or invalid background knowledge that they possess compared to average or good comprehenders (McMaster et al., 2012; Williams, 1993).

Following the findings of types of poor comprehenders, the developers of MOCCA originally designed four alternate responses in their pilot study: one correct response (i.e., causal inference), two incorrect responses that inform the type of a poor comprehender (i.e., paraphrase and elaboration), and a third incorrect response to reduce the guessing effect (i.e., lateral connection) (Carlson, Seipel et al., 2014). Subsequently, MOCCA developers excluded the fourth incorrect response because the response was regarded as uninformative and thus redundant. The version of MOCCA used in this dissertation research includes items that consist of a seven-sentence narrative text and three response types that represent each specific cognitive process that an examinee as a reader resorts to during text comprehension in taking the MOCCA assessment, including

causal inference as a correct response, and paraphrase and elaboration as two possible incorrect responses (Davison, Biancarosa, Seipel et al., 2018).

MOCCA identifies the examinee as a good or poor comprehender from the results of the assessment, and diagnoses the examinee according to the response types that the examinee selects dominantly during the comprehension processes (Carlson, Seipel et al., 2014; Davison, Biancarosa, Seipel et al., 2018). An examinee who fails to choose causal inference as a correct response and chooses more paraphrase responses than elaboration responses will be diagnosed as a paraphraser. Likewise, an examinee who fails to choose causal inference as a correct response and chooses more elaboration responses than paraphrase responses will be diagnosed as an elaborator.

3.1.2 Item format

MOCCA is administered to young readers enrolled in grades 3, 4, and 5, and has three forms per grade, with a pool of 9 forms in total across grades. One form is randomly assigned to an examinee out of three forms pertaining to the grade. Each form uniformly consists of 40 items including 10 common linked items across forms and across grades and 30 items specific only to the form. Each item consists of a unique narrative text and three responses. The narrative text has a main goal that a main character intends to achieve and consequently causes subgoals, actions, and events to happen (Biancarosa et al., 2018). The text consists of a title and seven sentences among which the sixth sentence is missing. An examinee needs to select the best response out of three given responses to fill in the missing sixth sentence for coherently connecting the events to complete the text. The seven sentences of a text are constructed as follows (the sentences in the parentheses are from the practice item in Figure 1):

- The first sentence presents the main character ("Jimmy's best friend had a new bike.");
- The second sentence presents the main goal ("Jimmy wanted a new bike too.");
- The third, fourth, and fifth sentences present the subgoal(s) and develop the story in details ("He got a job. Jimmy earned a lot of money. He went to the store.");
- The sixth sentence is deleted as an item for an examinee to select a causal inference response out of three responses discussed above (here, the correct response is "Jimmy bought a bike."); and
- The seventh sentence ends the story ("Jimmy was happy.")

(from Carlson, Seipel et al., 2014).

Figure 1 presents a sample item from MOCCA website that students access to take the assessment online (MOCCA, https://mocca.uoregon.edu/#/). On the screen, the item starts with a title (in the example, "Jimmy and the New Bike"), followed by five sentences, a missing sentence in the sixth, and the final sentence. And below the text, three responses are presented with a prompt: "Select the best sentence to complete the story." An examinee needs to make causal inferences to select the best response to fill in the sixth blank sentence for successful comprehension. Among the responses in this example, "Jimmy bought a bike." is the correct response because the sentence coherently connects the previous events and the last sentence and explains why Jimmy feels "happy" in the ending. Two wrong responses include one paraphrase distractor repeating the sentence (in this example, "Jimmy wanted a new bike"), and an elaboration distractor

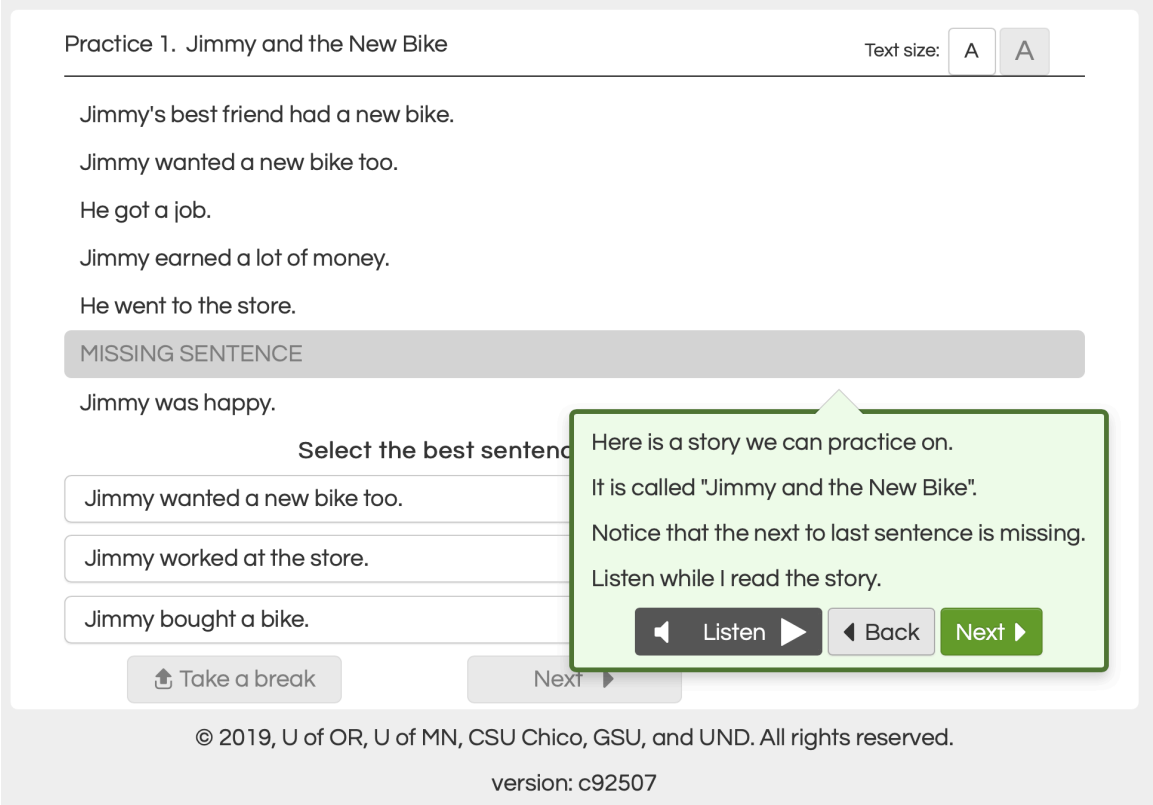

Figure 1. *Screenshot of MOCCA Item in Directions*.
Note. Screenshot of Practice item 1 "Jimmy and the New Bike" from https://mocca.uoregon.edu/.

connecting with invalid or inappropriate background knowledge (in this example, "Jimmy looked at the candy"). An examinee clicks one sentence as a response and then goes on to the next item by clicking the "Next" button. Examinees cannot skip an item without clicking any response.

3.1.3 Administration

MOCCA is a computer-administered online assessment. Students can take the test using computers or tablets while a teacher or test administrator supervises their test taking. They need to access the MOCCA website (https://mocca.uoregon.edu/) over the school network with session code and access code to start the test. Once they have access,

they can read, listen to, or skip the instructions by clicking a preference before taking the test (see Figure 1).

MOCCA includes three forms per grade, and one form is randomly assigned to a student. Each form has 40 items and is developed to have similar psychometric characteristics across forms within the same grade. The measure is administered without time limit, but the amount of time can be adjusted depending on the proctor.

3.1.4 Scoring and interpretations

Each item is scored 1 if a test taker selects a correct response, i.e., a causal inference and scored 0 if a test taker selects either of the other responses. The maximum raw score is 40 points, and the minimum raw score is 0 points. MOCCA score reports provide various types of information including Scaled Score, Risk Status, Error Propensity, and Comprehension Efficiency (Davison, Biancarosa, Seipel et al., 2018). Scaled Score, converted score of raw score (i.e., the number of correct items), is reported in a scale of 50 to 950, with higher numbers indicating better performance as more correct choices are selected. Risk Status is reported with respect to *At risk*, *Some risk*, and *Minimal risk*, indicating the degree of at-risk status that predicts end-of-year achievement of set goals for English Language Arts. Error Propensity is reported with respect to *Elaboration*, *Paraphrase*, *Indeterminate*, and *Not Applicable* based on the dominant response types of the examinee. Comprehension Efficiency is reported with respect to *Fast and accurate*, *Fast and inaccurate*, *Moderate and accurate*, *Moderate and inaccurate*, *Slow and accurate*, and *Slow and inaccurate*, based on the average minutes per correct answer.

3.1.5 Psychometric features

Scaling and equating. MOCCA includes ten common items in all forms across grades. The original MOCCA linked the common items across forms, across grades and within grades through the concurrent and fixed item parameter equating (FIPE) to scale scores (Davison, Biancarosa, Seipel et al., 2018). For the current study, I anchored 10 common items across three forms within grade 3. Thus, the results of item parameters were not identical with those of the original MOCCA study.

Differential item functioning. In the process of MOCCA development, Mantel-Haenszel differential item function (DIF) was conducted to examine whether there were items working against according to two group status, gender and ethnicity (Hispanic vs. White). Eight items displayed significant DIF—four for gender and four for ethnicity—and were dropped and replaced. But they did not conduct DIF for English learners and non-English learners.

3.1.6 MOCCA data set in the study

This study focused on the text factors including text difficulty and story structures that might be associated with the item difficulty estimated from the correct responses that an examinee selected. For this reason, the type of a poor comprehender such as a paraphraser or an elaborator was not considered in this dissertation study, and accordingly the items were calibrated dichotomously in this study. If a causal inference choice was selected, it was scored 1 as correct, and the other choices were scored 0 as incorrect, with the maximum raw score of 40 and the minimum score of 0.

The current study used the data set for grade 3, for which three forms were administered. Thus, linking 10 common items and differential item functioning analysis

were conducted with three forms within the grade, and the results might not be consistent with those of MOCCA analysis by the developers.

3.2 Text Features as Item Property Predictors

An item in MOCCA consists of one seven-sentence text passage with three choices. To explore the effects of text difficulty as item properties on selecting a correct choice, this study selected text features as item predictors and coded for analyses. The text features selected were calculated or coded in two ways: quantitative text complexity and qualitative structure of the story. The complexity was calculated by means of a web-based text difficulty tool, and the story structure was coded by human raters in relation to a main character, the goal, and ending. This study used the data set of the third-grade participants, and three forms administered for the grade were analyzed in terms of text features. Each form had 10 common items across forms and 30 specific items, resulting in a number of total items or texts analyzed of 100.

3.2.1 Text complexity indicators

Text complexity was quantitatively measured via a text difficulty analysis tool, Coh-Metrix 3.0 (University of Memphis), which is a computational web-based tool developed by a team of researchers at the University of Memphis (Graesser et al., 2004). The tool analyzes text in terms of 106 indices ranging from the measures of linguistic features and readability to cohesion characteristics of text at a range of word, sentence, paragraph, and discourse dimensions (Graesser et al., 2004; McNamara et al., 2010). Among the 106 indices, this study selected five indicators that might be related to the extent of young readers' comprehending the text: word count, word familiarity, word

concreteness, type-token ratio, and narrativity. Brief descriptions of each indicator are as follow:

- Word count indicates the total number of words in the text. It is related to syntactic complexity of the text because more words make a text longer, and a longer text makes comprehension more difficult (Freedle & Kostin, 1993).
- Word familiarity indicates the extent to which the words are familiar to a reader. A reader will process faster a text with more familiar words and comprehend it with more easiness. Coh-Metrix rates the word familiarity ranging from 100 (least familiar) to 700 (most familiar), by which the higher number indicates more familiarity. More familiar and common words will enable a reader to process easier and faster comprehension (Cain, Oakhill, & Bryant, 2004; Leroy & Kauchak, 2013; Perfetti & Stafura, 2014)
- Word concreteness indicates the extent to which a word is concrete or abstract to a reader. A reader will process faster a text with more concrete or non-abstract words that are used to describe the objects you can feel by senses, and comprehend it with more easiness. Like word familiarity, Coh-Metrix rates the word concreteness ranging from 100 (least concrete) to 700 (most concrete), with the higher number indicating more concreteness or non-abstractness. More concrete words help a reader process comprehension easier and faster (Feng et al., 2011; Gilhooly & Logie, 1980; Graesser et al., 2011).
- Type-token ratio (TTR) indicates the number of unique words (called types) divided by the number of tokens of the words (Templin, 1975). For example, if the word “cat” appears 5 times in the text, its type value is 1, whereas its token

value is 5. Then, its TTR is 0.2 (1/5). When the TTR of a text approaches 1, it implies that the text has lexically rich variation because words in the text are seldom repeated. Readers will have more difficulty with comprehending a text with more lexically rich variation because they need to decode more words and integrate them into the context (Graesser, McNamara, Louwerse, & Cai, 2004).

- Narrativity indicates the degree of familiarity of the text in terms of characters, goals, settings, events, places, and times to a reader. A higher narrativity score implies the text is more associated with common, oral conversation, whereas the non-narrative text is associated with less familiar topics (e.g., Graesser et al., 1994). This indicator is particularly related to word familiarity, world knowledge, and oral language. MOCCA intends to measure the ability of causally inferential comprehension from narrative texts. When a text passage is more narrative, it implies that the topic is more familiar and delivered in plain language, and readers are more likely to select correct choices to construct the situation from more narrative text (Carlson, Seipel et al., 2014; McNamara et al., 2011).

For this study, text complexity indicators were calculated for 100 third-grade MOCCA stories via Coh-Metrix. The texts were included with the full text including the sixth sentence filled with a correct, causal inference choice, but excluding titles and the two incorrect answer options.

3.2.2 Story feature coding

In the MOCCA pilot study, a total of 480 stories were developed and administered. After the first-year pilot study was complete, eleven features related to the

story structure were selected and coded to examine whether they might have a relationship to readers' making coherent inferences, comprehending the stories, and finally selecting a correct choice. The eleven story features coded were whether the title of the story included animal types; whether the story was written in the first person; what the main character's gender/sex was; in which sentence the main character was clearly identified; whether the story was about a child, an anthropomorphic animal, or an adult; whether there existed a secondary agent whose goal was in conflict with that of a main character; whether the story was realistic or fantastic; in which sentence the main goal of the story was clearly identified or inferable; whether the goal of the story was explicitly or implicitly stated; whether the main goal was met, not met, or remains unresolved in the passage; and whether the main character's emotion in the final sentence was positive, negative, neutral, or mixed (Biancarosa & Yoon, 2015).

Process of story feature coding. The story features were coded by four human raters, who were graduate students enrolled at the University of Oregon, under the supervision of a leading researcher on the MOCCA development team. In the first phase, the supervisor trained the students with the *Story Coding Guidelines* (Biancarosa & Yoon, 2015) about the characteristics of each story feature and the way to code them. Then, the human raters coded the same 180 out of 480 stories (37.5%) in common as a group. If the raters coded differently, they had discussions to reach agreements. The *Story Coding Guidelines* (Biancarosa & Yoon, 2015) were also revised to reflect the discussions and make the coding accurate and straight-forward.

In the second phase, the four students were paired up to code the remaining 300 stories (62.5%). If a pair coded a story differently, they discussed until they reached

agreement. If the pair did not reach to agreement, the story was coded by the whole four-person team. In the pilot study, MOCCA consisted of four forms within each grade, with each form consisting of 40 stories. In coding the story features, the 480 stories from three grades were not differentiated by grade or form to increase reliability and to minimize the impact of infrequent story features (Seipel, Biancarosa, Carlson, & Davison, 2015).

Interrater reliability. The interrater reliability was examined for human coding by means of the interrater agreement percentage, Gwet's AC1, and Fleiss' Kappa by code and by story. Gwet's AC1 was examined because it provides a more stable inter-rater reliability coefficient than Cohen's Kappa and is less affected by prevalence and marginal probability (Wongpakaran, N., Wongpakaran, T., Wedding, & Gwet, 2013). Fleiss' Kappa, an adaptation of Cohen's Kappa for 3 or more raters, was also calculated as the most conservative inter-rater reliability coefficient (McHugh, 2012; Seipel et al., 2015).

For the total stories, the mean inter-rater agreement percentage was 92%; the mean Gwet's AC1 was 91%; the mean Fleiss' Kappa was 88%. By code, the inter-rater agreement percentage ranged from 77% (i.e., Goal explicit) to 100% (i.e., Animal title); Gwet's AC1 ranged from 66% (i.e., Goal met) to 100% (i.e., Animal title); Fleiss' Kappa ranged from 50% (i.e., Goal met) to 100% (i.e., Animal title) (Seipel et al., 2015; Yoon et al., 2017).

Five story features. For this dissertation study, I initially selected nine out of eleven features and then recategorized them into five features with relevant associations. Using the new features, a total of 100 stories used for the third graders' forms were recoded as following:

- ***Child-centered by Realistic*** (coded as Child & Real): This feature integrated whether the story was about a child, and anthropomorphic animal, or an adult with whether the story was realistic or fantastic. This variable is associated with the background knowledge that a young reader possesses. If the story is constructed from a child's perspective with possible and plausible events to a child, a young reader may have better and faster understanding of the story. However, young readers may have more difficulty with comprehending the text if the story is not related to a child's familiar world, and/or the setting is set beyond their understanding of the situation. In the original study, child-centeredness was coded into three levels (Yes, No, and Anthropomorphized animal child), and being realistic was coded into two levels (Yes and No). Only a few stories had an anthropomorphized animal as a main character in the sample. Thus, the anthropomorphized stories were recoded as "child-centered" when the animal was obviously designated as a young animal, and otherwise as "not child-centered." This study coded this feature, Child & Real, into four values as YY if the story was child-centered (Y) and realistic (Y), YN if the story was child-centered (Y) but not realistic (N), NY if the story was not child-centered (N) and realistic (Y), and NN if the story was not child-centered (N) and not realistic (N).
- ***Secondary agent*** (coded as SecondAge): This feature was used as is. A story may be more difficult to comprehend if it has a secondary agent whose goal does not agree with a main character than if it does not have a secondary agent. This variable was coded as binary, Y if a secondary agent was present, and N if none.

- ***Goal sentence number by Explicitness*** (coded as Location & Explicit): This feature combined goal explicitness with the sentence number in which the main goal of the story was first clearly identified or inferable. If a reader identifies the explicit goal of the story as early as in the first or second sentence, it will guide a reader to attend to the main topic from the beginning and comprehend the story faster and more clearly than if the story has an inferable goal in the third or later sentence. It was coded as AY if the goal was located in the first or second sentence (A) with explicitness (Y), BY if the goal was located in the third or later sentence (B) with explicitness (Y), AN if the goal was located in the first or second sentence (A) without explicitness (N), BN if the goal was located in the third or later sentence (B) without explicitness (N)
- ***Position of Goal and Main Character*** (coded as Goal & Character): This variable collapsed the sentence number where the goal was first identified or inferable explicitly or implicitly and the sentence number where the main character first appeared. The item format of MOCCA indicates that the main character and the goal are intended to appear in different sentences (see 3.1.2) for step by step development of events in the narrative story. If the goal and the main character are identified in the same sentence, the reader will be more likely to have difficulty following the story to establish a situation because having both in one sentence may give a cognitive burden on a reader in processing the information at the same time. This variable was coded as S if both the main character and the goal appeared in the same sentence, and D if they appeared in the different sentences.

- ***Goal met by Emotion at ending*** (coded as GoalMet & Emotion): This feature integrated whether the main goal was met or not with what the main character's emotion was in the final sentence. A fulfilled goal will make the main character feel positive in the end of the story, and a failed or unresolved goal will make the main character feel negative. Happy endings and happy feelings may make a reader feel satisfied, and unhappy endings and negative feelings may make a reader feel uncomfortable and even confused. And the conflicting feeling against the goal achievement such as the main character's positive emotion despite an unfulfilled goal or vice versa might make a reader confused when asked to make a causal inference and more difficult or slower to comprehend the story. This feature is related to understanding how the success of goal achievement and related emotional valence affects the item difficulty in a reading comprehension assessment. This feature was coded MP if the goal was met (M) and the main character was in a positive mood in the end (P), MNP if the goal was met (M) but the main character was in a negative, neutral, or mixed mood in the ending (NP), NMP if the goal was not met or unresolved (NM) but the main character was in a positive mood in the ending (P), and NMNP if the goal was not met or unresolved (NM) and the main character was in a negative, neutral, or mixed mood in the ending (NP).

3.3 Participants and the Dataset

The data were collected in the U.S. across two academic years, 2016-2017 and 2017-2018. In the original data set, the students enrolled at grades 3, 4, and 5 participated in the MOCCA study from 52 schools within 33 districts. For this dissertation study, I

used and analyzed the data set of the third-grade participants (n = 1,569) from the original data set by collapsing the data across the two years. The five demographic categories such as gender, race/ethnicity represented as white, SES represented as free and reduced meals, special education status, and EL status were used as predictors of person properties to see whether there were any different responses to the MOCCA items depending on the person characteristics.

Three forms (i.e., Form 3.1, Form 3.2, and Form 3.3) that were administered for the third graders in MOCCA study were included in this study. I analyzed three forms separately in this study to consider the large disparity of sample sizes between common items and unique items and to see whether each form was consistently constructed in terms of text features. One out of three forms assigned to the grade was randomly given to each participant. Each form had 40 items including 10 common items across forms and 30 unique items within forms. A total of 100 items in the grade were analyzed.

3.4 Data Analytic Methods

Multiple psychometric methods were used to analyze the data set to answer the research questions. Explanatory Item Response Modeling (EIRM) was used to investigate how a range of item features and person characteristics related to the item difficulty and different responses on the MOCCA assessment. Differential Item Functioning (DIF) was also conducted to examine whether there were items that displayed collectively different responses between English learners and non-English learners. The following sections provide descriptions of the EIRM and DIF used in this study.

3.4.1 Explanatory item response models

Item Response Theory (IRT) has been increasingly used to assess individuals' abilities, skills, or characteristics on the construct measured through a test. IRT is a model-based measurement to analyze assessment in assessing the person and item parameters of the assessment by describing the relationship between the item properties and the examinees' responses (De Ayala, 2013; Embretson & Reise, 2013). However, traditional IRT models describe the locations of the item and person parameters on the same scale, but do not explain why they are located at a particular location on the scale (De Ayala, 2013). To explain why an item is more difficult than others and why an examinee responds to an item in a certain way, an explanatory approach is added to the descriptive IRT models. Explanatory item response modeling (EIRM) intends to explain what item or person variables are related to the difficulty of or responses to items on the basis of person and item parameters that are estimated by means of IRT models (De Boeck & Wilson, 2004).

Descriptive versus explanatory item response models. IRT measurement models measure the outcome variables in relation to persons, estimate the parameters of both the persons and items, and describe the relations between the persons and items through estimation (ibid). IRT models are understood as doubly descriptive in terms of persons and items. To this descriptive approach of the measurement models, an explanatory approach is added in order to explain how responses to items are related to external variables such as item properties (e.g., item formats, question types, etc.) and person properties (e.g., learning style, language proficiency, and intervention types) (Wilson & Moore, 2012).

In the descriptive models, one fixed measurement location is estimated for each item and person. The person's location on the latent trait continuum is represented by the person parameter (θ_p) which implies the person's ability on the scale. The item's location is represented by the item parameter (β_i) which implies the difficulty of the item. In the IRT models, the person parameter, i.e., person ability (θ_p), and the item parameter, i.e., item difficulty (β_i), are estimated and described on the same scale. By adding parameters of the external variables such as item properties and person properties in explanatory models, the effects of the added properties are estimated on how responses to items are generated. The parameters in the explanatory models are interpreted in a similar way to regression weights (De Boeck et al., 2011; De Boeck & Wilson, 2004; Wilson & Moore, 2012).

Wilson and his colleagues (e.g., De Boeck & Wilson, 2004; Wilson & Moore, 2012) present four item response models with respect to whether they are descriptive or explanatory on the person and the item sides. Table 1 present the four models according to the presence of person or item properties. The Rasch model in the upper left cell of the table describes each person's and item's location on the responses. The Linear Logistic Test Model (LLTM) is explanatory on the item side, and the latent regression model is explanatory on the person side. The latent regression LLTM is explanatory on both the item and the person sides.

Table 1

Models as a Function of the Predictors

	Person predictors	
Item predictors	Absence of properties (person indicators)	Inclusion of properties (person properties)
Absence of properties (item indicators)	Doubly descriptive (Rasch Model)	Person explanatory (Latent Regression Rasch Model)
Inclusion of properties (item properties)	Item explanatory (Linear Logistic Test Model: LLTM)	Doubly explanatory (Latent Regression LLTM)

Note. Adapted from "Explanatory Item Response Models: A brief introduction" by Wilson, M., De Boeck, P., and Carstensen, C.H., 2008, In Hartig, J., Klieme, E., & Leutner, D. (Eds.), *Assessment of Competencies in Educational Contexts* (pp.83-110). USA: Hogrefe & Huber Publishers.

A doubly descriptive model: The Rasch model. The Rasch model describes the data structure by means of the person and item parameters from the assessment (Wilson & Moore, 2012). It describes how persons' responses to item differ through a person ability parameter (θ_p), which is considered a random effect. At the same time, the model describes how persons' responses to item vary through an item difficulty parameter (β_i), which is considered a fixed effect. The model indicates that the difficulty of each item is consistent across the persons, but the different responses by each person depends on the person's ability. The equation for the Rasch model is the following:

$$\eta_{pi} = \theta_p - \beta_i \qquad (1)$$

where η_{pi} is the logit of the probability of person p' success on item i, θ_p is the latent trait level for person p as the intercept varying at random over person, $\theta_p \sim N(0, \sigma_\theta^2)$, and β_i is the difficulty of item i (De Boeck & Wilson, 2004). This equation implies that the probability of responding successfully to a given item (η_{pi}) is estimated in terms of the difference between person ability (θ_p) and item difficulty (β_i).

An item explanatory model: the LLTM. The Linear Logistic Test Model (LLTM) explains the effects of different item properties on the probability of responses (η_{pi}) on the basis of the Rasch model (Fischer, 1973). When the Rasch model estimates item parameters individually, the LLTM estimates the effects of item properties and explains the contribution of each item by means of the item properties included (Wilson, De Boeck, & Carstensen, 2008; Wilson & Moore, 2012). For example, if the Rasch model simply describes certain items are more difficult than the others, the LLTM explains why the items have such difficulty by analyzing the item properties (e.g., linguistic features or topic areas) and estimating the effects of the properties on the item difficulty. The interactive effects of item properties can also be estimated by means of adding the product of two or more item property variables to the model. The equation of the LLTM is the following:

$$\eta_{pi} = \theta_p - \sum_{k=o}^{K} \beta_k X_{ik} \quad (2)$$

where X_{ik} is the value of item i on item property k ($k = 0, \ldots, K$), and β_k is the regression weight of item property k (De Boeck & Wilson, 2004; Wilson et al., 2008).

The model in Equation 2 is referred to as the LLTM because it is based on a logit link and on a linear combination of multiple item properties for a linear equation for the logit of the probability of correct response (η_{pi}) (De Boeck, Cho, & Wilson, 2016; Wilson et al., 2008). No error term is added to Equation 2, suggesting that the item effects can be perfectly explained from the item properties and that the β_i in the Rasch model in Equation 1 is the same as the $\sum_{k=o}^{K} \beta_k X_{ik}$ from Equation 2.

A person explanatory model: the latent regression Rasch model. The person explanatory model, the latent regression Rasch Model, models person properties (e.g., socioeconomic status, English proficiency) as predictors in order to explain differences among persons with respect to the latent trait on the assessment. In a person explanatory model, the item side is left unexplained because it is treated as descriptive (Wilson & Moore, 2012). The equation of the latent regression Rasch model is the following:

$$\eta_{pi} = \sum_{j=1}^{J} \vartheta_{ij} Z_{pj} + \theta_p - \beta_i \qquad (3)$$

where Z_{pj} is the value of person p on person property j ($j = 1, \ldots, J$), ϑ_j is the (fixed) regression weight of person property j, θ_p is the remaining person effect after the effect of the person properties is accounted for, $\theta_p \sim N(0, \sigma_\varepsilon^2)$, which may be considered as the random effect Z_{p0}, the new random intercept (De Boeck & Wilson, 2004; Wilson et al., 2008).

The ϑ_j is not the same as θ_p in the equation. The ϑ_j indicates the regression weight of a person property. The θ_p represents the person parameter that is regressed on person

properties such as gender and that is explained with respect to person properties (the *Z*s) and their effects (the ϑs) (Wilson et al., 2008).

A doubly explanatory model: the latent regression LLTM. A final model is the latent regression LLTM, doubly explanatory in that both the person side and the item side are added together for Equations 2 and 3 on the basis of the Rasch model (Equation 1). The equation for the model is the following:

$$\eta_{pi} = \sum_{j=1}^{J} \vartheta_{ij} Z_{pj} + \theta_p - \sum_{k=o}^{K} \beta_k X_{ik} \qquad (4)$$

As with the previous models, the model in Equation 4 has two parts: a person side and an item side. The person side is explained by means of person properties and has a random effect term while the item side is explained by means of item properties (Wilson et al., 2008).

Table 2 presents a summary of the four models to be explained with equations. The notation used in the table is as following. θ_p indicates the random person parameter which is normally distributed with mean zero and variance σ_θ^2 , $\theta_p \sim N(0, \sigma_\theta^2)$. Z indicates the person properties as the predictors. The subscript *j* is used for these predicators, $j = 1, \ldots, J^2$. X indicates the item properties as the predictors, with subscript *k*, $k = 1, \ldots, K$. ϑ_j indicates the effects of person predictors which are considered fixed, and β_k indicates the fixed effects of item predictors (Wilson et al., 2008).

Table 2

Summary of the Four Models

Model	$\eta_{pi} =$ Person side	Item side	Random effect	Model Type
Rasch model	θ_p	$-\beta_i$	$\theta_p \sim N(0, \sigma_\theta^2)$	Doubly descriptive
LLTM	θ_p	$-\sum_{k=o}^{K} \beta_k X_{ik}$	$\theta_p \sim N(0, \sigma_\varepsilon^2)$	Item explanatory
Latent regression model	$\sum_{j=1}^{J} \vartheta_{ij} Z_{pj} + \theta_p$	$-\beta_i$,	$\theta_p \sim N(0, \sigma_\varepsilon^2)$	Person explanatory
Latent regression LLTM	$\sum_{j=1}^{J} \vartheta_{ij} Z_{pj} + \theta_p$	$-\sum_{k=o}^{K} \beta_k X_{ik}$	$\theta_p \sim N(0, \sigma_\varepsilon^2)$	Doubly explanatory

Note. Adapted from "Explanatory Item Response Models: A brief introduction" by Wilson, M., De Boeck, P., and Carstensen, C.H., 2008, In Hartig, J., Klieme, E., & Leutner, D. (Eds.), *Assessment of Competencies in Educational Contexts* (pp.83-110). USA: Hogrefe & Huber Publishers.

3.4.2 Differential item functioning (DIF)

Differential Item Functioning (DIF) is a statistical method used for examining whether items are equivalent or comparable for different groups such as English learners and non-English learners (Ercikan, 2002). The groups may respond to a certain item differently as a whole because each group actually has different abilities on the construct measured. But if the groups respond differently to a certain item which measures secondary traits not directly relevant to the construct being measured, this becomes a validity issue and a threat to a fair test by disadvantaging one group over another on the measurement. Clauser and Mazor (1998) discussed use of statistical procedures to identify test items functioning differently and stated that a test item is biased when it

unfairly favors one group over another. If one group performs worse than the other because language is difficult in a math calculation test, the differences between the groups are likely due to a secondary trait, language proficiency, rather than the construct being assessed, math calculation ability (Buzick & Stone, 2011; Clauser & Mazer, 1998).

DIF analysis only identifies which items show DIF, but does not tell us why they occur. If DIF items are identified with extreme gap, it is suggested that a panel of experts review the items to decide the extent to which DIF items are relevant to the construct intended to measure (Ayala, 2009). If the DIF items are identified as measuring the secondary traits, they are irrelevant to the intended construct in the test and considered biased. Item bias should be avoided because it is a measurement error to validity and fairness in testing and test use (AERA, APA, & NCME, 2004). However, identifying sources of DIF is not easy because multiple factors affect responses to test items.

This dissertation study used DIF to detect test items in a reading comprehension assessment on which English learners might show significantly different performance in comparison to non-English learners. The items found to show substantial DIF were analyzed to identify the sources of different performance with respect to text features such as text complexity indicators and story features coded for Research Question 2.

CHAPTER IV

RESULTS

The current chapter describes the analyses and findings under the phases and related research questions. The analytic methods were conducted by each form because the participants took one of three forms which was randomly distributed to them, and thus the sample was different by form.

First, to answer RQ1 under Phase I, a descriptive statistical analysis was conducted to understand the distribution and associations among the variables in the study of the MOCCA assessment, and an IRT Rasch model was run as a first step for the explanatory approach to understand the variation of persons and items of the MOCCA assessment.

Second, to answer RQ2 and RQ3 under Phase II, two explanatory item response models were conducted to explain the differences on the item difficulty and responses of MOCCA with respect to the external variables. To answer RQ2, the linear logistic test model (LLTM) was conducted on the basis of a selected model (Rasch model) by means of text complexity indicators and story features to examine the effects of textual features on the item difficulty in MOCCA. To answer RQ3, the latent regression analysis was conducted by means of person properties such as gender, white, FARMs, special education participation and EL status to investigate the effects of the person properties on the responses.

Third, to answer RQ4-a under Phase III, Differential Item Functioning (DIF) was conducted to detect whether ELs showed different responses on the MOCCA items as a

group than non-ELs. To answer RQ4-b under Phase III according to the findings of RQ4-a, the items detected with DIF were investigated to see what textual features might have a relation with the group differences.

4.1 Phase I: Understanding of the MOCCA data

This section describes the results of the statistical and psychometric analyses conducted to evaluate how the Multiple-choice Online Causal Comprehension Assessment, MOCCA, performed. As described previously, the current study included three forms within the grade in separate data analyses considering the sample size per item and possibility of different form composition. First, the statistical descriptive analyses provided the information about the composition and performances of the participants as a group and by subgroups on the MOCCA assessment, and the way that texts in MOCCA were constructed by means of the text complexity indicators and story features. Second, two models, the Rasch model and 2PL model, were compared to identify the one with better fit to use in subsequent analyses, and the Rasch model was selected to analyze the MOCCA data. Third, the Rasch model provided psychometric information about how MOCCA worked as an assessment.

4.1.1 Descriptive analyses

Descriptive analyses provided three results about the composition and performance of the participants and subgroups on MOCCA, the construction of items by means of the text factors, and the item difficulty of MOCCA items calculated by means of classical test theory.

Understanding composition and performance of the participants on MOCCA.

Table 3 presents the demographic information for the participating sample in the study. Not all participants provided demographic information, so the percentages were calculated from valid information. Among a total of 1,569 participants enrolled in the third grade, about 79% (n = 1,237) provided a whole or partial demographic information for gender, race, free and reduced-price meals (FARMs) status, special education participation (sped), and English learner status (EL).

In terms of gender composition, of the participants who provided demographic information, 51% were male and 49% female students. Non-binary gender choices were not collected in the data. In terms of race/ethnicity, the majority of the valid participants were white (53%), followed by Hispanic (27%), Black or African American (11%), two or more races (4%), Asian (3%), American Indian or Alaska Native (2%), and Native Hawaiian or Other Pacific Islander (<1%). Considering the large number of parameters and small number of some racial groups, the race/ethnicity variable in this study was collapsed and reclassified as two categories, white (52%) and non-white (48%). Among the valid participants, 60% were qualified for free and reduced-price meals, and about 10% were eligible for special education services. About thirteen percent of the participants were English learners. Although English learner classifications can vary along a variety of dimensions, the dataset only classified the students as EL and non-EL, which will make it difficult to generalize any findings about ELs who have different English proficiency ranging from high proficiency very similar to native speakers to low proficiency barely understanding English alphabets.

Table 3

Demographic Information by Subgroup

	Form 3.1 (*n* = 490)	Form 3.2 (*n* = 549)	Form 3.3 (*n* = 540)	Total (*n* = 1,569)
Gender				
Male, *n* (%, valid %)	217 (44.3%, 52.5%)	234 (42.6%, 50.3%)	225 (42.5%, 50.2%)	676 (43.1%, 51%)
Female, *n* (%, valid %)	196 (40%, 47.5%)	231 (42.1%, 49.7%)	223 (42.1%, 49.8%)	650 (41.4%, 49%)
NA, *n* (%)	77 (15.7%)	84 (15.3%)	84 (15.5%)	243 (15.5%)
White				
Yes, *n* (%, valid %)	201 (41%, 54%)	216 (39.3%, 50.7%)	209 (39.4%, 50.9%)	626 (39.9%, 51.8%)
No, *n* (%, valid %)	171 (34.9%, 46%)	210 (38.3%, 49.3%)	202 (38.1%, 49.1%)	583 (37.2%, 48.2%)
NA, *n* (%)	118 (24.1%)	123 (22.4%)	119 (22.5%)	360 (22.9%)
Free and reduced price meals				
Yes, *n* (%, valid %)	148 (30.2%, 60.4%)	165 (30.1%, 60.9%)	147 (27.7%, 59%)	460 (29.3%, 60.1%)
No, *n* (%, valid %)	97 (19.8%, 39.6%)	106 (19.3%, 39.1%)	102 (19.2%, 41%)	305 (19.4%, 39.9%)
NA, *n* (%)	245 (50%)	278 (50.6%)	281 (53%)	804 (51.2%)
Special education status				
Yes, *n* (%, valid %)	27 (5.5%, 8.6%)	39 (7.1%, 10.7%)	34 (6.4%, 9.6%)	100 (6.4%, 9.7%)
No, *n* (%, valid %)	286 (58.4%, 91.4%)	327 (59.6%, 89.3%)	320 (60.4%, 90.4%)	933 (59.5%, 90.3%)
NA, *n* (%)	177 (36.1%)	183 (33.3%)	176 (33.2%)	536 (34.2%)
EL status				
Yes, *n* (%, valid %)	47 (9.6%, 13.3%)	48 (8.7%, 11.9%)	49 (9.2%, 12.5%)	144 (9.2%, 12.5%)
No, *n* (%, valid %)	307 (62.7%, 86.7%)	355 (64.7%, 88.1%)	344 (64.9%, 87.5%)	1006 (64.1%, 87.5%)
NA, *n* (%)	136 (27.8%)	146 (26.6%)	137 (25.8)	419 (26.7%)

Table 4 presents the results of the descriptive statistics about how the subgroups performed on each form of MOCCA. The table lists the mean differences by subgroup according to the person characteristics: gender as collected self-declared male/female in the date set, race represented as white/non-white (white), free and reduced-price meals (farms) representing SES status, special education participation (sped), and EL status (EL). Because one form consists of 40 items and each item correct is scored as 1, the maximum score is 40 and the minimum is 0.

Table 4

Descriptive Statistics by Subgroup

		Form 3.1	Form 3.2	Form 3.3
Gender	Male *(n)*	217	234	225
	mean (sd)	24.38 (11.68)	25.24 (10.66)	24.53 (11.39)
	Female (*n*)	196	231	223
	mean (sd)	24.95 (11.01)	26.49 (10.83)	25.7 (11.14)
White	Yes (*n*)	201	216	209
	mean (sd)	25.74 (11.29)	28.22 (10.71)	27.67 (10.88)
	No (*n*)	171	210	202
	mean (sd)	21.62 (10.81)	22.67 (10.06)	21.29 (10.6)
Farms	Yes (*n*)	148	165	147
	mean (sd)	22.99 (11.26)	24.05 (10.77)	22.93 (11.2)
	No (*n*)	97	106	102
	mean (sd)	29.03 (10.53)	28.63 (10.34)	28.4 (10.03)
Sped	Yes (*n*)	27	39	34
	mean (sd)	17.44 (12.57)	17.38 (9)	20.15 (10.23)
	No (*n*)	286	327	320
	mean (sd)	25.84 (11.09)	26.92 (10.55)	25.82 (11.06)
EL	Yes (*n*)	47	48	49
	mean (sd)	17.87 (9.11)	17.54 (8.88)	17.67 (8.78)
	No (*n*)	307	355	344
	mean (sd)	25.02 (11.32)	26.66 (10.66)	25.85 (11.27)

With respect to gender, girls and boys had similar means across forms. It needs to be noted again that this data only included binary gender choices. Race/ethnicity was dichotomously categorized as white and non-white because of the proportion of sub-racial demographic compositions. The two groups showed significant mean differences at .001 level across forms when a two sample t-test was conducted, $t(370) = 3.58$, $p < .001$ for Form 3.1, $t(424) = 5.52$, $p < .001$ for Form 3.2, and $t(409) = 6.02$, $p < .001$ for Form 3.3. The results indicated that white students performed better than non-white students on all forms.

With respect to socio-economic status, there were also significant mean differences at the .001 level in all forms, $t(243) = 4.19$, $p < .001$ for Form 3.1, $t(269) = 3.46$, $p = .001$ for Form 3.2, and $t(247) = 3.96$, $p < .001$ for Form 3.3. The students who did not qualify for free and reduced-price meals performed better than those who were qualified for subsidized meals on all forms.

With respect to special education participation, significant mean differences were also detected at the .01 level in all three forms, $t(311) = 3.72$, $p < .001$ for Form 3.1, $t(364) = 5.42$, $p < .001$ for Form 3.2, and $t(352) = 2.86$, $p = .004$ for Form 3.3. The results need to be interpreted with caution due to the small sample size and lack of specificity about student's special education designation. The number of special education participants is so small on each form that caution is warranted, and the type of special education services for which students qualified was not included in the data.

With respect to EL status, significant mean differences were found in all forms, $t(69.75) = 4.84$, $p < .001$ for Form 3.1, $t(66.75) = 6.51$, $p = .001$ for Form 3.2, and $t(72.61) = 5.86$, $p < .001$ for Form 3.3. The results indicated that English learners

performed worse on all forms of the reading comprehension assessment than non-English learners. As mentioned previously, caution needs to be used in generalizing the results because nuanced EL classification status and their different proficiency level were not reported in the data.

The overall results suggested that students who were white, from higher SES backgrounds, not qualified for special education services, and non-English learners performed better on the MOCCA assessment than students who were non-white, from lower SES backgrounds, qualified for special education services or who were ELs.

Understanding MOCCA text construction by means of the text factors. Table 5 presents the results of the statistical descriptive analysis about how MOCCA texts were constructed by means of the text factors including the text complexity and story structures by form. Text complexity and story structures were separately coded because the first was quantitatively calculated via a text analysis tool, Coh-Metrix, and the latter was coded by human raters. Coh-Metrix 3.0 (University of Memphis) analyzed 100 texts used in MOCCA and provided numerical results, and five indicators out of 106 indicators were selected for this study. Story structure was analyzed by human raters with respect to five categorical story features.

Five text complexity indicators included word count (i.e., number of words in the text; the more, the longer text), word familiarity (i.e., how familiar the words are rated ranging from 100 to 700; the higher, the more familiar), word concreteness (i.e., how concrete/non-abstract the words are rated; the higher, the more concrete), type-token ratio (i.e., ratio of unique words and their tokens ranging from 0 to 1; the higher, the more

uniquely used), and narrativity (i.e., degree of how narrative the text is in percentage; the higher, the more story-like).

Word count ranged from 52 words to 116 words per text passage with a mean of 82.9 words for 100 text passages. The forms did not have significantly different number of word counts, m = 83.9 for Form 3.1, m = 85.1 for Form 3.2, and m = 83.75 for Form 3.3. Word familiarity ratings for 100 texts ranged from 538 to 607, with a mean of 577.5. Higher word familiarity means more familiar words to the reader with a maximum of 700 and a minimum of 100. The forms had similar mean ratings for familiarity for content words, 577.37 for Form 3.1, 579.78 for Form 3.2, and 574.84 for Form 3.3. Word concreteness ratings for 100 texts ranged from 338 to 501, with a mean of 412.1. Higher rating for word concreteness means the words are more concrete and less abstract to the reader with a maximum of 700 and a minimum of 100. The forms also had almost similar mean ratings for concreteness for content words, 411.22 for Form 3.1, 415.15 for Form 3.2, and 412.79 for Form 3.3. Type-token ratios (TTR) for 100 texts ranged from 0.52 to 0.77 with a mean of 0.64. Higher ratio between 0 and 1 means the text has less repeated and more unique words, implying more cognitive burden on the reader to extract meanings from more unique words. The forms had similar mean TTR, 0.63 for Form 3.1, 0.65 for Form 3.2, and 0.64 for Form 3.3. Narrativity for 100 texts ranged from 12.7% to 99.8%. Higher narrativity between 0 and 100 means the text has more familiar topics to the reader with simpler language. The forms also had similar mean narrativity, 62 for Form 3.1, 66.6 for Form 3.2, and 65.8 for Form 3.3.

Table 5 *Descriptive Statistics of MOCCA Text Construction*

		Form 3.1 (*n*=40)	Form 3.2 (*n*=40)	Form 3.3 (*n*=40)	Total (*n*=100)
Text complexity indicators					
Word count, *m* (*sd*) [min, max]		83.9(10.3)[62, 107]	85.1(11.1)[61, 116]	83.8 (11.8)[52, 107]	82.9(10.8)[52, 116]
Word familiarity, *m* (*sd*) [min, max]		577.4(14.3)[538, 595]	579.8(13.8)[538, 595]	574.8(13.6)[538, 594]	577.5(13.1)[538, 607]
Word concreteness, *m* (*sd*) [min, max]		411.2(36.9)[338, 501]	415.2(33.8)[357, 501]	412.8(34)[357, 501]	412.1(32.8)[338, 501]
Type-token ratio, *m* (*sd*) [min, max]		0.63(0.05)[0.54, 0.74]	0.65(0.05)[0.53, 0.77]	0.64(0.04)[0.52, 0.75]	0.64(0.05)[0.52, 0.77]
Narrativity, *m* (*sd*) [min, max]		62(25.7)[12.7, 97.5]	66.6(18.4)[24.2, 97.2]	65.8(21.7)[18.7, 99.8]	64.8(21.8)[12.7, 99.8]
Story features					
Child & Real	YY, *N* (%)	24 (60%)	28 (70%)	29 (72.5%)	67 (67%)
	YN, *N* (%)	3 (7.5%)	2 (5%)	2 (5%)	3 (3%)
	NY, *N* (%)	7 (17.5%)	4 (10%)	6 (15%)	15 (15%)
	NN, *N* (%)	6 (15%)	6 (15%)	3 (7.5%)	15 (15%)
SecondAge	Yes, *N* (%)	9 (22.5%)	9 (22.5%)	12 (30%)	22 (22%)
	No, *N* (%)	31 (77.5%)	31 (77.5%)	28 (70%)	78 (78%)
Goal & Character	Same, *N* (%)	13 (32.5%)	13 (32.5%)	13 (32.5%)	35 (35%)
	Different, *N* (%)	27 (67.5%)	27 (67.5%)	27 (67.5%)	65 (65%)
Location &	AY, *N* (%)	14 (35%)	15 (37.5%)	18 (45%)	37 (37%)
Explicit	AB, *N* (%)	8 (20%)	7 (17.5%)	5 (12.5%)	18 (18%)
	BY, *N* (%)	13 (32.5%)	11 (27.5%)	10 (25%)	28 (28%)
	BN, *N* (%)	5 (12.5%)	7 (17.5%)	7 (17.5%)	17 (17%)
GoalMet &	MP, *N* (%)	23 (57.5%)	19 (47.5%)	21 (52.5%)	57 (57%)
Emotion	MNP, *N* (%)	3 (7.5%)	2 (5%)	4 (10%)	7 (7%)
	NMP, *N* (%)	3 (7.5%)	5 (12.5%)	4 (10%)	10 (10%)
	NMNP, *N* (%)	11 (27.5%)	14 (35%)	11 (27.5%)	26 (26%)

Note. YY=child-centered & realistic. YN=child-centered & not realistic. NY=not child-centered & realistic. NN=not child-centered & not realistic. AY=explicit goal in the early sentence. AN=not explicit goal in the early sentence. BY=explicit goal in the later sentence. BN=not explicit goal in the later sentence. MP=goal met & emotionally positive. MN=goal met & emotionally not positive. UP=goal unmet & emotionally positive. UN=goal unmet & emotionally not positive.

The findings from the text complexity analysis suggested that the three forms in the third grade, one of which was randomly assigned to each examinee during testing, had quite uniform text complexity on average in terms of text length, word characteristics, and text structure. This may reflect a restriction of range in MOCCA because much cannot be expected to be explanatory due to little variation across forms. At the same time, it may be evidence that supports the validity of MOCCA forms within the grade.

Five story features included whether the story was child-centered and realistic (coded as Child & Real), whether the story had a second agent (coded as SecondAge), whether the goal and main character appeared in the same sentence (coded as Goal & Character), whether the goal was explicitly suggested in the early sentences (i.e., first or second sentence) (coded as Location & Explicit), and how the goal ended and how the main character emotionally accepted the ending (coded as GoalMet & Emotion).

Similar to text complexity, the overall story structures across the forms were not much different by form. With respect to whether they were child-centered and real, more than half of the stories, (67 out of 100), were coded as child-centered and realistic with 24 stories in Form 3.1, 28 stories in Form 3.2, and 29 stories in Form 3.3 classified in this way. With respect to the presence of a second agent, about ¾ of the stories per form did not have a second agent (n = 31 in Form 3.1, n=31 in Form 3.2, and n=28 in Form 3.3). With respect to the position of the goal and a main character, 27 stories per form respectively had the goal and a main character in different locations. With respect to the location and explicitness of the goal, the most stories within the form had an explicit goal in the first or second sentence (n = 14 in Form 3.1, n = 15 in Form 3.2, and n = 18 in

Form 3.3). But similarly slightly smaller number of stories within the form had an explicit goal in the third or later sentence (n = 13 in Form 3.1, n = 11 in Form 3.2, and n = 10 in Form 3.3). With respect to goal ending and emotional acceptance, most stories within the form had the goal met or achieved in the end and accordingly positive emotional valence (n = 23 in Form 3.1, n = 19 in Form 3.2, and n = 21 in Form 3.3). And the second most stories had the goal unmet or unresolved in the end and not positive emotional valence (n = 11 in Form 3.1, n = 14 in Form 3.2, and n = 11 in Form 3.3).

The findings from the story feature analysis indicated that in addition to text complexity, the three forms had overall quite uniform story structures in terms of character identification, goal identification, and story ending. Within the story features, many stories were constructed to have a similar structure, meaning that the majority of the stories had a child as the main character, were realistic, had no second agent, had an explicit goal located in the early sentences of the story, and had the goal achieved or resolved in the ending with positive emotional acceptance.

4.1.2 Item analysis by means of classical test theory

Item analysis was conducted by means of classical test theory framework to understand how the items of MOCCA performed. CTT item analysis involved calculating the percentage of the relative frequencies of item. The percentage known as item difficulty indicates the proportion of the examinees correctly answering each item, and thus it is conceptually similar to item easiness (Desjardins & Bulut, 2018). Thus, an item with higher percentage through CTT calculation means the item is answered correctly by more examinees and considered an easier item.

Appendix A presents the results of item difficulty analysis of MOCCA under classical test theory. The item numbers in the table are randomly assigned for this study and do not correspond to those of the actual assessments. The first 10 items (Item 1 to Item 10) were identical across all forms as common items, but the other 30 items were all unique to the given form. In Form 3.1, Item 25 was calculated as the most difficult, as only 44.9% of the examinees answered it correctly while Item 40 was calculated as the easiest, as 78.8% of the examinees answered it correctly. In Form 3.2, Item 7 was the most difficult, as 53.7% of the participants answered it correctly while Item 40 was the easiest as 80.1% of the participants answered it correctly. It is worth noting that Item 40 in Form 3.1 was not the same as Item 40 in Form 3.2, as the number was experimentally assigned for confidentiality. In Form 3.3, Item 7 was the most difficult, as 52.7% of the examinees answered it correctly to the item while Item 11 was the easiest, as 78.5% of the examinees answered it correctly. Item 7 on both Form 3.2 and Form 3.3 was the most difficult item on the two forms, respectively. The item was identical on both forms, as it was a common item linked across forms. It was also identified as a difficult item when it was used on Form 3.1, with only 51.8% of the examinees answering it correctly.

The average item difficulty was 63.9% ranging from 78.8% to 44.9% for Form 3.1, 67.36% ranging from 80.1% to 53.7% for Form 3.2, and 65.72% ranging from 78.5% to 52.7% for Form 3.3. The findings indicated that Form 3.1 had the lowest average item difficulty among the three forms, meaning the form was the most difficult on average, and Form 3.2 was the easiest on average among the three forms.

When examining the descriptive statistics for the set of common items, the mean and range of CTT item difficulty for the common items slightly varied depending on the

form. The average item difficulty of ten common items was 63.3%, (ranging from 72.1% to 51.8%) for Form 3.1, 65.7% (ranging from 72.2% to 53.7%) for Form 3.2, and 64.5% (ranging from 72.1% to 52.7%) for Form 3.3. Among the common items, Item 7 was the most difficult item administered across all forms. The easiest items varied by form, but had similar item difficulty between 72.1% and 72.2%. The remaining 30 items, unique on each form, had slightly higher item difficulty on average than the common items, indicating on average they were slightly easier for the examinees to answer correctly than the common items, but without statistically significant differences.

4.1.3 Selecting a psychometric model

Two IRT models were compared to select a model with relatively better fit to analyze the MOCCA data in this study and estimate the item and person parameters of the assessment. The specific research question for the analyses is:

RQ 1. Are item response models well-fitting for generating the parameter estimates of young readers' reading comprehension ability proficiency on the MOCCA assessment? If so, what do the results indicate regarding reading ability in the data set for this study?

To answer the research question, the model fits were compared between two item response models, Rasch model and 2 PL model, to select a better fit model as a baseline for further analyses. The Rasch model was applied to the data set to test the assumption about how the latent trait, reading comprehension ability on MOCCA, was associated with item responses in the assessment (Embretson & Reise, 2013). This model served as

a baseline to test whether item discrimination (2PL model) was needed to analyze this assessment. The R package Test Analysis Modules (**TAM**; Robitzsch, Kiefer, & Wu, 2019) was used to fit a marginal maximum likelihood estimation of a set of unidimensional item response models.

***Parameter estimation*.** First, **TAM** by means of `tam.mml` function estimated 41 parameters of Form 3.1, one form of three third-grade forms of MOCCA, including 40 item threshold parameters, no item slope parameters (all fixed to 1), no regression parameters, and 1 covariance parameter. Constraint on persons was specified for this estimation. Then, **TAM** by means of `tam.mml.2pl` function estimated 80 parameters of Form 3.1 for 2PL model, including 40 item threshold parameters and 40 item slope parameters.

Reliability. To examine how closely the items were related, Cronbach's alpha was estimated using classical test theory. For Form 3.1, Cronbach's alpha was 0.947. The IRT EAP reliability was also estimated as 0.909 for both Rasch and 2PL models. The findings indicated MOCCA approached a high overall instrument reliability because 0.8 or higher would be preferred. Both Rasch and 2PL models were highly reliable in estimating the parameters of the assessment.

Item fit statistics. Item fit for the Rasch model was good because the fit ranged from 0.83 to 1.34, within 3/4 – 4/3 mean square weighted fit for parameters in which no weighted fit T was greater than |2| (Ayala, 2009; Wu, Adams, & Wilson, 1998). The item fit was also good for 2PL model, ranging from .98 to 1.03, and no weighted fit T for the parameters was greater than |2|.

Standard Errors. Figure 2 shows the standard errors of measurement of the latent traits for the examinees for the Rasch model. The standard errors ranged from a high of about 1.35 at the high extreme of the latent trait to a low of about 0.31. The average of the standard errors approximated 0.47 logits. The plots had larger gaps in the far left but smaller gaps in the middle and in the far right, overall appearing stable and going up gradually to the right end. Figure 3 shows the standard errors for the 2PL model. The standard errors of the 2PL model ranged from a high of 0.06 to a low of 0.82, with an average of 0.28 logits. Though the 2PL model had a smaller average with a narrower range of standard errors than the Rasch model, the 2PL model showed more fluctuating plots than the Rasch model, repeatedly up and down, and rapidly increasing to the right end.

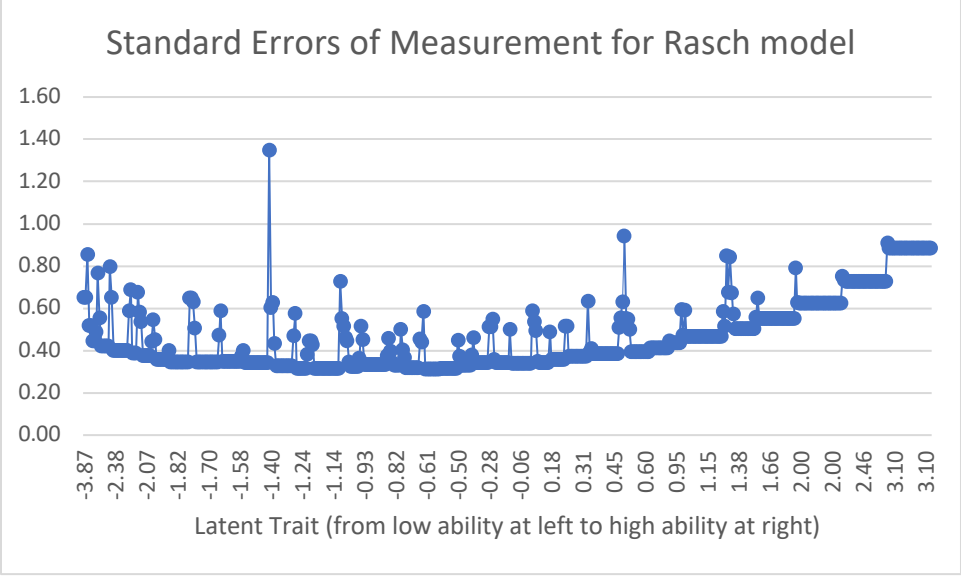

Figure 2. *The Plots of Standard Error of Measurement of Rasch Model.*

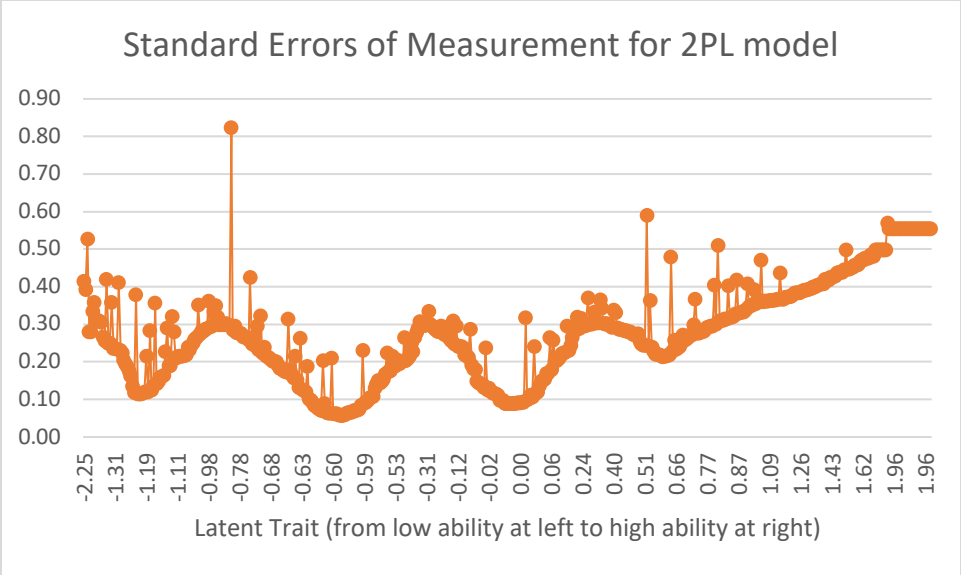

Figure 3. *The Plots of Standard Error of Measurement of 2PL Model*

Model fits between Rasch model and 2PL model. Table 6 provides the results of the comparison of the two models, Rasch and 2PL, with respect to AIC, BIC, and a loglikelihood ratio test. The comparison was made to select a relatively better-fitting model. The 2PL model had a better AIC (smaller is better), the Rasch model had a better BIC (smaller is better), and the 2PL model had a statistically significantly better fit than the Rasch model based on a Chi-square test ($\chi^2 = 176.09$, $df = 39$, $p < .001$).

Table 6

Model Fit Comparison Indices of the Relative Fit

Model	loglike	deviance	parameters	AIC	BIC	Chisq	df	p
Rasch model	-9025.53	18051.05	41	18133.05	18305.02			
2PL model	-8937.486	17874.96	80	18034.96	18370.51	176.09	39	0

Selecting a model. In selecting a better-fitting model, the indices did not clearly prefer one model to the other. The results of Chi-square test, AIC, and the average and range of the plots of SEM supported the 2PL model with slightly better fits, while BIC and the stability of the plots of SEM supported the Rasch model. Both models had identical EAP reliabilities. Item parameters of both models were good, all within 3/4 – 4/3 tolerance in infit statistics. Based on the evidence accumulated (AIC, BIC, Chi-square, SEM plots, reliability, and Infit), the Rasch model was selected as a reasonably and acceptably parsimonious model in using value 1 for estimating fixed item discrimination parameters, for the purposes of the research questions here. Thus, further analyses were conducted on the basis of the Rasch model.

4.1.4 Anchoring

Because ten common items were linked across three forms within the grade, the items were anchored to estimate the item parameters as well as person abilities. The item parameters estimated with Form 3.1 through the Rasch model were fixed in Form 3.2 and Form 3.3 with `tam.mml.mfr` function from R package **TAM**, having the same range of item locations across the forms. With the common items fixed in their estimation, the person and other item parameters of Form 3.2 and Form 3.3 were estimated with the Rasch model. The common items were anchored in differential item functioning, but not in evaluating the effects of item and person properties on the responses (LLTM and latent regression Rasch modeling) due to the limitation of the modeling and analysis package, which is discussed in the limitations section of this study.

4.1.5 Item estimation results by form

Table 7 presents the results of Rasch modeling estimation of the person and item parameters according to the forms. All the forms had high reliability, ranging from 0.901 to 0.909. The average item difficulty and standard deviation of the forms were also somewhat consistently estimated, *M*= -0.8562, *SD* = 0.4774 for Form 3.1, *M* = -0.9275, *SD* = 0.4102 for Form 3.2, and *M* = -0.9005, *SD* = 0.4026 for Form 3.3. Although there is no reason to assume that these are randomly equivalent samples, I do note here that the estimation of the average person ability on the latent trait varied according to the forms and sample used here, *M* = -0.001, *SD* = 1.582 for Form 3.1, *M* = 0.2504, *SD* = 1.609 for Form 3.2, *M* = 0.1809, *SD* = 1.6712 for Form 3.3. The average standard deviation of person ability through EAP estimation was similar for Form 3.2 and Form 3.3, *M* = 0.4961 (*SD* = .1913), and *M* = 0.4992 (*SD* = 0.2012), respectively, but Form 3.1 had slightly lower SD.EAP, *M* = 0.4702 (*SD* = 0.1662).

Table 7

IRT Estimation of the Parameters

	Form 3.1	Form 3.2	Form 3.3
EAP reliability	0.909	0.901	0.906
Item difficulty, *M (SD)*	-0.8562 (0.4776)	-0.9275 (0.4102)	-0.9005 (0.4026)
Person ability, *M (SD)*	-0.0010 (1.5820)	0.2504 (1.609)	0.1809 (1.6712)
SD.EAP	0.4702 (0.1662)	0.4961 (0.1913)	0.4992(0.2012)

Compared to Form 3.1, the parameters of Form 3.2 and Form 3.3 were estimated consistently, having lower item difficulty and higher person ability. In IRT, the probability of answering an item correctly is estimated by the difference between the person ability and the item difficulty. For Form 3.2, the estimated mean person ability is the highest and the mean item difficulty is the lowest among three forms, indicating the probability of the participants responding correctly to the items is the highest on average. For Form 3.1, the mean person ability is estimated as the lowest and the mean item difficulty as the highest, indicating the overall probability of correctly responding to the items is the lowest among the three forms. The results imply Form 3.2 was the easiest and Form 3.1 was the most difficult among the three forms. These results are consistent with those found in the classical test theory item analysis.

Appendix B presents the estimation of item parameters by form under the Rasch model. The item numbers in the table are not the same as the actual numbers in the assessment. The first 10 items as common items were linked across forms, having the same item difficulty estimated in three forms. The parameters of the linked items were fixed with the results of Form 3.1 estimation, and the parameters of the non-linking items were estimated under the Rasch model by form. As expected, the patterns of the estimation were similar to those of the CTT analysis. The most difficult items were Item 25 on Form 3.1 ($\beta = 0.34$), and Item 7 on Form 3.2 ($\beta = -0.0707$) and Form 3.3 ($\beta = -0.0707$), in which Item 7 on Form 3.2 was the same item as Item 7 on Form 3.3 as a common item. The easiest items were Item 40 on Form 3.1 ($\beta = -1.8787$), Item 40 on Form 3.2 ($\beta = -1.8514$), and Item 11 on Form 3.3 ($\beta = -1.8206$), in which Item 40 in Form 3.1 were not the same item as Item 40 on Form 3.2.

4.2 Phase II: Explanatory Item Response Modeling

This section presents the results of explanatory IRT modeling (EIRM) analyses including the linear logistic test model (LLTM) and latent regression analysis. The LLTM was conducted to examine how the item difficulty of MOCCA was associated with the text factors such as text complexity and story structure. The latent regression analysis was conducted to explain how the responses on MOCCA were associated with the personal properties of the subgroups of participants.

4.2.1 Linear Logistic Test Model (LLTM) and text features

On the basis of a selected Rasch model, an item explanatory approach by means of the linear logistic test model was used to analyze the data set in this study. The specific research question for the analyses is:

RQ 2. How are text features such as text complexity indicators and story features associated with explaining the item difficulty of MOCCA?

To answer Research Question 2, the linear logistic test model (LLTM) was conducted by means of the R package **lme4** (Bates, Maechler, Bolker, & Walker, 2019). The text features included five text complexity indicators and five story features coded. The text complexity indicators included word count, word familiarity, word concreteness, type-token ratio, and the narrativity, all of which were calculated numerically via a text analysis tool. The five story features included whether a story was realistically child-centered (Child & Real), the presence of a second agent (SecondAge), the position of goal and main character (Goal & Character), the location of explicit or inferable goal

(Location & Explicit), and emotional acceptance of the goal ending (GoalMet & Emotion).

A set of story features and a set of text complexity indicators were separately included in the LLTM analyses when the effects of the text features as the item properties were examined. When all ten predictors were included in the LLTM modeling, the model caused convergence problems because there were as many as 10 predictors (a total of five text complexity indicators and five story features) compared to the number of items (n = 40 per form), and the text complexity indicators and story features had different scales. The story features were categorical, and the text complexity indicators were numerically continuous. For these reasons, I conducted the LLTM models with one set of predictors at a time: one model with a set of text complexity indicators and another model with a set of story features, by which the effects of categorical story features were estimated separately from the effects of numerical text complexity indicators which were centered on grand mean and rescaled. This modeling may restrict the ability to compare the estimated effects of the predictors on the item difficulty.

The LLTM explains the item difficulty by means of item properties with a person effect (θ_p) set random (see Equation 2 in 3.4.1). There is no error term included in the model, and thus the prediction is assumed to explain the item effects perfectly from the item properties (De Boeck & Wilson, 2004;). The LLTM plus error adds an error term to the original LLTM. Like a regression model, the LLTM has all variance explained while the error term added to the LLTM allows for imperfect predictions (De Boeck et al., 2011). Both LLTM and LLTM plus error were conducted for each set of the text features. First, the LLTM was conducted to investigate the effects of each set of the features if one

set were the only properties exclusive to the other. Then, the LLTM plus error was conducted to examine the effects allowing for unexplained variance because there were more item properties that needed to be explored, as well as in acknowledgement that the set might not be exhaustive, and/or not fully predictive.

The **lme4** package in R was used to estimate LLTM and LLTM plus error by means of `glmer` function. The **lme4** package does not provide anchoring functions across forms in estimating the effects and parameters and hence, the common items were not linked across forms in conducting the LLTM models.

Item property effects of text complexity indicators on the item difficulty of MOCCA. The LLTM explored the extent to which text complexity indicators or story features within the set of item properties affected the item difficulty of the MOCCA assessment. Table 8 presents the estimated coefficients of the fixed effects of text complexity indicators from the results of the LLTM analyses by form. The estimated item parameters from the `glmer` function in the **lme4** package represent item easiness, meaning lower values are related to making items more difficult, and higher values are related to making items easier (Desjardins & Bulut, 2018). The directions, magnitudes, and statistical significances of the effects of the item properties on the item difficulty varied according to the forms. Form 3.2 and Form 3.3 presented the similar directions and magnitudes on the effects of the item properties, but Form 3.1 showed somewhat opposite directions.

Table 8

Estimates of the Effects of Text Complexity Indicators

	Form 3.1		Form 3.2		Form 3.3	
	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
LLTM model						
Word count	0.0699*	0.023	-0.1103**	0.026	-0.2062**	0.024
Word familiarity	-0.0423	0.024	0.0908**	0.023	0.0209	0.022
Word concreteness	-0.0206	0.024	0.1380**	0.024	0.1482**	0.024
Type-token ratio	-0.0055	0.021	0.0737*	0.024	-0.0545*	0.021
Narrativity	0.0842**	0.024	-0.0251	0.02	0.0739*	0.024
LLTM plus error model						
Word count	0.0768	0.146	-0.1145	0.237	-0.2169	0.192
Word familiarity	-0.0416	0.154	0.0976	0.208	0.0232	0.178
Word concreteness	-0.0299	0.154	0.1421	0.224	0.1467	0.189
Type-token ratio	-0.0102	0.133	0.0710	0.218	-0.0577	0.167
Narrativity	0.0861	0.157	-0.0260	0.184	0.0792	0.188

Note. * $p < 0.01$, ** $p < 0.001$

The effect of word count was statistically significant at the 0.01 level across all forms, but the direction varied by form. When an item had a text with more words, the item difficulty parameters in Form 3.2 and Form 3.3 were lower, meaning the item was more difficult, while the item difficulty parameter in Form 3.1 was higher, meaning the item was easier. The assumption often is that a longer text, meaning a text that has more words, is more difficult to comprehend (Davey, 1988; Freedle & Kostin, 1993), which was seen in Form 3.2 and Form 3.3. But a longer text was easier in Form 3.1. The findings suggest that considering the mixed effects of text length on the item difficulty, a longer text can give a cognitive workload reduction by providing more information and clues to help a reader construct a situation as in Form 3.1.

When the text had more familiar words, the items in Form 3.2 only were found to be easier at a level of statistical significance. The effect of word concreteness was statistically significant for Form 3.2 and Form 3.3 where the items were easier when the

text had more concrete words. The effect of type-token ratio was statistically significant at the 0.01 level in Form 3.2 and Form 3.3 but not significant in Form 3.1. The directions of the effects were different according to the forms. The items in Form 3.3 were more difficult with the higher type-token ratio, meaning the words were less repeated and more unique, while the items in Form 3.2 were easier with the higher TTR. These three predictors are usually considered to be associated with word difficulty. Words are considered more difficult if they are less frequent, more abstract, and more lexically rich, and such words make comprehension slower and more difficult (e.g., Graesser et al., 2011; Graesser et al., 2004). The results indicated that word difficulty also had a mixed effect on the item difficulty depending on the forms. Unlike the assumption, more familiar and concrete words were found on items classified as more difficult for Form 3.1, and more lexical variation (i.e., higher TTR) was found on items classified as easier for Form 3.2. Restriction of range is likely part of what is taking place in this analysis because of the degree of similarity intentionally designed in to the assessment. Also, the findings need to be interpreted with the context because the effects may vary depending on what types of words addressed the main idea of the text.

For the effect of narrativity based on this data set, when an item had more narrative text, the item for Form 3.1 and Form 3.3 was easier with statistical significance at 0.01 level.

The results from the LLTM modeling suggested that the items tended to be easier when words in the text were more familiar and included more concrete words, vocabulary in the text had little variation, and the text was more story-like. However, the findings suggested that the effects of the text complexity indicators on the item difficulty were

mixed with different significance and varying directions and magnitude depending on the forms. The findings suggest that individual effects of each indicator have different effects depending on other features of the text, and interpreting the results needs to go beyond explaining the effect of a single predictor, at least in this analysis where the items were designed to be similar and therefore had restriction of range designed into the results.

The original LLTM models estimated the effects of the text complexity indicators without residual terms. While these regular models imply perfect prediction based on the regression, LLTM with error models having residual terms added allows for imperfect prediction in the estimation process (Desjardins & Bulut, 2018). As this study included as predictors not only text complexity indicators but story features that might explain the item difficulty of MOCCA, the error term was added to the original LLTM model, allowing for unexplained predictions with room for explanation from other external variables such as story features. In the LLTM with error modeling, results were essentially the same as those of the original LLTM analysis but with substantial error variance remaining.

The results from the LLTM plus error modeling suggested that the item difficulty parameters of the MOCCA assessment could not be solely predicted by means of text complexity indicators including text length, word difficulty, and the features of a narrative text. To better explain how the MOCCA items performed, it may be necessary to consider not only text complexity indicators but other external variables, which might include the story features that are related to specific structures of a story.

Item property effects of story features on the item difficulty of MOCCA. In addition to the LLTM including text complexity indicators as predictors, another LLTM

analysis was conducted to explore the extent to which story features as the item properties were associated with the item difficulty of the MOCCA assessment. As mentioned earlier, due to the number of parameters and different scales, the effects of text complexity and story features were estimated in different LLTM modeling. Again, the LLTM analysis was first conducted with five story features, and the LLTM plus error analysis was conducted allowing for unexplained variance. Table 9 presents the estimates of the fixed effects of story features by means of the LLTM by form. Similar to the LLTM with text complexity predictors, the regular LLTM allowed for perfect explanation of the effects of story features on the item difficulty of MOCCA. The intercept was the effect of the story which was child-centered and realistic when controlled for other predictors, i.e., no second agent appearing in the text, the goal and the main character positioned in the different sentence, the explicit goal located in the early part of the story, and the goal met and positively accepted emotion in the ending. The `glmer` function from the lme4 package in R estimated the parameters, and thus higher values imply easier items, similar to the LLTM with text complexity indicators (Desjardins & Bulut, 2018).

The estimated effects of story features on predicting item difficulty of MOCCA also varied depending on the forms like those of text complexity indicators. The text being child-centered and realistic was a baseline in estimating the effects of the relationship of child-centeredness and being realistic on the item difficulty. When the text as an item was child-centered but not realistic, it was statistically significant only in Form 3.2 and Form 3.3, with the direction of more difficult.

Table 9

Estimates of the Effects of Story Features

	Form 3.1		Form 3.2		Form 3.3	
	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
intercept	1.0228***	0.0841	1.2314***	0.0826	1.1294***	0.0879
Child & Real						
YN	0.1446	0.0855	-0.5133***	0.0900	-0.5471***	0.0933
NY	-0.1684**	0.0556	-0.0415	0.0692	-0.1191*	0.0570
NN	0.0020	0.0587	-0.1107*	0.0536	0.0440	0.0766
SecondAge						
Yes	-0.1677***	0.0491	-0.0403	0.0487	0.0920*	0.0431
Goal & Character						
Same	-0.0707	0.0441	-0.3297***	0.0443	-0.1649***	0.0478
Location & Explicit						
AN	-0.4054***	0.0564	0.2195***	0.0598	-0.1649*	0.0677
BY	-0.2168***	0.0484	0.2141***	0.0481	-0.0342	0.0590
BN	-0.0033	0.0635	-0.1459**	0.0540	-0.1918***	0.0572
GoalMet & Emotion						
MNP	0.5017***	0.0773	-0.0939	0.0866	0.1746*	0.0759
NMP	0.6810***	0.0789	0.1661**	0.0644	0.2086**	0.0706
NMNP	-0.1603**	0.0519	0.0029	0.0438	0.0753	0.0476

Note. * $p < .05$ ** $p < .01$ *** $p < .001$. YN=child-centered & not realistic. NY=not child-centered & realistic. NN=not child-centered & not realistic. AN=not explicit goal in the early sentence. BY=explicit goal in the later sentence. BN=not explicit goal in the later sentence. MN=goal met & emotionally not positive. UP=goal unmet & emotionally positive. UN=goal unmet & emotionally not positive.

The estimated effects of story features on predicting item difficulty of MOCCA also varied depending on the forms like those of text complexity indicators. The text being child-centered and realistic was a baseline in estimating the effects of the relationship of child-centeredness and being realistic on the item difficulty. When the text as an item was child-centered but not realistic, it was statistically significant only in Form 3.2 and Form 3.3, with the direction of more difficult. When the text was not child-centered but realistic, it was more difficult with statistical significance in Form 3.1 and Form 3.3. When the text was not child-centered and not realistic, it was more difficult in

Form 3.2 with statistical significance. Considering the effects with significance on the item difficulty, the text in which a child was not a main character and/or the topic or setting was not realistic made the item more difficult. The results suggested that the text might be more difficult when it was not relevant to background knowledge of the target readers. In addition, most stories on each form were child-centered and realistic (n = 24 in Form 3.1, n = 28 on Form 3.2, and n = 29 on Form 3.3) while only two stories in each form were child-centered but not realistic. The small sample size substantially reduces the power of the interpretations, so these results should be interpreted cautiously.

When the text had a second agent who had an opposing intention to the main character, the item was more difficult with statistical significance only on Form 3.1, but easier on Form 3.3 with statistical significance, compared to the text not having a second agent. The text tends to be more difficult when a reader has to simultaneously follow the intentions of a main character and an opponent, but the results indicated that consideration might be needed to interpret the effect of a second agent in the story as for Form 3.3 where only the presence of a second agent did not seem to have a relationship with making the text difficult.

When the text had a main character and the goal in the same sentence, it was more difficult, with statistical significance on Form 3.2 and Form 3.3, compared to the text having a main character and the goal positioned in different sentences. The results confirm that processing two pieces of information at the same time tends to be related with making comprehension of text more difficult.

The text having the explicit goal located in the first or second sentence was a baseline in estimating the effects of where the explicit or inferable goal was located in the

text. When the text had the explicit goal located in the third or later sentence, it was more difficult with statistical significance on Form 3.1, but easier on Form 3.2 with statistical significance. When the text had the inferable goal implied in the first or second sentence, it was more difficult on Form 3.1 and Form 3.3 with statistical significance, but easier on Form 3.2 with statistical significance. When the text had the inferable goal located in the third or later sentence, it was more difficult with statistical significance on Form 3.2 and Form 3.3. The effects of goal location and explicitness were varying depending on the forms, all negative for Form 3.1 and Form 3.3, but positive except for the inferable goal in the later part for Form 3.2.

The text having the goal met or resolved with positive emotion in the ending was a baseline in estimating the effects of how the goal ended and how a main character emotionally accepted the ending. When the text had the goal met or resolved in the end with non-positive emotional acceptance, it was easier on Form 3.1 and Form 3.3 with statistical significance. When the text had the goal unmet or unresolved in the end but with positive emotional acceptance, it was easier on all forms with statistical significance. When the text had the goal unmet or unresolved in the end with non-positive emotional acceptance, it was more difficult on Form 3.1 with statistical significance. The results suggested that considering the varying effects of the goal met and emotion depending on the forms, consideration may be needed to interpret the results in relation to the other factors.

The results from the LLTM modeling with story features suggested that the effects of each story feature varied depending on the forms which had slightly different average item difficulty, but still considerable restriction of range. The findings implied

that like the results from the LLTM analysis with text complexity indicators as predictors, individual effects of each story feature would have different effects depending on other features that the text possessed given the restricted nature of the assessment, and therefore interpreting the results may need to go beyond explaining the effect of a single predictor, or include more varied material (not already corrected for maximal traits by reading experts, so this might not be a desirable assessment).

Similar to the LLTM analysis with text complexity indicators, this LLTM analysis only estimated the effects of story features with perfect consideration. Thus, the error term was added to the LLTM with story features to allow for unexplained predictions, as had been done with the LLTM with text complexity indicators. Table 10 presents the estimates of the fixed effects of story features by means of LLTM plus error by form. The error term was added to the original LLTM to explore the effects of story features on item difficulty, allowing for imperfect explanation of the estimation. When the effects of story features as predictors were estimated through LLTM plus error, almost no statistical significances were found in the effects of the story features on the item difficulty of MOCCA, but the magnitude and direction of the effects of the predictors remained similar in the LLTM plus error as those of the original LLTM. This result was almost identical with the LLTM and LLTM plus error modeling with text complexity indicators.

The results from the LLTM plus error analysis suggest that like the results from the LLTM plus error with text complexity predictors, the item difficulty of MOCCA could not be sufficiently predicted only with respect to story features for the restriction of range here, including the way to perceive the story, identifying the character, identifying the goal, and the story ending. Understanding the text difficulty in association with

comprehension may be needed to interpret with a combined effect of features from various sources, and more variation in the content of the assessment (again, this might not be desirable for the assessment).

Table 10

Estimates of the Effects of Story Features via LLTM Plus Error

	Form 3.1		Form 3.2		Form 3.3	
	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
intercept	1.0455***	0.1493	1.2492***	0.1378	1.1517***	0.1667
Child & Real						
YN	0.1384	0.2739	-0.5305*	0.2644	-0.5764	0.3106
NY	-0.1811	0.1789	-0.0432	0.2000	-0.1338	0.1879
NN	-0.0101	0.1869	-0.1115	0.1558	0.0354	0.2534
Second agent						
Yes	-0.1702	0.1563	-0.0370	0.1422	0.0939	0.1421
Goal & Character						
Same	-0.0663	0.1407	-0.3346**	0.1288	-0.1632	0.1571
Location & Explicit						
AN	-0.4099*	0.1798	0.2214	0.1726	-0.1684	0.2228
BY	-0.2194	0.1533	0.2128	0.1387	-0.0213	0.1925
BN	-0.0062	0.2027	-0.1506	0.1571	-0.1936	0.1884
GoalMet & Emotion						
MNP	0.5004*	0.2408	-0.1013	0.2507	0.1627	0.2486
NMP	-0.1669	0.1663	0.0069	0.1268	0.0822	0.1569
NMNP	0.6966**	0.2428	0.1688	0.1856	0.2104	0.2297

Note. * $p < .05$ ** $p < .01$ *** $p < .001$. YN=child-centered & not realistic. NY=not child-centered & realistic. NN=not child-centered & not realistic. AN=not explicit goal in the early sentence. BY=explicit goal in the later sentence. BN=not explicit goal in the later sentence. MN=goal met & emotionally not positive. NMP=goal unmet & emotionally positive. NMNP=goal unmet & emotionally not positive.

Estimates of variance parameters by model. Table 11 presents the estimated variance for person by model by form, which specifies the variation among examinees with respect to their latent trait levels. The estimated variance and standard deviation of the person random effect accounted for latent traits to item difficulty. Person variance through Rasch model was 2.72 (*SD* = 1.65) in Form 3.1, 2.83 (*SD* = 1.68) in Form 3.2,

and 3.05 (*SD* = 1.75) in Form 3.3. The variance increased in all forms when the text complexity indicators as the item properties were included in the fixed LLTM, 3.2 (*SD* = 1.79), 4.06 (*SD* = 2.01), and 4.05 (*SD* = 2.01) respectively. The variance was reduced again to the level of Rasch model when error term was added to the LLTM, 2.75 (*SD* = 1.66) in Form 3.1, 2.83 (*SD* = 1.68) in Form 3.2, and 3.07 (*SD* = 1.75) in Form 3.3. When the story features were included in the LLTM, person variance was reduced than that of Rasch model, 2.56 (*SD* = 1.6) in Form 3.1, 2.71 (*SD* = 1.65) in Form 3.2, and 2.89 (*SD* = 1.69) in Form 3.3. The variance was increased to the level of Rasch model when error term was added to the LLTM, 2.7 (*SD* = 1.64) in Form 3.1, 2.82 (*SD* = 1.68) in Form 3.2, and 3.04 (*SD* = 1.74) in Form 3.

Table 11

Estimates of Variance Parameters by Model by Form

	Form 3.1		Form 3.2		Form 3.3	
	σ_θ^2 (sd)	σ_ε^2(sd)	σ_θ^2(sd)	σ_ε^2(sd)	σ_θ^2(sd)	σ_ε^2(sd)
Rasch model	2.715 (1.65)	—	2.827 (1.68)	—	3.05 (1.75)	—
LLTM_text	3.197 (1.79)	—	4.057 (2.01)	—	4.05 (2.01)	—
LLTM_text_error	2.75 (1.66)	.562 (.75)	2.829 (1.68)	1.083(1.04)	3.066 (1.75)	.82 (.91)
LLTM_story	2.563 (1.6)	—	2.713 (1.65)	—	2.887 (1.69)	—
LLTM_story_error	2.7 (1.64)	.127 (.36)	2.816 (1.68)	.095 (.31)	3.037 (1.74)	.129 (.36)

Note. σ_θ^2 indicates person variance. σ_ε^2 indicates the variance of error term

The findings suggested when item properties were added to the Rasch model to explain the effects of the properties on the item difficulty of MOCCA, the variance of

person estimates was affected, depending on types of item properties (text complexity vs. story features) and types of models (LLTM vs. LLTM plus error). Text complexity indicators did not affect the estimates of person variance, but story features explained more item variance in LLTM and LLTM plus error by decreasing person variance estimates.

Comparing model fits. Table 12 presents the comparison of the model fits among the Rasch model, LLTM and LLTM plus error with text complexity indicators, and LLTM and LLTM plus with story features. The results were the same for all three forms.

Table 12

Model Comparisons: Rasch, LLTM, and LLTM Plus Error

	Form 3.1	Form 3.2	Form 3.3
Rasch			
No. of parameters	41	41	41
AIC	18143	19719	19247
BIC	18462	20043	19570
LLTM_text			
No. of parameters	6	6	6
AIC	18736	20232	19707
BIC	18782	20280	19754
LLTM_text_error			
No. of parameters	7	7	7
AIC	18272	19873	19393
BIC	18326	19928	19448
LLTM_ story			
No. of parameters	13	13	13
AIC	18470	19980	19600
BIC	18572	20082	19703
LLTM_ story_error			
No. of parameters	14	14	14
AIC	18219	19788	19326
BIC	18328	19899	19437

For the models with text complexity indicators as the item predictors, the Rasch model had a comparatively better fit by means of AIC (smaller is better) than LLTM and LLTM plus error, but LLTM plus error had a comparatively better fit by means of BIC (smaller is better) than the Rasch model and LLTM. For the models with story features, the results were the same: the Rasch model was comparatively fit better with respect to AIC, but LLTM plus error was fit better with respect to BIC.

Table 13 presents the results of Chi-square tests for three models with each text feature set. The results revealed that LLTM had a statistically significantly better fit than the Rasch model, and LLTM plus error had a statistically better fit than LLTM. The goodness-of-fit indices suggested that no one model was distinctively fit better than other models. The Rasch model was better in terms of AIC, but LLTM plus error was fit better in terms of BIC but without much difference and of Chi-square test.

Table 13

Results of Chi-Square Tests

	Form 3.1	Form 3.2	Form 3.3
Rasch vs. LLTM_text	$\chi^2(35, N=40) = 662.9, p < .001$	$\chi^2(35, N=4\ 0) = 583.3, p < .001$	$\chi^2(35, N=40) = 529.7, p < .001$
Rasch vs. LLTM_text_error	$\chi^2(34, N=40) = 196.6, p < .001$	$\chi^2(34, N=40) = 222.1, p < .001$	$\chi^2(34, N=40) = 213.2, p < .001$
LLTM_text vs. LLTM_text_error	$\chi^2(1, N=40) = 466.31, p < .001$	$\chi^2(1, N=40) = 361.13, p < .001$	$\chi^2(1, N=40) = 316.6, p < .001$
Rasch vs. LLTM_story	$\chi^2(28, N=40) = 383.51, p < .001$	$\chi^2(28, N=40) = 316.82, p < .001$	$\chi^2(28, N=40) = 408.78, p < .001$
Rasch vs. LLTM_story_error	$\chi^2(27, N=40) = 130.2, p < .001$	$\chi^2(27, N=40) = 123, p < .001$	$\chi^2(27, N=40) = 132.9, p < .001$
LLTM_story vs. LLTM_story_error	$\chi^2(1, N=40) = 253.39, p < .001$	$\chi^2(1, N=40) = 193.89, p < .001$	$\chi^2(1, N=40) = 275.94, p < .001$

4.2.2 Latent trait regression analysis and person properties

On the basis of the Rasch model, a person explanatory approach was used by means of the latent regression analysis to examine the effects of person properties on the item responses. The specific research question for the analyses is:

RQ 3. How are person properties associated with young readers' performance on MOCCA for this data set?

To answer Research Question 3, the `glmer` function in the **lme4** package (Bates, Maechler, Bolker, & Walker, 2019) in R was again applied to conduct the latent trait regression analysis. The latent regression Rasch model provided an estimate of the fixed effects of person indicators and the variance of random effect by form. The person properties in the analysis included gender, white (representing race/ethnicity), free or reduced-price meals (FARMS) representing socioeconomic status, special education participation (sped), and English learner status (EL). The sample sizes for each demographic variable by form were described in Table 3 (see 4.1.1). The samples in all demographic variables included missingness which was considered random in this study.

Person property effects on the responses of MOCCA. Table 14 presents the estimated coefficients of the fixed effects of the person characteristics by form. The reference group was composed of those students who were male, non-white, not eligible to receive farms, not qualified to receive special education services, and non-English learners. There were no statistically significant differences between male and female students in all forms. There were no statistically significant differences between white

and non-white students on Form 3.1 and Form 3.2, but white students performed better than non-white students with statistical significance at 0.001 level on Form 3.3. Socioeconomic status, represented as free and reduced-price meals (farms) showed as significant at least at the 0.01 level on all forms. The students eligible for farms performed worse than those who ineligible on all forms. The students qualified to receive special education services performed worse than those who did not qualify for special education services, at the 0.05 level of statistical significance on Form 3.1 and Form 3.2, but with no statistically significant difference on Form 3.3. English learners performed worse than non-English learners at a 0.01 level of statistical significance on Form 3.1 and Form 3.2, but with no statistically significant difference on Form 3.3.

Table 14

Estimates of the Effects via Latent Regression Analysis

	Form 3.1		Form 3.2		Form 3.3	
	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
gender	0.047	0.242	0.107	0.239	0.159	0.241
white	0.172	0.274	0.311	0.266	1.066***	0.255
farms	-0.808**	0.257	-0.988***	0.261	-0.658*	0.259
sped	-1.086*	0.536	-1.300**	0.459	-0.687	0.514
EL	-0.892**	0.344	-0.987**	0.375	-0.517	0.376

Note. * $p < .05$ ** $p < .01$ *** $p < .001$.

The findings confirmed some of the assumptions about the relation between the personal characteristics and the responses of the examinees. The examinees who had higher SES, did not qualify for in special education services, and were non-English learners performed better on the MOCCA assessment than those who had lower SES,

participated in special education, and were English learners. These might be construct relevant differences so this would need to be compared in equivalent samples and with a research design intended to examine these differences, which was not the case here. Also, the sample size needs to be considered in interpreting the results of special education status and English learner status. The number of special education participants was very small (n = 27 on Form 3.1, n = 39 on Form 3.2, and n = 34 on Form 3.3), and the number of ELs was also quite small (n = 47, n = 48, and n = 49, on each form, respectively). Another consideration in interpreting the results is a possibility of double counting the demographic status. For example, the participants who received farms might include English learners or special education participants. Thus, the low-ses participants' lower performance might be partly associated with their poor language proficiency.

Estimates of variance parameters by model. Table 15 provides the estimated variance of the latent trait estimates by model across forms. Each form had larger variance and standard deviation estimated from the Rasch model than those estimated by the latent regression model. The estimated variances of the intercepts were 2.715 (SD = 1.65) for Form 3.1, 2.827 (SD = 1.68) for Form 3.2, and 3.05 (SD = 1.75) for Form 3.3, respectively, based on the Rasch model. However, they decreased to 2.226 (SD = 1.49), 2.438 (SD = 1.56), and 2.287 (SD = 1.51), respectively ,when estimated through the latent regression model.

Table 15

Estimates of Variance Parameters

		Form 3.1	Form 3.2	Form 3.3
Rasch	σ_θ^2	2.715 (1.65)	2.827 (1.68)	3.05 (1.75)
Latent regression	σ_θ^2	2.226 (1.49)	2.438 (1.56)	2.287 (1.51)

Comparing model fits. Table 16 presents the goodness-of-fit indices of the Rasch and latent regression models by form. Considering smaller AIC and BIC are better, the latent regression model improved the fit of the Rasch models on all forms. Table 17 provides the results of chi-square tests. Again, the chi-square fits suggested the latent regression models improved the fit of the Rasch models across the forms. The findings confirmed that the latent person abilities were related to person properties such as SES, special education status, and English language proficiency, although the degree of construct relevance is not possible to detect here due to the design (test impact as compared to test bias).

Table 16

Model Comparisons

	Form 3.1	Form 3.2	Form 3.3
Rasch			
No. of parameters	41	41	41
AIC	18142.9	19718.8	19247.4
BIC	18462.2	20043.3	19570.4
Latent regression			
No. of parameters	46	46	46
AIC	6763.6	7031.8	6867.5
BIC	7078.2	7351.2	7183.3

Table 17

Results of Chi-Square Tests

	Form 3.1	Form 3.2	Form 3.3
Rasch vs. Latent regression	$\chi^2(5, N = 40) = 11389.3, p < .001$	$\chi^2(5, N = 40) = 12697, p < .001$	$\chi^2(5, N = 40) = 12389.9, p < .001$

4.3 Phase III: English learners and Differential Item Functioning

The relationship of person properties to the item responses was explored in RQ 3, and the findings showed English proficiency as well as free lunch and special education participation was significantly associated with the responses in a reading comprehension assessment. Reading comprehension assessment requires the examinees to have a certain level of language proficiency because the examinees are required to answer the questions based on text comprehension. However, a subgroup of the participants such as English learners in this study might respond to some items differently as a group compared to the reference group due to features inherent in the items beyond the construct being measured, which was explored to some degree next in the current study. The specific research question is:

RQ 4-a. Do MOCCA items exhibit Differential Item Functioning (DIF) between English learners and non-English learners?

To explore whether there were any items displaying statistically different properties between ELs and non-ELs, and what unexpected or biased features were involved in the items, if any, I conducted an IRT-based differential item functioning

(DIF) analysis. For the analysis, the data from three forms were collapsed into one dataset to avoid noise from much missingness, indicating that 27% of the participants did not report EL status and their data were not included in DIF analysis in this study. English learner status (variable *EL*) was set up as a facet, and the Rasch model was executed by means of `tam.mml.mfr` function in **TAM** package.

4.3.1 IRT-based differential item functioning

To detect DIF items, I conducted a significance test by dividing the interaction term (i.e., item*EL) by the standard error and presented z-values of each item in each form. The criteria for DIF items were two: one was the z-value larger than |2| and the other was the difficulty difference of 0.5 logits between two groups. As a result, a total of 14 items including one common item (Item 4) out of 100 items across forms were detected to show DIF between ELs and non-ELs. Table 18 presents the items that were flagged as displaying different statistical properties with respect to IRT-based DIF calibrations. The table displays the item difficulty for DIF items, organized by three groups: the whole sample (ALL), non-ELs, and ELs, with respect to the estimated IRT item difficulty (xsi). A higher item difficulty indicates a less probability to answer the item correctly.

The findings suggest the DIF items existed on all forms and were significantly favoring non-ELs over ELs. For example, for Item 11 on Form 3.1, the estimated item difficulty for non-ELs (n = 288) was -1.8124, and the difficulty for ELs (n = 39) was -0.2193 when the average item difficulty for all the participants was -1.0159. The results indicated that non-ELs had higher probability of responding correctly to the given item than ELs as well as average participants, and that ELs had lower probability than

Table 18

DIF Items Across Forms

Form	Item no.	xsi for ALL	xsi for non-EL (*N*)	xsi for EL (*N*)
common	4	-0.0296	-0.8560 (952)	0.7968 (124)
f3.1	11	-1.0159	-1.8124 (288)	-0.2193 (39)
f3.1	18	0.3051	-0.5059 (293)	1.1161 (40)
f3.1	27	0.1722	-0.6173 (293)	0.9618 (43)
f3.1	32	-0.9729	-1.9022 (294)	-0.0436 (43)
f3.1	33	-0.0274	-0.9813 (292)	0.9266 (43)
f3.1	34	-1.2581	-2.1225 (292)	-0.3938 (43)
f3.1	37	-1.2904	-2.1209 (291)	-0.4599 (42)
f3.2	26	-0.0474	-0.9332 (343)	0.8384 (42)
f3.2	27	-0.2472	-1.4069 (340)	0.9125 (41)
f3.2	28	-0.3486	-1.2389 (339)	0.5417 (41)
f3.2	34	-0.1221	-0.9408 (335)	0.6966 (44)
f3.3	30	-0.9034	-1.7560 (326)	-0.0508 (43)
f3.3	35	-0.3613	-1.2575 (327)	0.5348 (43)

Note. xsi means item difficulty estimated for each item.; The item number is not the same as the actual assessment, but the same as in the tables in APPENDIX A and APPENDIX B.

non-ELs and average participants. Interpretations will be further discussed in the next chapter (see section 5.3.3). These results need to be interpreted with caution because the number of the focal group (ELs) was small on average per item (42 ELs, as compared to an average of 312 non-ELs) excluding the common item having 952 non-ELs and 124 ELs. And the proficiency difference among ELs was not considered in the analysis because ELs' proficiency level was not reported.

4.3.2 Reviewing DIF items through text features

DIF analysis only identified the items that function differently between the reference and focal groups and did not explain why these items function differently. Fourteen items in MOCCA were identified as providing different statistical properties between ELs and non-ELs. To explore why the items showed DIF between two groups on a reading comprehension assessment, the items were investigated with respect to the textual features used as the predictors for the LLTM models in Phase II. The research question is connected with the previous research question as following:

RQ 4-b. What is the relationship between textual features presented in the Research Question 2 and the items functioning differently between ELs and non-ELs on the MOCCA assessment?

Table 19 presents the means and standard deviations of text complexity indicators for items showing no-DIF and items showing DIF across forms after collapsing all the forms. To examine whether there were any mean differences between two types of items, independent t-test was conducted by text complexity indicator.

The results indicated that there were no statistically significant mean differences in terms of word count, word familiarity, and word concreteness. But the results showed that there were significant mean differences between no-DIF items and DIF items in terms of type-token ratio, $t(98) = 3.5$, $p < .001$, and narrativity, $t(98) = 2.1$, $p < .05$. The findings imply that the text passages functioning differently between two groups did not favor one group over another with respect to syntactic complexity represented as text

length and word difficulty represented as familiarity and concreteness, but favored non-ELs over ELs with respect to lexical variation represented as TTR and text genre represented as narrativity.

Table 19

Comparison of Text Complexity for DIF and no-DIF Items

	Word count	Word familiarity	Word concreteness	Type-Token ratio	Narrativity
No-DIF items					
mean (*sd*)	82.1 (11.1)	576.5 (13.2)	413.8 (33.9)	0.65 (0.05)	63.0 (22.1)
DIF items					
4	101	586.7	432.9	0.65	90.82
11	86	574.8	400.2	0.54	89.44
18	75	599.4	406.6	0.68	74.86
27	93	595.5	393.2	0.61	71.23
32	92	582.8	394.5	0.62	90.82
33	92	563.7	405.9	0.60	68.79
34	84	581.7	397.7	0.67	93.7
37	90	588.9	345.9	0.57	96.08
26	89	558.9	446.2	0.64	64.8
27	87	578.6	393.2	0.68	62.55
28	84	591.4	424.6	0.63	45.62
34	87	587.0	391.1	0.69	80.78
30	79	586.8	401.5	0.60	46.02
34	95	590.8	389.4	0.58	84.61
mean (*sd*)	88.1 (6.6)	583.4 (11.4)	401.6 (23.4)	0.62 (0.05)	75.7 (16.7)

The results indicated that there were no statistically significant mean differences in terms of word count, word familiarity, and word concreteness. But the results showed that there were significant mean differences between no-DIF items and DIF items in terms of type-token ratio, $t(98) = 3.5$, $p < .001$, and narrativity, $t(98) = 2.1$, $p < .05$. The findings imply that the text passages functioning differently between two groups did not

favor one group over another with respect to syntactic complexity represented as text length and word difficulty represented as familiarity and concreteness, but favored non-ELs over ELs with respect to lexical variation represented as TTR and text genre represented as narrativity.

Next, I investigated whether there were any different characteristics between DIF items and no-DIF items in terms of story features. Table 20 presents the descriptive results of story features for no-DIF items and DIF items. Because the number of DIF items was small, the difference between two types of items was reviewed in terms of the proportion. When the proportion of composition was compared by each story feature between no-DIF items and DIF items, the most items in both DIF and no-DIF were child-centered and realistic (64% vs 67%), had the goal and a main character at the same sentence (33% vs 50%), had an explicit goal in the early sentence (29% vs 38%), and had the goal met with positive emotion (43% vs 59%). But more DIF items (50%) had a secondary agent in the passage than no-DIF items (15%).

To examine whether the association of the DIF classification and story features was statistically significant, the chi-square tests for independence were conducted using SPSS. There were no statistically significant association between DIF identification and whether a passage was child-centered and/or realistic, and whether the goal and a main character appeared at the same sentence, where an explicit or inferable goal was detected, and how the goal ending was emotionally accepted. But there was a statistically significant association between DIF identification and the existence of a secondary agent, $\chi(1)=7.44$, $p < 0.01$; that is, DIF items had more passages having a secondary agent than no-DIF items.

Table 20

Comparison of Story Features for DIF and no-DIF Items

		No DIF (n = 86) n (%)	DIF (n = 14) n (%)
Child & Real	child-centered and realistic	58 (67%)	9 (64%)
	child-centered and not realistic	2 (2%)	1 (7%)
	not child-centered and realistic	13 (15%)	2 (14%)
	not child-centered and not realistic	13 (15%)	2 (14%)
Second agent	Yes	15 (17%)	7 (50%)
	No	71 (83%)	7 (50%)
Goal & Character	Same location	28 (33%)	7 (50%)
	Different location	58 (67%)	7 (50%)
Goal & Explicit	early explicit goal	33 (38%)	4 (29%)
	later explicit goal	25 (29%)	3 (21%)
	early inferable goal	13 (15%)	5 (36%)
	later inferable goal	15 (17%)	2 (14%)
Goal & Emotion	goal met and positive	51 (59%)	6 (43%)
	goal met and not positive	5 (6%)	2 (14%)
	goal unmet and positive	8 (9%)	2 (14%)
	goal unmet and not positive	22 (26%)	4 (29%)

In sum, the items detected DIF were more narrative with more repeated words having a secondary agent in the story than no-DIF items. The results imply that some indicators (TTR and narrativity) were not met with the assumption for which higher TTR and lower narrativity make a passage more difficult to comprehend. The findings are more explored in the Discussion section (see 5.3.3).

CHAPTER V

DISCUSSION AND CONCLUSIONS

The current research intended to validate a reading comprehension assessment by explaining differences in the item difficulty and responses with respect to textual factors and person characteristics. It was expected that textual features would be related to the item difficulty in a reading comprehension assessment that requires the examinees to read a text and make a causal inference to answer questions; however, in this case, the design of the assessment was intended to remove such variations as much as possible (introducing restriction of range). Along the same lines, textual factors if substantially different enough, would affect the difficulty of the text, which would be associated with the item difficulty in the reading assessment. At the same time, it was expected that the characteristics that an individual examinee possesses would have an influence in responding to the items. It was also expected that English learners were likely to perform differently on the reading comprehension assessment than non-ELs. ELs' performance might be related to their language proficiency because the reading assessment demands a certain level of language proficiency. But there might be above and beyond construct-relevant language factors in explaining Els' different performance on some items.

The subsequent sections summarize the findings according to the phase, address the limitations of this study, and discuss how the findings of this study can be interpreted in light of textual features, person properties, and English learner status in a reading comprehension assessment. Finally, this chapter concludes with the implications of the study.

5.1 Summary

This section summarizes the results of the study by phase. Phase I summarizes the statistical and psychometric understanding of the MOCCA assessment, Phase II summarizes the effects of the text and person factors on the parameters of the MOCCA assessment by means of the explanatory approaches, and Phase III summarizes the results of DIF analysis between ELs and non-ELs on the MOCCA items and reviews of the features used to explain the DIF results.

5.1.1 Phase I: Understanding of the MOCCA assessment

An examinee was randomly assigned to one of three forms which were linked and equated with 10 common items and 30 unique items per form. This study analyzed the three forms through descriptive statistical analyses to see how each form was constructed by means of the text factors. The forms did not show any statistically significant difference in terms of text complexity including word count, word familiarity, word concreteness, type-token ratio (TTR), and narrativity. Nor did they provide much difference in terms of the story features. For example, the number of the texts which were realistically child-centered ranged from a low of 24 on Form 3.1 to a high of 28 on Form 3.3. In sum, the forms were quite uniformly constructed text-wise, as an intentional aspect of the design. But the responses by subgroups of the participants showed some differences, as expected, since such differences are already well documented in the research literature. Examinees ineligible for free and reduced-price meals, those not qualified to receive special education services, and non EL students performed better than their counterparts, but there were no statistically significant differences in performance based on gender (male as baseline) or ethnicity (white as baseline).

IRT models were applied to the MOCCA data in an attempt to psychometrically understand how MOCCA worked for the young examinees. The item and person parameters of MOCCA data set were estimated under Rasch model, following model comparisons. The Rasch model estimated the item and person parameters by form with 10 common items linked. The estimation found on average, Form 3.1 was more difficult than Form 3.2 with the highest item difficulty and lowest person ability. Note that the persons were not designed as equivalent groups. The overall difficulty of Form 3.3 was located between the other two forms with no statistically significant difference between the forms. These result were consistent with the Classical Test Theory analysis which calculated the item difficulty by means of raw scores.

5.1.2 Phase II: Explanatory item response modeling and external variables

Explanatory item response modeling (EIRM) including the linear logistic test model (LLTM) and latent regression analysis was conducted with the MOCCA data set in an attempt to explain differences in item difficulty and responses on the assessment. LLTM included textual features such as text complexity indicators and story features as predictors in separate analyses, and latent regression analysis included person properties as predictors.

The effects of textual features on the item difficulty. The text variables were divided into a set of numerical text complexity indicators and another set of categorical story features due to the nature of the features. Their effects were separately estimated by means of the LLTM analyses.

The text complexity indicators as the predictors. The text complexity indicators included word count, word familiarity, word concreteness, type-token ratio (TTR), and

narrativity, all of which were calculated via a text difficulty measure, Coh-Metrix (University of Memphis). Each predictor showed different statistical significance, direction, and magnitude according to the forms.

Word count had a statistically significant effect on all forms, but with different directions. The longer text with more words was more difficult on Form 3.2 and Form 3.3, but easier on Form 3.1. With respect to word familiarity, more familiar words were statistically significantly related to easier text on Form 3.2. With respect to word concreteness, more concrete, non-abstract words made the text easier in Form 3.2 and Form 3.3, and this difference was statistically significant. With respect to Type-token ratio, more unique, less repeated words made the text easier in Form 3.2 but more difficult in Form 3.3, with statistically significant differences. With respect to narrativity, the stories more relevant to everyday activities made the text easier in Form 3.1 and Form 3. 3, again with statistically significant results.

Through CTT item difficulty calculation (see Appendix A) and IRT parameter estimation (see Table 7), Form 3.1 was found to be the most difficult form while Form 3.2 was the easiest form, and the mean item difficulty difference between the two forms was statistically significant. Considering the average form difficulty, more words made the difficult Form 3.1 easier but the easy Form 3.2 more difficult. Easier words (i.e., more familiar, more concrete, and less varying words) made the easy Form 3.2 easier.

Story features as the predictors. The effects of story features on the item difficulty of MOCCA were also estimated by means of the separate LLTM analysis. The features included whether the story had a child as a main character and was accordingly realistic (Child & Real), whether there was a second agent (SecondAge), whether a main

character and the goal appeared in the same sentence (Goal & Character), whether the goal was explicit and where it was located (Location & Explicit), and how the goal ended and how the ending was emotionally accepted (GoalMet & Emotion).

Like the text complexity indicators, story features showed different statistical significance, direction, and magnitude depending on the forms. Compared to the text in which a child was a main character and the story was realistically set in an everyday situation, when the story was child-centered but not realistic, it was more difficult on Form 3.2 and Form 3.3 with statistical significance. When the story was not child-centered but realistic, it was more difficult on all three forms with statistically significant differences for Form 3.1 and Form 3.3. When the story was not child-centered nor realistic, it was more statistically significantly more difficult on Form 3.2.

Compared to the story without a second agent, when the story had a second agent, it was more difficult on Form 3.1 but easier on Form 3.3 with the differences both statistically significant. Compared to the story having a main character and the goal in the different sentences, when the story had both in the same sentence, it was more difficult on all three forms with statistically significant differences for Form 3.2 and Form 3.3. Compared to the text having the explicit goal in the first or second (early) sentence, when the story had the inferable goal in the early sentence, it was statistically significantly more difficult in Form 3.1 and Form 3.3 but statistically significantly easier in Form 3.2. When the story had the explicit goal in the later (i.e., third or later) sentence, it was statistically significantly more difficult on Form 3.1 but statistically significantly easier on Form 3.2.

When the story had the inferable goal in the later sentence, it was more difficult on all three forms with the difference reaching statistical significance for Form 3.2 and Form 3.3. Finally, compared to the story having the goal met in the end and a main character feel positively about the ending, when the story had the goal met with not positive feeling in the ending, it was statistically significantly easier on Form 3.1 and Form 3.3, and when the story had the goal not met but with positive feeling in the ending, it was statistically significantly easier on all forms. However, when the story had the goal not met with not positive feeling in the ending, it was statistically significantly more difficult on Form 3.1.

When the difficult Form 3.1 was compared to the easy Form 3.2, both forms were more difficult if the stories in a given form were not child-centered and/or not plausible, having a second agent, and having the goal and main character in the same sentence. But Form 3.1 was more difficult and Form 3.2 was easier when the stories in a given form had an explicit goal in the later sentence or an inferable goal in the early sentence.

The results imply that the relations between text complexity indicators and story features with item difficulty was varied within forms and across forms. Especially when a difficult Form 3.1 was compared to an easy Form 3.2, some interesting results were found. The findings including both text complexity and story structure are discussed with detail in the Discussion section. However, when error terms were added to each of the LLTM models to allow for unexplained variance, no statistically significant relation between any text property and item difficulty was found. In model comparisons, the Rasch model had a comparatively better fit than LLTM and LLTM plus error models on

all forms, where LLTM plus error models were fitted better than LLTM (see Table 12 and Table 13).

Table 21 summarizes the direction and statistical significance of the relationships of the textual factors on the item difficulty by form without the coefficients to understand how each value under each feature worked on the MOCCA items. In the table, *easy* indicates the given feature made items in a given form easier, and *difficult* indicates the feature made items more difficult in a given form with asterisk representing significance.

The effects of person properties on the responses. The results of the latent regression analysis provided information about how each person factor was related to the responses on the MOCCA items. Gender (i.e., female vs. male) was not statistically significant in any form. Race was significant on Form 3.3, meaning white test takers responded correctly to more of the items on Form 3.3. SES and special education status were statistically significant on all forms, meaning students eligible for free or reduced-price meals or participating in special education responded more incorrectly to the items on all forms than their non-FARM and non-SpEd peers. EL status was also a statistically significant predictor in the analysis, with ELs performing worse on all forms compared to non-Els, and the difference statistically significant for Form 3.1 and Form 3.2. The directions and significance of the relations from the latent regression analysis were almost the same as those from the descriptive statistics by subgroups except for whiteness for Form 3.1 and Form 3.2 and EL status for Form 3.3.

Table 21

Descriptive Summary of Text Features on Item Difficulty of MOCCA

		Form 3.1 (n = 490)	Form 3.2 (n = 549)	Form 3.3 (n = 530)
Mean of item difficulty (CTT)		63.9%	67.4%	65.7%
Mean of item difficulty (IRT)		-0.856	-0.928	-0.901
Text complexity indicators				
Word count		easy*	difficult* *	difficult**
Word familiarity		difficult	easy**	easy
Word concreteness		difficult	easy**	easy**
TTR		difficult	easy*	difficult*
Narrativity		easy**	difficult	easy*
Story features				
Child & Real	Child & Real	(intercept)	(intercept)	(intercept)
	Child & Not real	easy	difficult***	difficult***
	No child & Real	difficult**	difficult	difficult*
	No child & Not real	easy	difficult*	easy
SecondAge	Yes	difficult***	difficult	easy*
Goal & Character	Same	difficult	difficult***	difficult***
Location & Explicit	Early & Implicit	difficult***	easy***	difficult*
	Later & Explicit	difficult***	easy***	difficult
	Later & Implicit	difficult	difficult**	difficult***
GoalMet & Emotion	Met & Not positive	easy***	difficult	easy*
	Not met & Positive	easy***	easy**	difficult**
	Not met & Not positive	difficult**	easy	easy

Note. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.1.3 Phase III: English learners and differentially functioning items

To explore the extent to which English language proficiency was related to the MOCCA items and whether there were items functioning differently for a subgroup with limited English proficiency, a follow-up IRT-based DIF analysis was conducted after

collapsing all the items across forms with the samples having EL status information (i.e., missing data were excluded). This analysis identified 14 out of 100 items including one common item that showed DIF between ELs and non-ELs. These 14 items favored non-ELs at a statistically significant level than ELs.

The characteristics of these DIF items were explored afterwards with respect to the text complexity indicators and the story features that were used as external variables in the linear logistic test modeling analysis to answer Research Question 2. Compared to no-DIF items, DIF items had a secondary agent in more narrative stories with less lexical variation with a statistical significance.

5.2 Limitations

This study has several limitations. First, the length of the text passages used in this study was short, ranging from 52 words to 116 words per text including all seven sentences without a title. Coh-Metrix, a tool used to measure the text complexity, can reliably analyze texts that range from 200 to 1000 words (Nelson et al., 2012). Thus, the results of the Coh-Metrix analysis for MOCCA text with fewer than 200 words may be less reliable than is ideal. Thus, a caution may be needed to interpret the inferences to the indicators of MOCCA texts measured by Coh-Metrix.

Second, I was unable to include all the text factors (five numeric text complexity indicators and five categorical story features) in a single linear logistic test model due to the nature of the variables. Including all the text factors in one LLTM as predictors caused convergence issues as reported in the Methods section. The text indicators needed centering, as they were numeric, while the story features did not need centering, as they were categorical. Ultimately, I divided the text factors into two sets of predictors by

variable type, and conducted the LLTM analyses separately, one model with text complexity indicators and another with story features, without convergence issues. Thus, this condition made it difficult to directly compare the effects of all the text features used in this study and impossible to detect which text features had the largest effect among the text factors including text complexity indicators and story features. Future studies incorporating fewer but strongly predictable features of both linguistic features and story structures should be conducted to explore which features are more strongly associated with item difficulty in reading assessments.

Third, another limitation put on comparing and generalizing the effects of the text may be related to the characteristics of the text passages. The text passages used in this dissertation study had the syntactic structures and words that did not differentiate with each other due to the restrictions on the length and difficulty set by item design specifications. Consequently, selecting the types of the textual factors was limited for text analysis, and the values of the selected factors did not show much variation. The descriptive statistics revealed many stories dominantly belonged to concentrated values under each story feature. For example, out of 100 stories across all forms, 67 stories were child-centered and realistic while only 3 stories were child-centered but not realistic. To sum, when variables were factored in, great care must be taken in interpreting the findings from the analyses.

Fourth, this study included an examination of the relations of text to the item difficulty on the assessment to focus on the text effects. Texts on the MOCCA assessment are restricted through specifications related to text genre, text length, vocabulary range, and syntactic range. However, the types of questions (e.g., multiple

choice vs short answers), or language of questions are associated with item difficulty on assessments (Alderson, 2005; Freedle & Kostin, 1993). MOCCA was developed to diagnose the nature of poor reading comprehension by means of students' propensity to select incorrect responses. Accordingly, each MOCCA item consists of a text passage with one correct response, one paraphrase incorrect response, and one elaboration incorrect response. I did not include the effects of those incorrect choices on the responses, but rather limited my study to an examination of the text features. However, considering the nature of assessment, it may be limited to explaining the different item difficulty only with respect to text factors.

Fifth, the IRT calibration and estimation of the parameters in this study may not be consistent with those of the MOCCA team. MOCCA is administered for Grades 3 to 5 with three forms per grade, in which the common items across all forms and across all grades are anchored with the parameters estimated for the fourth-grade data. However, this dissertation study analyzed the data set for Grade 3 for which the item parameters were estimated freely with the common items anchored and fixed across the forms with the item difficulty estimated at Form 3.1. The common items were anchored to have the same item difficulty across forms when Rasch modeling and DIF were conducted, but were not anchored when the explanatory models were conducted because the **lme4** package in R did not provide anchoring functions. Thus, IRT estimation of the parameters for this study should not be interpreted in relation to the previously-published MOCCA data results.

Sixth and last, the interpretation of DIF analysis and related analysis is limited to the sample size and data analytic methods. The number of EL participants in the sample

was small in this study, raising concerns about the ability to appropriately interpret the DIF analysis by means of IRT methods. IRT-based DIF requires more than 200 participants for each control group (i.e., non-ELs) and reference group (i.e., ELs) to be effective (Clauser & Mazor, 1998). However, the size of EL participants in this data excluding a common item was 42 on average ranging from 39 to 43 compared to an average of 312 non-ELs. One common item was flagged as showing DIF with 124 ELs and 952 non-ELs in the sample. To make appropriate uses of DIF items, it is recommended to conduct purification (Clauser & Mazor, 1998) or review the items by content experts. However, such further analyses were not conducted in the current study except for comparing the features of the items with respect to the text features used in Phase II. And the methods by which ELs were classified was unknown in this study, and only whether to be an EL was reported without proficiency level. When classifying ELs, Home Language Surveys and English Proficiency Tests are commonly used, but both have validity and reliability concerns which might impact the interpretation and application of the results, and the methods for classification vary by state, with different standards and cut-scores. (Abedi, 2008). The MOCCA demographic data about EL status were gathered from schools without reporting on the way of classification and varying English proficiency levels among ELs. As such, conclusions regarding the DIF analysis and the causes cannot be generalized beyond the context on this dissertation study.

5.3 Discussion

The current study explored how responses to the items of a reading comprehension assessment were affected by both item and person properties based on the findings from a descriptive measurement approach. The study also investigated how

English learners responded to the items on a reading comprehension assessment when they were measured on the same scale with non-English learners and what made them respond differently to items than their EL peers. The subsequent sections discuss the findings, organized by the research questions under the phases.

5.3.1 Phase I: Statistical and psychometric understanding of MOCCA

Descriptive analysis, classical test theory (CTT) item analysis, and Rasch modeling were conducted to understand how MOCCA performed as a reading comprehension assessment. The three forms within the grade were not statistically significantly different from each other in text construction with respect to text complexity and story structure. However, the subgroups showed significant mean differences in terms of race/ethnicity, SES status, special education participation, and EL status. Through CTT item difficulty analysis, on average Form 3.1 was the most difficult while Form 3.2 was the easiest of the three forms, and the differences were statistically significant.

The Rasch model was selected to estimate the item and person parameters of the MOCCA data psychometrically. The results showed that MOCCA was a reliable measure with a high reliability (0.9 or higher), and the estimated item difficulty showed the same results as CTT item analysis, with Form 3.1 the most difficult and Form 3.2 the easiest by means of the mean item difficulty estimation. The Rasch model was subsequently used as a baseline for further analysis such as the explanatory item response modeling and differential item functioning analyses, considering parsimony although some comparative model fits slightly preferred 2PL model. MOCCA is a multiple-choice measure in which a guessing effect may be embedded, and further psychometric-loaded analysis may

provide different psychometric information for the full MOCCA data. This study used partial data from the third grade and proceeded with the decision of Rasch modeling following the results of model-fit comparison.

An interesting finding summing up the results from all the analyses for Phase I was that the average item difficulty as a form difficulty was slightly different between the forms although 10 common items were linked to account for the uniform form difficulty, and although the forms had similar text constructions. The mean item difficulty of Form 3.1 was more difficult than that of Form 3.2 and this difference was statistically significant under both classical test theory and Rasch modeling, while the mean item difficulty of Form 3.3 was located in between the other two forms, but was not statistically significantly different than either Form 3.1 or 3.2. This difference in mean item difficulty may come from the difference in participants even though a form was randomly assigned. It might also come from differences in items, in that texts on one form might have different item properties than another form's texts. An explanatory approach was completed to explain the differences that were found through descriptive Rasch modeling. The meaning of this finding is discussed in connection with the findings of Phase II and Phase III in the next section.

5.3.2 Phase II: EIRM and external variables

Explanatory item response modeling (EIRM) was conducted to explain why some items were more difficult than others and how subgroups of participants showed different responses on the MOCCA assessment beyond the estimation of item difficulty and person ability from the Rasch model. This section discusses the findings about the effects of those properties on MOCCA performance.

The effects of text factors as item properties on the item difficulty. Text plays an important role in reading comprehension assessments because questions on the assessment are strongly related to the text that a test taker reads during testing. The test taker's ability to correctly answer a given item depends on the extent to which the test taker comprehends the text. Thus, the item difficulty of the reading assessment may be associated with the text difficulty. This dissertation study investigated the text factors associated with text difficulty that might be associated with the item difficulty on the reading comprehension assessment.

I explored the relations between the item properties and item difficulty of the MOCCA assessment through the linear logistic test model as an explanatory approach. The LLTM modeling included text factors as the predictors such as text complexity indicators and story features. As explained in the previous section (see section 4.2.1), due to scaling and convergence issues, the factors were divided into one set of text complexity and one set of story features, and entered into the LLTM modeling separately. This method solved the technical problems in conducting modeling, but put some limitations on interpreting the results because it was not able to select a best-fitting model among a variation of LLTM models, and two sets of text factors could not be compared concurrently.

First, LLTM modeling was conducted with the text complexity indicators such as syntactic complexity represented as word count, word difficulty represented as word familiarity, word concreteness, and type-token ratio (TTR), and the text structure represented as narrativity. The assumptions were that longer text having more words would make the item on the reading assessment more difficult (Freedle & Kostin, 1991,

1992, 1993); more familiar and concrete words would make the text easier to comprehend and consequently the item easier on the assessment (Cain, Oakhill, & Bryant, 2004; Gilhooly & Logie, 1980); higher TTR, which implies the text has rich vocabulary because the words are less repeated with a unique meaning, would make the text more difficult to comprehend and consequently the item more difficult on the assessment (Graesser, McNamara, Louwerse, & Cai, 2004); a story narrated with more familiar topics in everyday language would be easier to comprehend, making the items easier (Carlson et al., 2014; McNamara et al., 2011).

The findings were that the directions and significances of the effects of the features were not consistent across the forms with or without statistical significance (see Table 21). Some of the assumptions were met and others were not, depending on the forms. As reported in the descriptive analysis, Form 3.1 was the most difficult and Form 3.2 was the easiest of the three forms in terms of CTT and IRT mean item difficulty, and the difference was statistically significant. The mean difficulty of Form 3.3 fell in between both forms, without statistical significance.

With respect to syntactic complexity represented as word count, a text passage in Form 3.2 and Form 3.3 was more difficult as assumed, but it was easier in Form 3.1. This finding was interesting in that a longer text was estimated to be easier in a difficult form, Form 3.1. This finding implies that more words make a text passage difficult to comprehend with more cognitive workload such as decoding and information construction, but they may provide more information and clues to readers to build a coherent understanding of a text, as in Form 3.1.

With respect to word difficulty represented as word familiarity and word concreteness, the assumptions were met in Form 3.2 and Form 3.3 with statistical significance. Although the effects were not significant, however, it may be meaningful to point that a text passage with more familiar and concrete words in the difficult Form 3.1 was more difficult. In terms of type-token ration (TTR), the assumption (i.e., higher TTR is more difficult) was met in Form 3.3 but not in Form 3.2. Higher TTR implies that the text has rich vocabulary, which gives cognitive workload for readers to decode, but the mixed effects also need to be noted.

With respect to text structure represented as narrativity, the assumption was met in Form 3.1 and Form 3.3, where a text passage with higher narrativity was easier to comprehend. Although it was not statistically significant, however, a text passage with higher narrativity in the easiest Form 3.2 was more difficult.

The results suggest that the relations between item difficulty and text features is relative rather than absolute, implying one text feature does not have a consistent effect on item difficulty across forms. The text features expected to make text easier made text easier in the easiest Form 3.2, but made text more difficult in the most difficult Form 3.1, except for the narrativity effect. The findings suggest that though readers tend to struggle with comprehending a long text containing more words, sometimes readers can gather more information from a text with more words to build a mental representation rather than having more cognitive load to retrieve meanings from the words. This suggestion may explain why longer texts made the items easier on the most difficult Form 3.1.

The findings about word difficulty also suggest that text features may need to be interpreted in association with other factors in understanding their relation with

differences in item difficulty. Words are difficult when they have little frequency (i.e., less familiar), little concreteness, and higher lexical variation, and tend to make a reader struggle with processing comprehension (Feng et al., 2011; Gilhooly & Logie, 1980; Graesser, McNamara, Louwerse, & Cai, 2004). The effects of the word difficulty including familiarity, concreteness, and richness (TTR) may be better interpreted interconnectedly rather than individually. A reader can struggle with comprehending a text passage if a main idea or goal is delivered in one or a few difficult, (i.e., less familiar or less concrete), words (Freedle & Kostin, 1993). In this regard, the important information such as a main character's goal might be addressed in a few less familiar, less concrete, and unique words in the difficult Form 3.1. This finding also implies that word knowledge alone cannot ensure full text comprehension even at the surface level from the text base, and that more than word knowledge such as syntactic complexity knowledge, knowledge of story structure, and appropriate background knowledge may be required to integrate information from the text for understanding and to construct an appropriate situation model.

The finding about narrativity was interesting because more story-like texts with familiar everyday topics were more difficult in the easiest Form 3.2, but easier in the more difficult Forms 3.1 and 3.3. The topic familiarity in narrativity includes word familiarity, background knowledge, and language feature, and higher narrativity indicates the topic is more familiar to the reader with more familiar words and more acceptable background knowledge. But there is a concern about the extent to which a topic is familiar to a reader. For example, a story in Form 3.2 had one item with high narrativity (90%) but a low CTT item difficulty as 59%, meaning only 59% of the respondents

answered this item correctly, while the mean item difficulty of the form was 67.4%. The story was about a boy going to a "barber" shop to have a haircut. In this story, the difficulty might be related to background knowledge rather than narrativity. As the world is changing, the words used to signify the world are also changing, dying, or emerging. I wonder how familiar the word "barber" and the context "barber shop" would be to young readers in today's world. The rating of familiarity measured by Coh-Metrix is about how familiar a word or context seems to an adult (Dowell et al., 2016; Graesser et al., 2011), which implies higher narrative stories are familiar to adults, but may not be familiar to children, as is exemplified by the barber shop story.

When an error term was added to the original LLTM modeling which allowed for perfect prediction only with text complexity indicators, the statistical significance of the predictors disappeared. And LLTM plus error was a better-fitting model compared to the original LLTM. This result implies that only text complexity indicators could not appropriately explain the different item difficulty of MOCCA, and that other features such as story structure features might explain the differences of item difficulty that the first LLTM could not explain.

The effects of the features of story structure were explored in a separate LLTM analysis in terms of the way to accept the world in the story, the presence of an antagonist, the positioning of a main character and the goal, the identification of the goal, the goal fulfilment, and emotional acceptance of the goal ending. The findings of the story features were interesting because the directions and significance of the features were varied depending on the forms, which was similar to the analyses with text complexity indicators. The assumptions were that a text passage would be easier to

comprehend if it was about a child in the realistic context, if there was no second agent in the text, if the goal and the main character were addressed one after another in the different sentences, if the explicit goal was introduced in the early sentence of the story, and if the goal was achieved or resolved, and a main character accepted the goal ending with positive feeling. However, the effects of the features were sometimes inconsistent with the assumptions across the forms and within the forms (see Table 21).

Whether the story was about a child and/or realistic is related to the way that a main character perceives the context, much associated with background knowledge. If a story realistically centers on a child's perspective, it implies that the topics, actions, events, or settings are pertinent to the way of the young readers' perceiving the world. The child-centered but not realistic texts were more difficult in the easy Form 3.2 (n = 2) and Form 3.3 (n = 2) compared to child-centered and realistic texts in both forms whereas they were easier in the difficult Form 3.1 (n = 3) but these differences were not statistically significant. The results were interesting because two items in each form were common items linked across forms. Considering the small number of samples for some features, a careful interpretation is needed to ascertain the relationship of the story features with the item difficulty. The stories in which a main character was not a child and the context was not realistic were easier in the difficult Form 3.1 (n = 6) but more difficult in the easy Form 3.2 (n = 6) with no overlapping stories. Looking into the contents of the stories carefully, most of the text passages under this category had anthropomorphized animals who could not be identified as a child as the main character. Among six texts in Form 3.1 coded as not child-centered and not realistic, two passages with higher CTT item difficulty than the form mean (i.e., easier items than average) were

explored further, and turned out to have anthropomorphic animals as a main character (a frog and a dragon) coded as not children because no evident identity was found to classify them as a child. The finding can be explained with respect to familiarity. Stories of frogs and dragons may be familiar to children although the frog and dragon characters were not described as children in a fantastic world. A frog wanted to fly, and friends helped him to have wings. A frightening dragon was defeated by a knight. Such narrative structures are familiar to young readers as test takers because many story books have such creatures as protagonists or antagonists (e.g., The Frog Prince). The young readers who answered the stories during testing might not have difficulty with constructing an appropriate situation model in the context regardless of who the main characters were and what the context was like, such as the stories in Form 3.1.

Text passages having no second agent were easier in Form 3.1 and Form 3.2, but more difficult in Form 3.3 compared to the text having a second agent whose intention was conflicting with a main character's. Among the stories without a second agent in Form 3.3, the most difficult one was about a girl waiting for her dad at the airport. It may be worth being aware of the structure of the sentences in which the first subject was 'Virginia,' the second subject was 'He,' the third subject was 'Virginia,' the fourth subject was again 'He,' and the fifth and seventh subjects were 'Virginia.' The alternate subjects by sentence might confuse the young readers about who was the main character and whether both Virginia and 'he' (i.e., Virginia's dad) were conflicting with each other to generate an appropriate situation. This ambiguity might make the story more difficult to comprehend even though there was no conflicting interest between characters.

The text having the goal and a main character in the same sentence was more difficult in all forms compared to the text having both in different sentences. The finding was interesting because the test specifications of MOCCA intended to have a main character introduced in the first sentence and the goal introduced in the second sentence (see Section 3.1.2), but about 33% of the items in each form had both the main character and the goal in the same sentence. And as assumed, these stories having a main character and the goal at the same sentence were more difficult compared to the stories having them in different locations. This result suggests that having both information at the same location might give test takers more cognitive workload or confusion in processing the identity of the goal and main character at the same time

Understanding whether the goal is explicit or inferable and where such goal is located is related to identifying the main idea of the story. The goal explicitly introduced in the early sentence (i.e., the first or second sentence) will help a reader to follow the events or actions in the story with obvious expectations and to comprehend the text more quickly (Freedle & Kostin, 1993; Graesser et al., 1994). With respect to the goal location and its explicitness, the effects showed different directions and significance of the estimated coefficients across the forms. A text passage having an inferable goal in the third or later sentence was difficult in all forms, and this finding confirmed the assumption. However, texts having an inferable goal in the first or second sentence (coded as AN) and having an explicit goal in the third or later sentence (coded as BY) were more difficult in Form 3.1 and Form 3.3, but easier in Form 3.2 compared to the text having the explicit goal in the first or second sentence. One story, coded as AN and easier than average item difficulty in Form 3.2, was about a driver and a car out of gas.

Though the goal was not clearly introduced in the early sentence, the title of the story was 'Out of Gas', which might prepare a reader for the development of the story. And the context about a car out of gas might be common, familiar to the readers, who might have no difficulty with constructing an appropriate situation with relevant background knowledge to the context. Another story, coded as BY and easier than average item difficulty in Form 3.2, was about a crab's moving with the title of 'Moving Day.' The story had an anthropomorphic animal as a main character, and the straight goal was introduced in the fifth sentence. But the title also might provide a strong clue about what the crab as a main character needed to do in addition to some textual clues in the first sentence ('home') and the second sentence ('find'). These clues might make a reader prepared for what would be expected in the story. These devices such as title and textual clues will help readers comprehend stories better despite the delayed appearance of the goal in the text.

Understanding the emotional reaction by a main character in the story is significant for understanding organization of the story leading to generating global goal coherence (Graesser et al., 1994). The results of how the goal ending was emotionally accepted revealed inconsistency between the assumptions and the estimation across forms and within forms. In interpreting the results, the number of the sample needs to be considered: the number of text coded as MNP (i.e., the goal met but not positive emotion accepted in the ending) was 3 items for Form 3.1, 2 items for Form 3.2, and 4 items for Form 3.3; the number of text coded as NMP (i.e., the goal unmet but positive emotion in the ending) was 3 items, 5 items, and 4 items respectively. Considering the small number of samples, a careful interpretation is needed to ascertain the effects of the two categories

under the feature on the item difficulty. But it will be meaningful to examine the effect of the features coded as such because the items coded as MNP and NMP were easier in Form 3.1 not as assumed. One story coded as MNP in Form 3.1 was about the same frog who wanted to fly. The story was easier with 74% of CTT item difficulty compared to average (about 64%) though the story ended with not positive emotion though the frog's goal was fulfilled, the main character was not a child, and the story was not realistic. Another story coded as NMP in Form 3.1 was also easier with 79% of CTT item difficulty though the story ended with the goal unfulfilled. In the second story, the main character's goal was not to be greeted on school bus because she was new. That her goal was not met meant someone talked to her and they became friends, and such ending gave a positive feeling to not only the main character but the readers.

The current study included text complexity indicators and story features as item properties to explain the differences of the item parameters of the MOCCA assessment. Many findings suggest that holistic interpretation of the effects of multiple text features might provide more appropriate explanation of the variation of the item difficulty on the reading comprehension assessment. However, because text complexity indicators and story features were not included simultaneously at one model, it was not able to interpret which feature might be more plausible in explaining the findings, and it needed some inferences to consider the effects of both types of predictors. Moreover, the findings of different effects of a same feature on the item difficulty also suggest that there might exist the effects of other features in the text that were not included in this study. Further study can include the predictors above and beyond the text difficulty and story structure features in addition to the features found meaningful in this study, and explore the

relationship between those features and parameters on MOCCA or other reading comprehension assessment.

The effects of person factors on the responses. The latent regression analysis made it possible to analyze the group differences in the responses on the MOCCA items depending on the group properties. The findings of the effects of person properties on the responses confirmed that non-white students, students with low SES, participants in special education, and ELs performed worse than their counterparts. My study is insufficient to make claims about the reasons for the differences in performance between students from these groups. Performance differences may be related to differences in accessibility to environmental factors that help promote learning, learning disabilities, or English language proficiency. However, the sample size needs to be considered in interpreting my results. The extant data set I used for this study included a substantial amount of missing demographic information, with the exception of gender, and the way in which EL status was classified during data collection was not clearly reported. And considering the analytic methods, there were chances that the counts in the study were overlapped across the parameters. The participants who received farms consisted of those who received farms regardless of gender, race/ethnicity, sped participants, and EL status. Thus, the group's performance might be affected by other demographic factors such as English language proficiency or by combined effects of more than one demographic effects. Further study is needed to understand more specifically how each person factor is associated with responses on the assessment. In the next phase, I discuss how ELs performed on the MOCCA assessment compared to non-ELs, focusing specifically on DIF.

5.3.3 Phase III: ELs and DIF items

English learners have idiosyncratic characteristics attributed to their variety in languages, cultures, and education backgrounds. Prior studies have attributed EL students' poor performance on standardized comprehension assessments to their limited English proficiency (e.g., Abedi, 2002, 2004; Solano-Flores, Sexton, & Navarrete, 2001). The argument is made that one should expect ELs to perform poorly on a reading comprehension assessment which is heavily loaded with a non-native language. However, as the findings from the linear logistic test model analysis revealed in this study, test takers' performance was associated with not only linguistic difficulty of the text but understanding of the story structures such as topic familiarity, background knowledge, and the goal identification. If a certain story structure is identified with having association with the different responses of ELs on the MOCCA items than non ELs, the structure may be related to the construct of reading comprehension assessment that ELs need to obtain but fail to achieve to the level with poor ability. Or the structure may be not related to the construct that is required to comprehend the passages in the assessment. If the latter case, the ability may be an unintended second trait that the assessments is measuring and as such should be considered as measurement error.

I examined whether there were items that ELs performed differently than non-ELs on the MOCCA assessment by means of IRT-based DIF analysis. The methods for DIF analysis The flag criteria between two groups in a given item were z-value and 0.5 logits of difference, which were substantial values for detecting DIF. Accordingly, a total of 14 different items including one common item across from three forms were detected as having DIF between the two groups, implying the items favored no-ELs than ELs.

Then, I reviewed the items showing DIF in terms of text features that might be associated with such different responses, and explored whether the features were related to the construct or a secondary trait in comprehending a passage. The text factors used to evaluate DIF items included the same text complexity indicators and the story features used as the predictors in the LLTM analyses under Phase II. Among the text complexity indicators, there were no significant differences between DIF items and no-DIF items in terms of word count, word familiarity and word concreteness, but there were significant differences in terms of Type-Token Ratio and narrativity. Considering the assumptions in Phase II, the findings were interesting because the DIF items had less lexical variation and more narrative than no-DIF items in opposition to the assumptions. It can be understood that though the words were not repeated with lower TTR, the kinds of words used in the passage might be related to the DIF items. Similarly, though the passage was more narrative with everyday topics in common language, how the story was constructed, for example, where and how the goal was introduced, might be associated with the DIF items. Overall understanding is that the MOCCA items did not function significantly differently between non-ELs and ELs in terms of linguistic complexity, and English learners were not disadvantaged in comprehending the text answering the items because of overly difficult words or sentences. This will be a part of validity evidence of MOCCA assessment. Nevertheless, further studies need to evaluate the linguistic and semantic features of the DIF items that will provide more insight into whether the DIF items are related to the ability of ELs' performance on the constructs or bias not intended to measure on MOCCA.

When I reviewed the DIF items with respect to the story features, the composition of DIF items by category under each story feature was not significantly different than that of no-DIF items except for the existence of a secondary agent. There was significant association between EL status and the existence of a secondary agent, meaning more DIF items had a secondary agent in the story than no-DIF items. The assumption was that a passage is more difficult to comprehend when there exists a secondary agent who has a conflicting interest with a main character. The result suggests that ELs had more difficulty with answering correctly an item having conflicts between characters. The results also suggest that any MOCCA items did not favor one group over the other in terms of the composition of the story, and MOCCA items function validly between ELs and non-ELs.

It is hard to generalize the results because of the limitations on the analyses. I need to point out the limitation on interpreting the results of DIF analysis because the sample size was small for the reference group (42 on average), and warranting for the 14 items as DIF because I did not conduct further analyses for implementation decisions such as purification (Clauser & Mazor, 1998). And it may need to compare the performance of poor comprehenders and ELs to examine whether items on a reading comprehension assessment are more associated with the construct that the assessment intends to measure because poor comprehenders are likely to struggle with comprehending passages having the features that showed statistical differences between ELs and non-ELs. It is also possible that other features not used in this study such as cultural background knowledge will offer more explanatory power of the DIF items on the MOCCA assessment. The results from the further analysis may confirm whether the

DIF items in this study are more associated with the construct that MOCCA intends to measure or a bias irrelevant for the construct. Thus, future analyses should include a larger sample with diverse features. Incorporating these design considerations could provide better understanding of the response patterns of ELs, and the findings could better contribute to ensuring that the MOCCA is a valid and fair assessment of inference making in reading comprehension for all test takers.

5.4 Implications

The current study intended to explore how text construction in designing a reading comprehension assessment was associated with test performance on the assessment. The results from the study using the explanatory approach may serve to increase our attention to the effect of item design on how to interpret test scores and make inferences from scores. There are several important implications that emerged from the current study. The first implication is the application of the explanatory approach to support the needs of developing and validating a reading comprehension assessment. When IRT estimates the parameters for persons and items, Explanatory Item Response Modeling complements the measurement information to model the responses and the items by a range of item properties and person properties and to explain variation across persons, items, or both. (De Boeck, Cho, & Wilson, 2017; Desjardins & Bulut, 2018; Wilson & Moore, 2012). And because the explanatory item response modeling is model-based, a better-fitting model can be selected to estimate the parameters of the assessment data.

EIRM in this study provided some positive findings to improve the understanding of the effects of text factors such as text length, word difficulty, and topic familiarity on

the variation of item difficulty, and the understanding of the mixed effects of one factor intertwined with other factors. It also provided findings about the different responses on the reading comprehension assessment depending on person characteristics. No findings or the unexpected findings from EIRM of this study would lead to investigate further alternatives by including different properties or excluding non-significant properties. This way, the explanatory approach can be further applied to test design and validation by appropriately constructing items controlling for application of text factors. MOCCA is intended to not only identify whether a reader is a good or poor comprehender but also to diagnose the type of a poor comprehender a person is (paraphraser or elaborator). In this study, the type of comprehender was not considered. Future studies should explore the interaction effects of the item properties and the type of a poor comprehender by means of the latent regression LLTM. This explanatory approach will provide more understanding of the weaknesses in the approach used by poor comprehenders when they attempt to construct global meanings from text.

A second implication is identifying text features in relation to the degree of text comprehension that may lead to item difficulty in the reading comprehension assessment. Text features will serve to provide information and cues in comprehending a text passage, and understanding the effects of text features on the degree of comprehension will be related to test performance on a reading comprehension assessment. Reading comprehension is completed with constructing a mental representation of a coherent situation model by means of generating and integrating meanings from the text. During the process of comprehension, readers must not only possess relevant language knowledge to construct meanings from the local and surface text level but also to

understand the structures of a text to integrate meanings. But it is important to keep in mind that no individual text feature had unchangingly significant effect on the item difficulty. Appropriate interpretation of the estimation demands holistic understanding of the effects of textual features on comprehension. Gaining a more nuanced understanding of text features might provide test developers with suggestions for evidence of construct validity and with information about item design and test specifications for the development of reading comprehension assessments. This understanding will also provide educators with insight useful for developing curriculum and improving instructional strategies, insofar as assessments can have a backwash effect on teaching and learning (Hughes, 2005; Messick, 1996).

A third implication of this study is its contribution to our understanding the variation of responses to items on a reading comprehension assessment and how this variation might depend on the properties that individual examinees possess, such as socioeconomic status, special education participation status, and English proficiency status. Through DIF analysis, a few items were identified with functioning differently as a group depending on EL status. However, the statistical approach did not provide explanation of the group differences. Thus, it is important to involve experts who can evaluate items with diverse and fair points of view during the early stages of constructing and reviewing items during test development processes (AERA, APA, & NCME, 2014; Lane et al., 2016). Because still group responses differed among subgroups depending on variables, it is necessary to have more diverse inputs and reviews to consider such group differences in the processes of a reading comprehension assessment development.

The fourth and last implication of my study is that it provides a guide to attempting to explore the effects of textual features on ELs' performance compared to non-ELs on a reading comprehension assessment. Few recognizable differences were identified in this study with respect to the text features that might be related to the responses of ELs on MOCCA. My results indicate the existence of different responses between ELs and non-ELs on certain items which did not have particular linguistic or semantic differences compared to other items. These findings suggest it will be meaningful to keep questioning what makes ELs perform differently than their counterparts, beyond the linguistic features of the tests, as long as the differences exist.

5.5 Conclusions

In my dissertation study, I examined the effects of text and person factors that might affect item difficulty and responses on a reading comprehension assessment, the Multiple-choice Online Causal Comprehension Assessment (MOCCA). Text structure and features are essential elements in reading assessments because an examinee is required to answer questions after reading and comprehending the text passage(s) that make up the assessment. Thus, understanding the factors that make text difficult is a substantive process in item design and validity evidence gathering in support of reading comprehension assessments.

I found that the explanatory approach to analyze the MOCCA data provided explanation of how the external variables such as text features and person properties were associated with the processes of generating responses to items on the MOCCA assessment. This approach provided insight into why certain items might be more difficult in terms of text features such as text complexity and story features, and how the

subgroups performed differently on MOCCA. In addition to these explanations, I identified a few items on the MOCCA assessment that were functioning differently between ELs and non-ELs despite having no distinctly different text features when compared to non-DIF items. Some findings are still inconclusive. The effects of text features were varied in association with the overall test difficulty, and I was unable to definitively identify the causes of DIF. These topics need to be examined further in future studies.

Future investigation, allowing for the interaction effect of text and person features or adding the features not addressed in this study, are needed to examine their possible relation to variation in item difficulty on a reading comprehension assessment. In addition, extending the study to grades other than third-grade, will provide a better understanding of the developmental growth of comprehension and if interactions between text features by person properties such as age exist when the sample population is broadened. Most importantly, additional studies are needed to gain a better understanding of the ways in which responses vary depending on the propensity of comprehension. Moreover, more samples and mixed research methods will provide more powerful explanatory information about the MOCCA items. Continued exploration of text features in the reading comprehension assessment will improve our understanding of what works in reading comprehension, how an assessment measuring reading comprehension can be developed and validated, and how the findings will be related to teaching and learning in classroom.

APPENDIX A

CTT ITEM DIFFICULTY BY ITEM BY FORM

Item no.	Form 3.1	Form 3.2	Form 3.3
1	66.7%	70.3%	72.1%
2	58.2%	62.8%	59.3%
3	67.7%	72.2%	71.8%
4	57.5%	62.2%	59.4%
5	61.9%	67.0%	65.8%
6	67.0%	67.7%	64.3%
7	51.8%	**53.7%**	**52.7%**
8	72.1%	70.8%	70.6%
9	68.5%	68.0%	68.0%
10	61.8%	62.2%	60.6%
11	73.7%	76.9%	**78.5%**
12	71.4%	68.5%	70.2%
13	61.6%	60.3%	66.4%
14	69.7%	71.1%	77.9%
15	59.5%	58.2%	62.9%
16	67.9%	73.9%	65.8%
17	65.6%	75.7%	69.7%
18	50.6%	60.5%	62.5%
19	70.1%	69.0%	73.6%
20	65.6%	69.6%	68.9%
21	57.3%	69.7%	71.6%
22	62.9%	73.0%	66.5%
23	57.1%	65.9%	66.2%
24	65.2%	63.0%	59.8%
25	**44.9%**	59.0%	56.9%
26	57.9%	58.9%	58.4%
27	54.8%	67.9%	64.7%

28	58.1%	64.2%	65.6%
29	59.7%	74.3%	69.2%
30	64.3%	59.8%	72.0%
31	64.3%	68.2%	71.0%
32	73.2%	65.7%	57.6%
33	58.6%	67.5%	54.9%
34	75.3%	61.4%	61.8%
35	63.0%	64.9%	65.3%
36	58.3%	70.4%	56.5%
37	75.5%	71.5%	66.5%
38	68.3%	75.1%	65.6%
39	69.5%	73.5%	70.1%
40	**78.8%**	**80.1%**	67.3%
Average of 10 common items	63.3%	65.7%	64.5%
Average of unique items	64.1%	67.9%	66.1%
Average of all items	63.9%	67.4%	65.7%

Note. Item numbers are different from actual item numbers of the MOCCA assessment. Items 1 to 10 are identical items across forms as anchored common items, but the other items are not the same item across forms despite the same item number.

APPENDIX B

IRT ESTIMATED ITEM DIFFICULTY BY ITEM BY FORM

Item no.	Form 3.1	Form 3.2	Form 3.3
1	-1.0385	-1.0385	-1.0385
2	-0.4954	-0.4954	-0.4954
3	-1.0788	-1.0788	-1.0788
4	-0.4255	-0.4255	-0.4255
5	-0.7082	-0.7082	-0.7082
6	-1.0163	-1.0163	-1.0163
7	-0.0707	**-0.0707**	**-0.0707**
8	-1.3671	-1.3671	-1.3671
9	-1.1286	-1.1286	-1.1286
10	-0.6934	-0.6934	-0.6934
11	-1.5289	-1.6258	**-1.8206**
12	-1.3646	-1.0226	-1.2223
13	-0.7273	-0.4913	-0.9696
14	-1.2525	-1.1992	-1.7808
15	-0.5930	-0.3539	-0.7414
16	-1.1275	-1.3962	-0.9290
17	-0.9805	-1.5273	-1.1894
18	-0.0450	-0.4888	-0.7148
19	-1.2656	-1.0475	-1.4607
20	-0.9758	-1.0875	-1.1416
21	-0.4411	-1.0847	-1.2915
22	-0.7835	-1.3066	-0.9408
23	-0.4169	-0.8163	-0.9102
24	-0.9091	-0.6265	-0.4896
25	**0.3400**	-0.3530	-0.2813
26	-0.4537	-0.3395	-0.3877
27	-0.2578	-0.9182	-0.7967

28	-0.4586	-0.6961	-0.8514
29	-0.5631	-1.3761	-1.0926
30	-0.8636	-0.4046	-1.3048
31	-0.8683	-0.9576	-1.2452
32	-1.4626	-0.8032	-0.3734
33	-0.5105	-0.9186	-0.2030
34	-1.6084	-0.5212	-0.6453
35	-0.7805	-0.7450	-0.8703
36	-0.4894	-1.1184	-0.3077
37	-1.6196	-1.2003	-0.9489
38	-1.1323	-1.4574	-0.8872
39	-1.2075	-1.3414	-1.1917
40	**-1.8787**	**-1.8514**	-1.0076
Average of 10 common items	-0.8023	-0.8023	-0.8023
Average of unique items	-0.8742	-0.9692	-0.9332
Average of all items	-0.8562	-0.9275	-0.9005

Note. Item numbers are different from actual item numbers of the MOCCA assessment. Items 1 to 10 are identical items across forms as anchored common items, but the other items are not the same item across forms despite the same item number.

REFERENCES CITED

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment*, *8*, 231-257.

Abedi, J. (2004). The no child left behind act and English language learners: Assessment and accountability issues. *Educational Researcher*, *33*, 4-14.

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice, 27*, 17-31.

Abedi, J. (2010). *Performance assessments for English language learners.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2005). The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives. CSE Report 663. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *25*, 36-46.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*, 1-28.

Abedi, J., Leon, S., & Mirocha, J. (2005). Chapter 1: Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analyses of extant data. In Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives*. *CSE Report 663* (pp. 1-46). National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*, 219-234.

Abedi, J., Lord, C., & Plummer, J. (1997). Language background as a variable in NAEP mathematics performance. *CSE Tech. Rep. No. 429*. Los Angeles: University of California. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Afflerbach, P. (2007). *Understanding and Using Reading Assessment, K-12*. International Reading Association.

Alderson, J. (2005). *Assessing reading* (Cambridge language assessment series). Cambridge, UK: Cambridge University Press.

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

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

Barnitz, J. G., & Speaker Jr, R. B. (1991). Second language readers' comprehension of a poem: Exploring contextual and linguistic aspects. *World Englishes, 10*, 197-209.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2018). lme4: Linear Mixed-Effects Models using 'Eigen' and S4. R package version 1.1-19. Retrieved from https://cran.r-project.org/web/packages/lme4/lme4.pdf.

Biancarosa, G., Kennedy, P. C., Carlson, S. E., Yoon, H., Seipel, B., Liu, B., & Davison, M. L. (2019). Constructing subscores that add validity: A case study of identifying students at risk. *Educational and Psychological Measurement*, *79*, 65-84.

Biancarosa, G., & Yoon, H., (2015). [Story Coding Guidelines]. Unpublished raw document. University of Oregon.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717-726.

Bransford, J. D., & Johnson, M. K. (1973). Considerations of some problems of comprehension. In W. G. Chase, *Visual information processing* (pp. 383-438). New York: Academic Press.

Brantmeier, C. (2003). Does gender make a difference? Passage content and comprehension in second language reading. *Reading in a Foreign Language, 15*, 1.

Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension in Spanish. *The Modern Language Journal, 89*, 37-53.

Buzick, H., & Stone, E. (2011). Recommendations for conducting differential item function (DIF) analyses for students with disabilities based on previous DIF studies. *ETS Research Report Series, 2011*, I-26.

Caccamise, D., & Snyder, L. (2005). Theory and pedagogical practices of text comprehension. *Topics in language disorders, 25*, 5-20.

Cain, K. (2003). Text comprehension and its relation to coherence and cohesion in children's fictional narratives. *British Journal of Developmental Psychology, 21*, 335-351.

Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology, 76*, 683-696.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*, 31.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology, 96*, 671.

Carlson, S. E., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences, 32*, 40-53.

Carrell, P.L. (1981). Culture-specific schemata in L2 comprehension. In R. Orem & J. Haskell (Eds.), *Selected papers from the Ninth Illinois TESOL/BE Annual Convention, First Midwest TESOL Conference* (pp. 123-132). Chicago: Illinois TESOL/BE.

Carrell, P.L. (1983). Some issues in studying the role of schemata, or background knowledge, in second language comprehension. *Reading in a Foreign Language*, *1*, 81-92.

Carrell, P. L. (1987). Content and formal schemata in ESL reading. *TESOL quarterly, 21*, 461-481.

Clauser, B., & Mazor, K. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*, 1227-1237.

Davey, B. (1988). Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *The Journal of Experimental Education*, *56*, 67-76.

Davison, M. L., Biancarosa, G., Carlson, S. E., Seipel, B., & Liu, B. (2018). Preliminary findings on the computer-administered multiple-choice online causal comprehension assessment, a diagnostic reading comprehension test. *Assessment for Effective Intervention*, *43*, 169-181.

Davison, M. L., Biancarosa, G., Seipel, B., Carlson, S. E., Liu, B., & Kennedy, P. (2008). "Technical manual 2018: multiple-choice online comprehension assessment (MOCCA)," MOCCA Technical Report MTR-2018-1.

De Ayala, R. (2013). *The theory and practice of item response theory* (Methodology in the social sciences). New York: Guilford Press.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1-28.

De Boeck, P., Cho, S. J., & Wilson, M. (2016). Explanatory item response models. In Rupp, A.A. & Leighton, J.P., *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 249-266). John Wiley & Sons.

De Boeck, P. & Wilson, Mark. (2004). Descriptive and explanatory item response models. In De Boeck, P. & Wilson, M., *Explanatory item response models: A generalized linear and nonlinear approach* (Statistics for social science and public policy) (pp. 43-74). Springer, New York, NY.

Desjardins, C., & Bulut, Okan. (2018). *Handbook of educational measurement and psychometrics using R* (Chapman & Hall/CRC the R series (CRC Press)). Boca Raton, FL: CRC Press.

Dowell, N. M., Graesser, A. C., & Cai, Z. (2016). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, *3*, 72-95.

Embretson, S., & Reise, Steven Paul. (2013). *Item response theory for psychologists* (Multivariate applications book series). Mahwah, N.J.: L. Erlbaum Associates.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*, 175-193.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*, 199-215.

Feng, S., Cai, Z., Crossley, S., & McNamara, D. S. (2011, March). Simulating human ratings on word concreteness. In *Twenty-Fourth International FLAIRS Conference*.

Fitzgerald, J. (1995). English-as-a-second-language learners' cognitive reading processes: A review of research in the United States. *Review of Educational research*, *65*, 145-190.

Freedle, R., & Kostin, I. (1991). The prediction of SAT reading comprehension item difficulty for expository prose passages. *ETS Research Report Series*, *1991*, i-52.

Freedle, R., Kostin, I., & Educational Testing Service, P. (1992). The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: main ideas, inferences and explicit statements. *GRE board Professional Report No.87-10P*, i-53.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: main ideas, inferences and supporting idea items. *ETS Research Report Series*, *1993*, i-48.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*, 395-427.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement, 42*, 351-373.

Gorin, J. S., & Svetina, D. (2012). Cognitive psychometric models as a tool for reading assessment engineering. In Sabatini, J., Albro, EL., & O'Reilly, T., *Reaching an understanding: Innovations in how we view reading assessment* (pp. 169-183). New York: R&L Education.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*, 223-234.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, *48*, 163-189.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In Sweet, A. P., & Snow, C. E. (Eds.), *Rethinking reading comprehension: solving problems in the teaching of literacy* (pp. 82-98). New York: Guilford Publications.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 3*6, 193-202.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371.

Harris, T. L., & Hodges, R. E. (1995). *The literacy dictionary: The vocabulary of reading and writing*. Newark: International Reading Association.

Hughes, A. (2007). *Testing for language teachers*. UK: Cambridge University Press.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*, 53.

Johnson, P. (1981). Effects on reading comprehension of language complexity and cultural background of a text. *TESOL Quarterly*, *15*, 169-181.

Johnson, P. (1982). Effects on Reading Comprehension of building background knowledge. *TESOL Quarterly*, *16*, 503-516.

Just, M., Carpenter, P., & Woolley, J. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General, 111*, 228-238.

Kena, G., Musu-Gillette, L., Robinson, J., Wang, X., Rathbun, A., Zhang, J., Wilkinson-Flicker, S., Barmer, A., and Dunlop Velez, E. (2015). *The condition of education 2015* (NCES 2015-144). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch.

Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders, 25*, 51-64.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY, USA: Cambridge University Press.

Lane, S., Raymond, Mark R., & Haladyna, Thomas M. (2016). *Handbook of test development* (Second ed.). New York: Routledge, Taylor & Francis Group.

Leroy, G., & Kauchak, D. (2013). The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association*, *21*, e169-e172.

Logan, S., & Johnston, R. (2009). Gender differences in reading ability and attitudes: Examining where these differences lie. *Journal of Research in Reading*, *32*, 199-214.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, *14*, 160-179.

Mau, W. C., & Lynn, R. (2001). Gender differences on the Scholastic Aptitude Test, the American College Test and college grades. *Educational Psychology, 21*, 133-136.

McFarland, J., Hussar, B., Wang, X., Zhang, J., Wang, K., Rathbun, A., Barmer, A., Forrest Cataldi, E., and Bullock Mann, F. (2018). *The condition of education 2018* (NCES 2018-144). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*, 276-282.

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

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247-298.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1-43.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*, 292-330.

McNamara, D. S., Ozuru, Y., & Floyd, R. G. (2011). Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge. *International Electronic Journal of Elementary Education, 4*, 229-257.

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

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241-256.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. National Institute of Child Health and Human Development, National Institutes of Health.

Nelson, K. (1996). *Language in cognitive development: The emergence of the mediated mind*. New York: Cambridge University Press.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). Measures of text difficulty: Testing their predictive value for grade levels and student performance. *Council of Chief State School Officers, Washington, DC*.

Olson, M. W. (1985). Text type and reader ability: The effects on paraphrase and text-based inference questions. *Journal of Reading Behavior, 17*, 199-214.

Perfetti, Charles, & Stafura, Joseph. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*, 22-37.

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., & Ginsburgh, M. (2009). Guidelines for the assessment of English language learners. *Princeton, NJ: Educational Testing Service*.

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

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

Robitzsch, A., Kiefer, T., & Wu, M. (2019). Test Analysis Module (TAM). R package version 3.3-10. Retrieved from https://cran.r-project.org/web/packages/TAM/TAM.pdf.

Seipel, B., Biancarosa, G., Carlson, S., & Davison, M. (2015). Analysis of textual features of a new reading comprehension assessment: MOCCA. *Society for Research on Educational Effectiveness.*

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*, 105-126.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*, 147-170.

Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*, 100-107.

Sireci, S. G., & Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review of Research in Education*, *39*, 215-252.

Solano-Flores, G., Lara, J., Sexton, U., & Navarrete, C. (2001). Testing English language learners: A sampler of student responses to science and mathematics test items. *Washington, DC: Council of Chief State School Officers.*

Tabors, P. O., & Snow, C. E. (2001). Young bilingual children and early literacy development. *Handbook of Early Literacy Research*, *1*, 159-178.

U.S. Department of Education. (2016). Non-regulatory guidance: English learners and Title III of the Elementary and Secondary Education Act (ESEA), as amended by the Every Student Succeeds Act (ESSA).

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

Williams, J. P. (1993). Comprehension of students with and without learning disabilities: Identification of narrative themes and idiosyncratic text representations. *Journal of Educational Psychology, 85*, 631.

Williams, J. P., Hall, K. M., Lauer, K. D., Stafford, K. B., DeSisto, L. A., & deCani, J. S. (2005). Expository text comprehension in the primary grade classroom. *Journal of Educational Psychology, 97*, 538.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In Hartig, J., Klieme, E., & Leutner, D. (Eds.), *Assessment of competencies in educational contexts* (pp.83-110). USA: Hogrefe & Huber Publishers.

Wilson, M., & Moore, S. (2012). An explanative modeling approach to measurement of reading comprehension. In Sabatini, J., Albro, EL., & O'Reilly, T., *Reaching an understanding: Innovations in how we view reading assessment* (pp. 147-168). New York: R&L Education.

Wongpakaran, N., Wongpakaran, T., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology, 13*, 61.

Zwaan, R. A. (1999) Embodied cognition, perceptual symbols, and situation models, *Discourse Processes, 28*, 81-88, DOI: 10.1080/01638539909545070

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.