

**TOWARD MORE VALID AND RELIABLE CLASSROOM ASSESSMENT OF
DIGITAL MULTIMODAL COMPOSITION PERFORMANCES:
DEVELOPING A SCORING RUBRIC TO ASSESS UPPER ELEMENTARY
STUDENTS' DIGITAL MULTIMODAL BOOK REVIEWS**

by

Sohee Park

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Education

Summer 2018

© 2018 Sohee Park
All Rights Reserved

**TOWARD MORE VALID AND RELIABLE CLASSROOM ASSESSMENT OF
DIGITAL MULTIMODAL COMPOSITION PERFORMANCES:
DEVELOPING A SCORING RUBRIC TO ASSESS UPPER ELEMENTARY
STUDENTS' DIGITAL MULTIMODAL BOOK REVIEWS**

by

Sohee Park

Approved: _____
Chrystalla Mouza, Ed.D.
Director of the School of Education

Approved: _____
Carol Vukelich, Ph.D.
Dean of the College of Education and Human Development

Approved: _____
Douglas J. Doren, Ph.D.
Interim Vice Provost for the Office of Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Rachel Karchmer-Klein, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Elizabeth Farley-Ripple, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

William Lewis, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Joshua Wilson, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Valerie Harlow Shinas, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

This dissertation would not have been possible without valuable interactions with my great people whom I sincerely appreciate.

First of all, I would like to express my sincere gratitude and appreciation to my adviser, Dr. Rachel Karchmer-Klein. She saw my potential as a literacy scholar and provided me with a variety of research and teaching opportunities. I must thank her for her numerous hours in guiding my work and providing constructive feedback on my writing. I would not have accomplished this work without her criticism, encouragement, and patience.

I also owe special gratitude to the rest of my dissertation committee—Dr. Elizabeth Farley-Ripple, Dr. William Lewis, Dr. Valerie Harlow Shinas, and Dr. Joshua Wilson—for their constructive criticism and thoughtful suggestions that have greatly influenced my dissertation.

I am also grateful to learn from other great literacy faculty outside my committee: Dr. Steve Amendum, Dr. David Coker, Dr. Charles MacArthur, Dr. Adrian Pasquarella, Dr. Sharon Walpole, and Dr. Carol Vukelich. I am indebted to them for their generous support and encouragement.

I would like to thank the participating teachers who helped me create, revise, and validate the rubric for my dissertation study. My study is built upon their hard works and sincere feedback.

Many thanks to my previous adviser, Dr. Myoungwan Noh, and colleagues in Moonyoen, the community of Korean language research. Dr. Noh and my colleagues in the community guided me to read up-to-date literacy journal articles and pose critical questions during my undergraduate and master's programs. The discussions I had with them became a great foundation for my doctoral study. Their academic advice and emotional support were strong motives of my doctoral journey.

DEDICATION

This dissertation is dedicated to my families in South Korea.

My father, Jongok Park, who passed away in 2005 was my best life adviser. He encouraged me to pursue the highest degree in the United States when I was at college. I could successfully finish this long journey because of your unconditional love. My mother, Younghee Yoon: you are a strong woman. I truly appreciate your wholehearted love and support. My younger sister, Sora Park, and younger brother, Hyowon Park: I could not have dreamed of studying in the United States without your consistent encouragement.

I also want to send my heartfelt appreciation to my mother-in-law, Chunglan Yon, father-in-law, Whoyoung Chang, and brother-in-law, Hyosik Chang. Since I became a member of this family, all of you showed what real love is. Your all-around support encouraged me to go through this journey.

I sincerely appreciate Hyo Joon Chang who is my dear love, best friend, and best father of our son. Meeting and getting married to you is the best luck of my life. You have been and are always a good consultant and critique. Getting doctoral degrees at the same time was not a usual life event, but we finally made it.

Lastly, my dear son, Hajoon Chang: You made me experience amplified happiness in my life. There were some moments that I struggled to balance works with caring of you. However, your smiles and giggles washed away all my worries and led me to have new energy. I will do my best to be your best mother!

TABLE OF CONTENTS

LIST OF TABLES	xi
LIST OF FIGURES	xii
ABSTRACT	xiii
Chapter	
1 INTRODUCTION	1
Goals for the Study and Research Questions	4
Relevant Terms and Definitions	6
Terms Related to Multimodality	6
Terms Related to Rubric-Based Performance Assessments in Classrooms	8
2 CONCEPTUAL FRAMEWORK	11
Multimodality	11
Multimodality: A Broad Concept	11
Universal Theoretical Assumptions on Multimodality	14
Digital Multimodal Composition Performances of Upper Elementary Grades Students	17
Rubric-Based Formative Assessment of Performances	25
Performance Assessment and Digital Multimodal Composition	26
Formative Assessment in Classrooms	27
Reliability and Construct Validity of Rubric-Based Performance Assessment	30
3 METHODS	38
Participants	40
A Collaborator	41
Raters	42
Data for Scoring: Fourth-Grade Students' Digital Multimodal Book Reviews	44
Research Purpose of the Larger Study	44
Fourth-Grade Students	45
Characteristics of the Task	46
Procedures to Collect Students' Digital Multimodal Book Reviews	47

Procedures and Data Analyses	47
Phase 1: A Systematic Literature Review	47
Phase 2: Rubric Creation	51
Phase 3: Rubric Validation (Convergent Parallel Design)	58
Phase 4: Rubric Revision (Explanatory Sequential Design)	65
4 RESULTS.....	68
Results for Research Phase 1: Identified Domains and Criteria.....	68
Criteria for the Artifact Domain.....	68
Criteria for the Context Domain	72
Criteria for the Substance Domain.....	73
Criteria for the Process Management and Technique Domain	75
Criteria for the Habits of Mind Domain	75
Results for Research Phase 2: Rubric Creation Procedures.....	79
Step 1: Decisions on Four Aspects of Scoring Rubrics.....	79
Step 2: Reflecting.....	80
Step 3: Listing.....	81
Step 4: Grouping and Labeling.....	84
Step 5: Writing Definitions and Descriptors.....	85
Step 6: Revising with a Collaborator	88
Results for Research Phase 3: Rubric Validation.....	94
Quantitative Evidence on Inter-Rater Reliability of the Use of the Rubric	94
Qualitative Evidence on the Reliability of the Use of the Rubric.....	99
Quantitative Evidence on the Structural Aspect of Construct Validity.....	101
Qualitative Evidence on the Structural Aspect of Construct Validity	106
Qualitative Evidence on the Content Aspect of Construct Validity ..	109
Results of Research Question 4: Rubric Revision.....	117
Problematic Criteria that were Identified in Research Phase 3	117
Problems and Solutions for C3-Conventions of Oral Language	118
Problems and Solutions for C4-Conventions of Written Language and C6-Relationship Between Oral and Written Language.....	121
Problems and Solutions for Criterion 10-Quality of Summary	124
Problems and Solutions for C11-Quality of Opinion.....	127
5 DISCUSSION AND IMPLICATIONS	129

Overview of the Study	129
Inter-Rater Reliability	131
Consensus Agreements.....	131
Consistency Estimates	134
Rater Training and Inter-Rater Reliability.....	135
Construct Validity.....	139
Content Aspects of Construct Validity.....	139
Structural Aspects of Construct Validity.....	141
The Rubric as a Formative Assessment Tool.....	142
Strategy 1: Clarifying Learning Intentions and Criteria for Success .	143
Strategy 2: Engineering Effective Classroom Discussions and Other Learning.....	144
Strategy 3: Providing Feedback that Moves Learners Forward.....	145
Limitations	147
Implications	149
Future Research.....	149
Classroom Application of the Rubric.....	150
Conclusion.....	153
REFERENCES.....	155
Appendix	
A PRESENTATION SLIDES USED FOR THE OFFLINE SCORING TRAINING.....	170
B INITIAL SCORING PROTOCOL	174
C LIST OF ONLINE DIGITAL MULTIMODAL BOOK REVIEWS ABOUT A BOOK, FRINDLE	180
D REVISED SCORING PROTOCOL.....	181
E A SAMPLE OF THE SCORING SHEET INCLUDING THE SPACE FOR SCORING LOG	189
F A SEMI-STRUCTURED INTERVIEW QUESTIONNAIRE.....	190
G CODEBOOK (35 CODES AND 11 CATEGORIES).....	191
H RELATIONSHIPS AMONG DOMAINS, NEW CRITERIA AND ORIGINAL CRITERIA IN BOTH THE RUBRIC AND NON-RUBRIC LITERATURE.....	201

I	LISTED CHARACTERISTICS OF EACH STUDENT’S PERFORMANCE RELATED TO THE RELATIONAL RELEVANCE OF WRITTEN LANGUAGE MODE.....	205
J	GROUPED CHARACTERISTICS ON EACH CRITERION.....	208
K	FIRST DRAFT OF THE RUBRIC (RUBRIC #1).....	211
L	A METARUBRIC FOR REVISION OF THE RUBRIC (ARTER & CHAPPUIS, 2006).....	217
M	RUBRIC ANALYSIS FORM (ARTER & CHAPPUIS, 2006).....	220
N	FINAL VERSION OF THE RUBRIC USED FOR INDEPENDENT SCORING (RUBRIC #7).....	221
O	MAJOR CHANGES ON THE DESCRIPTORS OF EACH CRITERION IN DIFFERENT VERSIONS OF THE RUBRIC	228
P	REVISED SCORING RUBRIC	252
Q	INSTITUTIONAL REVIEW BOARD APPROVAL LETTER.....	260

LIST OF TABLES

Table 1	Descriptions of Modes and Modal Resources	13
Table 2	Relationship between Five Strategies of Formative Assessment and Their Agents (Black & Wiliam, 2009, p. 8)	29
Table 3	Modified Steps of Rubric Creation	57
Table 4	Research Questions, Purposes, Tasks and Data.....	66
Table 5	Finalized Domains and Criteria of Digital Multimodal Composition Assessment	77
Table 6	Initial Descriptors for the Technical Aspects of Visual Mode.	86
Table 7	Domains, Categories, and Criteria included in the Rubric.....	87
Table 8	Exact and Adjacent Agreement Rates Between the Researcher and Collaborator on 8 Digital Multimodal Book Reviews using Rubric #3..	90
Table 9	Exact and Adjacent Agreement Rates Between the Researcher and Collaborator on 25 Digital Multimodal Book Reviews using Rubric #4	91
Table 10	Percent Exact and Adjacent Agreement Among Raters ($n = 25$).....	95
Table 11	Cohen's Kappa Per Criterion Among Raters ($n = 25$).....	97
Table 12	Spearman's <i>Rho</i> Per Criterion Among Raters ($n = 25$).....	99
Table 13	Spearman's <i>Rho</i> Correlations Among Experts' Scores on 11 Criteria .	102
Table 14	Spearman's <i>Rho</i> Correlations Among Kristen's Scores on 11 Criteria	103
Table 15	Spearman's <i>Rho</i> Correlations Among Lindsey's Scores on 11 Criteria	104
Table 16	Pearson Correlations Among the Experts' Scores on 6 Categories.....	105
Table 17	Pearson Correlations Among the Kristen's Scores on 6 Categories.....	105
Table 18	Pearson Correlations Among the Lindsey's Scores on 6 Categories....	106
Table 19	Criteria Meeting the Target Levels of Consensus Agreements	132

LIST OF FIGURES

Figure 1 Flowchart of the Implementation of Four-Phase Mixed-Methods Study 39

ABSTRACT

Digital multimodal composition has drawn attention from literacy scholars and educators as a new type of writing activity that leverages digital technologies and their affordances. However, there is limited research on what and how to assess elementary students' task-specific digital multimodal composition performances. This mixed-methods study aims to explore the following four research questions in four phases: 1) what are the key domains and criteria that represent the construct of digital multimodal composition? 2) how might these domain and criteria be structured in a task- and grade-specific rubric to evaluate upper-elementary-grade students' digital multimodal book reviews? 3) to what extent does the use of the proposed rubric display evidence of inter-rater reliability, construct validity and usability, as indicated by raters' scores and interview feedback? and 4) how does the evidence of inter-rater reliability, construct validity, and raters' feedback inform the process of rubric revision? In the first phase, a systematic literature review was conducted and a total of five domains and 19 distinguishable criteria were identified from 111 criteria in pre-existing 17 studies. In the second phase, a rubric was created in six steps. In the third phase, rubric validation was conducted by training two raters and collecting both quantitative and qualitative evidence on inter-rater reliability, construct validity, and usability of the developed rubric. The analyses revealed that there were five criteria (i.e., C3-conventions of oral language, C4-conventions of written language, C6-relationship between oral language and written language, C10-quality of summary, and C11-quality of opinion) that seemed to be problematic. In the final phase, the rubric was revised by closely examining the problematic criteria and discussing possible solutions

for those through a focus group interview. The findings could contribute to our understanding of digital multimodal composition, rubric development procedures, and the role of the rubric for formative assessments in classrooms.

Keywords: digital multimodal composition, digital multimodal book review, scoring rubric, inter-rater reliability, construct validity, formative assessment, upper-elementary grades

Chapter 1

INTRODUCTION

The purpose of this mixed-methods study was to develop a scoring rubric for valid and reliable formative assessment of upper-elementary students' digital multimodal book reviews (DMBRs). A digital book review is a type of digital multimodal composition (DMC) that combines oral, written, visual and audio representations (Kalantzis, Cope, & Cloonan, 2010) to present book summaries and recommendations to others (Bates, 2012; Dalton & Grisham, 2013; Ehret, Hollett, & Jocius, 2016). A digital book review is an important form of DMC for three reasons. First, it provides students opportunities to practice persuasive writing, which is emphasized in the Common Core State Standards (CCSS) (Dalton & Grisham, 2013; National Governors Association Center for Best Practices [NGACBP] & Council of Chief State School Officers [CCSSO], 2010). Second, it motivates students to actively engage in the book review because they can expect a larger audience when it is published online (Butler, Leahy & Mc Cormack, 2010). Finally, creating digital book reviews allows students to utilize technology and different modes of representation to produce and publish their writing (Ehret et al., 2016; Phillips & Smith, 2012), which is another requirement of curriculum standards (NGACBP & CCSSO, 2010).

DMC has drawn attention from literacy scholars and educators as a new type of writing activity that leverages digital technologies and their affordances (Dalton, 2012;

DeVoss, Eidman-Aadahl, & Hicks, 2010; Miller & McVee, 2012). In addition to digital book reviews, digital stories, digital poems, and music videos are other examples of DMC (Smith, 2014). Given increased access and ease of publishing (U.S. Department of Education, Office of Educational Technology, 2017), many scholars and professional organizations in education (e.g., International Society for Technology in Education [ISTE] and National Council of Teachers of English [NCTE]) have pointed out the critical need to prepare students in their earliest years to compose these types of texts (Bearne, 2009; ISTE, 2017; NCTE, 2016; United Kingdom Literacy Association [UKLA]/Qualification and Curriculum Authority [QCA], 2004, 2005). Recent studies report that upper-elementary-grade students are capable of composing digital multimodal texts, although few are adequately equipped with the necessary skills (Dalton, Robinson, Lovvorn, Smith, Alvey, Mo, Uccelli, & Proctor, 2015; Karchmer-Klein, 2007; Phillips & Smith, 2012; Shanahan, 2012, 2013a, 2013b). Therefore, more investigations on both instruction and assessment are needed to raise upper-elementary students as the well-prepared multimodal communicators.

Educational assessments play important roles by determining how well students are learning (National Research Council, 2001). Performance assessments are greatly needed to properly measure how students acquire and apply knowledge to complete complex and real-world tasks that are increasingly needed in the 21st century (Darling-Hammond, 2014). Based on assessment purposes, performance assessments can be categorized into two types: summative assessment and formative assessment (Taras, 2005). Summative assessment aims to evaluate whether or not students reached the final

learning outcomes or benchmarks. High-stakes writing tests in National Assessment of Educational Progress (NAEP) or final exams including multiple choice and essay items are examples of summative assessment. On the contrary, formative assessment is used to share success criteria with students, provide constructive feedback, and activate students as the owners of their learning or as useful resources for their peers (Black & Wiliam, 1998; 2008). Rubric-referenced self-assessments of students' essays in the classroom are an example of formative assessment (Andrade & Boulay, 2003). Particularly, in order to teach students how to compose digital multimodal texts and provide them adequate feedback, formative assessment of DMC is in great needs (Hung, Chiu, & Yeh, 2013; Silseth & Gilje, 2017).

Despite the growing importance of DMC and needs for its assessment, how to assess DMC has been understudied, especially in elementary contexts (Bearne, 2009; McGrail & Behizadeh, 2017). The majority of studies have investigated the college composition field by discussing some evaluative criteria (e.g., Adsanatham, 2012; Yancey, 2004; Sorapure, 2006). However, there is a huge gap between college students' DMC performances and elementary students' DMC performances, so it is not certain if the findings of college compositions studies can be applicable to elementary school settings. Although there are a few studies discussing frameworks and metalanguages of DMC assessment in K-12 settings (Bearne, 2009; Eidman-Aadahl et al., 2013; Hicks, 2015; Levy & Kimber, 2009; McGrail & Behizadeh, 2017), the discussions are still theoretical, which calls on researchers to examine the actual negotiation between teachers and students on DMC assessment practices in classrooms (Silseth & Gilje, 2017).

Thus, it is necessary to gain more knowledge about what and how to assess grade- and task-specific DMCs. For this study I chose to develop a rubric for formative assessment of students' DMBRs. At a glance, the task seems to be similar to the print-based book review, a type of writing output familiar to upper-elementary students. However, in reality, DMC is relatively new and complex task that requires students to acquire knowledge on multiple aspects such as rhetorical purpose, modes, and technical functions of digital tools. Therefore, before students are evaluated in summative way, they need to learn how to compose digital multimodal text coherently through several formative assessments. A task- and grade-specific rubric is appropriate to achieve the purpose, and it would enable teachers to provide more detailed feedback to students on multiple traits without extra effort.

Goals for the Study and Research Questions

With the above-stated rationales in mind, this mixed-methods study aimed to develop a scoring rubric for the evaluation of upper-elementary-grade students' DMBRs in classrooms. The following four research questions guided this study:

1. What are the key domains and criteria that represent the construct of digital multimodal composition?
2. How might these domains and criteria be structured in a task- and grade-specific rubric to evaluate upper-elementary-grade students' digital multimodal book reviews?

3. To what extent does the use of the proposed rubric display evidence of inter-rater reliability, construct validity and usability, as indicated by raters' scores and interview feedback?
4. How does the evidence of inter-rater reliability, construct validity, and raters' feedback inform the process of rubric revision?

The first research question is designed to identify the domains and criteria that are important for any task of DMCs. A systematic review of literature on the assessment of DMC will reveal some key domains and criteria. The second research question guides the rubric-development procedures. Reasons for selecting certain domains and criteria among those identified with the first research question and the collaborating procedures of rubric development are shared as the answers to this question.

Within the convergent parallel mixed-methods design (Creswell & Plano Clark, 2010), the third research questions seek to explore the reliability, validity and usability of the developed rubric. The quantitative analyses and comparison of the scores that were obtained from the developer and the two raters reveal if the rubric can consistently measure different and separate criteria/traits of the DMBR. The results of qualitative analysis of two raters' interviews shed light on the appropriateness of content and the usefulness of the rubric. Rubric development accompanies recurring processes that use, validate, and revise a rubric. The final research question is set to detect the problems of the scoring rubric, generate some ideas for rubric revision, and present the revised version of the rubric. Focus group interview data are collected and analyzed to further explain the findings of quantitative and qualitative data analyses at the previous phase.

Relevant Terms and Definitions

This study includes terminology from the studies on multimodality and rubric-based performance assessment in classrooms. Studies from both fields use many different terms to indicate the same or similar concepts. For example, some scholars use the term dimensions instead of criteria/traits in order to indicate the important characteristics of student work that need to be separate components in an analytic rubric. Definitions provided below will be applied throughout this study.

Terms Related to Multimodality

Digital multimodal book reviews (DMBRs). A type of DMC that intends to persuade audiences to read the introduced book. The format of DMBRs is very similar to that of movie trailers, which convey a brief portion of the story and the reviewer's opinion by using oral and written language, visual, audio, and spatial modes. DMBRs are usually created with digital photo/video editing tools such as Animoto, iMovie, Movie Maker, Photo Story, et cetera (Bates, 2012).

Digital multimodal composition (DMC). Composing digital texts using a combination of modes of representation (e.g., written and oral language and visual, audio, tactile, gestural, and spatial representations) to communicate a coherent message (Burke & Hammett, 2009; Eidman-Aadahl et al., 2013; McGrail & Behizadeh, 2017). Some examples of DMCs include DMBRs, digital stories, presentation slides, blog, glog, or vlog posts, wiki posts, posts on social media such as Facebook, Instagram, and Twitter.

Intermodal relationships. Relationships between meanings of different modes, such as writing and images, in a multimodal text. More broadly, the term refers to the

comprehensive understanding of all intermodal identifications between two modalities in a multimodal text.

Modal resources. All aspects of a mode that affect meaning of the multimodal text. Bezemer and Kress (2008) provide examples of modal resources by analyzing the resources of writing: font type, size, and color are graphic resources, and punctuation is a resource to frame writing. Regarding a static image, size and shape of the image, and size, shape, color, and position of the elements included in the image are its modal resources.

Modes. Basic units of multimodal communication, which are socially shaped and culturally given (Kress, 2010). In this study, categories of modes from the multiliteracies perspective—audio, linguistic (i.e., written and oral language), visual representations—will be used to describe different levels of multimodality-related criteria in the rubric (Kalantzis et al., 2010).

Multimodal ensemble. A multimodal text or artifact that conveys meaning through interrelated and co-presented modes (Bezemer & Kress, 2008; Jewitt, 2014; Serafini, 2014). In this study, the term *multimodal ensemble* will be used to indicate a multimodal text that clearly displays one of the following interrelationships between modes: corresponding, complementary, augment, and dissonant/irrelevant (Jewitt, 2014).

Transmediation. Selecting available modes and modal resources from an original text and then rearranging and reconnecting them in a different type of text to convey different meanings and/or the sign makers' interest (Gilje, 2010; Mills, 2011; Siegel, 1995; Suhor, 1984).

Terms Related to Rubric-Based Performance Assessments in Classrooms

Analytic rubric. A type of rubric that includes several criteria or traits to help in the assessment of different domains of a performance. It is usually used to indicate the opposite of a holistic rubric.

Construct validity. A unified view on validity that “integrates considerations of content, criteria, and consequences into a construct framework” (Messick, 1995, p. 741). Six distinguishable aspects of construct validity are “content, substantive, structural, generalizability, external, and consequential” (p. 744). Content and structural aspects of construct validity are highly related to this study. The content aspect of construct validity examines if the content of an assessment tool is relevant to and representative of the construct domain. The structural aspect of construct validity evaluates whether the structure of scoring tool represents the structure of the construct domain.

Inter-rater reliability. “The level of consistency in rank ordering of ratings across raters” (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014, p. 220). There are three different ways of measuring inter-rater reliability: (a) consensus agreements, (b) consistency estimates, and (c) measurement estimates (Stemler, 2004; Jonsson & Svingby, 2007).

Rubric components. The developed rubric includes several components in order to describe and assess different aspects and levels of DMBR performances: domains, categories, criteria, and descriptors.

Domain. The largest and separable dimension of student performances that need to be assessed. For the assessment of DMBR, the current study adopted five domains that were suggested by Eidman-Aadahl and colleagues (2013): artifact, context, substance, process management and technique, and habits of mind. In general, each domain may include a few categories and each category contains some criteria that can represent different aspects of the domain.

Category. The larger unit encompassing a few similar criteria. In the current study, for example, two criteria—technical aspects of visual modes and technical aspects of voice—are included in the technical aspects of modes category since the two criteria aim to assess technical aspects of modes that students utilized in their DMBRs.

Criterion. The smallest unit of assessment in the rubric. In this study, each criterion represents a unique and distinguishable characteristic of the DMBR. Technical aspects of visual mode and quality of summary are the examples of criteria.

Descriptors. Descriptions of each performance level under each criterion in analytic rubrics. For example, if an analytic writing rubric includes four different levels of performance (e.g., weak, fair, good, and excellent) under each criterion, the descriptors will provide more detailed information about how performances can be categorized under each level.

Rubric Development. The entire process of developing a scoring rubric, which involves the phases of rubric creation, rubric validation, and rubric revision. Rubric development is not finished in one cycle of the phases; rather, it involves multiple cycles of the rubric validation and revision phases until the rubric reaches the desirable levels of

reliability and validity. Each phase of rubric development will be explained further in the methods section.

Chapter 2

CONCEPTUAL FRAMEWORK

Given that the purposes of this study are to identify domains and criteria of the DMC performances as the construct and to develop a scoring rubric for upper-elementary grade students' digital book reviews, this chapter is intended to construct a conceptual base for the study.

The two dimensions guiding this study are *multimodality* and *rubric-based performance assessment*. I begin this section by defining multimodality and its underlying theoretical assumptions (Jewitt, 2014). Findings from existing studies on upper-elementary students' DMCs are synthesized under each assumption. With regard to the rubric-based performance assessment, I first define DMC as performance assessment and explain why formative assessment and rubrics matter for the assessment of DMC. Then, I discuss issues around the reliability and validity of rubric-based performance assessment.

Multimodality

Multimodality: A Broad Concept

This study considers multimodality to be the core characteristic of DMC because the digital texts are inherently multimodal in terms of the used modes of representations and their interactions (Jewitt & Price, 2012). Multimodality is a broad concept indicating

multiple layers: the act of meaning-making using multiple modes (Kress & van Leeuwen, 2001), the condition of an event or product involving multiple modes (Kress, 2003, 2010) or, a field of inquiry (Jewitt, 2014). Simply put, multimodality refers to almost everything related to meaning-making practices using several modes of representations such as oral, written, visual, audio, spatial and gestural representations (Kalantzis et al., 2010).

Modes, modal resources and multimodal ensembles. In order to fully understand the multimodality of digital texts, modes, modal resources, and multimodal ensembles found in multimodal artifacts, these terms need to be defined.

Mode is “a socially shaped and culturally given resource for making meaning. Image, writing, layout, music, gesture, speech, moving image, soundtrack are examples of modes” (Kress, 2014, p. 60). Modal resources refer to the “potentials and constraints” (Bezemer & Kress, 2008, p. 172) of a mode when it is used to convey and represent meanings (Jewitt, 2014; Kress, 1993). With regards to the modal resources of written language, words, sentences, font type, size, spacing, color, and punctuation are examples of its modal resources (Bezemer & Kress, 2008; Kress, 2014). When composing multimodal artifacts using digital tools, selecting and utilizing modal resources can be constrained by the affordances of the digital tool. For example, the iMovie program is not appropriate to write an essay with several paragraphs in the *title* section because its affordance constrains written language to be used as a title, not a paragraph. On the contrary, iMovie affords some modal resources of written language such as font style (e.g., boldface and italic), font size and colors to highlight the written language in iMovie products.

There is no common list of modes. It seems different theorists and researchers use different terms. For example, Kalantzis, Cope, and Cloonan (2010) reconfigured “the range of possible modalities” (p. 66) and categorized the modes that were originally discussed by Kress and his colleagues (Kress, 2010, 2014; Bezemer & Kress, 2008) into distinguishable groups. For the purposes of this study, I created a list of modes, skills related to the modes, examples, and modal resources by synthesizing information provided in Bezemer & Kress, 2008; Kalantzis et al. (2010), Kress (2014), Ostenson (2012) and Wise (2016). See Table 1.

Table 1 Descriptions of Modes and Modal Resources

Modes	Examples	Modal Resources
Written language	<ul style="list-style-type: none"> • Related skills: reading and writing • Examples: hand writing, the printed book, written language in eBooks or movies 	<ul style="list-style-type: none"> • Graphic: Font size, type, shape (cursive/print, italic, bold); underline; spacing • Words, phrases, sentences, paragraphs, and genres
Oral language	<ul style="list-style-type: none"> • Related skills: listening and speaking • Examples: live or recorded speech 	<ul style="list-style-type: none"> • Pitch: low ~ high • Tempo: slow ~ fast • Tone: soft ~ sharp • Inflection: monotone ~ animated • Words, phrases, sentences, paragraphs, and genres
Visual representation	<ul style="list-style-type: none"> • Related skills: watching or viewing • Examples: still or moving image (e.g., pictures, illustrations, cartoons), sculpture 	<ul style="list-style-type: none"> • Color • Perspective • Angles • Lighting • Emphasis

Audio representation	<ul style="list-style-type: none"> • Related skills: listening and playing instruments • Examples: instrumental or vocal music, sound effects, noises, and silence 	<ul style="list-style-type: none"> • Layout • Volume: quiet ~ loud • Beat: slow ~ fast • Harmony
Tactile representation	<ul style="list-style-type: none"> • Related skills: touching, smelling, cooking, and eating • Examples: physical contact, kinesthesia, skin sensation, grasp, foods, and aromas 	<ul style="list-style-type: none"> • Temperature: cold ~ hot • Pressure • Texture • Flavors • Ingredients
Gestural representation	<ul style="list-style-type: none"> • Related skills: moving, gazing, exercising, dressing • Examples: body movements (e.g., dance, posture, gesture), facial expressions, eye movements and gaze, demeanors of the body, fashion, hairstyle 	<ul style="list-style-type: none"> • Body parts: hands, arms, legs, face and eyes • Action sequences • Clothes and accessories

Universal Theoretical Assumptions on Multimodality

Jewitt (2014) presents four theoretical assumptions that are universal for any theoretical perspective toward multimodality. These universal theoretical assumptions are useful to understand the multimodality as a broad and unconstrained concept. The assumptions are also suitable to serve as the theoretical framework of the current study because they can be the key components to be assessed through any rubric of DMC

The first theoretical assumption on multimodality is that “language is part of a multimodal ensemble” (Jewitt, 2014, p. 15). A multimodal ensemble—a multimodal text or artifact that conveys meaning through interrelated and co-presented modes—may include both linguistic and non-linguistic modes. Regardless of their theoretical

perspectives, multimodal theorists such as Gunther Kress and Len Unsworth agree that traditional linguistics cannot fully explain the making of meaning in the real world because it does not consider other signs and symbols as being resources for the process of meaning-making. Hence, semiotics, the study of signs and symbols in society, is often pressed into service as a meaningful lens through which to interpret meaning-making in society, with the mode being the key unit of these studies. For example, in DMBRs of the current study, students used oral language, written language, static and/or moving images, sound effects and theme music. Therefore, the rubric of DMBRs should not privilege oral and written language; Rather, it should consider all different modes used in students' DMBRs equally important.

The second assumption shares the idea that each mode is interpreted by readers or writers by realizing a unique communicative purpose. For instance, the STOP traffic sign on road has two notable modes: written language and the background color. The all capitalized word STOP delivers the meaning, *cease your drive at this point*, to drivers, to drivers and the red background color adds the meaning of warning or urgency (Kress, 2010). The assumption and its example inform that the rubric which will be developed in the current study needs to consider a unique communicative purpose of each mode.

The third assumption is related to the sign-maker's selection and configuration of modes. It means that although all people have exactly the same modal resources available for their DMCs, their resulting multimodal ensembles will be different due to their different choices and combinations of modes. In addition, when sign-makers select and combine modes, they interrelate each mode and sometimes the new relationship between

modes create a new meaning by expanding the meaning of each mode (Martinec & Salway, 2005; Unsworth, 2006). This assumption needs to be reflected to the rubric with the consideration of the intermodal relationships.

The final assumption concerns the social aspects of meanings. When sign-makers create signs using multiple modes, they follow “norms and rules” (Jewitt, 2014, p. 17) that are consensual in that society. These norms and rules allow audiences to read or interpret the signs in context. That is, although sign-makers have their own intentions and interests when creating DMCs, the social conventions and principles related to the signs should be considered as a part of that communication. Bringing back the STOP sign example, the social rule surrounding the traffic sign adds the meaning of MUST onto it. The complete meaning of the sign becomes the following: drivers MUST come to a complete stop at the point, if violate, drivers will pay the penalty. In the current study, the six components of DMBRs (i.e., introduction, brief summary of the book, one thing I liked, one thing I did not like, recommendations, and conclusion) are the examples of the norms and rules related to DMBRs because they were taught by the teacher using a graphic organizer and are commonly found from other online book reviews or book trailers.

Taken together, these four assumptions provide a lens through which multimodality of DMC can be understood. With this lens in mind, findings of empirical studies examining upper-elementary students’ DMC performances will be summarized under each assumption.

Digital Multimodal Composition Performances of Upper Elementary Grades

Students

It is important to understand the level of upper-elementary-grade-students' performances of DMC to reflect it into the descriptors of the rubric; Reviewing empirical studies on upper-elementary-grade-students' digital-multimodal-composition-performances provides a benchmark of the rubric descriptors. In this section, I synthesize the evidence from the empirical studies related to each assumption.

Assumption 1: Language is part of a multimodal ensemble. The first assumption emphasizes the importance of both linguistic and non-linguistic modes in multimodal ensembles as active conveyers of meaning. This assumption implies that only teaching and evaluating the affordances and meanings of language in students' DMCs is not adequate and that these students should also learn and be assessed on the affordances and meanings of the non-linguistic modes. Several empirical studies shared this assumption and reported two phenomena observed from upper-elementary classrooms: (a) teachers privileged linguistic modes over the non-linguistic modes, therefore, (b) they did not provide appropriate instruction on the types and roles of non-linguistic modes for DMC (Shanahan 2012, 2013a, 2013b).

Shanahan (2012, 2013a, 2013b) conducted interpretive case studies on the relationship between a fifth-grade teacher's instruction on multimodal composition and her students' use of multiple modes while they were creating multimodal presentation slides using the *Hyperstudio* program in a science class. These three studies collected and analyzed the same classroom data but focused on students' use of different modes:

Shanahan (2012) investigated students' use of sound, whereas Shanahan (2013a) concentrated on their use of language and visual modes. In addition, Shanahan (2013b) focused on the relationship between visual and linguistic modes and examined how the teacher's instruction of multimodal composition with visual and linguistic modes reflected to the fifth-grade students' meaning-making with the two modes.

The three studies reported that the teacher's emphasis on linguistic modes in multimodal texts influenced students' perspectives on the role of visual and audio representations. More specifically, Shanahan's (2013b) analyses of the teacher's whole-group conversations on multimodal composition and conversations between the teacher and students revealed that due to the teacher's limited knowledge on the functions and terminologies of multiple modes, the teacher provided language-centered perspective and instruction. As a result, her students considered the linguistic modes to be the primary conveyers of meaning in multimodal texts, relegating the visual and audio modes to secondary roles. These findings suggest that the teacher's instruction on and stance toward DMC may have substantial impact on upper-elementary grades students' preferences and usages of modes.

Assumption 2: Each mode reflects a unique communicative purpose. The second assumption emphasizes a unique communicative purpose of each mode. There are several studies focused on the role of one mode in upper-elementary students' multimodal ensembles (Pantaleo, 2012; Phillips & Smith, 2012; Shanahan, 2012; Shanahan, 2013a). Pantaleo (2012) and Shanahan (2013a) both focused on the analysis of fifth and sixth grade students' use of visual modes (e.g., image) and modal resources

(e.g., color), whereas Phillips and Smith (2012) and Shanahan (2012) examined students' use of sound in multimodal texts.

Pantaleo (2012) reported findings from a case study of a 11-year-old female student who participated in a larger classroom-based research project which aimed to explore how students' knowledge and understanding of visual grammar influenced on their own print-based multimodal texts. Analyses of the student's multimodal text and interview transcript found that the student purposefully chose colors to convey distinctive meanings or intentions. For example, on the first page of her graphic novel, the student used "complementary colors" (p. 150) to represent the main character she created and her house where her mother and father lives together. The doors of mother's and father's rooms were differentiated by the two complementary colors: blue and orange. She stated that the blue color reflected her mother's calm and strict personality, whereas the orange color shows her father's fun and energetic characteristics. To highlight the main character, she used another set of complementary colors: yellow and green. Then she associated the green color to the nice summer weather and the main character's outgoing personality.

Shanahan (2013a) also reported two students' intentional use of color: in their presentation slides about acid rain, they represented a fish damaged by acid rain using a lighter pale orange color in order to signify the fish's ill health. Interestingly, these purposeful decision-makings contradict with findings of other research studies that student preference outweighs other factors when choosing the aesthetics for digital magazine layout. Zammitt (2009), Callow (2003), and Karchmer-Klein (2007) found that

even after explicit instruction on how to use color in digital multimodal texts, children chose colors based upon personal interest rather than cohesiveness, audience, or content.

Phillips and Smith's (2012) study examined how 50 students ages 5-18 used sound (e.g., music, sound effects), speech, and silence in their StoryTube videos. Among all students, there were 33 students 10-years-old and younger. All of the 10-year-old and below group students included speech and 22 out of 33 used music. Only 15 students inserted sound effects in their videos. The authors also explored the sound spaces, which is constructed through "the interactions among video producers, elements of the video, and audiences" (p. 89). Three distinctive sound spaces were identified: book sound space, report sound space, and kid sound space. Among the spaces, the report sound space is related to the social conventions and rules on the book report or trailers and the other two spaces show how students use speech, music and sound effects to realize unique communicative purpose of the mode. For example, in the book sound space, a student spoke English with British accent to introduce *Harry Potter and the Sorcerer's Stone*. Another student inserted moaning noises of a zombie to introduce the book *Zombies*. In addition to this, students used music and sounds to represent the physical settings of the book or main characters' actions. These findings revealed that students who created multimodal digital book trailers purposefully layered different kinds of sound such as music, sound effects, and voice in order to evoke unique and deeper feelings in their audiences.

Based on these findings, it can be assumed that upper-elementary students may be able to use each mode and modal resources purposefully to achieve certain

communicative purpose. Therefore, the rubric for the evaluation of upper-elementary students' DMBRs should include criteria measuring if their use of each mode and modal resources is achieving the intended communicative purposes.

Assumption 3: Interplay between modes. In this section, I focus on the sign-maker's selection and configuration of modes, that is, the connections upper-elementary students made between different modes they chose in their multimodal products (Dalton et al., 2015; Mills, 2011; Ranker, 2012).

Dalton et al. (2015) reported findings from two studies on fifth-grade students' overall use of multiple modes and their design intentionality in their digital retellings. The first study analyzed 83 students' uses of modes in their folktale retellings that were created using a PPT story frame. The authors provided the pre-structured frame, which included directional slides on how to add sound and color and eight content slides with pre-added illustrations. The analysis of the modal design revealed that all students included visual design features such as background color, PPT design template and text fonts and written language. In terms of the sound mode, 85 used music, audio-recording or character's dialogue or monologue, narration of the written text, and sound effects. 66 students used animation, transitions within and between slides, even though it was not instructed and not the original feature of the e-folktale.

The second study conducted in the subsequent year with 14 students who participated in the first. The researchers explored students' intentionality of their design with multiple modes through the retrospective interviews. Findings from the interview analysis showed how the fifth-grade students configured different modes together. They

tried to make connections between sound and visual modes, between sound and story events, and between written language and other modes. For example, a student applied a dark color and an animation effect to a portion of text to show the sadness of the scene. The student's decision on each mode was the results of careful consideration of relationships among modes in and across scenes.

Ranker (2012) conducted a cross-case analysis study on how elementary school students in kindergarten, first and fifth grades use multiple modes and composing media while they compose multimodal texts in English Language Arts classrooms. Among three different cases, the case of two fifth grade students, Kendra and Nicole, composed multimodal text by using old and new media. The multimodal task was about three prominent females in US African-American civil rights history. To locate appropriate images of these women, two students typed keywords, searched images, and imported them into the digital video-production program. Then they edited the video by matching the imported images and the content they brainstormed. Searching for images and combining them to different modes also can be considered as an example of transmediation, which was the complex process considering interplay between modes within and across media. In sum, Ranker (2012) revealed the fifth-grade students' ability to consider intersemiotic relationships between a visual mode and other modes during the video production.

Mills (2011) studied young children's transformation of modes between and within print-based and digital media. Three eight-years-old students transformed pictures from a picture book into digital environment. The results showed that students'

background knowledge on the topic of the picture book influenced students to transform illustrations in the picture book differently. Mills (2011) reported “transmediation involved a process of continual adaptation of intentions for representing knowledge in response to the possibilities and limitation of the sign-making systems, including those embedded in digital software” (p. 64).

In sum, findings around the third assumption indicate that upper-elementary grade students have the ability to configure different modes by making connections between two or more different modes or transmediating one mode in one media into a different mode in another media. The take-away lesson from the findings for the rubric development would be the inclusion of criteria evaluating the relationship between at least two modes.

Assumption 4: Multimodality considers social conventions. The fourth assumption stresses the role of social conventions or rules related to the sign-making in multimodal compositions. Two empirical studies reported how teachers taught the social conventions or principles related to modes in multimodal texts to upper-elementary students (Mills & Exley, 2014; Shanahan, 2012).

Mills and Exley (2014) conducted a design-based study on teaching digital composition to 85 Year 4 students in Australian elementary classrooms and analyzed the time, space and text in the digital composition program by applying Bernstein’s pedagogic device as the theoretical lens. Although the focus of the study was not the teaching content, descriptions on the instructional content and evaluative methods used in the design-based study revealed that the teachers taught conventions related to composing

different digital texts such as web pages and storyboarding and conventions of linguistic modes. In this project, teachers explored several different types of digital texts and later focused on the web-page compositions since it includes a considerable amount of written language, which enables the web-page creations to be a part of the ELA writing instruction.

Shanahan (2012) investigated the relationship between the teacher's instruction on the various conventions of sound and her fifth-grade students' actual use of the sound conventions in their documentaries. Analyses of discussion between the teacher and students and students' documentaries identified that the teacher failed to provide adequate instruction on the conventions using sounds in documentaries. The teacher already knew how to use sounds such as narration, voice-over and music in the documentary genre to achieve communicative functions of them, but she did not teach them explicitly to students. As the results, most students did not include sound in their final documentaries.

The results of Mills and Exley (2014) and Shanahan (2012) indicate that social conventions around multimodality include conventions on genres, text types, use of each mode. If the conventions are not explicitly taught or reinforced through discussions or practices, upper-elementary grade students are less likely to apply the them into their multimodal compositions. Therefore, when we assess upper-elementary grade students' DMCs, we should not insist on students applying all conventions or principles regarding modal use unless they have previously been taught in class, as students' personal and subjective preferences for various modes can be a major source of creativity. Therefore,

the current study selectively included some conventions around composing DMBRs as the rubric criteria or descriptors after reviewing instructional contexts and students' products.

In summary, Jewitt's (2014) four assumptions provide a useful lens through which the empirical findings on upper-elementary grade students' digital-multimodal-composition performances can be reviewed. The assumptions also play a critical role to identify what should be considered in the digital-multimodal-book-review rubric for upper-elementary-grades students. The rubric should assess both linguistic and non-linguistic modes and their communicative purposes. It also need to capture the level of connections between at least two different modes. Finally, the rubric should include reasonable and realistic expectations on the students' knowledge on the conventions and principles governing modal use. In the method section, it will be addressed further how these characteristics were considered in the development of suitable rubrics for the assessment of fourth-grade students' DMBRs.

Rubric-Based Formative Assessment of Performances

Another key strand of this study is the rubric-based formative assessment of student performances. In this section, first, I will define DMC as a type of performance assessment and provide examples of formative assessments of performances. Then, I will introduce strategies of formative assessment and explain how the rubric can be used with the strategies. Finally, I will introduce two psychometric properties of performance

assessment, reliability and construct validity, and then issues around the inter-rater reliability and construct validity of rubric-based performance assessment will be explored.

Performance Assessment and Digital Multimodal Composition

Performance assessments have a comparatively shorter history in the United States than other educational assessments. They have been actively employed in classroom instruction since the late 1980s and early 1990s (Stecher, 2014, p. 17). The AERA, APA, and NCME (2014) define performance assessments as “assessments for which the test taker actually demonstrates the skills the test is intended to measure by doing tasks that require those skills” (p. 221). Darling-Hammond (2014) delineates them as “authentic assessments that require students to develop a product, response, analysis or problem solution that reflects the kind of reasoning or performance required beyond the classroom setting” (p. 12). For example, products of performance assessments include any artifacts that were created through certain performances. Essays, term papers, laboratory reports, drawings, and dance performances are some examples of these products (Johnson, Penny, & Gordon, 2009). On the contrary, performances encompass the process of performing some tasks such as oral assessments and demonstrations. That is, performance assessments evaluate actual skills or knowledge that the test-takers have by evaluating the product, the problem-solving process or both (Fitzpatrick & Morrison, 1971; Johnson et al., 2009).

Digital multimodal composition as performance assessment tasks. A DMC can be considered to be a task of performance assessment because its processes and products reflect the composer’s knowledge and skills, thus, the notion of performance

assessment is useful to understand the entire scope of DMC assessment. In detail, performance assessment of DMC tasks should be able to observe the composition process of digital multimodal texts as a way of checking the composer's individual or collaborative decisions on some design features and technical skills. Performance assessment of DMC tasks also evaluates final products; it allows the evaluator to measure the composer's understanding of modes and their affordances, the substance, and technical skills related to the digital multimodal texts. As Zammit (2014) argues, assessment of DMC needs to move beyond written-language-centered and high-stakes testing and employ more classroom-level performance or project-based assessments. Therefore, the current study aims to develop a scoring rubric as an in-class formative assessment tool. In the following section, five ways that formative assessments can be used in classrooms and the roles of a scoring rubric for formative assessment will be examined more closely.

Formative Assessment in Classrooms

The importance of formative assessment in classrooms was highlighted after a seminal review of empirical studies by Black and Wiliam (1998). Unlike summative assessment, which aims to evaluate student learning outcomes as measured against some standards or criteria, formative assessment works toward monitoring individual student's learning and providing feedback in order to improve their learning. For this reason, formative assessment is often called *assessment for learning* whereas summative assessment is called *assessment of learning* (Black & Wiliam, 1998; Black & Wiliam, 2009; Wiliam, 2011).

Five strategies of formative assessment. According to Black and Wiliam (2009, p. 8), classroom formative assessment consists of five key strategies:

1. Clarifying and sharing learning intentions and criteria for success;
2. Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding;
3. Providing feedback that moves learners forward;
4. Activating students as instructional resources for one another; and
5. Activating students as the owners of their own learning.

The first strategy—clarifying and sharing learning intentions and criteria for success—is related to the teacher and students. In classroom, the teacher explicitly presents the learning outcomes, or a scoring rubric including criteria for success, and explains their meaning. As the results of the teacher’s clarification, students understand and share the same learning intentions and criteria for success. The second and third strategies are teacher-centered activities. While using the second strategy, engineering effective classroom discussions and other learning tasks, the teacher provides questions for discussion, reading, and other in-class tasks to facilitate student achievement of the target level of performance or learning outcomes. The third strategy, providing feedback, allows teachers to constructively respond to students’ work with written feedback or checks on rubrics to help them improve their works. The fourth and fifth strategies, activating students either as instructional resources for one another or as the owners of their own learning, are related to either peer or individual student-centered activities. By evaluating their peers’ or their own works, students can reflect and gain a better understanding of

their learning. These five strategies provide a useful lens through which to view the uses of rubric in formative assessment. Table 2 presents the relationship between the five strategies and their agents (i.e., teacher, peer and individual learner) (Black & Wiliam, 2009, p. 8).

Table 2 Relationship between Five Strategies of Formative Assessment and Their Agents (Black & Wiliam, 2009, p. 8)

	Where the learner is going	Where the learner is right now	How to get there
Teacher	Strategy 1. Clarifying learning intentions and criteria for success	Strategy 2. Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding	Strategy 3. Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	Strategy 4. Activating students as instructional resources for one another	
Individual Learner	Understanding learning intentions and criteria for success	Strategy 5. Activating students as the owners of their own learning	

The uses of rubrics for formative assessment. Scoring rubrics have been widely used in formative assessment to promote student learning as well as to improve teacher instruction (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013). With regard to the five strategies of formative assessment, teacher-oriented rubrics help teachers clarify the content or criteria to be assessed and deliver their meanings to students. In order to explain the definitions and specific descriptors of the criteria, the teacher can invite

students to participate in classroom discussions on those components. If a teacher uses a student-friendly rubric, students can understand the content or criteria to be assessed during the formative assessment better and get feedback from the teacher in more structured ways. If the student-friendly rubric is used for peer- or self-evaluations, it can be a tool to activate students as instructional resources for one another or as the owners of their own learning (Andrade & Boulay, 2003). In sum, rubrics are useful for teachers to implement the five strategies of formative assessment.

The uses of rubric in the current study. The current study develops a rubric to help teachers to implement formative assessment in upper-elementary classrooms. Since the agent of the formative assessment in the current study is a teacher, three aforementioned strategies should be considered. However, this study only uses the rubric to evaluate the performance based on the final products and does not observe teachers' actual uses of the rubric for instructional purposes. Therefore, this study examines whether or not the rubric clarifies and shares learning intentions and criteria for success adequately.

Reliability and Construct Validity of Rubric-Based Performance Assessment

Psychometric properties of evaluative instruments. Reliability and validity are important psychometric properties of evaluative instruments. According to Johnson and colleagues (2009), reliability addresses “the consistency of examinees’ scores across such facets as occasions, tasks, and raters” (p. 22); validity refers to “the accuracy of our inferences based on our interpretation of the performance assessment scores” (p. 23). Validity thus concerns the nature of the trait supposedly assessed by the performance

assessment and the interpretation and use of that measure's scores, whereas reliability relates to the consistency of the measure; "That is, reliability is a relatively simple quantitative property of test responses, but validity is an issue more tied to psychological theory and to the implications of test scores" (Furr & Bacharach, 2014, p. 219). Since the current study aims to develop a rubric as an evaluative instrument of upper-elementary students' DMBRs, the consistency across raters (i.e., inter-rater reliability) and construct validity of the rubric should be examined carefully.

Inter-rater reliability of rubric-based assessment. Inter-rater reliability concerns the extent to which two or more raters assign the same score to the same criteria specified by a rubric or scoring criteria (Moskal & Leydens, 2000). According to Stemler (2004) and Jonsson and Svingby (2007), there are three different means of estimating inter-rater reliability of rubric-based assessment: estimates of consensus agreements, estimates of consistency, and estimates of measurement. The first two methods can be applied when there are two raters of the measurement, and the last one is used for the calculation of the inter-rater reliability between three or more raters. Since the current study deals with the consistency between two raters, in the following section, estimates of consensus agreements and consistency will be closely examined.

Estimates of consensus agreements. Among several ways of calculating consensus between raters, the most widely used approach is to use the percentage of total agreement. For example, if an analytic rubric contains five criteria that yield five separate scores and two raters grade 20 products independently, each rater would assign 100 scores. In this case, the percentage of the exactly agreed-upon scores between the two

raters is the inter-rater reliability. This is an easy way to calculate inter-rater reliability. However, obtaining a good level of agreement of above 70% is difficult (Stemler, 2004). For this reason, some scholars prefer to instead calculate the percentage of adjacent agreement between raters.

The second measure of consensus agreement is the percentage of adjacent agreement. This method considers all the discrepancies within one score point as representing agreement between raters and then calculates the consensus agreement between raters in the same way as the first method. If the adjacent agreement exceeds 90%, it is considered as a good level of consistency. One benefit of using this method is that higher inter-rater reliability is obtained than by the first method. However, discrepancies within one score point still indicate that raters interpreted the rubric slightly differently, so in that sense the method alone does not accurately represent inter-rater reliability. Reporting the exact and adjacent consensus agreements is useful “in diagnosing problems with judges’ interpretations of how to apply the rating scale” (Stemler, 2004, p. 5). In particular, a visual analysis of the crosstab table allows the researcher to identify the exact data of discrepancy.

The final method used to calculate consensus agreements is Cohen’s kappa. According to Hallgren (2012), “Cohen’s (1960) kappa and related kappa variants are commonly used for assessing inter-rater reliability for nominal variables” (p. 5). The scores generated by scoring rubrics can be considered as nominal data if points assigned to each scale play roles as categories. For example, if a rubric developer uses score one through four onto four scales—weak, moderate, good, strong—the numbers become

nominal data. Cohen's kappa ranges from -1 to +1, where +1 represents perfect agreement between the raters and 0 indicates that the amount of agreement is the result of random chance. If the kappa value is between 0.21 and 0.40, the level of agreement between raters is fair; 0.41-0.60 indicates moderate agreement, 0.61-0.80 means strong agreement, and 0.81-1 means almost perfect agreement (McHugh, 2012).

Estimates of consistency. Consistency estimates of inter-rater reliability are measured to test if the two raters interpret the phenomenon consistently based on their own understanding of the scale (Stemler, 2004). Although raters do not agree upon the scores, if the difference between their scores is predictable throughout similar cases, the two raters' gradings can be interpreted as consistent. The Pearson correlation coefficient is the most commonly-used statistic for assessing inter-rater reliability for interval and ratio variables (Jonsson & Svingby, 2007; Stemler, 2004), and Spearman's correlation coefficient, *rho*, is an appropriate means for the estimation of correlation between two raters' ordinal and rank-ordered scores assigned to student performances (Johnson, Penny, & Gordon, 2000). That is, in order to estimate Spearman's correlation coefficient between two raters, rank-ordered total scores of student performances are used. Interpretations of Spearman's correlation coefficients are similar to those of Pearson's correlation coefficients: they range from -1 to +1, where -1 represents perfect negative correlation, 0 refers to no correlation, and +1 means perfect positive correlation between two raters' scores. The effect sizes of the correlation coefficient are interpreted as follows: .10-.29 is small/weak, .30-.49 is medium/moderate, and .50 and above is large/strong effect sizes (Cohen, 1988; Field, 2009).

To sum up, in this study, percentages of exact and adjacent agreements, Cohen's kappa, and Spearman's *rho* correlation coefficient will be utilized in order to provide a more accurate picture of the reliability of the newly developed rubric. Although the above values are considered as the rules of thumb, Jonsson and Svingby (2007) emphasized that reliability in classroom assessments "is not of the same crucial importance as in large-scale assessments" because classroom assessments can easily be changed if they appear to be wrong (p. 135). On the basis of this logic, they concluded that "at least when the assessment is relatively low-stakes, lower levels of reliability can be considered acceptable" (p. 135). In this study, therefore, lower thresholds will be applied to make decisions on the acceptable inter-rater reliability and these will be presented in the methods section.

Argument-based approach to construct validity. This study conceptualizes validity in terms of Kane's (1992, 2006, 2013) argument-based approach to validity and Messick's (1989, 1995) comprehensive theory of construct validity. From these perspectives, validity is "associated with the interpretation assigned to test scores rather than with the test scores or the test" (Kane, 1992, p. 527; Messick, 1995).

The main point of Kane's (1992) argument-based approach to validity is that the validation of score interpretations involves a range of interpretive arguments associated with multiple inferences and assumptions. Inferences from test scores related to theoretical constructs depend on assumptions that are included in the theory regarding constructs. These assumptions are backed up by evidence; in order to validate a score interpretation, a rater or someone using a scoring tool needs to generate clear and

coherent arguments based on plausible assumptions. These plausible assumptions are supported by the use of multiple independent sources of evidence, a process known as triangulation. Possible counterarguments also need to be considered to ensure that the main argument is robust.

Messick (1995) suggested the use of construct validity as a unified conception of traditional content, criterion, and construct validities, arguing that “[the traditional] view is fragmented and incomplete, because it fails to take into account both evidence of the value implications of score meaning as a basis for action and the social consequences of score use” (p. 741). His proposed framework for construct validity introduced six distinguishable features of construct validity: content, substantive, structural, generalizable, external, and consequential aspects. Messick (1995) connected these six aspects of the construct validity of psychological assessment to Kane’s (1992) argument-based approach by pointing out, “one can prioritize the forms of validity evidence needed according to the points in the argument requiring justification or support” (p. 747). In other words, interpretive arguments around scores can be supported by providing evidence for these six aspects of construct validity, but the priority assigned to each may differ somewhat depending on the focus of the argument.

Construct validity of the rubric-based assessment. Given the discussion surrounding the argument-based approach to construct validity, this study anticipates that the scores that will be produced by the proposed rubric for DMBRs will provide an adequate measure of upper elementary-grade students’ competence in creating DMBRs. In order to support this argument, content and structure-related evidence of construct

validity is both critical and relevant in this context because the domains and criteria invoked in the scoring rubric will represent specific areas of competence. Hence, the content and structural aspects of construct validity must be subjected to an exhaustive examination.

Content aspect of construct validity. The content aspect of construct validity includes “evidence of content relevance, representativeness, and technical quality” (Messick, 1995, p. 745). To obtain such evidence, the accurate specification of the boundaries of DMC as a construct represents the key step in determining the success of the whole process. Hence, the knowledge, skills, attitudes, motives, and other attributes that will be assessed through the rubric must first be determined. This depends on the experts’ professional judgment and documentation of prior research in this area. In this study, a systematic literature review of the domains and criteria related to assessments of DMCs will serve as the content-related evidence for the construct validity.

Structural aspect of construct validity. The structural aspect of construct validity evaluates if the structure of scoring tool represents the structure of the construct domain (Messick, 1995). Either exploratory or confirmatory factor analysis can provide evidence of the assessment tool’s structural aspect of construct validity. In order to make the factor analysis results valuable, however, the structure of the assessment tool should be based on “a well-articulated theory of the classificatory relations among the different elements” (Morey, 2003, p. 402). Due to the limited number of students’ DMBRs, this study cannot perform any factor analysis. However, by checking the patterns of correlations between

any two criteria through Spearman's *rho* and between any two categories using Pearson correlation, internal structure of the developed rubric can be estimated.

For example, if two criteria under the same category are identified as weakly to moderately correlated, it can be interpreted that the two criteria share some commonalities under the same category, but also still measure different characteristics of the performance. On the other hand, if there are two criteria under different categories correlated from moderately to strongly, this can be problematic for two reasons. First, the rubric's category may not represent the actual structure of the construct. Second, the criteria may measure somewhat similar characteristics, so revision of the descriptors is required.

Chapter 3

METHODS

The purpose of this study is to identify key domains and criteria that comprise the construct of DMC and to develop a rubric for the reliable scoring and valid interpretation of scores for formative assessment of fourth-grade students' DMBRs. This mixed-methods study includes four distinctive phases of data collection and analysis: (1) systematic literature review, (2) rubric creation, and (3) rubric validation and (4) rubric revision. Specifically, the systematic literature review identified key domains and criteria of DMC assessment. At the second phase, a rubric has been collaboratively created by considering the identified key domains and criteria, genre characteristics of DMBR, and upper-elementary students' performances of DMC. During the third phase, inter-rater reliabilities—the level of agreement and consistency among raters—were analyzed quantitatively and concurrently with raters' opinions on the validity and usability of the developed rubric. Finally, analysis of the focus group interview detected possible threats to reliability and validity of the rubric and informed the rubric's revisions. Figure 1 presents the flowchart of the procedures in implementing this four-phase mixed-methods study.

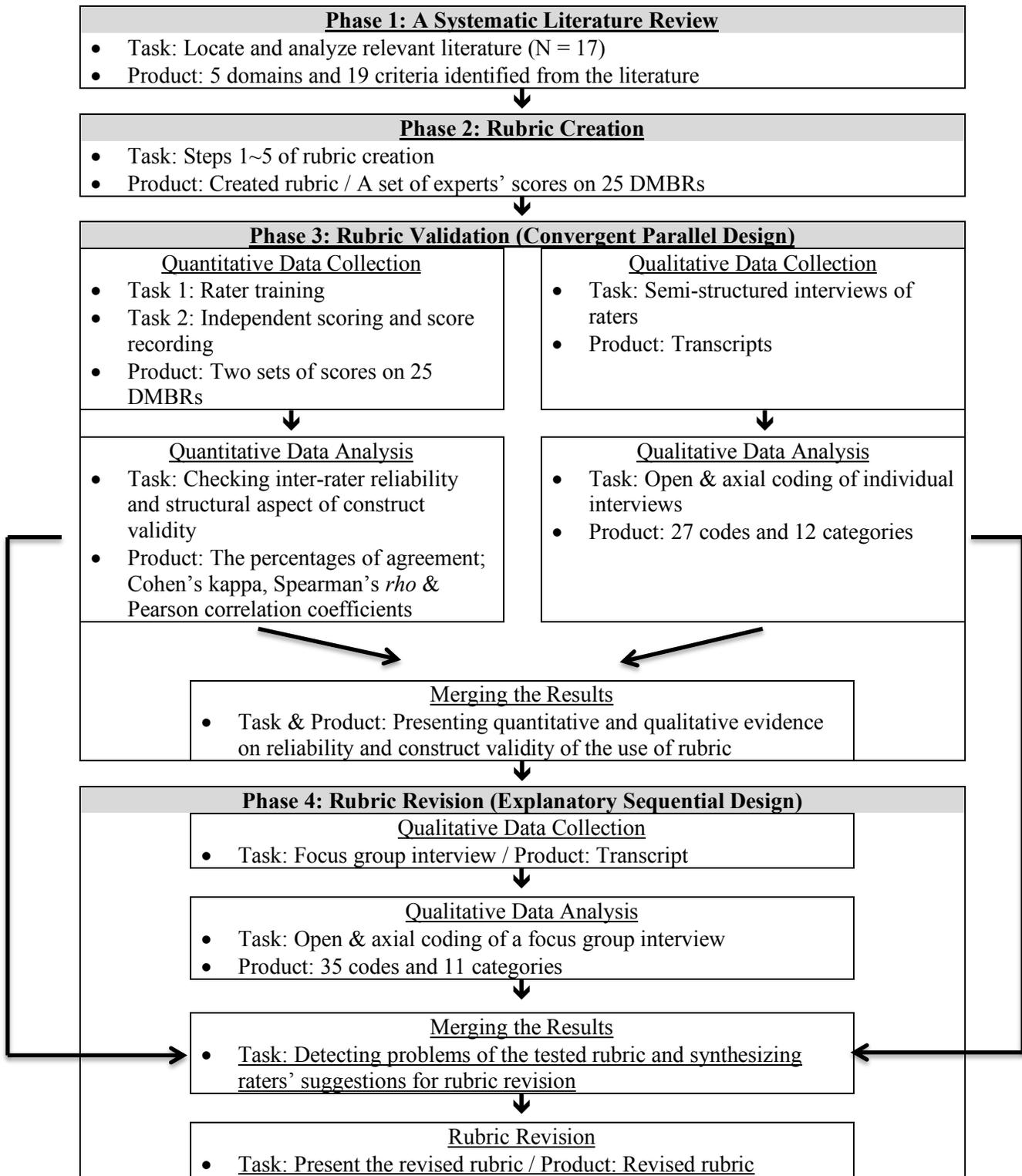


Figure 1 Flowchart of the Implementation of Four-Phase Mixed-Methods Study

The following research questions reflecting the phases guided this work:

1. What are the key domains and criteria that represent the construct of digital multimodal composition?
2. How might these domains and criteria be structured in a task- and grade-specific rubric to evaluate upper-elementary-grade students' digital multimodal book reviews?
3. To what extent does the use of the proposed rubric display evidence of inter-rater reliability, construct validity and usability, as indicated by raters' scores and interview feedback?
4. How does the evidence of inter-rater reliability, construct validity, and rater's feedback inform the process of rubric revision?

Participants

The participants of this rubric development study are one collaborator who helped me draft the rubric and two raters. Since I have developed the rubric for upper-elementary school teachers who would teach and assess students' DMBRs, the rubric was best drafted with a collaborator who has in-depth knowledge on the students during the second phase, rubric creation. On the other hand, the raters participated in the study as both research staffers and subjects during the third and fourth phases. Once they finished grading students' DMBRs, they provided feedback through individual and focus group interviews.

A Collaborator

In this study, the role of a collaborator was a research staff member who helped me refine the language in the rubric and grades students' digital multimodal book review products to generate the experts' scores. My decision to involve a collaborator was inspired by Stevens and Levi's (2013) advocacy for the collaborative construction of rubrics with others, such as teaching assistants, tutors or colleagues. Regardless of the characteristics of the collaborator, collaborative rubric construction ensures the new rubric will be a clearer and more effective tool for both assessment and instruction of target performances. This collaborative refining process for the draft rubric was essential because it provided a preliminary indication of the level of inter-rater reliability and the appropriateness of the rubric descriptors as the representative levels of upper-elementary-grades students' performances.

To recruit the collaborator, I set three requirements. The collaborator should (a) be a native English speaker with an advanced level of English reading and writing skills, (b) be a current educator with more than 5 years of teaching experience with upper-elementary-grade students, and (c) have previous experience with multimodal literacy practices. In order to recruit the collaborator, I created a flyer and sent it to several special interest groups (SIGs) such as the Technology in Literacy Education (TILE) of the International Literacy Association (ILA). After careful reviews of the applicants' resumes, I selected Abigail (pseudonym) as a collaborator.

At the time of recruitment, she was a K-12 Reading/Language Arts and ESL supervisor at a public-school district located in a mid-Atlantic state. As a native English

speaker who majored in English during college and Reading and Language Arts Leadership during her doctoral program, she began her career as a middle and high school English and writing teacher in 1995. In 2006, she changed her career path and became a literacy coach and reading specialist. Since then, she has had more than five years of experience with a wide variety of upper-elementary-grade teachers and students. In addition, as an active member of the regional literacy association and also the TILE-SIG of ILA, she has conducted and presented projects related to multimodal literacy practices in K-12 classrooms.

Raters

Since the rubric was developed for upper-elementary-grade teachers, recruiting teachers who were teaching upper-elementary students at the data collection moment was ideal. In order to be able to meet the raters face-to-face, the raters needed to be found in the local area. I established two qualifications for interested teachers: the teacher should (a) currently teach upper-elementary-grade students, and (b) have extensive experience using or integrating technologies for in-class literacy practices, including DMC.

Two teachers who have taken a course or have participated in research projects for faculty members at the University of Delaware expressed their willingness to participate in the current study. After careful reviews of their resumes, I decided to include Lindsey and Kristen (both pseudonyms) in this study as raters.

Rater 1, Lindsey, is currently a third-grade teacher and a professional educator of technology to teachers in a mid-Atlantic state. She has at least 10 years of teaching experience as a reading specialist and a general and special education teacher of

elementary grades. She also has extensive professional development experience in technology implementation and certificates for various technological tools such as SMART Board, Scratch, and the Chromebook. While collecting the data, she was working in a public elementary school where a 1:1 initiative with Chromebooks was implemented, so she was integrating them into her teaching on a daily basis. In addition to her experience with digital technology, she had participated in meetings for development and validation of state-level writing rubrics. Even though she had not assigned any tasks similar to the DMBRs to her students yet, the genre was very familiar to her, and she was eager to learn more about teaching and assessing of DMBRs.

Rater 2, Kristen, is currently a second-grade teacher in a public school located in a mid-Atlantic state. While the data was collected, she was working as a middle-grade special education teacher in a school for students with learning difficulties where she has worked for 8 years. She also has extensive experience with technology integration into classroom teaching because her school emphasizes the uses of digital devices such as iPads and laptops for classroom teaching and learning. She is particularly familiar with the Apple iMovie program and has experience creating digital multimodal texts with her students.

Even though the two teachers' teaching experiences are different, both understand upper-elementary-grade students well and also have enough experience and background knowledge to understand and use the rubric for the evaluation of DMBRs.

Data for Scoring: Fourth-Grade Students' Digital Multimodal Book Reviews

This dissertation study used fourth-grade students' DMBRs ($N = 30$) that were obtained from a larger study as the data for rubric development on the genre. To understand why the data were used to develop a new rubric for assessment of DMBRs, several contexts of the larger study need to be discussed: research purposes, participants, characteristics of the task, and procedures to collect DMBRs.

Research Purpose of the Larger Study

The purpose of the larger study was to explore fourth-grade students' transmediation of content from a traditional graphic organizer to DMBRs. The teacher was a former masters level-student at the university where the principal investigator (PI) of the larger study was working. She had taken a course on multimodality before the data collection. The PI provided four iPads to the teacher in the fall semester of 2012. Before the previous study's data were collected, the teacher utilized one iPad for lesson planning and class demonstrations, and the remaining three iPads were used by small groups of four or five students. As a research assistant, I observed 10 occasions of classroom instruction that integrated iPads over the course of the fall 2012 semester, and the DMBRs were collected in spring 2013.

Three research questions guided this study: (1) How do students represent content on a print-based graphic organizer knowing they will later represent the content in an iMovie?; (2) How do students transmediate the modes represented in the print-based graphic organizer to the iMovie?; and (3) What are the advantages and challenges associated with using print-based and digital writing tools to compose multimodal texts?

Fourth-Grade Students

Fourth-grade students in a female teacher's English Language Arts classroom ($N = 30$) in a public elementary school located in a Mid-Atlantic state participated in this study. Data on the racial/ethnic composition of the classroom was not collected. In terms of the racial/ethnic composition of the school, 46.2% of students were White, 32.9% African American, 12.4% Hispanic/Latino, 3.4% Asian, 0.2% American Indian, and 4.9% of students were multi-racial. Among these students, 48.3% were from low-income families, 7.8% were receiving special education and 3.4% of them were English language learners (ELLs). The total number of fourth-grade students in the school in spring 2013 was 80.

Regarding fourth-grade students' academic achievements during the 2012-13 academic year, approximately 65% of fourth-grade students in the school were meeting the state reading standards, whereas 72% of fourth-grade students in the state were meeting state reading standards. The school-specific national assessment data was not available, but the 30 fourth-grade students' average academic proficiencies in reading could be assumed by comparing the state average and national average on NAEP Reading tests administered in 2011. Students attending fourth grade in the state schools performed slightly better than the average of all national public school fourth-graders. In detail, 28% of the state's students were below the basic level, whereas 34% of the students across the nation were below the basic level. Since the elementary school where the DMBRs were collected was almost at or slightly below the state average achievements in fourth grade in reading during 2012-13, it can be assumed that the group

of 30 students who composed the book reviews might have literacy proficiencies at levels around the national average.

Characteristics of the Task

A DMBR was chosen as the target task for two reasons. First, the book review was a familiar genre to the majority of students in the class. During the previous semester, they already gained some experience in writing book reviews with graphic organizers. The graphic organizer for the DMBR project had six components: introduction, brief summary, one thing I liked, one thing I did not like, recommendation, and conclusion. Graphic organizers with this structure are easily found in educational resources such as resource books published by textbook companies and websites listing classroom resources for teachers. Second, the teacher had created a multimodal book review for *James and the Giant Peach* in her master's level course on multimodality that she could use as a model or an example for her students.

At the beginning of the spring 2013 semester, the teacher provided instruction on the DMBR task by following the below mentioned procedures. Before data collection, the teacher read aloud a book called *Frindle* to students for several days. After the teacher had finished reading the book to her students, she distributed a print-based graphic organizer for the multimodal book review task and provided a mini lesson explaining types of modes, different resources that each mode had and the importance of using multiple modes for the creation of DMBRs. Then she demonstrated how to use the graphic organizer to plan the use of modes and to organize the content of the book review. She also demonstrated how to use the iMovie app to create the digital book

review. Finally, she presented her DMBR for the book *James and the Giant Peach* as a model. Students watched the model book review two times and listened to the teacher's explanations about the roles of modes and the teacher's intention of using specific modes in certain scenes.

Procedures to Collect Students' Digital Multimodal Book Reviews

Students in one fourth-grade classroom ($N = 30$) created DMBRs independently using iMovie on the iPad. Each student was brought from the classroom to a quiet room. To help make the students' composition of the digital book reviews easier, the teacher downloaded some images from the internet and saved them on the iPad. Students could also access to the internet to find different images if they desired. While a student created his or her book review on iMovie while considering the information on the graphic organizer, I observed the composing processes of the student. When the student had difficulties or asked questions, I occasionally guided the student in the usage of the iPad and iMovie app. Students spent time ranging from 30 minutes to 1 hour to create their DMBRs. Students' finished products were saved on the iPad and also immediately exported to an external hard drive.

Procedures and Data Analyses

Phase 1: A Systematic Literature Review

The main task of this phase was to identify key domains and criteria of DMC assessment from existing literature. This phase was essential to obtain content-related evidence of validity of DMC assessments. Domains and criteria of DMC assessment that

will be shared at the first section of the results were identified by going through the two steps: (1) locating literature and (2) analyses of literature.

Step 1: Locating literature. To locate relevant empirical studies on domains and criteria of DMC assessment, systematic searches were made of five different databases: Education Full Text, Education Resources Information Center (ERIC), Library and Information Science Abstracts (LISA), Linguistics and Language Behavior Abstracts (LLBA), and PsycINFO. Combined searches in Education Full Text using the search terms “multimodal composition” AND “assessment,” “multimodal composing” AND “assessment,” OR “digital writing” AND “assessment” yielded 44 publications. By reading the abstracts of each, eight relevant studies were identified. A search of the remaining four databases using the ProQuest search engine and using the same search terms and logic retrieved a further 86 empirical studies. Based on the information in the abstracts, 10 new relevant studies were added to the list. Reviewing the reference lists in the 18 studies and manual searches of book chapters added 26 more studies to the final list for a final total of 44 relevant papers in the prior literature.

Step 2: Analyses of literature. Among the 44 papers identified in the literature, only one study systematically established the domains of DMC assessment (Eidman-Aadahl et al., 2013). Five domains—artifact, context, substance, process management and technique, and habits of mind—and their definitions were used as the *a priori* code for the categorization of criteria in this study.

For the first stage of the categorization, the remaining 43 studies were carefully reviewed to find relevant studies discussing evaluative criteria. After this process, only 18

studies remained on the list. Specifically, 10 studies (rubric literature) presented criteria in scoring rubrics for DMC (i.e., Borton & Hout, 2007; Brown, 2013; Burnett, Frazee, Hanggi, & Madden, 2014; Howell, Reinking, & Kaminski, 2015; Hung et al., 2013; Husbye & Rust, 2014; Morain & Swarts, 2012; Ostenson, 2013; Towndrow, Nelson, & Yusuf, 2013; Vassilikopoulou, Retalis, Nezi, & Boloudakis, 2011), while the other eight studies (non-rubric literature) discussed one or more evaluative questions or criteria for DMC without presenting scoring rubrics (i.e., Adsanatham, 2012; Anderson et al., 2006; Levy & Kimber, 2009; Selfe & Selfe, 2008; Sorapure, 2006; Wierszewski, 2013; Yancey, 2004; Yu, 2014).

Next, Dinsmore, Alexander, and Loughlin's (2008) framework was applied to evaluate the definitional clarity of the concepts. An explicit (E) code was assigned to a criterion if the author provided exact wording in the definition of each term, while an implicit (I) code was assigned if words, phrases, or references that alluded to the meaning of a criterion was used in the text, and, if no definition of the criterion was provided, an absent (A) code was assigned. Most criteria presented in scoring rubrics were categorized as explicitly clear concepts except for the ones in the rubric of Vassilikopoulou et al. (2011). Unlike other studies presenting their rubrics with clear definition of criteria and descriptors of performance levels, the Vassilikopoulou et al. (2011) rubric listed only criteria without any explanations. For this reason, 16 criteria in this rubric were coded as absent. Implicit codes were assigned to 33 criteria presented in Adsanatham (2012), Anderson et al., (2006), Selfe and Selfe (2008), Wierszewski (2013) and Yancey (2004). For example, Yancey (2004) implicitly defined the meaning of coherence by explaining

its relationship with patterns: “Patterns are one way to talk about coherence in digital texts. Another way to think about this patterning and how the pieces within a pattern connect ... is to think in terms of weaving” (p. 90). To sum up, a total of 129 criteria were categorized into 80 explicit, 33 implicit, and 16 absent criteria. This resulted in the removal of the Vassilikopoulou et al. (2011) study since it did not present either explicit or implicit definitions of the criteria. Thus, a total of 111 criteria from 17 studies remained on the list.

In the third stage, the constant comparative method (Strauss & Corbin, 1998) was applied to group the criteria according to their different names and categorize them under appropriate domains. First, the definition of each domain was carefully considered, after which an explicit or implicit definition was determined. Wherever similarities were found among the definitions provided for different criteria, these were grouped together. This was an essential stage of the research process because scholars often used different terms to represent similar concepts. The explicit definitions of domains and examples provided by the Eidman-Aadahl et al., (2013) were then applied, and explicit or implicit definitions of criteria were categorized into individual groups. This constant comparison of the definitions of domains and criteria in some cases led to the renaming or relocation of certain criteria. Finally, the definitions of the criteria in each group were synthesized, and a new explicit definition assigned to each new criterion. The finalized set of 19 criteria will be introduced in the first part of the results section.

Phase 2: Rubric Creation

In the second phase, a rubric was created in six steps, which have been modified from the steps of rubric development that were previously suggested by Arter and Chappuis (2006), Arter and McTighe (2001), and Stevens and Levi (2013). Since I will present actual tasks that I completed in each step at the second section of the results, I, here, explain which aspects I considered to create a rubric and why and how I modified the existing steps.

Decisions on four aspects of the scoring rubric. Several literature on scoring rubric (e.g., Arter & Chappuis, 2006; Arter & McTighe, 2001; Moskal, 2000) highlighted that those developing new scoring rubrics for assessment must make decisions on four aspects: the type (holistic vs. analytic), the specificity (general vs. genre or task-specific), the number of score points possible, and the audience (teacher vs. students).

Type. The first choice to be made about the scoring rubric is its type. There are two distinctively different types of scoring rubrics: holistic and analytic. I explain characteristics of each based on a consideration of writing rubrics. A holistic writing rubric asks a rater to assign a single holistic score to a student's writing performance after considering various criteria such as content, organization, sentence structure, word choice, and mechanics (Camara, 2003; Myers, 1980; Spandel, 2009; Wolcott & Legg, 1998). When raters read a student's writing product, they compare it to descriptions or anchor papers for each score and then assign a single overall score based on their "general impression" (Wolcott & Legg, 1998, p. 71). Using a holistic scoring rubric is beneficial for teachers because it is the most efficient way of scoring students' writing

products in a limited time (Arter & McTighe, 2001; Stevens & Levi, 2013). However, holistic scores do not allow teachers to diagnose each student's specific writing weaknesses; students who receive the same score for their writing performance may suffer from different weaknesses, but a holistic rubric cannot diagnose them, so providing specific feedback on those weaknesses is not possible.

Unlike holistic rubrics, analytical scoring rubrics include the consideration of several traits or criteria. Using this rubric, raters can assign a score for each trait or criterion, as well as a composite score covering all the traits for each student's writing product. One of the most widely used analytical scoring rubrics for writing is Spandel's (2009) six-traits writing rubric, which is based on six traits for teaching and assessing writing: ideas, organization, voice, word choice, sentence fluency, and conventions. Each trait is described in detail for each score level. The advantage of using analytical scoring rubrics is that they facilitate a teacher's ability to diagnose students' writing weaknesses and provide feedback. However, they also require raters to devote much more time to assessing each student's writing, so most summative, high-stakes and standardized writing assessments do not use this method. This rubric is more appropriate for formative and instructional assessments in the classroom.

Specificity. The second decision that must be made relates to the scoring rubric's specificity. A rubric can be either generic, for any DMC task, or task-specific. Using generic rubrics is beneficial when the DMC task shares some similarities with other DMC tasks. By using the generic rubric for several similar DMC tasks, teachers can provide consistent feedback on students' DMC performance and save time since they do

not need to develop new rubrics for each task. However, there are some drawbacks to using generic rubrics. First, the descriptions of domains or criteria in the rubric are necessarily more abstract, so raters need more time to interpret and apply the abstract meaning to each specific task. On the other hand, using a task-specific rubric offers two major advantages, as it saves raters' time and there is likely to be greater consistency between raters. These advantages are due to the inherent specificity and straightforwardness of the task-specific rubric.

Scoring points. The third choice concerns the number of scoring points. Arter and McTighe (2001) recommends using 3 to 6 points for rubrics, arguing that this is sufficient to describe student achievement within a single grade level. Fewer than 3 points makes it difficult to distinguish quality, while more than 6 is usually too complex. However, if the rubric is to be utilized to track student development for a skill over time, the use of 6 points or more is appropriate. Some scholars prefer avoiding use of an odd number of score points because this can encourage a tendency to regress to the mean. For example, if there are five points in a scoring rubric, raters are highly likely to choose the third point rather than the second or fourth.

Audience. The final decision needs to be made on the audiences of the rubric. In the classroom formative assessment setting, the audiences of the rubric can either be teachers or students; depending on the audience, the language of the rubric (e.g., tone and word choices) will be different. If the rubric is only for the teacher's evaluation of students' performances, it can contain some jargon and adult-level words. On the

contrary, if the rubric is developed for peer- or self-evaluation, it should be written in plain language and include grade-appropriate words.

To sum, the four aspects listed above should be considered at the very beginning of the rubric creation. Decisions about the four aspects of scoring rubrics were made and presented in the results section.

Reviewing existing procedures for rubric creation. Arter and McTighe (2001) and Stevens and Levi (2013) both proposed a set of procedures for rubric development. Even though their procedures may appear different at first glance, they actually share more similarities than differences. For example, Arter and McTighe (2001) suggested six steps for rubric development. Step 1 *gathers* samples of student performance, then Step 2 *sorts* the student work into groups and states the rationale for each decision. In this step, the rubric developer places the student work into three piles—strong, medium and weak—and notes why he or she has placed a piece of work in a particular pile. If the developer creates a 3-point scale rubric, these reasons will be used as the basis of each scale descriptor. If the rubric developer instead creates a 4-point scale rubric, the work in the middle pile will need to be subdivided into upper-middle (stronger) and lower-middle (weaker) piles. Step 3 is to *cluster* the reasons into the traits or important dimensions of the performance. In this step, the developer lists all the reasons he or she gave for placing the student work in each pile. By reading these carefully, the developer can identify key words that can be applied to all the student work in each pile. In step 4, the developer writes a value-neutral *definition* of each trait. Considering the reasons given in step 2 and the key words identified in step 3 enables the developer to write a considered definition

of each trait. In this step, the developer does not take into account any descriptors for each score point on each trait. In step 5, the rubric developer finds samples of student performance that *illustrate* each score point for each trait. For example, if the developer wants to describe a score of 4 for coherence, he or she rereads the definition of coherence and finds the best possible sample from the strong pile and uses this to write a descriptor of a 4-point coherence score. In the sixth and final step, the developer uses the preliminary or draft rubric to score the student work, continuing to *refine* it by adding or modifying descriptors.

Although the six steps described above inform the actual procedures used to review the student work and highlight the importance of the revision process, they do not include any review of the related literature and fail to explicitly consider any learning objectives. These drawbacks can be addressed by incorporating several of the stages suggested by Stevens and Levi (2013). Unlike Arter and McTighe (2001), Stevens and Levi (2013) begin their rubric development procedure with a consideration of learning objectives, for which they suggest four stages. In stage 1, *reflecting*, the rubric developer reflects on both the learning objectives for the performance task and the related course instruction. For this study, the characteristics, research and instructional contexts of the DMBRs, and the domains and criteria identified from the literature on DMC assessment can be applied to inform the learning objectives of the task. In stage 2, *listing*, the developer writes down the learning objectives that he or she expects to see in students' performances. By considering the students' levels of prior knowledge or abilities, the developer adds descriptors of the highest level of performance that can be expected under

each learning objective. In stage 3, *grouping and labeling*, the developer groups similar performance expectations together and labels each group accordingly. This process is inherently iterative because sometimes performance descriptors will be related to more than one group or not be included any of the existing groups at all (Stevens & Levi, 2013). Once the developer has grouped all the performance descriptors stably, he or she must label them carefully as these labels will become the domains or criteria of the final rubric. In the final stage, *application*, the rubric developer transfers the groups and labels onto the rubric grid. Labels then become the domains or criteria of the rubric and the grouped descriptors of the highest-level performances become the descriptors of the domains or criteria.

Connecting and modifying existing procedures of rubric creation. Although Stevens and Levi's (2013) four stages were both abstract and difficult to replicate, all could be made to work relatively well if applied in conjunction with steps 1 to 5 of Arter and McTighe's (2001) model. In particular, Stevens and Levi's (2013) Stage 4, application, could be specified with steps 4, 5, and 6 of Arter and McTighe's (2001) model. The need for a collaborative revision in the final stage of the rubric development was not mentioned in the existing literature, but a revision step was clearly invaluable and so was added to the present study. In particular, I thought that beginning the collaborative refining by using a metarubric (Arter & Chappuis, 2006; Arter & McTighe, 2001) would be effective to evaluate the coverage/organization and clarity of the rubric. For this reason, refining with a collaborator included as the last step. Before going through the steps, decisions on the four aspects of scoring rubrics had to be made first.

Table 3 summarizes each step of the modified rubric creation procedure. Detailed rubric creation procedures followed by the modified steps will be shared in the second part of the results section.

Table 3 Modified Steps of Rubric Creation

Modified Steps	Original steps or stages in the Literature	Specific Tasks
Step 1: Making decisions on four aspects of scoring rubrics	<ul style="list-style-type: none"> Chapter 2: Choices, Choices (Arter & McTighe, 2001) 	<ul style="list-style-type: none"> Type (holistic vs. analytic) Level of specificity (generic vs. task-specific) The number of scoring points Audience (teacher vs. students)
Step 2: Reflecting	<ul style="list-style-type: none"> Stage 1: Reflecting (Stevens & Levi, 2013) Step 1: Gather samples of student performance (Arter & McTighe, 2001) 	<ul style="list-style-type: none"> Reflect learning objectives for the performance task for DMBRs Review domains and criteria identified from the literature Prepare 30 products (4th grade students' DMBRs)
Step 3: Listing	<ul style="list-style-type: none"> Stage 2: Listing (Stevens & Levi, 2013) Step 2: Sort student work into groups and write down the rationale guiding the choices (Arter & McTighe, 2001) 	<ul style="list-style-type: none"> List relevant learning objectives and related domains and criteria Write down characteristics of each student's DMBR under each criterion
Step 4: Grouping and labeling	<ul style="list-style-type: none"> Stage 3: Grouping and labeling (Stevens & Levi, 2013) Step 3: Cluster the reasons into traits or important dimensions of performance (Arter & McTighe, 2001) 	<ul style="list-style-type: none"> Group the characteristics Assign labels to each group (use domains and criteria from literature. If necessary, create new criteria)
Step 5: Writing definitions	<ul style="list-style-type: none"> Step 4: Write a value-neutral definition of each trait 	<ul style="list-style-type: none"> Use the label as a domain or criteria for the rubric Define each domain or criteria

and descriptors	<ul style="list-style-type: none"> • Step 5: Find samples of student performance that illustrate each score point on each trait (Arter & McTighe, 2001) • Stage 4: Application (Stevens & Levi, 2013) 	<ul style="list-style-type: none"> • Write descriptors of each level under each criterion by reviewing the grouped characteristics • Transfer the domain and criteria, definitions, and descriptors on the rubric grid
Step 6: Refining with a collaborator	<ul style="list-style-type: none"> • Step 6: Continue to refine (Arter & McTighe, 2001) • Using a metarubric (Arter & Chappuis, 2006; Arter & McTighe, 2001) 	<ul style="list-style-type: none"> • Evaluate the coverage/organization and clarity of the rubric using a metarubric with the collaborator • Refine the structure and/or language of the rubric • Grade two products separately and compare the scores awarded by the collaborator and me. If there are criteria that demonstrate a score difference of more than 1 point, discuss the reasons for this and resolve the problem. • Continue the process until we grade all DMBRs ($N = 30$) and the online book reviews for training ($n = 8$). • Record agreed scores (the experts' scores) on the 38 products.

Phase 3: Rubric Validation (Convergent Parallel Design)

The rubric-validation phase included tasks related to checking inter-rater reliability and construct validity of the use of rubric, and usability of the rubric. The convergent parallel design of mixed methods (Creswell & Plano Clark, 2010) was used in this phase to “triangulate” some aspects of reliability and validity of the use of rubric. According to this design, “the researcher collects and analyzes both quantitative and

qualitative data during the same phase of the research process and then merges the two sets of results into an overall interpretation” (Creswell & Plano Clark, 2010, p. 77). To be specific, quantitative data—two sets of scores on 25 students’ DMBRs—were gathered and analyzed to explain the rubric’s inter-rater reliabilities and structural aspect of construct validity. At the same time, qualitative data—semi-structured individual interviews of raters—informed the content and structural aspects of construct validity and the utility of the developed rubric. The following section presented the procedures related to the quantitative and qualitative data collections and analyses.

Quantitative data collection 1: Rater training. Training raters is an essential part of rubric development (Knoch, Read, & Randow, 2007). Extensive rater training helps them avoid capricious subjectivity and improves agreement and inter-rater reliability among a group of raters (Graham, Milanowski, & Miller, 2012; Wren, 2009). According to Wren (2009, p. 7), rater training involves the following steps:

- Step 1: Orientation to the assessment task
- Step 2: Clarification of the scoring criteria
- Step 3: Practice scoring
- Step 4: Protocol revision
- Step 5: Score recording
- Step 6: Documenting rater reliability

Although Wren (2009) outlines rater training in six steps, the rater training in the current study only indicated steps 1 through 4. Wren’s (2009) step 5 described works completed at the level of *quantitative data collection 2: independent scoring and score*

recording of the current study. Step 6 was related to the *quantitative data analysis: documenting inter-rater reliability and structural aspect of construct validity* of the current study. A more detailed discussion of the tasks involved in each training step was provided below.

Step 1: Orientation for the assessment task. Steps 1 through 3 were completed during a 3-hour offline training session. At the beginning of the offline scoring training, the raters were introduced to the purpose of the dissertation study and the concepts of mode, multimodality, and DMC. Then, the DMBR task and instructional setting of the task in the fourth-grade classroom were explained.

Step 2: Clarification of the scoring criteria. In this step, the definitions of each domain and criterion, and descriptors of each level in the rubric #6 were explained. Presentation slides used for the explanation of the above-mentioned information were included in Appendix A. To help the raters understand all the components of the rubric, one DMBR was selected and the video played twice, with the researcher providing a running commentary to demonstrate how the scoring decisions on each criterion were made.

Step 3: Practice scoring. Raters watched two students' book reviews using separate computers. To help with their interpretations of each criterion, a scoring protocol and supplemental materials for scoring (Appendix B) were provided. After they finished grading, their scores on each criterion for the two book reviews were discussed, criterion by criterion. Any discrepancy between the experts' scores and those of each rater, or between two raters' scores, was resolved through discussion. Five of the 30 book reviews

created in the fourth-grade classroom and eight online book reviews (Appendix C) were prepared for this practice scoring. The initial plan was to repeat the guided practice until the exact consensus agreements reached 60% or greater, and the adjacent agreements were at or above 80% between the experts and each rater, and between raters (Jonsson & Svingby, 2007). Due to the time limit of the offline training session, however, we moved to the independent practice scoring. Based on our discussion of the three products during the offline training session (i.e., a modeled product and two guided practiced products), two raters graded five products independently in three weeks. Three weeks later, we had a second one-hour training session via Skype. Comparisons of the experts' and two raters' scores indicated that all criteria except criterion 4 and 10 met the above-presented agreement rates. In case of criterion 4, two raters misinterpreted the first descriptors presented in each level. When grading criterion 10, rater 2 consistently graded it 1 point lower than the experts and rater 1. These differences between raters were resolved through discussion.

Step 4: Protocol revision. After the second scoring training session, the scoring protocol and the rubric #6 were revised in order to reflect discussion with the raters during the scoring practice. The major changes made to the revised scoring protocol (Appendix D) are as follows. First, a suggested schedule of grading was included. Second, more detailed examples about resolution and Ken Burns under the criterion 1 were added. For more accurate interpretation of criterion 4, examples of the types of errors that might be found in students' written language in the DMBRs were presented. A descriptor of level 1 of criterion 6 was clarified. Finally, examples for each level of

criterion 11 were added. Before the revised scoring protocol was used, raters were asked to review it in order to confirm that the document adequately included all discussed changes.

Quantitative data collection 2: Independent scoring and score recording.

Each rater graded 25 student products over a three-week period using rubric #7. The revised protocol was provided. It included detailed directions regarding the independent scoring process, such as the number of book reviews that they need to grade at once, the maximum interval between grading, and so on. They were received print- and electronic versions of the worksheet that include spaces for score recording and a scoring log (Appendix E). The raters were asked to complete the scoring log space by recording frequency counts, their analysis of the product on each criterion and any other thoughts they had during the grading procedures. These logs would be used as prompts during the subsequent rater interviews.

Quantitative data analysis: Documenting inter-rater reliability and the structural aspect of construct validity.

Inter-rater reliability. The inter-rater reliability between the experts' scores and each rater's scores, and between the two raters' scores, was calculated with a focus on two different aspects of the inter-rater reliability: consensus agreement and consistency estimates. Rather than comparing total scores given by each rater, the scores given for each of the 11 criteria for each DMBR were compared. This method helps identify more specific agreements and disagreements among raters. Three methods were used for the consensus agreement calculation—the percentage of exact agreement, the percentage of

adjacent agreement within ± 1 -point differences, and Cohen's kappa—to examine interrater reliability between the experts' scores and each rater's scores, and between the two raters' scores. To measure the consistency estimates, Spearman's *rho* correlation coefficient for rank-order items was calculated. The following thresholds for each measurement were applied. First, exact agreement exceeding 60% and adjacent agreements exceeding 80% were considered acceptable. These thresholds were 10% lower than those of high-stakes standardized tests (Brown, Glasswell, & Harland, 2004; Stemler, 2004). In case of Cohen's kappa, .40, the minimum for agreement beyond chance, was used as the benchmark, as Jonsson and Svingby (2007) used it. Finally, a Spearman's *rho* correlation coefficient lower than .50 was considered problematic as Gearhart, Herman, Novak, and Wolf (1995) stated.

Structural aspect of construct validity. To investigate the structural validity of the developed rubric, the relationship among the criteria per rater were examined using Spearman's *rho* correlation coefficient, and the relationship among the categories per person were presented through Pearson correlation coefficients. Generally, Spearman's *rho* and Pearson correlation coefficients are not the ideal options to examine the structural aspect of the construct validity; exploratory or confirmatory factor analyses are better ways for the aspect. However, since factor analyses require at least 100 participants (or artifacts) to have a statistical power (Gorsuch, 1983; Klein, 1979), only Spearman's *rho* and Pearson correlation coefficients are appropriate to be performed for this study with 25 participants. For both coefficients, an effect size .50 was used as the threshold of the strong correlation (Cohen, 1988; Field, 2009). To be specific, two criteria or

categories that are correlated more than .50 were considered problematic because strong correlation between the two criteria or categories indicated that they might measure similar characteristics of certain performances beyond chance.

Qualitative data collection: semi-structured individual rater interviews. As a way of obtaining social validity of the developed rubric, two raters were interviewed after their independent scorings of students' DMBRs. In order to obtain the raters' unbiased opinion based on the quantitative analyses, I conducted interviews one week after they completed the independent grading. A semi-structured interview questionnaire (Appendix F) was used to ask their opinions about the utility and appropriateness of the developed rubric. Raters referred to their scoring logs to recall their thoughts while using the rubric.

Qualitative data analysis: open and axial coding of the interviews. All interviews were recorded using a digital voice recorder; all audio and transcription files were stored on a password-protected hard drive and stored in a locked filing cabinet. Recorded semi-structured interviews were transcribed and names of the participants were replaced with pseudonyms. To code the transcribed interviews, two levels of analyses involved in the constant comparative method, open coding and axial coding (Strauss & Corbin, 1998), were applied. During the open coding phase, I read the transcript of each rater's interview and assigned codes by continually thinking of the semi-structured interview questions. In the axial coding phase, I compared my codes from the two raters' interviews and grouped several relevant codes into new categories. Then, the core categories were compared to the quantitative results from the inter-rater reliability and

structural validity of the rubric. As a result of both open and axial coding, 25 codes and 7 categories were identified.

Merging the results. Findings from both quantitative and qualitative analyses of the data were compared to identify patterns. Then they were laid out to display the evidence of inter-rater reliability, construct validity, and usability of the developed rubric.

Phase 4: Rubric Revision (Explanatory Sequential Design)

In this phase, the explanatory sequential design of mixed methods was applied. This design refers to the studies that begin with a quantitative phase and follow up with a qualitative phase on specific results (Creswell & Plano Clark, 2010). The data for the fourth phase came from a focus group interview with the raters; the main purpose of this focus group was to explore the reasons for some problematic quantitative results and generate some ideas for rubric revision.

Qualitative data collection: Focus group interview. The focus group interview was held after the synthetic analyses of both quantitative and qualitative data on the reliability, construct validity and usability of the rubric. The focus group interview was crucial in order to understand the causes of inconsistent interpretations of the rubric by raters and generate ideas for rubric revision. Instead of preparing semi-structured interview questions, I presented the quantitative analyses results and asked the raters to explain their interpretations of criteria 10 and 11 while they referred to their scoring logs. I also asked them to suggest ideas for revisions of criteria 3, 4 and 6.

Qualitative data analysis: Focus group interview. Open and axial coding procedures were applied to the analysis of the focus group interview among the three

levels of coding in the constant comparative method (Strauss & Corbin, 1998). The findings from the open coding were compared to the codes and categories obtained from the analyses of individual rater interviews and resulted in the revision of codes and categories. A total of 35 codes and 11 categories were identified through the analysis. Appendix G presents the codebook, including definitions of codes and categories and related examples.

Merging the results. Findings from both quantitative and qualitative data analyses conducted in phase 3 were combined with the findings of the focus group interview to explore the causes of different interpretations of the rubric, the rubric’s content and structural problems, and raters’ suggestions for the revision of the rubric.

Rubric revision. I incorporated the merged results and presented a newly revised rubric. More details about actual revisions will be presented in the results chapter.

Table 4 shows the relationships among the research questions, purposes, tasks, and data of this study.

Table 4 Research Questions, Purposes, Tasks and Data

Procedures for Each of the Study’s Phases			
Phase-Research Questions	Purposes	Tasks	Data
Phase 1: What are the key domains and criteria that represent the construct of digital multimodal composition?	<ul style="list-style-type: none"> • To identify domains and criteria that will be included in the rubric • To establish content-aspect of construct 	<ul style="list-style-type: none"> • Literature review (Steps 1-2) 	<ul style="list-style-type: none"> • 5 domains and 19 criteria

	validity of the scoring rubric		
Phase 2: How might these domains and criteria be structured in a task- and grade-specific rubric in order to evaluate upper-elementary-grade students' digital multimodal book reviews?	<ul style="list-style-type: none"> To create a rubric for the assessment upper-elementary-grade students' DMBRs 	<ul style="list-style-type: none"> Rubric creation (Steps 1-6) 	<ul style="list-style-type: none"> Proposed rubric A set of experts' scores on 30 in-class and 8 online DMBRs
Phase 3: To what extent does the use of the proposed rubric display evidence of inter-rater reliability, construct validity and usability, as indicated by raters' scores and interview feedback?	<ul style="list-style-type: none"> To check the evidence of inter-rater reliability, content and structural aspects of construct validity and usability of the proposed rubric 	<ul style="list-style-type: none"> Rater training Quantitative data collection & analysis: Calculation of inter-rater reliabilities Qualitative data collection & analysis: semi-structured interviews of raters Merging the results of both quantitative and qualitative analyses 	<ul style="list-style-type: none"> Two sets of scores on 25 in-class DMBRs Transcripts of semi-structured individual interview
Phase 4: How does the evidence of inter-rater reliability, construct validity, and rater's feedback inform the process of rubric revision?	<ul style="list-style-type: none"> To detect problems of the tested rubric and synthesize raters' suggestions for rubric revision 	<ul style="list-style-type: none"> Qualitative data collection & analysis of focus group interview Merging the results of phase 3 analysis and the results of focus group interview analysis Rubric revision 	<ul style="list-style-type: none"> Transcript of focus group interview

Chapter 4

RESULTS

The goal of this dissertation study is to identify domains and criteria of DMC from the existing literature and develop a reliable and valid rubric to evaluate upper-elementary-grade students' DMBRs. This chapter presents results for each research phase presented in the methods section.

Results for Research Phase 1: Identified Domains and Criteria

“Effective performance assessment requires the identification of key student learning and related performances, aided by consultation with research related to the demands of the performance” (Afflerbach, 2012, p. 98). As a result of the step-by-step analyses described in the methods section, the 111 criteria identified in the literature (Appendix H) were reduced to 19 separate criteria. The following section explains in detail the 19 new criteria categorized under the five existing domains.

Criteria for the Artifact Domain

The *artifact* domain is linked to the finished digital multimodal product. This finished product incorporates elements related to multiple modes such as message, structure, medium, and technique (Eidman-Aadahl et al., 2013). For the artifact domain, 10 criteria emerged: (a) multimodal coherence, (b) organization of content, (c) conventions of linguistic modes, (d) relational relevance of linguistic modes, (e-g)

technical aspects of audio, visual, and spatial modes and (h-j) relational relevance of audio, visual, and spatial modes.

Multimodal coherence. This criterion is about the overall unity of a DMC. If a digital multimodal artifact consists of different modes that match, complement, or blend in with each other and the results of using mixed modes convey and support ideas and enhance the comprehensibility and usability of the artifact, it can be considered a highly coherent product. This criterion emerged from 13 different criteria, including coherence (Yancey, 2004), cohesion (Levy & Kimber, 2009), design for medium (Burnett et al., 2014), organization (Ostenson, 2013), and multimodality (Wierszewski, 2013). Although these 13 criteria were identified by different names in the literature, they were all defined in terms of the relationships between different modes for the evaluation of the unity of multimodal products. For example, Levy and Kimber (2009) explicitly defined the cohesion criterion as “the way in which the various elements of the text are drawn together to achieve unity” (p. 493). Although they used the term “cohesion,” using “coherence” is more appropriate to indicate the overall quality of a digital multimodal artifact as a united whole.

Organization of content. The organization of content criterion denotes a logical structure of messages or content conveyed by a DMC. This definition was drawn from reviews of the following criteria: cognitive design-completeness (Morain & Swarts, 2012) and organization (Burnett et al, 2014; Wierszewski, 2013). The organization criterion of a programmatic rubric suggested by Burnett et al. (2014) covers definitions related to both the multimodal coherence criterion presented above and the organization

criterion of traditional writing. Unlike the multimodal coherence criterion, which focuses on the relationships among modes, this criterion focuses on the connection among messages or content in different sections of a digital multimodal artifact. In a traditional writing assessment, the organization criterion targets the quality of connections between components of an essay such as the introduction, body and conclusion. In DMC, on the other hand, structures differ by their types or purposes. In this study, therefore, the organization of content criterion of DMC pays attention to the logical flow of messages conveyed by the artifact.

Conventions and relational relevance of linguistic modes. Linguistic modes included in digital multimodal artifacts are oral and written language. Preparing valid criteria for linguistic modes is crucial because different types of DMC, such as blog posts, presentation slides, and digital book reviews or trailers, still heavily rely on linguistic modes when they convey messages. A review of existing literature resulted in two criteria for linguistic modes: conventions and relational relevance. To be specific, most of the literature included criteria for linguistic modes that focused on English conventions such as grammar, mechanics, style, citation and genre (Borton & Hout, 2007; Burnett et al., 2014; Hung et al., 2013; Selfe & Selfe, 2008; Towndrow et al., 2013; Wierszewski, 2013; Yu, 2014). Among these studies, only Hung et al. (2013) attempted to consider the relationship between linguistic modes and other modes in a multimodal text. In order to put equal weight on both English conventions and the relationship between linguistic modes and other modes, I created two separate criteria under the linguistic mode: conventions and relational relevance.

Technical aspects and relational relevance of audio, visual, and spatial

modes. Previous studies on evaluative criteria of DMC attended to two different spheres of audio, visual, and spatial modes: technical aspects and relational relevance. The former refers to the effects of modal resources and technical skills related to the mode on its quality. The latter intends to consider the relationship between one mode and the other modes in a multimodal text.

Technical aspects of the audio mode encompass voice elements (e.g., fluency, articulation, intonation, volume), sound elements (e.g., pitch, volume, length), and general editing techniques (e.g., handling noises, cuts, fades) (Brown, 2013; Hung et al., 2013; Morain & Swarts, 2012; Ostenson, 2013; Selfe & Selfe, 2008; Towndrow et al., 2013). Technical aspects of the visual mode include camera shots and angles, lighting, color, size, movement, and sequencing (Brown, 2013; Hung et al., 2013; Levy & Kimber, 2009; Morain & Swarts, 2012; Ostenson, 2013; Selfe & Selfe, 2008; Yu, 2014). Lastly, the spatial mode is assessed by consideration of technical aspects including layout, alignment of modes and margins (Hung et al., 2013; Wierszewski, 2013). It should be noted that the specific technical aspects of each mode could differ depending on the type of DMC. For example, if a student created a music video in iMovie using only static images, the shots and camera angles might not be relevant aspects for the music video since he or she did not use any camera recording techniques.

The purpose of establishing the relational relevance criterion of each mode is to take a closer look at the intersemiotic relationships between and among modes. As Jewitt's (2014) second assumption on multimodality states, each mode in a multimodal

ensemble plays an important role in close connection with other modes. Evaluating only the overall coherence of digital multimodal texts cannot capture the unique contribution of each mode. In fact, evaluation of relational relevance between the target mode and the other modes helps us assess the overall coherence in the end. Beginning the evaluation by grasping the meaning of the most dominant mode in the digital multimodal text is the most effective way of examining relational relevance. For example, if a student's digital report on his/her community relies heavily on images, meanings in each image should be listed first. Then the meanings of the visual mode need to be compared to the gist of second dominant mode. If all or most of this one-to-one comparison shows a high relevance of meaning between modes, the digital multimodal artifact can be evaluated as a coherent one in general.

Criteria for the Context Domain

The context domain concerns purposes, audiences and tasks surrounding the creation and circulation of the artifact (Eidman-Aadahl et al., 2013). This domain included two criteria: rhetorical awareness-task and rhetorical awareness-audience.

Rhetorical awareness-task. The rhetorical awareness-task criterion was developed from seven related criteria, such as mode of presentation (Borton & Hout, 2007), following the assignment and purpose (Wierszewski, 2013), physical design: accessibility (Morain & Swarts, 2012), rhetorical awareness (Burnett et al., 2014), rhetorical context (Selfe & Selfe, 2008), rhetorical knowledge (Yu, 2014). This criterion emphasizes the composer's consideration of DMC task environments such as purposes, genres, directions and physical environments. In fact, rhetorical awareness-tasks and

rhetorical awareness-audience are difficult to be separate. In some cases, directions for DMC tasks require composers to be aware of all rhetorical components. For example, Burnett and colleagues (2014) defined rhetorical awareness as a “response to situation, considering elements such as context, purpose, audience, and register” (p. 57). On the contrary, students frequently compose digital multimodal texts without considering and motivating specific audiences. In this study, I purposefully separated the audience awareness component in order to underscore the importance of setting specific real or virtual audiences for students who are composing digital multimodal texts.

Rhetorical awareness-audience. The rhetorical awareness-audience criterion attends to the composer’s consideration of explicit or implicit audiences and their engagement with the artifact. This new criterion was drawn from six criteria from existing literature, including the following items: audience (Wierszewski, 2013), engagement (Morain & Swarts, 2012), rhetorical awareness (Burnett et al., 2014), rhetorical context (Selfe & Selfe, 2008), rhetorical knowledge (Yu, 2014), and voice (Howell et al., 2013). For example, the engagement criterion included in Morain and Swarts’s (2012) rubric was included here in order to supply the rhetorical awareness-audience criterion because it checks if the video created by students is “designed to interest and motivate users” (p. 24), and it checks whether or not the goal is directly related to the consideration of audiences.

Criteria for the Substance Domain

According to Eidman-Aadah et al. (2013), the substance domain “refers to the content and overall quality and significance of the ideas presented.” Credibility, accuracy

and significance of information presented in the artifact are also evaluated with the criterion. For this domain, two new criteria were identified: quality of ideas and quality of opinions/arguments.

Quality of ideas. This criterion was set to cover the goals of evaluating the quality of content in narrative or informative texts. Eleven existing criteria were categorized under the quality of ideas criterion: character analysis (Husbye & Rust, 2014), cognitive design: accuracy and pertinence (Morain & Swarts, 2012), content (Levy & Kimber, 2009; Towndrow et al., 2013; Yu, 2014), critical thinking skills (Borton & Hout, 2007), economy (Towndrow et al., 2013), interpretation (Husbye & Rust, 2014), movement (Wierszewski, 2013), theme (Husbye & Rust, 2014); theme/point of view (Towndrow et al., 2013). Although most of these criteria aimed to evaluate the quality of content presented in the DMC, three criteria on Husbye and Rust's (2014) rubric, character analysis, interpretation, and theme, were targeted to check the student multimodal composers' understanding of these key narrative text components. On the other hand, Towndrow et al.'s (2013) three criteria, content, economy, and theme/point of view, were measuring the interest, uniqueness, depth, length and focus of the DMC content.

Quality of opinions/arguments. This criterion was separated from the quality of ideas in order to emphasize the persuasive purpose of DMC content. Two existing criteria were considered to comprise this criterion: ideas and organization (Howell et al., 2015), and stance and support: argument, evidence, and analysis (Burnett et al., 2014). For example, the ideas and organization criterion of Howell and colleagues (2015) evaluates

the quality of the argument by considering the relationships among multiple modes.

Criteria for the Process Management and Technique Domain

The fourth domain, process management and technique, is related to technical and task management skills during the entire composing process of multimodal texts, from planning to composing to publishing. By reviewing existing literature, I identified three criteria: collaboration, technical skills, and writing processes and strategies.

First, the collaboration criterion was the only task management skill found in two studies (Howell et al., 2015; Yu, 2014). Although two examples are not enough to establish a new criterion, both defined collaboration in a substantively similar way and so it can be considered an important aspect of DMC. Second, the technical skills criterion was created from previous criteria such as development of new literacies (Brown, 2013), *ICT usage* (Towndrow et al., 2013) and technical execution (Wierszewski, 2013), in order to refer to the composer's ability to use both print-based and digital media. Finally, I set writing processes and strategies as an independent criterion encompassing writing process (Brown, 2013) and publication (Howell et al, 2015), both criteria from literature. Writing processes and strategies need to be taught and evaluated explicitly because engaging in various steps of the writing process, such as brainstorming, drafting, writing, image construction, revising, editing, and publishing, can facilitate seamless application of technical skills and collaboration.

Criteria for the Habits of Mind Domain

The final domain, habits of mind, included two criteria: creativity (Wierszewski, 2013; Yu, 2014) and self-efficacy (Morain & Swarts, 2012). While the definition of the

domain listed several behavioral or attitudinal characteristics such as creativity, engagement, mindfulness, and risk-taking, self-efficacy was newly identified from Morain and Swarts (2012). The creativity criterion refers to the uniqueness and originality of the composer's ideas and of the ways used to convey meaning using multiple modes. This criterion was mentioned several times in non-rubric literature, and not included in the rubric literature. This does not mean that creativity cannot be measured or is not important. Attempts to measure creativity in writing can be indeed found in the scholarship (e.g., Bear & McKool, 2009; Mozaffari, 2013). One feasible explanation is that a rubric may not be understood as an appropriate means of assessing the creativity of students' writing or DMC. More studies are needed to define creativity in DMC and how to assess it.

The other criterion, self-efficacy, indicates an individual's belief in self as a skilled and confident composer of digital multimodal texts. For example, two criteria from Morain and Swarts's (2012) rubric, confidence and self-efficacy, contained content related to the composer's belief in a knowledgeable and skilled self through use of a confident and persuasive voice.

To sum up, a total of 19 distinguishable criteria were drawn by reviewing 111 criteria from existing non-rubric and rubric literature. These 19 criteria might not be the exhaustive components of the DMC as a construct. However, the 19 criteria provide us with an overview of DMC and with ideas for what to teach and what to assess. Table 5 presents definitions of the 19 distinguishable criteria.

Table 5 Finalized Domains and Criteria of Digital Multimodal Composition Assessment

Domain 1: Artifact “is the finished product. Audiences expect artifacts to convey a coherent message with a clear focus created through an appropriate use of structure, medium, and technique. Artifacts incorporate elements from multiple modes, and are often digital, but do not have to be—they may be analog works (e.g., texts that incorporate both writing and drawing)” (Eidman-Aadah et al., 2013, para. 5).

Criteria	Definitions
1. Multimodal coherence	The overall unity of the digital multimodal product; to support unity, the different modes used in the multimodal product should match, complement, or blend in with each other (relationship among modes) and the results of using different modes should convey and support ideas and enhance the comprehensibility and usability of the multimodal product.
2. Organization of content	Logical structure of content or messages within and among frames or sections
3. Conventions of linguistic mode	The effects of grammar, mechanics, style, citation, and genre on the quality of written and oral language
4. Relational relevance of linguistic mode	The relationship between written or oral language and other modes
5. Technical aspect of audio mode	The effects of fluency, articulation, intonation, volume, pitch, length and editing techniques (e.g., cuts and fades) on the quality of audio mode such as voice, sound effects, and music
6. Relational relevance of audio mode	The relationship between the audio mode and other modes
7. Technical aspects of visual mode	The effects of camera shots and angles, lighting, color, size, movement, and sequencing on the quality of visual mode such as static or moving images
8. Relational relevance of visual mode	The relationship between the visual mode and other modes
9. Technical aspects of spatial mode	The effects of layout, alignment of modes, and margins on the quality of spatial design
10. Relational relevance of spatial mode	The relationship between the spatial mode and other modes

Domain 2: Context “is the world around the artifact, around the creation of the artifact, and how the artifact enters, circulates, and fits into the world. Authors attend to the context of a multimodal artifact when they make design decisions related to genre or to an artifact’s intended uses. Given their purposes, authors consider the affordances, constraints, and opportunities, given purpose, audience, composing environment, and delivery mode” (Eidman-Aadah et al., 2013, para. 5).

Criteria	Definitions
11. Rhetorical awareness-task	Consideration of specific purposes, genres, task directions, and physical environments of DMC
12. Rhetorical awareness-audience	Consideration of explicit or implicit audiences and their engagement with the artifact

Domain 3: Substance “refers to the content and overall quality and significance of the ideas presented. The substance of a piece is related to an artifact’s message in relationship to the contextual elements of purpose, genre, and audiences. Considering the substance of a piece encourages authors to think about elements such as quality of ideas, quality of performance, credibility, accuracy, and significance” (Eidman-Aadah et al., 2013, para. 5).

Criteria	Definitions
13. Quality of ideas	Clarity, credibility, significance (depth and length), and interest of the content and the pace of content progress or development
14. Quality of opinions/arguments	Clarity and persuasiveness of arguments and the use of analysis and evidence to support the argument

Domain 4: Process Management and Technique “refers to the skills, capacities, and processes involved in planning, creating, and circulating multimodal artifacts. Creating multimodal products involves the technical skills of production using the chosen tools, but it also includes larger project management skills as well as the ability to collaborate with others in diverse and often interactive situations. Over time, individuals learn to more effectively control the skills and manage the processes of producing and circulating digital content” (Eidman-Aadah et al., 2013, para. 5).

Criteria	Definitions
15. Collaboration	In the case of group projects, students work collaboratively by generating ideas together, dividing the labor fairly, and providing comments on each part of the project
16. Technical skills	The ability to use print-based and digital media and to export, import, modify, and switch between modes in the medium effectively
17. Writing processes and strategies	Engaging in various writing processes such as brainstorming, drafting, writing, image construction, revising, editing, and publishing and use of different writing strategies effectively

Domain 5: Habits of mind “are patterns of behavior or attitudes that reach beyond the artifact being created at the moment. They develop over time and can be nurtured through self-sponsored learning as well as teacher-facilitated activities throughout the process. Examples include creativity, persistence, risk-taking, mindfulness, and engagement. Habits of mind can also include an openness to participatory and interactive forms of engagement with audiences” (Eidman-Aadah et al., 2013, para. 5).

Criteria	Definitions
18. Creativity	Uniqueness and originality of ideas and of the ways used to convey meaning using multiple modes
19. Self-efficacy	An individual’s belief in self as a skilled and confident composer of digital multimodal texts

Results for Research Phase 2: Rubric Creation Procedures

The second phase of the current study was to create a draft rubric. In this section, I provided detailed descriptions on the steps I followed to create the rubric.

Step 1: Decisions on Four Aspects of Scoring Rubrics

As the first step of creating a draft rubric, I began by deciding on four aspects of scoring rubrics. In light of information on the types of scoring rubrics, developing an analytic rubric for the assessment of students’ DMBRs was deemed a more appropriate way to address the current study’s objectives because the rubric, once it is developed, will be used in a classroom setting to diagnosis weaknesses in upper-elementary-grade students’ DMC performance and provide instructional feedback. In addition, an analytic scoring rubric seemed to be a convenient approach to organizing the domains and criteria involved in DMC assessments in a systematic way.

In terms of the level of specificity, this study sought to combine the best aspects of both the generic and task-specific rubric by creating a rubric located in the middle of

the continuum. For example, many of the criteria related to students' use of modes and relationships between and among modes were able to be generic and applicable to across a range of DMCs. However, since the current study was seeking to develop an assessment tool specifically for DMBRs, the characteristics of the book review task and developmental characteristics that were relevant for DMBRs of upper-elementary-grade students had to be considered. For these reasons, I decided to create a partially generic but task-specific rubric.

Regarding scoring points, this study adopted a 4-point scale for the DMBR rubric in order to avoid the tendency to regress toward the mean and to ensure the scoring rubric is a sufficiently distinguishable tool for DMC performances in grades 3 to 5.

Finally, the current study determined the audience of the rubric to be teachers since teachers would be the best and most likely agents to use the rubric as a tool for both formative and summative assessments. This consideration of audience influenced the word choice on the rubric.

Step 2: Reflecting

In this second step of the rubric creation, task characteristics of book reviews and instructional components of DMBRs in upper-elementary-grade classrooms were considered. This involved examining existing literature and lesson plans on book reviews or book reports. I also reviewed the observation field notes of the larger study, which I recorded during the teacher's mini lesson on the DMBR.

Task characteristics of digital multimodal book reviews. DMBRs are similar to print-based traditional book reviews or reports in terms of its purpose and components.

To be specific, a print-based traditional book review is a type of persuasive writing (Graham, Bollinger, Booth Olson, D'Aoust, MacArthur, McCutchen, & Olinghouse, 2012). DMBR is also a type of persuasive writing, with the purpose of persuading audiences to read or not to read a certain book. Major differences between DMBRs and print-based book reviews are the audiences and the modal affordances. Digital composing environments allow students to engage a wider audience beyond the teacher, including other students, parents, and anyone who has access to the internet (Reinking & Watkins, 2000). In addition, the digital environment enables students to incorporate moving visuals and audio modes, which cannot be afforded in the print-based environment.

Instructional components of digital multimodal book reviews. In order to identify how the DMBRs are taught to or completed by upper-elementary-grades students, I searched for lesson plans and actual students' DMBR samples that were published on websites such as YouTube, Vimeo, and ShowMe. Lesson plans for print-based or digital book reviews included the following components: introduction, summary, what do you like or dislike, and recommendation (ReadWriteThink book review template, Scholastic website). The lesson on the DMBR that I observed in the fourth-grade classroom also included these components. In addition to these, the teacher added conclusion component.

Step 3: Listing

At this step, I selected domains and criteria that could be critical for the DMBR rubric and created an Excel spreadsheet. Then I watched 30 students' DMBRs and wrote

down characteristics that I observed in each student's DMBR and listed them under each criterion.

Selection of appropriate domains and criteria for the rubric. Artifact, context, and substance domains were included in the rubric. The other two domains, process management and technique and habits of mind, were excluded since these could not be evaluated by the finished products. Observation of composing procedures and processes and interviewing of students were essential to assess the two excluded domains.

Among the 14 criteria under the artifact, 9 criteria were selected to be included in the rubric: (a) technical aspects of the visual mode, (b) technical aspects of the audio mode, (c) conventions of the linguistic mode, (d) relational relevance of the linguistic mode, (e) coherence of the multimodal product, (f) organization of content, (g) rhetorical awareness-audience, (h) quality of ideas, and (i) quality of opinions/arguments. The other five criteria—relational relevance of the audio mode, relational relevance of the visual mode, technical aspects of the spatial mode, relational relevance of the spatial mode, and rhetorical awareness-task—were excluded due for the following reasons.

First, criteria related to visual and linguistic modes were included, but the criteria related to the spatial mode were excluded due to the modal affordances of the iMovie app and task characteristics of the DMBR. For example, without inserting any static or moving visuals, iMovie did not allow students to create their DMBRs. In addition to this, the teacher demonstrated how to use oral language, written language, and sound modes. Once she put visuals in the iMovie, she recorded a narration, inserted written language using the title function, and added other audio clips such as sound effects or theme music.

No specific instruction on the spatial mode was provided. To sum up, students' DMBRs were more strongly related to the visual, linguistic, and audio modes than the spatial mode. For this reason, two criteria related to the spatial mode were not included in the rubric.

Second, instead of including the three criteria (i.e., relational relevance of audio, visual, and spatial modes), the current study selected the relational relevance of linguistic modes and focused on the relationship between oral language and other modes. Examining the relationship between oral language and other modes was essential because most students' DMBRs conveyed most information using the oral language. Also, the "title" function of the iMovie app limited the application of written language conventions in English. For example, the title function did not support written language in paragraph form. When students typed lengthy sentences of more than 10 words, the font size became increasingly smaller until it ultimately became illegible. In an environment where students cannot fully apply the conventions of standard written English, the criteria for it should be different from the criteria for oral and written language. To reflect the relative importance of oral language, the linguistic modes criterion was divided into two criteria: conventions of oral language and conventions of written language; the conventions of linguistic modes served as a category, a higher level of the two criteria. Since the other modes could still be considered in relation with the oral language, criteria about the relational relevance of the three modes were excluded.

Finally, one of the two criteria under the context domain, rhetorical awareness-task, was excluded because the purpose, genre, and directions for the DMC task were clearly structured and articulated in the graphic organizer.

Other criteria were grouped under specific categories. For instance, technical aspects of the visual mode and audio-voice mode were located under the technical aspects of non-linguistic modes category. Two criteria under the substance domain, quality of ideas and quality of opinion/arguments, were renamed as quality of summary and quality of opinion. This change was to present the task characteristics of the DMBRs. As reviewed at the step 2, book reviews contained both the summary of the book and the reviewer's opinion on the book. These renamed criteria were grouped into the quality of content category.

Listing characteristics. Once I finalized the three domains and eleven criteria, I listed them in an Excel spreadsheet. Then I watched all students' DMBRs ($N = 30$) at least two times and wrote down the characteristics of each student's book reviews. Then I located them in each blank of the Excel spreadsheet that included the selected domains and criteria. As an example, Appendix I presents the listed characteristics of each student's performance related to the relational relevance of written language mode.

Step 4: Grouping and Labeling

In the fourth step, the listed characteristics were read carefully. Some look-alike characteristics were grouped together. These grouped characteristics and other characteristics notable for their repeated appearances in many students' DMBRs were

placed in a table for labeling. Appendix J presents the grouped characteristics of each criterion.

Domains and criteria included in the Excel spreadsheet were reexamined to verify appropriate labels for the grouped characteristics. While listing the grouped characteristics on the relationship between modes, it was noticed that oral language plays an important role in the relationship between the modes. This resulted in the revision of the three criteria about relationship between modes: relationship between oral language and visual, relationship between oral language and written language, and relationship between oral language and sound. Relationships between written language and sound and written language and visual were not included in the grouped characteristics because those did not stand out among the students' DMBRs.

Step 5: Writing Definitions and Descriptors

The domains and criteria selected and used as labels in the previous steps were transferred to the rubric grid. Based on readings of the grouped characteristics, descriptors of the four different levels of the analytic rubric were generated. For example, after I read the grouped characteristics about technical aspects of the visual mode, I decided to assign higher scores to the students who used more than one image per frame than to the students who used only one image per frame. The more students used images in the DMBR, the better they represented specific stories and their opinions on the book. I also found that including both videos and images required more technical skills than including either videos or images alone. For this reason, I assigned higher scores to the students who included both videos and images than to the students who used visuals in

only one way in their DMBRs. By considering the two aspects, I wrote down descriptors for four different levels: 4-excellent, 3-good, 2-fair, and 1-needs improvement. In addition to these, one more level was added to grade not-assessable cases. Table 6 shows the initial descriptors for the technical aspects of visual mode.

Table 6 Initial Descriptors for the Technical Aspects of Visual Mode.

Levels	Descriptors
4-Excellent	<ul style="list-style-type: none"> • The artifact includes both static images and videos. • All technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are of good quality such that visuals convey meaning clearly.
3-Good	<ul style="list-style-type: none"> • The artifact includes both static images and videos. • Three or four technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are of good quality such that visuals convey meaning clearly.
2-Fair	<ul style="list-style-type: none"> • The artifact includes only either static images or videos. • One or two technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are not of good quality, so they hinder the visuals' ability to convey meaning clearly.
1-Needs Improvement	<ul style="list-style-type: none"> • The artifact includes only/either static images or videos. • No technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are of good quality, so they hinder visuals conveying meaning clearly.
0-Not Assessable	<ul style="list-style-type: none"> • The artifact does not include any visual mode.

Table 7 Domains, Categories, and Criteria included in the Rubric

Domains	Categories	Criteria
Artifact	A. Technical Aspects of Modes	1. Technical aspects of visual mode: The effects of camera shots and angles, lighting, color, size, movement, and sequencing on the quality of visual modes such as static and moving images
		2. Technical aspects of audio mode-voice: The effects of fluency, articulation, intonation, volume, pitch, and length on the quality of voice
	B. Conventions of Linguistic Modes	3. Conventions of oral language (narration): The effects of grade-appropriate conventions of Standard English on the quality of the oral language
		4. Conventions of written language (titles): The effects of grade-appropriate grammar, mechanics, style, citation, and genre on the quality of written language
Context	C. Coherence of Multimodal Products	5. Relationship between oral language and visual
		6. Relationship between oral language and written language
		7. Relationship between oral language and audio mode-sound
Substance	D. Organization of Multimodal Content	8. Organization of multimodal content: Logical structure of content or messages within and among frames or sections
		E. Rhetorical Awareness
Substance	F. Quality of Content	10. Quality of summary: Clarity, credibility, significance (depth and length), and interest of the

summary and the pace of summary progress or development

11. Quality of opinion: Clarity and persuasiveness of opinion and the use of details to support the opinion

Table 7 presents the domains, categories, and criteria included in the rubric with numbers. Appendix K shows the first version of the rubric (rubric #1). In the following sections, names of criteria will be replaced by the numbers.

Step 6: Revising with a Collaborator

With the rubric #1, the collaborator and I graded all students' DMBRs and revised the rubric over the course of six meetings.

First meeting: Introducing the study and the metarubric. At the beginning of this step, the background information about the developed rubric was explained to the collaborator, namely, the purpose of the project, the definitions of modes and multimodality, and the characteristics of DMBR tasks. The steps 1 through 5 that I followed to create the rubric were explained briefly. Because the next meeting would cover the evaluation of rubric, the collaborator was asked to read Arter and Chappuis's (2006) metarubric, a rubric for rubrics, and their chapter about it.

Second meeting: Evaluating rubric #1 using the metarubric. In evaluating the rubric's validity of content and clarity of the language, we used the metarubric and its analysis form presented in Appendix L and M (Arter & Chappuis, 2006). According to the metarubric, rubric #1's descriptors of some criteria (e.g., the quality of summary)

were represented redundantly in more than one criterion by covering the content that would be evaluated by different criteria. Descriptors of some criteria such as the relationship between oral language and the visual mode and the relationship between oral language and written language contained vague quantitative words and some levels were not parallel in content. In the relationship between oral language and sound criterion, descriptors for levels 2 and 3 were reversed. The collaborator and I discussed how to revise problematic descriptors for criteria 5, 6, 7, and 10. Based on the conversation, I revised rubric #1 and sent the second version of the rubric (rubric #2) to the collaborator.

Third meeting: Grading six DMBRs using rubric #2. During this meeting, the collaborator and I graded a total of six DMBRs using the rubric #2. Five DMBRs were from the fourth-grade study data and the other book review was found on the internet. Specifically, we graded two products separately and compared the scores awarded for each of the criteria. We discussed the reasons for criteria scores that differed by more than ± 1 point and resolved any problems. We repeated these steps twice until we finish grading six DMBRs. The discussion revealed that our understandings of the descriptors included in criteria 5 and 7 were different and descriptors of criteria 1, 2, 3, 4, and 10 should be revised. Based on the discussion, I revised rubric #2 and sent the third version of the rubric (rubric #3) to the collaborator.

Fourth meeting: Grading eight digital multimodal book reviews from the internet using rubric #3. For this meeting, the collaborator and I used rubric #3 to independently grade eight DMBRs found online. Then we compared the exact and

adjacent agreement rates between our scores. Table 8 presents the exact and adjacent agreement rates on the eight DMBRs.

Table 8 Exact and Adjacent Agreement Rates Between the Researcher and Collaborator on 8 Digital Multimodal Book Reviews using Rubric #3

Criteria Agreement	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
Exact	62.5	50	87.5	75	87.5	62.5	75	100	37.5	50	87.5
Adjacent	100	100	100	87.5	100	75	100	100	100	100	100

C1: Technical aspects of visual mode, C2: Technical aspects of audio-voice, C3: Conventions of oral language, C4: Conventions of written language, C5: Relationship between visual and oral language, C6: Relationship between oral language and written language, C7: Relationship between oral language and sound, C8: Organization of multimodal content, C9: Audience awareness, C10: Quality of summary, C11: Quality of opinion. Unit: Percentage (%)

As shown on the Table 8, rubric #3 yielded 100% adjacent agreement on the revised criteria 3 and 10. However, revisions made on criteria 1, 2, and 4 were not effective and problems interpreting criteria 4, 6, and 9 were detected. Conversation with the collaborator revealed that low adjacent agreement rates on criteria 4 and 6 originated from our different and irregular interpretations of cases including only one or two titles. In addition, descriptors of criterion 6 did not address the cases where students included only one title in the entire book review. In case of criterion 9, more descriptors were necessary for levels 3 and 4. Even though the agreement rate on criterion 10 was not problematic, the collaborator and I were not satisfied with the descriptors about the inaccuracy of summary. Since only a few students included inaccurate information about

the story, the descriptors could not be applied to all students' products consistently. For this reason, the descriptors were replaced to address coherence of the summary. All things considered, revisions on rubric #3 resulted in the fourth version of the rubric (rubric #4).

Fifth meeting: Grading 30 digital multimodal book reviews using rubric #4.

Since most criteria in the proposed rubric were revised several times, reviewing the five fourth-grade students' DMBRs which were graded during the third meeting was necessary. Independent grading of the five artifacts using rubric #4 indicated that all criteria except criterion 10 achieved acceptable adjacent agreement rates (i.e., over 90%) between the collaborator and me. Due to schedule conflicts, the collaborator and I could not discuss the problem with criterion 10 immediately. Consequently, we proceeded to the independent grading of the rest of the fourth-grade students' ($n = 25$) DMBRs on all criteria except 10. Agreement rates on the independently graded products ($n = 25$) are shown in the Table 9.

Table 9 Exact and Adjacent Agreement Rates Between the Researcher and Collaborator on 25 Digital Multimodal Book Reviews using Rubric #4

Criteria \ Agreement	1	2	3	4	5	6	7	8	9	10	11
Exact	88	80	76	72	84	72	76	100	76	N/A	80
Adjacent	100	100	100	84	100	76	100	100	100	N/A	100

C1: Technical aspects of visual mode, C2: Technical aspects of audio-voice, C3: Conventions of oral language, C4: Conventions of written language, C5: Relationship between visual and oral language, C6: Relationship between oral language and written language, C7: Relationship between oral language and sound, C8: Organization of multimodal content, C9: Audience awareness, C10: Quality of summary, C11: Quality of opinion. Unit: Percentage (%)

During the discussion on the 25 products that were graded independently, we mainly talked about our different interpretations of the written language mode, which resulted in the low adjacent agreement rates for criteria 4 and 6. In the fourth meeting, we were not sure if images that included only written language should be counted as written language mode or not. When we graded 25 students' DBMRs using rubric #4, we decided to count them as written language. However, this caused different interpretations of images including written language as well as other visual modal resources. Consequently, we returned to our previous understanding of the written language mode: in this rubric, only the titles that students intentionally included in the artifact are counted as written language. Words in images or visuals are not considered as the written language mode.

We also made major changes to criteria 6, 7, 10, and 11. In terms of criterion 6, using percentage descriptors was not an effective way to represent the relationship between oral and written language since many students included only one or two titles. We concluded that reviewing all students' DMBRs was necessary in order to find the optimal number of titles for each level. In the case of criterion 7, original descriptors did not present any percentage or number of the sound. Our conclusion for revision of the criterion was to review all students' products and find optimal numbers of sound for each level. Reexamination of the descriptors in criterion 10 identified that descriptors about the disclosure of the story's ending were not useful to differentiate levels of the performance. We decided to delete the descriptor and to present number of details that should be included in each level. Criterion 11 had to be revised because the first descriptors in each level of the criterion were already covered by criterion 8. Other minor

changes on some words of criteria 2, 4, 5, and 8 were also made in the fifth version of the rubric (Rubric #5).

Final meeting: Re-grading 38 digital multimodal book reviews using rubric #5 and #6. The collaborator and I used rubric #5 for independent re-grading of 30 DMBRs created in the fourth-grade classroom. Since rubric #5 underwent major revisions on criteria 6, 7, 10, and 11 only, scores on the other criteria did not change much. Our agreement rates on the criteria 6, 7, and 11 were significantly high, exceeding 90% on both exact and adjacent agreement. However, our interpretations of descriptors included in criterion 10 were still substantively different. For more accurate evaluation of the summary section, I created a list of major events of the story, *Frindle*, and shared it with the collaborator. The number of details about the major events was counted by using the list. Although the agreement rates on criteria 1 and 5 were high enough, we agreed that the term included in the second descriptor, “almost all,” was significantly vague enough that it was not helpful in evaluating the technical aspects of visuals. In criterion 5, setting 80% as the reference point between level 2 and 3 was inconsistent with the reference point of other criteria on the same level (e.g., 75%) and seemed to be set arbitrarily. Therefore, the 80% included in levels 2, 3, and 4 was changed to 75%. As a result of this discussion, the sixth draft of the rubric (rubric #6) was created. The agreed scores (i.e. the experts’ scores) on a total of 30 fourth grade students’ DMBRs and 8 online book reviews were generated from rubric #6. Appendix N displays tables showing major changes on the descriptors of each criterion.

Results for Research Phase 3: Rubric Validation

This section presents the results of rubric validation. Quantitative and qualitative evidence on the inter-rater reliability, construct validity, and usability of the developed rubric was presented. First, I presented quantitative results on inter-rater reliability among the experts and the two raters (Kristen and Lindsey) and qualitative results of the two raters' individual interviews on the rubric #7's role for reliable scoring. Then both quantitative and qualitative evidence on structural aspect of construct validity of the rubric was displayed. Third, raters' statements from individual interviews on the rubric's content aspect of construct validity was presented. Finally, I presented qualitative evidence on the usability of the rubric.

Quantitative Evidence on Inter-Rater Reliability of the Use of the Rubric

Consensus estimates: Percent agreement among raters. The current study examined the exact and adjacent agreement rates among the three raters (the experts, Lindsey & Kristen). The exact percent agreement between any two among the three raters for all 11 criteria for 25 students was 60.4%, which meets the target exact agreement rate of this study (60%). In terms of the exact percent agreement per criterion, the lowest agreed criterion was criterion C3-conventions of oral language (37.3%) and the highest agreed criterion was C6-relationship between oral language and written language (92%). Only five out of eleven criteria were at or above 60 percent of exact agreement among raters (i.e., C4-conventions of written language, C5-relationship between visual and oral language, C6-relationship between oral and written language, C8-organization of multimodal content, and C11-quality of opinion).

The adjacent percent agreement among the three raters for all 11 criteria for 25 students was 91.4%, which is above the target adjacent agreement rate of this study (80%). All criteria exceeded 80% agreement rate on average except C4-conventions of written language. There were four criteria below 90% agreement: C3-conventions of oral language, C4-conventions of written language, C5-relationship between visual and oral language, and C11-quality of opinion. Raters' interpretations of the rubric perfectly agreed within 1 point for C6-relationship between oral language and written language, C8-organization of multimodal content, and C9-audience awareness. Table 10 shows all percent of exact and adjacent agreement among raters.

Table 10 Percent Exact and Adjacent Agreement Among Raters ($n = 25$)

Criteria	Exact and Adjacent Agreement Among Three Raters							
	Experts×Kristen		Experts×Lindsey		Kristen×Lindsey		Means	
	Exact	Adjacent	Exact	Adjacent	Exact	Adjacent	Exact	Adjacent
C1	48.0%	96.0%	56.0%	100%	40.0%	92.0%	48.0%	96.0%
C2	64.0%	96.0%	64.0%	92.0%	48.0%	96.0%	58.7%	94.7%
C3	52.0%	84.0%	40.0%	80.0%	20.0%	76.0%	37.3%	80.0%
C4	72.0%	76.0%	72.0%	80.0%	72.0%	80.0%	72.0%	78.7%
C5	68.0%	84.0%	80.0%	92.0%	56.0%	76.0%	68.0%	84.0%
C6	96.0%	100%	88.0%	100%	92.0%	100%	92.0%	100%
C7	48.0%	92.0%	40.0%	96.0%	48.0%	100%	45.0%	96.0%

C8	64.0%	100%	68.0%	100%	56.0%	100%	62.7%	100%
C9	64.0%	100%	80.0%	100%	52.0%	100%	53.0%	100%
C10	76.0%	92.0%	56.0%	96.0%	32.0%	88.0%	54.7%	92.0%
C11	52.0%	84.0%	64.0%	80.0%	64.0%	88.0%	60.0%	84.0%
Means	64%	91.3%	64.4%	92.4%	52.7%	90.5%	60.4%	91.4%

C1: Technical aspects of visual mode, C2: Technical aspects of audio-voice, C3: Conventions of oral language, C4: Conventions of written language, C5: Relationship between visual and oral language, C6: Relationship between oral language and written language, C7: Relationship between oral language and sound, C8: Organization of multimodal content, C9: Audience awareness, C10: Quality of summary, C11: Quality of opinion

Consensus estimates: Cohen’s kappa between raters. Cohen’s kappa is a statistical technique “to estimate the degree to which consensus agreement ratings vary from the rate expected by chance” (Jonsson & Svingby, 2007, p. 134). Generally, kappa values above .40 represent fair to moderate agreement (Fleiss, 1981; Jonsson & Svingby, 2007; Stoddart, Abrams, Gasper, & Canaday, 2000) and values above .80 indicate strong to almost perfect agreement beyond chance (Landis & Koch, 1977; McHugh, 2012; Shweta, Bajpai, & Chaturvedi, 2015). The Cohen’s kappa mean value for all 11 criteria of 25 students among three raters was .412, which indicates fair agreement beyond chance among raters. Except for C4-conventions of written language, C5-relationship between visual and oral language, C6-relationship between oral language and written language, and C9-audience awareness, Cohen’s kappa mean values for the other seven criteria indicated poor to slight agreement, ranging from .161 to .377. Table 11 presents all Cohen’s kappa values among raters.

Table 11 Cohen's Kappa Per Criterion Among Raters ($n = 25$)

Criterion	Cohen's Kappa Between Any Two Raters			Means
	Experts × Kristen	Experts × Lindsey	Kristen × Lindsey	
C1	.181	.336	.132	.216
C2	.396	.377	.173	.315
C3	.257	.199	.027	.161
C4	.649	.695	.635	.659
C5	.528	.691	.291	.503
C6	.948	.846	.898	.897
C7	.261	.176	.304	.247
C8	.422	.405	.306	.377
C9	.467	.676	.306	.483
C10	.644	.369	.036	.349
C11	.246	.420	.299	.321
Means	.454	.471	.309	.412

C1: Technical aspects of visual mode, C2: Technical aspects of audio-voice, C3: Conventions of oral language, C4: Conventions of written language, C5: Relationship between visual and oral language, C6: Relationship between oral language and written language, C7: Relationship between oral language and sound, C8: Organization of multimodal content, C9: Audience awareness, C10: Quality of summary, C11: Quality of opinion

Consistency estimates: Spearman's ρ between raters. Consistency estimates of interrater reliability concern the consistent grading pattern between raters. For example, if rater A assigns a higher score within 1 point than rater B for a certain

criterion, the difference in how they apply the rating scales is predictable (Stemler, 2004). The Pearson correlation coefficient is used when the rating scale is dealing with ratio or interval data. On the other hand, Spearman's *rho* coefficient works for ordinal data such as ranks and Likert scales with the values "strongly disagree," "disagree," "neutral," "agree," and "strongly agree." For both methods, coefficient values above .70 are deemed acceptable (Jonsson & Svingby, 2007, p. 134) and below .50 are problematic (Gearhart et al., 1995, p. 224). Since the data generated by this rubric are based on the Likert scale with values such as excellent, good, fair, needs improvement, and not assessable, Spearman's *rho* needs to be calculated to check the consistency between two raters.

Based on the above describe values, two criteria—C3-conventions of oral language and C11-quality of opinion—included values below .50 and seemed to be problematic. The values indicated that the scoring difference between two raters is difficult to predict. On the contrary, C4-conventions of written language, C6-relationship between oral language and written language, and C9-audience awareness displayed values above .70 across the raters. This means that the differences between any two raters are predictable and not generated randomly. Table 12 shows the Spearman's *rho* coefficients per criterion between any two raters.

After considering all analyses on inter-rater reliability of the rubric, I reached the conclusion that C3-conventions of oral language, C4-conventions of written language, and C11-quality of opinion were problematic and the reasons for different and inconsistent interpretations of those criteria needed to be further explored through a focus-group interview.

Table 12 Spearman's *Rho* Per Criterion Among Raters ($n = 25$)

Criterion	Spearman's <i>Rho</i> Between Any Two Raters		
	Experts × Kristen	Experts × Lindsey	Kristen × Lindsey
C1	.527**	.693**	.519**
C2	.677**	.600**	.573**
C3	.611**	.308	.300
C4	.822**	.803**	.755**
C5	.765**	.831**	.542**
C6	.993**	.976**	.987**
C7	.566**	.804**	.714**
C8	.582**	.569**	.556**
C9	.832**	.892**	.800**
C10	.877**	.787**	.593**
C11	.325	.518**	.347

** $p < .01$ (2-tailed).

C1: Technical aspects of visual mode, C2: Technical aspects of audio-voice, C3: Conventions of oral language, C4: Conventions of written language, C5: Relationship between visual and oral language, C6: Relationship between oral language and written language, C7: Relationship between oral language and sound, C8: Organization of multimodal content, C9: Audience awareness, C10: Quality of summary, C11: Quality of opinion

Qualitative Evidence on the Reliability of the Use of the Rubric

Before the quantitative analyses, I collected qualitative data by interviewing each rater individually. During the interviews, two raters stated that using the rubric was helpful for consistent and objective scoring, which refers to important characteristics of

reliability. However, their interpretations of that consistency and objectivity were different.

Internal consistency. For example, when I asked about the benefits of grading a criterion of all students' DMBRs at a time, Lindsey stated:

I'm sitting there trying to just focus on those things, I found that that was more useful because that, even though you don't compare one to the other, you're supposed to be looking straight at the rubric, you're not supposed to compare videos to each other, I found that it helped me kind of crystallize my understanding of the rubric.

The excerpt from Lindsey's interview showed that focusing on one criterion helped her have an internal consistency of criterion interpretation by crystalizing her understanding of the criterion across different students' products.

On the other hand, Kristen stated the importance of using the rubric for internal consistency of her grading. She said:

As a teacher if I'm grading half of the class, then that particular grading session I may be harsher in one particular element, then I come back to it and I grade the second half of the class, maybe I get a little more lax in terms of "well that's okay". But if I have a rubric, then there is no room for personal opinion.

Kristen pointed out that using the rubric to grade students' products helped her apply the criteria consistently.

External consistency. In addition to her recognition of the rubric's contribution for internally consistent grading, Kristen conceptualized the consistency in relation to the existence of other raters. She remarked:

I think just from a consistency standpoint, knowing that I wasn't the only grader, it helped me know I'm using the same information that other scorers would be using or other teachers or what have you. I think that helped with the consistency or at least to take out the subjective portions where as a teacher, if they were in my class, maybe I would have a little bit more knowledge or I would know more of the items not consistent and that would have affected my scoring.

From her point of view, consistent grading seemed to be more related to the “not subjective,” that is, objective, use of the rubric. In case of grading the same students' products with other raters, the objective grading seemed to be a matter to her. In other words, this might indicate that if she used the rubric in her classroom with her students without having any external rater, her interpretation of the rubric might be different.

Quantitative Evidence on the Structural Aspect of Construct Validity

Spearman's *rho* correlations among each rater's scores on 11 criteria. In order to check if each of the 11 criteria in the rubric represents one of the distinctive characteristics of the DMBR as a construct, Spearman's *rho* correlations among each rater's scores on those criteria were calculated. The correlation between C4-conventions of written language and C6-relationship between oral language and written language was statistically significant and it reflected strong effect sizes from all raters (ranging from $rho = .599 \sim .710, p < .01$). The associations between the following criteria were also

significant with medium to strong effect sizes from two out of three raters: C1-technical aspects of visual mode and C8-organization of multimodal content ($\rho = .429$ or $.446, p < .05.$); C2-technical aspects of audio-voice and C3-conventions of oral language ($\rho = .452$ or $.540, p < .05.$); C2-technical aspects of audio-voice and C10-quality of summary ($\rho = .524$ or $.465, p < .05.$); and C7-relationship between oral language and sound and C9-audience awareness ($\rho = .420$ or $.499, p < .05.$). The consistent and strong correlations between C4-conventions of written language and C6-relationship between oral language and written language among raters indicate that the two criteria may measure similar characteristics of students' DMBRs. Tables 13 through 15 provide the correlation matrix of each rater.

Table 13 Spearman's ρ Correlations Among Experts' Scores on 11 Criteria

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
C1	–										
C2	.307	–									
C3	-.046	.452*	–								
C4	-.364	-.191	-.227	–							
C5	.443*	.325	.380	-.024	–						
C6	-.103	.128	.061	.710**	.310	–					
C7	.218	.148	.311	.156	.233	.338	–				
C8	.429*	.268	.041	-.085	.102	-.020	.302	–			

C9	.401*	.072	.103	-.112	.114	.120	.420*	.256	–		
C10	.234	.335	.119	.115	.314	.373	-.011	-.045	.320	–	
C11	.335	.385	.003	.156	.286	.362	.237	.344	.167	.309	–

* $p < .05$. ** $p < .01$. (2-tailed).

C1: Technical aspects of visual mode, C2: Technical aspects of audio-voice, C3: Conventions of oral language, C4: Conventions of written language, C5: Relationship between visual and oral language, C6: Relationship between oral language and written language, C7: Relationship between oral language and sound, C8: Organization of multimodal content, C9: Audience awareness, C10: Quality of summary, C11: Quality of opinion

Table 14 Spearman's *Rho* Correlations Among Kristen's Scores on 11 Criteria

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
C1	–										
C2	-.067	–									
C3	-.061	.540**	–								
C4	-.296	-.189	-.351	–							
C5	.020	.500*	.452*	-.118	–						
C6	-.089	.079	-.147	.599**	.422*	–					
C7	.019	.186	.052	.272	.528*	.366	–				
C8	.446*	.191	.187	-.257	.202	-.143	.333	–			
C9	-.038	.483*	.210	-.151	.169	.061	.254	.291	–		
C10	.116	.524**	.185	.170	.490*	.257	.449*	.392	.259	–	
C11	-.102	.436*	.345	-.037	.430*	.308	.316	-.012	.142	.311	–

* $p < .05$. ** $p < .01$. (2-tailed).

C1: Technical aspects of visual mode, C2: Technical aspects of audio-voice, C3: Conventions of oral language, C4: Conventions of written language, C5: Relationship between visual and oral language, C6: Relationship between oral language and written language, C7: Relationship between oral language and sound, C8: Organization of multimodal content, C9: Audience awareness, C10: Quality of summary, C11: Quality of opinion

Table 15 Spearman's *Rho* Correlations Among Lindsey's Scores on 11 Criteria

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
C1	–										
C2	-.196	–									
C3	.148	.180	–								
C4	-.276	.203	.028	–							
C5	.359	.225	.299	.159	–						
C6	-.173	.174	-.046	.646**	.291	–					
C7	.183	.235	.190	.288	.103	.373	–				
C8	.216	.235	-.133	-.150	.261	.052	.317	–			
C9	.105	.261	.124	.137	.194	.147	.499*	.445*	–		
C10	-.040	.465*	.319	.067	.182	.149	.131	.127	.309	–	
C11	-.100	.281	.316	.084	.389	.318	.239	.397*	.407*	.247	–

* $p < .05$. ** $p < .01$. (2-tailed).

C1: Technical aspects of visual mode, C2: Technical aspects of audio-voice, C3: Conventions of oral language, C4: Conventions of written language, C5: Relationship between visual and oral language, C6: Relationship between oral language and written language, C7: Relationship between oral language and sound, C8: Organization of multimodal content, C9: Audience awareness, C10: Quality of summary, C11: Quality of opinion

Pearson correlations among each rater's scores on six categories. Eleven criteria of the rubric were grouped in six different categories. Across the raters, the following categories were statistically significantly correlated with medium to large effect sizes: CA1-technical aspect of modes and CA4-organization of multimodal product (ranging from $r = .466 \sim .598$, $p < .05$.), CA2-conventions of linguistic modes and CA3-coherence of multimodal product ($r = .559 \sim .660$, $p < .01$.), and CA3-coherence of

multimodal product and CA6-quality of content ($r = .431 \sim .572, p < .05$). This may indicate that unlike the distinctive criteria, the categories of the developed rubric may measure similar characteristics of students' DMBRs or do not reflect the structure of DMBR as a construct. Tables 16 through 18 provide the Pearson correlations among each rater's scores on the six categories.

Table 16 Pearson Correlations Among the Experts' Scores on 6 Categories

Categories	CA1	CA2	CA3	CA4	CA5	CA6
CA1	–					
CA2	-.029	–				
CA3	.421*	.660**	–			
CA4	.598**	.121	.256	–		
CA5	.449*	.046	.305	.412*	–	
CA6	.636**	.284	.530**	.349	.434*	–

* $p < .05$. ** $p < .01$. (2-tailed).

CA1: Technical aspects of modes, CA2: Conventions of linguistic modes, CA3: Coherence of multimodal product, CA4: Organization, CA5: Audience awareness, CA6: Quality of content

Table 17 Pearson Correlations Among the Kristen's Scores on 6 Categories

Categories	CA1	CA2	CA3	CA4	CA5	CA6
CA1	–					
CA2	.041	–				

CA3	.288	.650**	–			
CA4	.568**	.000	.267	–		
CA5	.304	.117	.259	.394	–	
CA6	.517**	.423*	.572**	.381	.360	–

* $p < .05$. ** $p < .01$. (2-tailed).

CA1: Technical aspects of modes, CA2: Conventions of linguistic modes, CA3: Coherence of multimodal product, CA4: Organization, CA5: Audience awareness, CA6: Quality of content

Table 18 Pearson Correlations Among the Lindsey's Scores on 6 Categories

Categories	CA1	CA2	CA3	CA4	CA5	CA6
CA1	–					
CA2	.190	–				
CA3	.398*	.559**	–			
CA4	.466*	-.052	.342	–		
CA5	.416*	.138	.315	.486*	–	
CA6	.347	.331	.431*	.348	.424*	–

* $p < .05$. ** $p < .01$. (2-tailed).

CA1: Technical aspects of modes, CA2: Conventions of linguistic modes, CA3: Coherence of multimodal product, CA4: Organization, CA5: Audience awareness, CA6: Quality of content

Qualitative Evidence on the Structural Aspect of Construct Validity

Raters provided comments on the rubric structure and the relationship between criteria and categories in the designed rubric. Both raters perceived that some criteria

overlapped each other. However, the criteria that each of them thought to be related were different.

For example, Lindsey thought that the descriptions of different levels of C11-quality of opinion seemed to include some content related to C9-audience awareness:

This quality of opinion, goes kind of into your audience awareness, too. Like how you're presenting your opinion, your thoughts. I found that sometimes the opinion was peppered throughout also.

The correlation matrix presenting relationships between criteria of each rater supported Lindsey's interpretation of overlap between C9-audience awareness and C11-quality of opinion. Unlike the very weak correlation coefficients between C9-audience awareness and C11-quality of opinion from the experts and Kristen, Lindsey's correlation matrix presented statistically significant correlation between the two criteria at .05 level ($r = .407$).

On the other hand, Kristen had difficulty separating the content related to the C10-quality of summary and C11-quality of opinion. She stated:

[Be]cause for certain students it seemed to be split in some ways than it always kept all the summary elements together. Sometimes they wrapped up certain pieces in the conclusion where they had all the elements, but they didn't keep all the summary together, so I just want to make sure it wasn't a support for their opinion or their likes and dislikes if it was connected in some ways. So that threw me off a little bit, so I would re-watch those videos several times to make sure the elements of the summary, that I was getting all of them clearly.

As Kristen stated, some students provided some parts of the *Frindle* story in the opinion sections (i.e., one thing I liked and one thing I did not liked about the story). For example, in his DMBR video, Ethan said:

One thing I did like is when Nick comes up with the word 'Frindle' for pen. One thing I do not like is when Miss Granger got mad and kept giving students detention.

The actual scoring result on Ethan's quality of summary criterion supports Kristen's difficulty with separating the content of C10-quality of summary and C11-quality of opinion. She included two story elements—that Nick created the new word *Frindle* and Miss Granger kept students in detention—as the parts of the summary. As the results, she gave higher scores on the quality of summary (4 points) than the experts and Lindsey (2 points). However, the quantitative relationship between Kristen's scores of criteria 10 and 11 did not support her difficulty.

In sum, both quantitative and qualitative evidence on the structural aspect of construct validity of the developed rubric exhibited that each criterion of the developed rubric measures discriminant characteristics of DMBRs except C4-conventions of written language and C6-relationship between oral language and written language. However, some categories of the rubric such as CA1-technical aspect of modes and CA4-organization of multimodal product, CA2-conventions of linguistic modes and CA3-coherence of multimodal product, and CA3-coherence of multimodal product and CA6-quality of content were highly correlated. This can refer to the fact that the categories may not reflect the structure of the DMBR as a construct.

Qualitative Evidence on the Content Aspect of Construct Validity

During their individual interviews, raters provided comments on the relevance and representativeness of the rubric content.

Relevance: No unnecessary criteria due to the multimodality. Regarding the relevance of the rubric content, Kristen reported that there was no unnecessary criterion in the rubric:

Well for me, I think one of the questions I had was because it's a multimodal product, it's difficult to remove because you're assessing content you know, which would be the summary, the opinion, all those things, organization. But then you're assessing the mode to which they're delivering that which is the iMovie in this case or the digital story telling basically. So if you're only going to score based on their use of technology, that would be its own rubric, if you're only looking for content, that would be its own rubric but since you're looking at both of them together, honestly, I don't know how you really would eliminate too many of these.

To Kristen, evaluating multimodal products was to consider the technological and content aspects of them together. She thought that if she used a scoring rubric for print-based book reviews and another rubric for iMovie products separately, the two rubrics would not be able to capture the multimodal aspects of the digital book reviews.

Lindsey also had similar thoughts on the relevance of the rubric's criteria. Particularly, she stressed that using the rubric was important and beneficial in order to evaluate multimodal aspects of the DMBRs and each mode and their relationships (i.e.,

C1 through C8). The following excerpt is Lindsey's statement on criterion 5, relationship between visual and oral language:

I actually enjoyed evaluating this part. Because, I liked seeing how well they supported themselves with that tool. Because that's what I thought was super important. I don't know if some of these things are included or could be included in this area, because when I liked being able to see, well, did that particular picture that they chose, did that really convey, (...) Or you saw some kids, who when they had the picture of the beach, they put the bird sound on the beach. Or the child who was saying she didn't like that Nick and his friend made bird sounds in the classroom, she put the bird sounds there while she had a picture of a desk. I thought that those were ways that they enhanced their message, and I liked being able to see that, and see that kids were able to do that. I don't know if you guys gave them much instruction to do that, or not. 'Oh, this is a little nugget. This is really neat.'

However, if she had to choose one criterion that needed to be removed from the rubric, she thought that it would be the C9-audience awareness. She said:

I was trying to think what you would cut out, because you have really good information in here. So it's hard. (...) Most of these kids have watched YouTube by now, and they know those YouTubers and they like Dan DTM or whatever, so they like these characters, and they would like to emulate them. So they'll say, 'Hey ...' Maybe just making sure that they're aware, so that might be like just something that the teacher tells them, 'You will have an audience for this. It's not

the book reports you're used to writing. You will have an audience.' Then maybe you don't have to measure that. Maybe you don't need.

In the statement presented above, Lindsey suggested that if the teacher would let the students know that they would have audiences for the project, the rubric would not need to include audience awareness criterion.

Relevance: Inappropriate label. On the rubric, C8-organization of multimodal content was intended to refer to the logical structure of content or messages within and among frames or sections. Specific descriptors of the criterion considered whether or not the student's artifact included six required sections (i.e., introduction, brief summary, one thing you liked, one thing you did not like, recommendation, and conclusion) and if the student weighed sections differently with the intention to emphasize specific content. From Lindsey's understanding of organization, the word seemed to be inappropriate for the content in the criterion. She stated:

I tried to look as if they had all the categories that were listed in your protocol, so that I could make sure that I was looking, and you can see on my rubric, like I stated writing introduction, summary, like, dislike, recommendation and conclusion. So I put those pieces there, so I could kind of check them off and see if they did it. I don't know here, where you're talking about organization, didn't necessarily say what order they should do it in, so I didn't know if it really fit into organization or not. (...) That might end up being ... Let's think. The idea was really that they included all of the artifacts. This doesn't really say that it has to

be in order. It could be emphasis of required elements, you know, or something like that.

The excerpt from Lindsey's interview revealed her understanding of organization as a similar criterion included in other writing rubrics. Usually, the organization criterion of writing rubrics evaluates overall coherence of writing by examining the inclusion and logical connection of key content in the written product. Lindsey seemed to have the impression that the descriptors of C8-organization of multimodal content required the rater to focus only on the inclusion of components and not to consider the logical connection between and among the components.

Relevance: Tough language. Lindsey stated that some words used in the descriptors of C3-conventions of oral language made it difficult to distinguish level 2 from level 3. She said "some of the language [is] still tough" due to the words included in the level 2 and 3 descriptors presented below:

C3-Level 3: Oral language in the artifact has sentences that are generally complete with sufficient variety in length and structure. Some sentences do not follow the conventions of Standard English suggested in [the CCSS](#).

C3-Level 2: Oral language in the artifact has short and simply structured sentences. Almost all sentences follow the conventions of standard English suggested in [the CCSS](#)

Although the first descriptors under each level differentiated levels 3 and 2, the words in the second descriptors (e.g., "some sentences do not follow the conventions" and "almost all sentences follow the conventions") indicated similar levels of conventional knowledge

of students. The raters' suggestions for the revision of the tough language will be discussed in the rubric revision section.

Relevance: Appropriate for teachers, not for students. In terms of the difficulty of language and content, the raters considered that the rubric would be appropriate only for teachers, not for students. Kristen stated:

I think the rubric is clear from a teacher's standpoint. I think if I were to give this to my students, it would be overwhelming. It would be difficult for them to process what they're being asked to do. So, I don't know if there would be a way to modify for kid-friendly language with this.

Representativeness: Mismatch between the raters' expectations and the described levels of performance. Raters reported that C4-conventions of written language and C10-quality of summary included some descriptors describing the levels of performance that do not correspond with the raters' expected levels of performance.

Regarding C4-conventions of written language, Lindsey reflected that the rubric "pushed" her to give a certain score that she did not want to. She stated:

I think that this (conventions of written language) was one that I sometimes was pushed, when I didn't really want to give a good, I might have to give a good, even though I didn't want to. Or, if I thought that it would have been a good, but I had to give it a fair, because it had more than 50% of the frames. I could have gone either direction.

The cause that pushed her to give unwanted scores was the quantitative descriptor that first led her identify whether the artifacts include written language in more than 50% of

frames or not. The rubric was created to give 4 points or 2 points, if more than 50% of frames included written language. It seemed that the frequency of written language in the artifact was not important for Lindsey to evaluate the conventions of written language.

Regarding C10-quality of summary, the descriptors made Lindsey assign higher points to students than she would assign without the rubric.

Lindsey: Even if I didn't agree as a teacher that I thought it was as good, but because of the tool that I was using, yes, they did what they were supposed to do, maybe ... I think we talked about that in the training, like the summary. When we looked at the summary, it wasn't a very good summary. However, it met all of the pieces that they needed for this. (...) when we look here, it really was doing what it was supposed to be doing according to this rubric.

When I asked Kristen about the quality and accuracy of the descriptors in the rubric, she also stated that her expectations for students' summaries and the scores generated by the rubric were different.

Kristen: When I was looking at 25 different products, there is human nature tendency that says, "wow this one was really great and this one was really poor." You know you really had to and I know myself with spending more time per student or per product was really over trying to eliminate some of that human error as it were to just look at 'Okay but that's what the criteria says,' and that was I thinking something that I had to fix in myself especially the quality of the summary. As the teacher, what I would expect of my students in terms of their summaries wasn't necessarily ... Like I was being too harsh of a grader but then I

had to really look at if they met a four and they met all those requirements then they deserved a four whereas maybe I would have given it a three just because I wanted some more.

Both Lindsey and Kristen considered that the content of C10-quality of summary did not correspond with their expected level of students' performances for each level. The only difference between the two raters was their attitudes toward the discrepancy. For example, Kristen expressed the difference as the result of human error or subjective interpretation, whereas Lindsey considered it as the result of misrepresentation of the desired performances.

In conclusion, the raters' feedback on the content of the rubric revealed that the rubric contains relevant content to evaluate upper-elementary students' DMBRs. The reason why they considered most of the criteria in the rubric as relevant was due to the multimodal aspect of the digital book reviews. However, the raters provided their opinions on the representativeness of the rubric content only for the quality of summary criterion.

Qualitative Evidence on the Usability of the Rubric

It is very important to create a useable assessment. In this study, the usability of a rubric is more related to the convenience of using it. Physical structures of the rubric such as length, layout, and page breaks can influence on the usability of the rubric. Regarding the usability of the developed rubric, the raters showed contradictory responses.

When I asked Kristen to evaluate the usability of the rubric with one to four points, one as most difficult to use and four as very easy to use, she stated, "It was easy, a

four. I didn't find any problems with it." She stated that the rubric was "perfectly suitable" to be used in her classroom and she "would love to use it again in the future. She especially liked the memo boxes that included in the scoring sheet. The memo boxes motivated her to justify her scores on each criterion.

On the contrary, Lindsey evaluated the rubric as less user-friendly. She particularly felt that searching for information from both the rubric and scoring protocol was burdensome. Lindsey realized that she did not refer to the scoring protocol once she internalized the information in the scoring protocol helping interpretation of rubric descriptors. To her, useful information in the scoring protocol was about C9-audience awareness and C10-quality of summary. Although she already mentioned that it was difficult to find unnecessary criteria from the rubric except the audience awareness criterion, she still thought that other teachers "would be overwhelmed to start with" by the length of the rubric.

An inappropriate page break on the rubric was another factor that led her to evaluate the usability lower than Kristen. Lindsey said:

And really, with it being on a page break like this, I had to keep paying attention to ... I couldn't just go five, five, three to four, two, because then I thought, "Oh, I have to turn the page. Oh no." And then I had to fix a couple of them, because like, "Okay, three to four, two, whatever, that's easy." (...) I didn't like it on the page break. Because it made it harder for me to remember that there was more to it.

The wrong page break was not intended by the researcher. Rather, it originated from the different technical setting of Lindsey's Microsoft word program. Due to the difference of the version of Microsoft word program, a few lines of descriptors included in criterion 7, relationship between oral language and sound, moved to the next page, and Lindsey found it later. Nonetheless, Lindsey expressed her intention to use the rubric after she figured out an appropriate tool for the composition of DMBRs.

Lindsey: Oh, I intend to [use it]. I just have to figure out my tool, because I don't have the iMovie thing. We have Chromebooks, so I have to figure out the tool that I'm going to use. Right now they use a lot of slides for Google Slides, and PowerPoint is maybe a little friendlier for that. There is the iMovie, but I don't know if that's on Chromebook or not.

Results of Research Question 4: Rubric Revision

In this section, the causes of different and inconsistent interpretations of the rubric, which were identified by the focus group interview, are discussed. Then the raters' suggestions for its revision that were identified from both the focus-group and individual interviews are exhibited. Finally, the revised rubric based on the raters' suggestions is presented. I also provide a rationale for why some suggestions were not adopted.

Problematic Criteria that were Identified in Research Phase 3

The quantitative analyses on inter-rater reliability and the qualitative analyses of rater interviews conducted for the results of the research question 3 section revealed that raters interpreted C3-conventions of oral language and C11-quality of opinion very

differently and more inconsistently than other criteria. In case of C10-quality of summary, the Cohen's kappa value obtained by the comparison between Kristen's and Lindsey's scores was much lower (.036) than the Cohen's kappa values of the other two comparisons: the experts vs. Kristen (.644) and the experts vs. Lindsey (.369). I wanted to understand why and how the raters interpreted the criterion differently.

In terms of the evidence of structural validity, C4-conventions of written language and/or C6-relationship between oral language and written language seemed to be revised since they were statistically significantly correlated with the moderate to large effect sizes from all raters (ranging from $\rho = .599 \sim .710$, $p < .01$). In the following section, the specific problems of the above-mentioned criteria (i.e., C3-conventions of oral language, C4-conventions of written language, C6-relationship between oral language and written language, C10-quality of summary, and C11-quality of opinion) and solutions to resolve the problems that were identified through the focus group or previous individual interviews are presented.

Problems and Solutions for C3-Conventions of Oral Language

C3-conventions of oral language was problematic in terms of its very low values on inter-rater reliability and the structural aspect of construct validity. To be specific, the mean exact agreement was 37.3%, the mean Cohen's kappa value was .161 and the lowest Spearman's ρ on the criterion between any two raters was .300.

The possible causes of the problems on inter-rater reliability and structural aspect of construct validity were detected during the individual and focus group interviews. As mentioned in the results for research question 3, the tough language in the descriptors of

the criterion gave Lindsey some difficulty when she had to assign a score of 2 or 3. During the focus group interview, Kristen also addressed the confusion caused by the words “some sentences do not follow” and “almost all sentences follow” in the second descriptors of levels 2 and 3 of the criterion. Additionally, Kristen stated the relative difficulty of grading oral language compared to written language. Since the teachers might not have time to make a script of students’ oral language, they might have to depend on what they heard. For teachers grading in real time, evaluating sentence lengths and structures based on what they heard would be more difficult than evaluating them from what they read. Moreover, sometimes the technical quality of audio might be poor, or less than ideal. On this difficulty, Kristen stated:

Even if they're complex or they're not either it's all or it's not, so I think too, part of it is you're getting discrepancies because of hearing. (...) You're not grading a written product, you can't break [it] down, this is a compound sentence, this is a simple sentence. (...) The one where it's about aspects of audio, voicing, number two, if they mumble, or if their volume isn't as easy to hear, you're going to miss certain things like other criteria impacting them, as opposed to a written product, you're just going to look at their written language. I think it would help though, I think it would maybe alter some of the discrepancy and get it closer to a one-point difference, either plus or minus.

Lindsey suggested two possible solutions for the revision of criterion 3. The first solution was to divide criterion 3 into two different criteria: 3A on sentence lengths and structures and 3B on conventions of standard English. She thought that if I provided four

different levels for both 3A and 3B criteria using quantifiable words, such as from “all” to “none” or from “almost all” to “a few,” future users may be less confused. However, she also could not come up with good descriptors for each level, considering sentence length and structure and conventions of standard English.

Lindsey: I know that when we talked, I made notes all over and thought about how you could either separate those two into a 3A and a 3B, as areas, I think that it made it too hard. We were talking about, for instance, in a one, in the second bullet, in a one, ‘very few’ ... it's still slightly subjective, but very few sentences do not follow the conventions, versus many sentences follow the conventions, to most sentences, to all sentences, or almost all. Which, between most and almost all, is it partly sunny or mostly cloudy?

The second and the easiest solution was to delete “do not” from the second descriptors of level 3 and 1. Lindsey stated:

I think also when we had talked about it, if you eliminate the ‘do not’ part of that, and if you're focusing on what it does, then that will help, too. If it's very few sentences follow, if it's many sentences follow, if it's most, or if it's all, you have to figure out the wording, but I think if you're focusing on what it should do instead of what it does not, I think that's also a little clearer.

Summing up the raters’ suggestions, I revised criterion 3 by using less confusing and more separable, quantifiable words and phrases (i.e., “almost all,” “more than half,” “some,” and “only a few” or “none”) and by deleting the descriptors on sentence lengths and structures. Appendix P includes the revised criterion 3.

Problems and Solutions for C4-Conventions of Written Language and C6-Relationship Between Oral and Written Language

Quantitative analyses of raters' scores identified that the correlations between C4-conventions of written language and C6-relationship between oral language and written language were statistically significant and they reflected moderate to large effect sizes from all raters (ranging from $\rho = .599 \sim .710$, $p < .01$). These correlation coefficients indicated that the two criteria might measure similar aspects or characteristics of the DMBRs. During the focus group interview, I asked the raters what they thought about the descriptors of and the relationship between criteria 4 and 6.

Lindsey restated what she mentioned during the individual interview about the first descriptor of criterion 4: the quantifiable phrases such as “more than 50% of frames” forced her to choose either level 4 or 2. In regard to revising the criterion 4 descriptor, I asked raters what they thought about deleting the first descriptor of criterion 4. First, I explained why I had to consider the proportion of frames including the written language first to grade conventions of written language. Then I asked raters how they would grade the examples I provided. The following excerpt shows the conversation:

Sohee: If I just assess the conventions of written language based on the number of errors that included, that means the students who include only two words without error, they will get the score four. (...) Then, how about these? They included just three words in only one frame of the entire video without any errors. Or, they included their written language in two frames. For example, in the first frame, they put the word 'Frindle,' and in the second frame, they just put the word

'summary.' In this case, they just used only two words, but they included the two words in two frames. So they will not get score one, because the number of frames is two and there are no grammatical and spelling errors.

Kristen: *Then they get a four?*

Sohee: *They get a four. Do you think that that's fine?*

Kristen: *I do, because it's a multi-modal product. Well ... let me think about that.*

The vast majority of the product is oral, you'd have to specify how much they'd have to have written. Grammar ... because most kids are going to write things like the title of the book, maybe this is the summary, maybe this is their opinion, but most students probably aren't going to caption pictures that they're using because they're probably viewing the oral language as sufficient to describe what's presented on the screen. This type of product, you're limiting the amount of writing they're doing anyway. To me, it's like an alternative form of assessing.

Kristen's response to my question revealed that the amount of included written language was less important in this case, because she evaluated a multimodal product, not a written essay. Moreover, she thought that the decision to include written language in the DMBR would be solely the decision of the student.

I also agreed with Kristen's arguments, so I asked the raters what they thought about deleting C4-conventions of written language from the rubric. Lindsey's response was not clear. It seemed that she would be okay with the deletion, "if we're not expecting much beyond a sentence or a subheading" in the DMBRs. However, she continuously

said, “I don't know.” On the other hand, Kristen said deleting the criterion would be okay because “we'll grade conventions on their written work [based on other writing tasks].”

After this conversation, I decided to keep the C4-conventions of written language but made revisions to the descriptors. I deleted the descriptor about the percentage of frames including written language in part because the proportion of frames that included written language was also evaluated under criterion 6. Then I changed the word “error” to “type(s) of error.” For example, if a student repeatedly does not capitalize the first letter in the word or sentence, that will be counted as one type of error. Appendix P contains the revised descriptors of the criterion.

Regarding the first descriptor of C6-relationship between oral language and written language, Lindsey said that presenting a specific number of frames that included written language would not make the rubric a universal tool. Kristen also agreed with Lindsey's opinion and suggested replacing the number of frames with percentages.

Kristen stated:

To me personally, I like the language more of the percentages better than a number of frames, because student work product is going to vary widely. Some of the sample ones we had done only had written language. So you're scoring based on what they're presenting, as opposed to I think as a teacher, the second you set a number, it's six, or you're always going to have the one or two overachievers who [say], 'Well I did 15,' well okay, they do that. But the vast majority of students do the bare minimum. You must have three paragraphs, as soon as they hit three they're finished, but they probably could have broken into four or five

because the assignment [inaudible]. I think the percentage is going to get you a stronger student work product than saying a number of frames.

As a classroom teacher, Kristen preferred descriptors with percentages rather than the number of frames because she knew that the quality and style of students' products varied, and some students have the tendency to meet only the minimum requirements as stated on the rubric.

Based on Kristen's suggestion, I decided to use percentage descriptors for C6-relationship between oral language and written language instead of the "number of frames" descriptors. Details of the changes can be found in Appendix P.

Problems and Solutions for Criterion 10-Quality of Summary

When I compared the scores assigned by the experts and Kristen (.644), and the experts and Lindsey (.369), Cohen's kappa values of the two pairs were acceptable. However, the Cohen's kappa value obtained by comparing Kristen's and Lindsey's scores was very low (.036). I needed to understand why and how the raters interpreted the rubric differently.

Raters identified problems that were linked to the criterion 10, quality of summary, during the individual and focus group interviews. When we discussed the reasons why we graded certain students' products differently on this criterion, Kristen reported that she counted the story elements that were presented in other parts of the book review such as "one thing I like," "one thing I did not like," and "recommendations," as major events of the story. Her perspective on evaluating the brief summary section was different from both my perspective and Lindsey's perspective. The following

conversation shows how Kristen dealt with the story elements included in the opinion sections (i.e., one thing I liked, one thing I did not like, and recommendations).

Kristen: Okay. Ethan ... quality of summary ... I gave Ethan a score on his summary because I wrote down things like he had the names of the main characters, the fact that he came up with this word 'Frindle,' which meant pen, his fellow students start saying it but then the teacher had them in ... gave the students detention. But then it kind of ... this is maybe where I could have misinterpreted. Then he comes back around to it towards the end in what he liked about it, then Miss Granger gets 'Frindle' into the dictionary. 10 years later he ends up donating this money back to the school. He kind of, I think, went from beginning, middle, end, but it was a little disconnected. Maybe I went too high, I don't know.

Sohee: I understand. When you graded the quality of summary from Ethan's product, you picked up some summary kind of content from recommendations section.

Kristen: Yeah.

Sohee: You counted that as parts of the summary, so that's why your score was higher than mine and Lindsey's I guess.

Kristen: Right. I took the whole work product instead of just that he put it in, but that he would have ended with his like, his dislike, and his recommendation. I felt like he came back around to his closure was finishing parts of the summary.

As the conversation exhibited, Kristen focused on counting the story elements appearing in “the whole work product.” This difference might originate from her different perspective on the role of summary in DMBRs. In the following excerpt that Kristen emphasized the importance of instructional context of the book reviews for grading, her point of view on the summary in DMBRs was also displayed.

Kristen: I understand the rubric is grading an assignment, but was the assignment presented as if they had to start with the summary of the book, then a like, then a dislike, then a recommendation? Or are we looking at the work product as a whole? Because I think if a student just ... ‘your product must include all of these things and order doesn’t matter,’ it’s very different than interpreting as a teacher, ‘but I told you this had to be first, second, third,’ and then they’re losing points because that content was presented incorrectly. I don’t know if that makes any sense, but I think it would be dependent upon what the student was instructed to do in terms of he did have all of these things, it just maybe wasn’t as logical given the fact that we would have preferred to see him start with the summary.

Kristen’s main argument was that if the teacher did not teach students to create their book reviews in order from the introduction to the conclusion sections, the teacher should not focus on just the brief summary section for grading of the quality of summary criterion.

Lindsey understood why Kristen interpreted the quality of summary criterion differently. Unlike Kristen, however, Lindsey stated that if the rubric was created to grade the quality of summary as a coherent and separate section, the rubric should be interpreted in that way.

Lindsey: *I understand what Kristen was talking about when she was counting the extra things that he said, also. ... But if the rubric is being developed for what you hope it would be used for, then maybe it isn't out of line to think that all of the summaries should be together. Because in the future with teachers instructing with the rubric in mind, they would be able to emphasize that for students.*

In sum, the descriptors of C10-quality of summary did not have critical problems. Rather, difference of scores between the raters originated from different beliefs on the role of the summary part in the DMBRs. Nonetheless, I made changes in the descriptors in order to clearly deliver my intention on the criterion and decrease the probability of misinterpretation: “The summary” was changed to “the brief summary section.”

Appendix P presents the revised rubric, including this revision.

Problems and Solutions for C11-Quality of Opinion.

Quantitative analyses of raters' scores on C11-quality of opinion revealed that patterns of grading between the experts and Kristen ($\rho = .325$) and Kristen and Lindsey ($\rho = .349$) were inconsistent. The focus group interview identified the reason for inconsistent grading: the raters' confusion about the definition of opinions and supporting details resulted in their inconsistent counting.

The descriptors of C11-quality of opinion are presented below:

Level 4: The author provides a total of three or more supporting details or reasons in the opinion sections (i.e., “1 thing I liked,” “1 thing I did not like,” “and “recommendation”)

Level 3: The author provides a total of two supporting details or reasons in the opinion sections.

Level 2: The author provides only one supporting detail or reason in the opinion sections.

Level 1: The author does not provide any supporting details or reasons in the opinion sections.

Level 0: The author does not provide any opinion about the book.

Kristen addressed that the “i.e.” part in the rubric descriptor (i.e., “1 thing I liked,” “1 thing I did not like,” and “recommendation”) made her confused, so she counted them as supporting details or reasons.

Kristen: To be honest, I wouldn't put in number four a rating of four ... or the reasons and opinion sections, i.e. and it is my semantics misreading, I looked at it as 'Okay, they have a thing they liked, a thing they didn't, a recommendation, they got all of them.' I look that a clarifier of the opinion sections would be the like, dislike, and recommendation. That was an oversight on my part, but because it didn't appear in three, or reasons and the opinion section. I would remove the i.e. part, unless you're going to have it here in all of the other (descriptors). When you use the expression 'opinion sections,' I wouldn't use the i.e. as a clarifier.

Lindsey agreed with what Kristen said. She said, “I think how Kristen was saying, I probably was distracted by one thing I liked, one thing I did not like, and here's my recommendation.” Revisions to the descriptors of criterion 11 were made to reduce the chances of misinterpretation (Appendix P).

Chapter 5

DISCUSSION AND IMPLICATIONS

Overview of the Study

This study sought to achieve four main goals: (a) identify domains and criteria that can be universally applied to assess any digital-multimodal-composition genre; (b) create a scoring rubric for evaluating upper-elementary students' DMBRs; (c) confirm the inter-rater reliability and the content and structural aspects of construct validity of the proposed rubric using both quantitative and qualitative analyses; and (d) revise the used rubric based on the evidence of inter-rater reliability, construct validity, and raters' feedback.

To identify key domains and criteria of DMC, I began by conducting a systematic review of previous empirical studies assessing DMCs. This systematic literature review identified a total of 19 criteria in 5 domains. In the second phase of the study, I created a first draft of the new scoring rubric by considering these domains and criteria as well as the characteristics of upper-elementary students' DMBRs. Working with my collaborator, I then went through several revisions to improve the coverage and clarity of the proposed rubric.

During the third phase of the research, two raters received one offline and one online training session to familiarize them with the content, structure, and usage of the new scoring rubric. After the raters independently graded 25 typical student DMBRs, I

conducted one-on-one interviews with the raters to learn more about their opinions on the proposed rubric's content, structure, and usability. Quantitative analyses using the raters' scores were performed to examine a variety of inter-rater reliabilities such as consensus agreement and consistency between raters and appropriate values for the structural aspect of the construct validity. Qualitative analyses, including both open and axial coding, were also conducted on transcripts of the individual rater interviews. The findings from both the quantitative and qualitative analyses highlighted the importance of achieving consensus agreements and consistency in this context, as well as examining the structural and content aspects of construct validity.

In the final phase of the study, a focus group interview with the two raters was held in order to explore the underlying causes of the inconsistent interpretations by the raters of parts of the rubric and generate ideas for appropriate revisions to address these issues. Rather than preparing a series of semi-structured interview questions, I instead presented the results of the quantitative analyses and asked the raters to explain their interpretations of criteria 10 and 11 while they referred to their scoring logs. I also asked them to suggest useful ways to revise criteria 3, 4 and 6. Based on the raters' responses, further revisions were then made to the rubric.

This chapter discusses the inter-rater reliability and validity of the proposed uses of the newly-developed rubric in more detail. After addressing the rubric's appropriateness as a formative assessment tool, limitations of the current study are presented. Finally, the chapter concludes by considering potentially fruitful directions for

future research and providing suggestions to help teachers use the revised rubric as a formative assessment tool for DMBRs in upper-elementary classrooms.

Inter-Rater Reliability

The research conducted for this study analyzed and presented three different measures for consensus agreements (the percentage of total agreement, the percentage of adjacent agreement, and the Cohen's kappa) and one measurement of consistency estimate (Spearman's *rho* coefficient) to examine whether the rubric was used reliably by two different raters.

Consensus Agreements

The results showed that the means for five of the criteria were consistently at or above 60 percent of exact agreement among raters, which was the target exact agreement rate for this study. These five criteria were the C4-conventions of written language, C5-relationship between visual and oral language, C6-relationship between oral language and written language, C8-organization of multimodal content, and C11-quality of opinion. Concerning the adjacent agreements, the means of all the criteria except the C4-conventions of written language exceeded the target agreement rate (80%), and the means of seven criteria exceeded 90% for agreement. These seven criteria were the C1-technical aspects of visual mode, C2-technical aspects of audio-video, C6-relationship between oral language and written language, C7-relationship between oral language and sound, C8-organization of multimodal content, C9-audience awareness, and C10-quality of summary. The results of Cohen's kappa, the most conservative measurement of

consensus agreement, revealed that only the C4-conventions of written language, C5-relationship between visual and oral language, C6-relationship between oral language and written language, and C9-audience awareness represented fair to strong agreement beyond chance. Table 19 shows the criteria that met or exceeded the target levels of consensus agreements for this study.

Table 19 Criteria Meeting the Target Levels of Consensus Agreements

Criteria	Exact Agreement	Adjacent Agreement	Cohen's Kappa
1. Technical aspects of visual mode		O	
2. Technical aspects of audio-voice		O	
3. Conventions of oral language		O	
4. Conventions of written language	O		O
5. Relationship between visual and oral language	O	O	O
6. Relationship between oral language and written language	O	O	O
7. Relationship between oral language and sound		O	
8. Organization of multimodal content	O	O	
9. Audience awareness		O	O
10. Quality of summary		O	
11. Quality of opinion	O	O	

Overall, all criteria met or exceeded the target level for the agreement between raters in at least one measurement. Although the exact agreement rates were in some cases below 60%, most criteria were over 80% or nearly so for adjacent agreement. One criterion, C4-conventions of written language, achieved an exact agreement rate higher than 60% on average (72%), but its average adjacent agreement rate was lower than 80% (78.7%). For this criterion, raters exactly agreed on many students' written language, but extremely disagreed on a few students' DMBRs. The Cohen's kappa results were closely related to the results for exact agreement. Most criteria that exceeded 60% for the agreement rates (namely criteria 4, 5, 6, 8, and 11) obtained .3 or greater values in Cohen's kappa analyses. In particular, two criteria, the C5-relationship between visual and oral language and the C6-relationship between oral language and written language, established good levels of inter-rater consensus agreements in all three measures. This result indicates that C5 and C6 may be clearly written and there is thus a high likelihood that they will be interpreted in the same or similar ways by different raters in the future.

These findings provide additional evidence that relying on a single measure for inter-rater reliability could be misleading and should, therefore, be avoided whenever possible (Jonsson & Svingby, 2007; Stemler, 2004). As the results of the current study show, different measurements of inter-rater reliability exhibit different results and researchers and educators seeking to develop a rigorous rubric for use in classroom assessments would be well advised to examine a variety of inter-rater reliability measures to increase a new scoring rubric's applicability with different raters. If rubric developers or educators are forced to rely on a single measure for inter-rater reliability, however,

Cohen's kappa is preferable as it provides conservative inter-rater reliability beyond chance (Stemler, 2004).

Consistency Estimates

This study measured the Spearman's *rho* correlation between raters as a consistency estimate for the raters' uses of the rubric. If the Spearman's *rho* correlation coefficient is high (e.g., $rho > .70$), this could be interpreted as indicating that the rubric is written in such a way that it helps raters to use the rubric more consistently (Jonsson & Svingby, 2007; Stemler, 2004). Here, the Spearman's *rho* analyses of the inter-rater scores revealed that the scores awarded for the C3-conventions of oral language and C11-quality of summary were less well correlated than the scores for other criteria.

The information that was gathered during the raters' individual and focus-group interviews also supports the less-consistent grading for these two criteria. In the case of the C3-conventions of oral language, the somewhat confusing terms included in the descriptors for levels 2 and 3 led the raters to assign scores subjectively. With regard to C11-quality of opinion, raters were confused about the intended meanings for opinion and supporting details. For example, the researcher and one rater (Lindsey) counted only the number of supporting details, whereas the second rater (Kristen) counted the number of opinions presented in the students' book reviews. The qualitative analyses of the focus-group interview revealed that the information was not explicitly delivered to the raters through the rubric, supplementary material, and training. These findings suggest that clarity of the rubric descriptors and the supplementary materials, as well as

transparency in the scoring training, are essential to ensure inter-rater consistency for grading.

Rater Training and Inter-Rater Reliability

As described earlier in the methods section, raters were trained in how to use and interpret the newly developed rubric for grading. Two different training sessions were held, one of which was face-to-face and the other synchronous online training. During the 3-hour offline face-to-face training session, I presented background information on DMC and multiple modes and explained each criterion in the rubric individually. I then provided an opportunity for the raters to engage in guided scoring practices. After the two raters had graded five students' DMBRs independently over a three-week period, the second session offered an hour of online synchronous training. During both rater training sessions, the newly developed rubric and a scoring protocol document explaining scoring procedures and supplementary details for the rubric interpretations were provided.

Several existing studies have highlighted the importance of rater training for inter-rater reliability (Dempsey, PytlikZillig, & Bruning, 2009; Knoch et al., 2007; Pufpaff et al., 2015; Rezaei & Lovorn, 2010). For example, Rezaei and Lovorn (2010) examined how students majoring in either education or business/marketing graded essays on two social science topics differently when they assigned scores without and with the assistance of a rubric. The authors hypothesized that using a rubric might increase the reliability of grading by lowering the score variability compared to grading without the guidance of a rubric. They also suggested that education major students may be less influenced by mechanical errors in the essays than business/marketing major students

when they graded essays using the rubric. Unlike their initial hypotheses, however, the results of their study revealed that the rubric did not appear to reduce the variability of the scores awarded. Instead, the students focused excessively on grammar and spelling errors when they used the rubric for grading, even though the rubric was designed to assign only 10% of scores on these mechanical aspects. In addition, the education students' scores on the essays were not significantly different from those awarded by the business/marketing students, which might indicate that the education students had not received enough training. Rezaei and Lovorn (2010)'s results showed that without adequate training on the assessment purpose and design of the rubric, using a rubric did not guarantee more reliable and valid assessments of students writing.

Dempsey et al. (2009) examined the effect of training on the way an online assessment tool for elementary students' writing was used by pre-service teachers. In this case, the training included guided and scaffolded practice in assessing multiple student papers and feedback from experts and peers. Their results revealed that the scaffolded training, in particular, enabled pre-service teachers to assess elementary students' writing more accurately with a greater self-efficacy. This result can be interpreted as indicating that scaffolded scoring training with feedback from experts and peers might be an effective way to improve the validity of such assessments.

Unlike the above two studies, Knoch et al. (2007) compared the effectiveness of online and face-to-face training for writing assessments. Their study divided 16 raters into two groups, one of which received face-to-face training and the other online training. A multi-faceted Rasch analysis of the results revealed that both training methods were

effective overall, but the raters' self-report data mentioned slightly different effects for the two, with the face-to-face training being considered more effective at reducing the halo effects arising from the raters' misunderstanding of a number of descriptors and therefore assigning scores on the basis of their overall impression on the performance or artifact. In contrast, the online scoring tutorial for scoring was found to be more effective regarding consistency in the grades awarded.

Most of these existing studies on training raters to use rubrics have simply confirmed the importance of providing well-scaffolded and sufficient training on the rubric's purpose, design, and content. In the current study, raters' feedback on the rater training of this study and their suggestions for future rater trainings supported the findings of previous studies. During the individual interviews, Lindsey mentioned the importance of scoring training and suggested some possible forms of training for more reliable scoring. She particularly wanted to have continued access to training materials such as tutorial videos to enable her to watch again the training material at any time.

Lindsey: I think for some teachers, they might really appreciate it if you did like an online learning kind of video, where they could pause it. And if you did it kind of like a lesson, where you say, "Okay, let's watch ..."
Who was one of the kids? Like Erin. "Let's watch Erin together, and let's count all of the frames first." And so you watch it, and say, "Let's count all the frames." And so they count all the frames, and then they can input their score. And they have those different things, right? Blackboard, or whatever.

This feedback, along with the findings of Knoch et al. (2007), implies that providing both face-to-face and online tutorial videos as part of the scoring training might increase inter-rater reliability by helping raters to understand the criteria and their descriptors better and thus grade students' artifacts consistently, not too harsh and not too leniently.

The scoring procedures that were emphasized by the scoring protocol may be a factor related to the consistent and discriminant scoring with less halo errors. To be specific, the current study trained and required raters to score students' DMBRs for a few criteria under the same category at once, which was not a typically used method by teachers. Generally teachers use a rubric to score all criteria of one student's performance and move to score next student's performance. This scoring method is susceptible to halo errors that raters score each student's performance based on their overall impression, not on their actual performance by various criteria (Knoch et al., 2007). On the contrary, the scoring protocol of the current research asked the raters to score C1-technical aspects of visual mode and C2-technical aspects of audio-voice under the CA1-technical aspects of modes together included in all 25 students' DMBRs by following the procedures listed in the Appendix B:

1. Open Dede's book review and count the number of frames, titles, and sounds (sound effects = SE & theme music = TM) while watching the review.
2. Read the descriptors for criteria 1 and 2 carefully.
3. Re-watch the review and evaluate technical aspects of visual and voice in each frame. If necessary, please refer to the page 3 of this scoring protocol.

4. Open next student's book review and count the number of frames, titles, and sounds (sound effects = SE and theme music = TM) while watching the review.
5. Re-watch the review and evaluate technical aspects of visual and voice in each frame. If necessary, please refer to the page 3 of this scoring protocol.
6. Repeat 6 and 7 until you complete scoring criteria 1 and 2 of all students' book reviews.

This training procedures might reduce the halo effect by forcing the raters to focus on each criterion and differentiate performances among students. However, the current study's data do not clearly explain the relationship between the training procedures and the halo effect. Therefore, further research is needed to determine the impact of different scoring training procedures on the halo errors.

Construct Validity

Content Aspects of Construct Validity

The content aspects of construct validity represent the most important part of the new rubric developed for this study. If the rubric measures skills or abilities that are irrelevant for the performance of the target task the rubric would not be valid, even if it generates scores that are consistent between raters. For this reason, examining if the rubric measures appropriate aspects of the construct is crucial.

The content-related aspects of construct validity consist of two main characteristics: relevance and representativeness (Messick, 1995). In this study, the

representativeness of the rubric's content was established through a systematic review of the literature on the DMC assessment, while the relevance of the rubric's content was obtained by analyzing DMBR tasks and upper-elementary grade students' DMC performances. In both their individual and focus-group interviews, the raters pointed out that the rubric was relevant in terms of evaluating the multimodality of students' DMBRs. Both raters provided positive feedback regarding their impression of the utility of the multimodality-related criteria (i.e., C1-technical aspects of visual mode, C2-technical aspects of audio-voice, C5-relationship between visual and oral language, C6-relationship between oral and written language, and C7- relationship between oral language and sound) because the rubric presented specific methods with which to evaluate new types of student products. However, the raters also thought that some criteria such as the C3-conventions of oral language and C10-quality of summary did not include descriptors that were relevant for the upper-elementary grade students they taught in their classes.

Overall, the domains and criteria included in the proposed rubric appear to represent the DMBR as a construct well. However, some content in the rubric may not be relevant for the evaluation of upper-elementary students' performances. Therefore, the rubric needs to be tested further by evaluating more student DMBRs to confirm its relevance and representativeness as an assessment tool for upper-elementary student DMBRs.

Structural Aspects of Construct Validity

Evidence of structural validity was obtained to show the (a) discriminant characteristics of each criterion and (b) stronger correlations among criteria within a category than criteria across categories. In the current study, the discriminant validity between criteria was measured by conducting the Spearman's *rho* analyses among criteria per rater. The latter was tested through both the Spearman's *rho* analyses among criteria and the Pearson's correlation analyses among categories per person.

The Spearman's *rho* analyses revealed that two criteria (C4-conventions of written language and C6-relationship between oral language and written language) were highly related, which meant that they were likely to measure similar characteristics of students' DMBRs. The raters' interview data also supported these quantitative findings. However, the Pearson correlation coefficients on six different categories did not provide much in the way of meaningful information about the structure of DMBR as a construct. Across the raters, CA2 and CA3 were highly correlated, but this could be because C4 and C6 were included in both these categories. CA1 and CA4 were correlated with each other with medium to large effect sizes for the experts and both raters. This might be because Criterion 8, organization of multimodal content, was included in CA4-organization, which looked at whether students organized their DMBRs coherently based on the inclusion of visuals in the final products. Since C1-technical aspects of visuals also focused on the visual mode, it was not surprising even though these two categories (C1 and C8) were highly correlated.

The final medium to strong correlation found was between CA3-coherence of multimodal product and CA6-quality of content. This correlation was expected because the criteria under CA3 were designed to evaluate the relationship between oral language and other modes. Criteria under CA6 such as C10-quality of summary and C11-quality of opinion are also evaluated based on information provided by the oral language mode.

Although the Pearson correlation coefficients per rater among the various categories provide some ideas of the structure of the rubric, the data collected for this study failed to statistically prove the established relationships between/among criteria under each category. For example, C1-technical aspects of visual and C2-technical aspects of audio-voice, both of which were included in the first category, technical aspects of modes, were rarely related to each other. This suggests that some of the new categories created for the proposed rubric might not reflect the aspects of the construct but instead simply represent convenient groupings of the various criteria. This is an important point and thus requires further statistical testing through exploratory or confirmatory factor analysis to determine whether the rubric structure does in fact adequately reflect the structural aspects of the construct.

The Rubric as a Formative Assessment Tool

The proposed rubric was developed to provide a tool for teachers seeking to conduct a formative assessment of the DMBRs of the upper-elementary students in their classrooms. Individual and focus group interviews were used to collect valuable feedback from the two teachers who tested the proposed rubric regarding its utility and

effectiveness as a formative assessment tool. Both raters provided meaningful comments. This feedback was organized based on the three strategies proposed by Black and Wiliam (2008) related to teachers' uses of formative assessment, as described below. This is followed by a discussion of whether the proposed rubric is suitable for teachers seeking to apply these strategies to enhance their upper-elementary students' DMBRs.

Strategy 1: Clarifying Learning Intentions and Criteria for Success

The first formative assessment strategy is related to the rubric's structure and content. Usually, scoring rubrics list the criteria students need to learn and descriptors explaining what would be considered a successful level of performance. As the two raters pointed out, there were no unnecessary criteria in the proposed rubric, which successfully listing almost all of the must know or must-be-assessed criteria. In particular, both raters mentioned that the criteria for evaluating multimodality (C1~C7) would become a useful guide when determining how to teach students best how to compose digital multimodal book reviews. Also, the descriptors for a score of 4, representing an excellent performance, for most of the criteria significantly clarified the success criteria. However, as the raters agreed during the focus group interview, the descriptors of some of the criteria such as the C4-conventions of written language and C10-quality of summary still did not accurately reflect what the teachers thought about the desired performance level. The revised descriptors for criteria 4 and 10 need to be confirmed by teachers if they adequately present the excellent performance.

Strategy 2: Engineering Effective Classroom Discussions and Other Learning

The second strategy, engineering effective classroom discussions and other learning tasks, is related to the discussion around the rubric between the teacher and students. Since the current study did not observe how the teachers used the rubric for instructional purposes, it was difficult to conclude whether the proposed rubric would be an appropriate tool for promoting active discussions of each criterion between teachers and students. However, the raters did consider that a teacher's level of knowledge regarding the rubric's content and the instructional context of the school where they work would influence the level of integration of the rubric for classroom teaching and assessment.

For example, Kristen pointed out that individual teachers' level of knowledge regarding multimodality and technology may influence their understanding of the rubric and the scoring training.

Kristen: I think with a teacher who wasn't trying to implement more of the digital text, they're not going to be as confident and comfortable with the technical aspects. So I think that could be a challenge because in that case what a teacher review as you know poor quality may or, you know as be even done well, may not be exactly what the rubric is discussing because the resolution, pixilation, angle, they may not understand. They won't even know how to change the Ken Burns effect so to them, they might feel like it's adequate. So I just think knowing ... The teacher would have to know something. They would have to have a foundation with technology.

She also mentioned that students' prior experience with tasks similar to the DMBRs and their level of knowledge regarding technology might influence the discussion.

Kristen: Especially that was the one question I think I had to ask you when we first met was how much technical training did the students have because if they just know iMovie in terms of just the very basics or just the very standards cause sometimes I think some of the technical errors or loss of points for students who didn't perform well in those sections could have been a lack of instruction for them.

To sum up, in order to use the rubric to facilitate effective classroom discussions around the DMBRs, the teacher and students should have knowledge on modes, modal affordances, and technical aspects of using the DMBR making tool. Without proper training or instruction on those aspects, the rubric may not work as a tool for effective classroom discussion and other learning.

Strategy 3: Providing Feedback that Moves Learners Forward

The third strategy, providing feedback, allows teachers to constructively respond to students' work with written feedback or checks on rubrics to help them improve their work.

Kristen thought that the content in the rubric could eventually facilitate effective classroom discussions on DMBRs and on each criterion by providing students with consistent feedback:

I think it definitely gives them more consistent approach to providing students with their score and feedback to be specific as a teacher instead of

saying that they needed to improve, you can say specifically where they needed to improve.

However, she also thought that the rubric's language would be too difficult for most upper-elementary students to understand. To be able to communicate the teacher's expectations on each criterion with the students effectively, she argued that a student version of rubric should also be provided that is written with more kid-friendly words.

Lindsey also mentioned the need of student-friendly rubric that both teachers and students can use together. She stated:

The students and the teachers can have the same scoring version. So the kids have access, they know what they're going to be scored on, and if they need further definition, the teacher can teach them that. But then they have that scoring. (...) If they would give themselves a four, three, two or one. And then they can have a student score and a teacher score, I've seen ones like that before too.

The raters feedback highlighted two important aspects of the rubric as a formative assessment tool. First, the rubric's language should be plain and easy to be understood by the students. Second, the rubric should be able to identify strengths and weaknesses of students' DMBR performances.

Overall, findings from this study indicate the new rubric developed for this study could serve as a good formative assessment tool for communicating learning intentions and the expected levels of DMBR performance for each criterion. However, since it requires teachers to have sufficient knowledge of and experience with multimodality and

digital tools for DMC, even with a well-structured and well-scaffolded training program on its use, it may be difficult for teachers who lack such knowledge and experience to use the rubric for the formative assessment of DMBRs in class as the developer intended. Therefore, professional development serves a critical role in its effective use. Particularly, teachers need to learn about the modes, modal resources, and relationship between modes as meta-languages on multimodality (New London Group, 1996; Kress, 2010; Unsworth, 2006). Moreover, because the rubric as written is aimed at the teachers, its language is not student friendly, hence the development of a student-friendly version is highly recommended to ensure smooth communication between teachers and students.

Limitations

Given the iterative nature of this study, the results of this research should be interpreted bearing in mind the following three limitations. First, the domains and criteria identified from the 111 criteria suggested in the existing literature may not represent an absolute and complete picture of the various aspects of the DMC as a construct. This limitation largely arises because the domains and criteria were not double-checked by another coder. It was also not possible to validate the relationships between the domains and criteria quantitatively via exploratory or confirmatory factor analyses in this study due to the small number of student products available ($n = 25$). The literature on factor analysis recommends the use of at least 100 participants (or artifacts) as a rule of thumb (Gorsuch, 1983; Klein, 1979). For this reason, it was not deemed appropriate to run either an exploratory or a confirmatory factor analysis for this current study.

Second, it is unclear whether the new rubric developed for this study will yield valid and reliable scores when it is used by other raters or for DMBRs on books other than the one utilized in this study. The two teachers who received the training and used the rubric were both tech-savvy and were already familiar with multimodal composition. It is therefore not possible to determine whether teachers who are less familiar with the technology and the concept of multimodality will be able to interpret the rubric in the same way as the two participating raters even if they receive identical training. Moreover, the new rubric created in this study was checked for inter-rater reliability and some aspects of construct validity based on 25 fourth-grade students' DMBRs that were originally collected for a larger study on another topic. This might mean that the rubric reflects particular task- and sample-specific characteristics, even though a considerable effort was made to consider more general aspects of upper-elementary students' DMBRs during the development process. For example, the descriptors for C10-quality of summary were based on the students' summary of the book *Frindle*. If students create DMBRs after they read stories that are either much shorter or much longer than *Frindle*, that might affect the number of details they need to include in the brief summary section of the book review.

Third, the revised version of the rubric that was presented in the fourth phase of this study was not validated. All the statistical analyses on inter-rater reliability and the structural aspects of construct validity were performed on the version of the rubric used for the independent scoring (rubric #7). Since the final version (rubric #8) was not validated, using the rubric to assess actual students' DMBRs is not desirable.

Implications

Despite the limitations described above, the results of the current study have useful implications for future research in this area and potential classroom applications of the rubric.

Future Research

Given the limitations of the current study, future research focusing on ways to improve the assessment of DMC is needed. First, a study on the inter-rater reliability and construct validity of the final rubric should be conducted by utilizing more than 100 students' DMBRs about the same book. Validating the final rubric with larger samples would resolve the most critical limitations of the current study. For example, the larger sample size would permit exploratory and/or confirmatory factor analyses on the scores of the rubric to be performed, thus providing more robust evidence related to the structural aspects of the construct validity of the proposed rubric. This would also increase the statistical power of analyses for the Spearman's *rho* and Pearson correlation coefficients.

Second, extending the applicability of the rubric to meet the needs of other teachers and other DMBRs should be investigated. By recruiting three or more raters who have different levels of background knowledge and experience regarding technology integration and multimodality, a range of interpretations of the rubric could be collected. In addition, since the rubric was intended for use as a generic tool for any DMBRs that created by upper-elementary students using the iMovie program, it would be interesting

to find out whether the proposed rubric continues to serve as a reliable and valid formative assessment tool for DMBRs on other books.

Third, studies on actual applications of the proposed rubric as a formative assessment tool in classrooms need to be conducted. Observing how teachers use the rubric in a real-world setting will reveal whether the rubric will in truth be a convenient tool for them to use when teaching DMBRs as well as assessing its ability as a way to provide useful feedback to students.

Finally, a student-friendly version of the rubric should be developed and validated once the teacher's version has been shown to provide the desired level of inter-rater reliability and validity. In addition to the teachers' feedback on the student-version rubric, using the rubric for peer feedback and asking students' opinions regarding the rubric will be especially helpful when developing the student-friendly version.

Classroom Application of the Rubric

The final version of the rubric (rubric #8) can be used by upper-elementary-grades teachers under the below-listed conditions.

What teachers should know. In order to use the rubric to teach and provide constructive feedback on students DMBRs, teachers should understand the four theoretical assumptions on multimodality (Jewitt, 2014). The first assumption emphasizes the recognition of modes beyond language as part of a multimodal ensemble. This implies that both linguistic and non-linguistic modes are active conveyers of meaning. As reported in Shanahan's series of studies (2012, 2013a, 2013b), teachers might tend to emphasize the role of linguistic modes in the DMBRs. Although oral

language plays a role of the primary mode in the iMovie-based DMBRs, it is coherently connected to the other modes. Also, unlike other writing tasks, DMBRs on iMovie or other equivalent tools do not privilege the written language. For this reason, teachers should consciously try not to privilege linguistic modes. Instead, they should consider the affordances of the digital tool and consider both linguistic and non-linguistic modes as parts of a multimodal ensemble.

The second assumption stresses modal resources of each mode that were used to achieve a unique communicative purpose. C1-technical aspects of visual and C2-technical aspects of audio-voice included in the rubric are reflecting this assumption. For in-class formative assessment of DMBRs, teachers should know about modal resources of each mode that are available in digital tools such as iMovie or equivalent tools. They also have to model how to use modal resources such as colors of visual mode (Pantaleo, 2012) and music and sound effects of audio mode (Phillips & Smith, 2012) purposefully.

The third assumption indicates that DMC composers select and configure modes with the consideration of intermodal relationships. In the rubric, intermodal relationships were considered in the criteria five through eight. Previous studies reported that upper-elementary grade students could make connections between two different modes such as between a visual and linguistic mode (Ranker, 2012) or between a sound and linguistic mode (Dalton et al., 2015). However, their selections of modal affordances and connections between modes were mostly based on personal preferences, rather than based on the social norms and conventions (Mills & Exley, 2014; Shanahan, 2012),

The final assumption emphasizes the social influences on reading/interpretation of multimodal texts. The social aspects around DMBRs of the current study include the classroom contexts, social norms and conventions on DMBRs such as knowledge of genres, text types, and use of each mode. Therefore, teachers need to explain composing DMBRs as a social activity that is influenced by the norms of conventions on DMBRS that were taught and discussed in the classroom and demonstrate how to apply social conventions to compose and interpret DMBRs.

Task format. To use the rubric, teachers should lead students to create DMBRs as described below. First, the teachers should teach students to use iMovie or other programs that have similar affordances. As mentioned at the results of research phase 2: rubric creation, iMovie requires users to include visuals. In addition to the visual mode, users can add other modes such as oral and written language and audio mode. Recording voice is easy, and the length of the oral language content is unlimited. However, the program limits spaces for written language, so usually oral language is the primary conveyer of meaning. Second, the DMBRs should be organized into six components: an introduction, a brief summary, one thing I liked, one thing I did not like, a recommendation, and conclusion. Except for the introduction and conclusion, students' DMBRs do not have to follow the suggested order of components if the components are connected coherently and logically. The length of the DMBRs is not strictly limited, but most students DMBRs did not exceed three minutes.

How teacher should use it. As pointed out throughout the current study, the rubric was developed to be used only for a formative assessment in classrooms. It cannot

be used to assess students' learning on DMBRs summatively. Since the rubric includes 11 criteria that are longer than usual writing rubrics with five or six traits (e.g., Spandel, 2009), focusing on all criteria in one lesson is impossible. I recommend teachers concentrate on a few criteria or one category in the rubric for instructional purposes. For example, if a teacher wants to teach the relationships between different modes and provide students with feedback on the category, he or she can only cover the criteria 5 through 7. Teachers can obtain the total score of students' DMBRs by using the rubric, but the scores should not be used to give any grades on students' performances. Before using the rubric, teachers should read this dissertation and the final version of the rubric (Appendix P) and the scoring protocol (Appendix D).

Conclusion

This study adds to the line of research in several ways. First, the current research attempted to systematically index the criteria that are related to the assessment of DMBRs. Those criteria were obtained as the results of a systematic literature review, which is considered as a rigorous literature review (Petticrew & Roberts, 2006). The field of multimodal assessment has researched it without comprehensive guidelines on what constitutes quality or proficient DMC. The identified domains and criteria through the systematic literature reviews can play a role as the roadmap of DMC assessment in general. In other words, the domains and criteria can be used as the base content to develop rubrics for different DMC tasks such as presentation slides and glogs.

Second, the current study provides step-by-step guides on the rubric development procedures. Although there was a study on developing a rubric to assess DMC (Burnett et al., 2014), this study mainly explained their final product, rather than the developing procedures. The phases and steps that presented in the current study can be replicated by other researchers to develop rubrics for different tasks and grades.

Finally, the current study's attempts to examine the inter-rater reliability and construct validity of the formative assessment rubric can have positive influences on other future studies on DMC assessment. Traditionally, educators rarely considered the inter-rater reliability and construct validity of rubrics in classroom assessments. However, reliability and validity of rubrics for a classroom use are also important since the rubrics are sometimes used for summative assessment. Even for formative assessment, if there are theory-driven and well-structured rubrics for teachers, the quality of classroom formative assessment can be improved.

To this end, this dissertation study informs other researchers and teachers what can be taught and assessed from DMCs. The rubric development procedures that I presented in the current study can be used to develop rubrics for different MDC tasks. I hope the arguments that I made on the importance of the inter-rater reliability and construct validity of classroom formative assessment rubric on DMBR have positive influences on other research studies on and classroom implementations of DMC assessment.

REFERENCES

- Adsanatham, C. (2012). Integrating assessment and instruction: Using student-generated grading criteria to evaluate multimodal digital projects. *Computers and Composition, 29*(2), 152-174.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Afflerbach, P. (2012). *Understanding and using reading assessment, K-12* (2nd ed.). Newark, DE: International Reading Association.
- Anderson, D., Atkins, A., Ball, C., Millar, K. H., Selfe, C., & Selfe, R. (2006). Integrating multimodality into composition curricular: Survey. *Composition Studies, 34*(2).
- Andrade, H. G., & Boulay, B. A. (2003). Role of rubric-referenced self-assessment in learning to write. *The Journal of Educational Research, 97*(1), 21-30.
- Arter, J., & Chappuis, J. (2006). *Creating & recognising quality rubrics*. Portland, OR: Educational Testing Services.
- Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press.
- Bates, N. (2012). Weaving a virtual story-creating book trailers 101. *Knowledge Quest, 40*(3), 72.

- Bear, J., & McKool, S. S. (2009). Assessing creativity using the consensual assessment technique. In C. S. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65-77). Hershey, PA: IGI Global.
- Bearne, E. (2009). Assessing multimodal texts. In A. Burke, & R. F. Hammett (Eds.), *Assessing new literacies: perspectives from the classroom* (pp. 15-54). New York, NY: Peter Lang.
- Bearne, E., & Wolstencroft, H. (2007). *Visual approaches to teaching writing: multimodal literacy 5-11*. Sage.
- Bezemer, J., & Kress, G. (2008). Writing in multimodal texts: A social semiotic account of designs for learning. *Written Communication, 25*(2), 166-195.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31.
- Borton, S. C., & Hout, B. (2007). Responding and assessing. In C. L. Selfe (Ed.), *Multimodal composition: Resources for teachers* (pp. 99-111). Cresskill, NJ: Hampton Press.
- Brown, S. (2013). A blended approach to reading and writing graphic stories. *The Reading Teacher, 67*(3), 208-219.
- Burke, A., & Hammett, R. F. (2009). *Assessing new literacies: Perspectives from the classroom*. New York City: Peter Lang.

- Burnett, R. E., Frazee, A., Hanggi, K., & Madden, A. (2014). A Programmatic ecology of assessment: Using a common rubric to evaluate multimodal processes and artifacts. *Computers and Composition, 31*, 53-66.
- Butler, D., Leahy, M., & Mc Cormack, C. (2010). Redefining book reviews for the digital age. *Contemporary Issues in Technology and Teacher Education, 10*(1), 80-99.
- Callow, J. (2003). Talking about visual texts with students. *Reading Online, 6*(8).
Retrieved from: <http://www.readingonline.org/articles/callow/>
- Camara, W. J. (July 2003). Scoring the essay on the SAT writing section. *Research Summary, RS-10*. New York: The College Board.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychosocial Measurement, 20*(1), 37-46.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Creswell, J., & Plano Clark, V. (2010). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Crusan, D. (2015). Dance, ten; looks, three: Why rubrics matter. *Assessing Writing, 26*, 1-4.
- Dalton, B. (2012). Multimodal composition and the common core state standards. *The Reading Teacher, 66*(4), 333-339.
- Dalton, B., & Grisham, D. L. (2013). Love that book. *The Reading Teacher, 67*(3), 220-225.

- Dalton, B., Robinson, K. H., Lovvorn, J. F., Smith, B. E., Alvey, T., Mo, E., ... & Proctor, C. P. (2015). Fifth-grade students' digital retellings and the Common Core: Modal use and design intentionality. *Elementary School Journal, 115*(4), 548-569.
- Darling-Hammond, L. (2014). Introduction: The rationale and context for performance assessment. In L. Darling-Hammond & F. Adamson (Eds.) *Beyond the bubble test: How performance assessments support 21st century learning*. (pp. 1-14). San Francisco, CA: Jossey-Bass.
- Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a web-based environment. *Assessing Writing, 14*(1), 38-61.
- DeVoss, D. N., Eidman-Aadahl, E., & Hicks, T. (2010). *Because digital writing matters: Improving student writing in online and multimedia environments*. San Francisco, CA: Jossey-Bass.
- Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20*(4), 391-409.
- Eidman-Aadahl, E., Blair, K., DeVoss, D. N., Hochman, W., Jimerson, L., Jurich, C., ... Wood, J. (2013). Developing domains for multimodal writing assessment: The language of evaluation, the language of instruction. In H. A. McKee & D. N. DeVoss (Eds.) *Digital writing assessment and evaluation*. Logan, UT: Computers and Composition Digital Press/Utah State University Press. Retrieved from http://ccdigitalpress.org/dwae/07_nwp.html

- Ehret, C., Hollett, T., & Jocius, R. (2016). The matter of new media making: An intra-action analysis of adolescents making a digital book trailer. *Journal of Literacy Research, 48*(3), 346-377.
- Field, A. (2009). *Discovering statistics using SPSS*. Thousand Oaks, CA: Sage Publications.
- Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 237-270). Washington, DC: American Council on Education.
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. (2nd ed.). New York: John Wiley.
- Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Gearhart, M., Herman, J. L., Novak, J. R., & Wolf, S. A. (1995). Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric. *Assessing Writing, 2*, 207–242.
- Gorsuch, R. L. (1983) *Factor analysis*. 2nd ed. Hillsdale, NJ: Laurence Erlbaum Associates.
- Graham, S., Bollinger, A., Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., & Olinghouse, N. (2012). Teaching elementary school students to be effective writers: A practice guide (NCEE 2012-4058). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

- Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. US Center for Education Compensation Reform (CECR).
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1), 23-34.
- Hicks, T. (Ed.). (2015). *Assessing students' digital writing: Protocols for looking closely*. Berkeley, CA: National Writing Project.
- Howell, E., Reinking, D., & Kaminski, R. (2015). Writing as creative design: Constructing multimodal arguments in a multiliteracies framework. *Journal of Literacy and Technology*, 16(1), 2-23.
- Hung, H., Chiu, Y. J., & Yeh, H. (2013). Multimodal assessment of and for learning: A theory-driven design rubric. *British Journal of Educational Technology*, 44(3), 400-409.
- Husbye, N. E., & Rust, J. (2014). Considering design: The challenges of assessing multimodal texts. In R. E. Ferdig & K. E. Pytash (Eds.), *Exploring multimodal composition and digital writing* (pp. 135-149). Hershey, PA: IGI Global.
- International Society for Technology in Education. (2017). The ISTE standards for educators. Retrieved from
- Jewitt, C. (Ed.). (2014). *The Routledge handbook of multimodal analysis* (2nd ed.). London, UK: Routledge.
- Jewitt, C., & Kress, G. (2003). *Multimodal literacy*. New York: Peter Lang.

- Jewitt, C., & Price, S. (2012). Multimodal approaches to video analysis of digital learning environments. *Proceedings of BCS HCI 2012 Workshops Video Analysis Techniques for Human-Computer Interaction*.
- Johnson, R. L., Penny, J. A., & Gordon, B. (2000). The relationship between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13*(2), 121-138.
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: Guilford Press.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review, 2*(2), 130-144.
- Kalantzis, M., Cope, B., & Cloonan, A. (2010). A multiliteracies perspective on the new literacies. In E. A. Baker (Ed.), *New literacies: Multiple perspectives on research and practice* (pp. 61–87). New York: Guilford Press.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.
- Kane, M. T. (2006). *Validation*. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Greenwood Publishing.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi:10.1111/jedm.12000
- Karchmer-Klein, R. (2007). Audience awareness and Internet publishing: A qualitative analysis of factors influencing how fourth graders write electronic text. *Action in Teacher Education, 29*(2), 39-50.

- Katz, S. M., & Odell, L. (2012). Making the implicit explicit in assessing multimodal composition: Continuing the conversation. *Technical Communication Quarterly*, 21(1), 1-5.
- Kline, P. (1979). *Psychometrics and psychology*. London: Academic Press.
- Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12(1), 26-43.
- Kress, G. (1993). Against arbitrariness: The social production of the sign as a foundational issue in critical discourse analysis. *Discourse and Society*, 4(2), 169-191.
- Kress, G. (2003). *Literacy in the new media age*. New York: Routledge.
- Kress, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. New York: Routledge.
- Kress, G. (2014). What is mode? In C. Jewitt (Ed.) *The Routledge handbook of multimodal analysis* (2nd ed.) (pp. 60-75). London, UK: Routledge.
- Kress, G. R., & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. London, UK: Edward Arnold.
- Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*. 33, 159-174.
- Levy, M., & Kimber, K. (2009). Developing an approach for comparing students' multimodal text creations: A case study. *Australasian Journal of Educational Technology*, 25, 489-508.

- Martinec, R., & Salway, A. (2005). A system for image-text relations in new (and old) media. *Visual Communication*, 4(3), 337-371.
- McGrail, E., & Behizadeh, N. (2017). K-12 multimodal assessment and interactive audiences: An exploratory analysis of existing frameworks. *Assessing Writing*, 31, 24-38.
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276-282.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education/Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
- Miller, S. M., & McVee, M. B. (Eds.). (2012). *Multimodal composing in classrooms: Learning and teaching for the digital world*. New York: Routledge.
- Mills, K. (2011). 'I'm making it different to the book': Transmediation in young children's multimodal and digital texts. *Australasian Journal of Early Childhood*, 36(3), 56-65.
- Mills, K., & Exley, B. (2014). Time, space, and text in the elementary school digital writing classroom. *Written Communication*, 31(4), 434-469.
- Morain, M., & Swarts, J. (2012). YouTutorial: A framework for assessing instructional online video. *Technical Communication Quarterly*, 21(1), 6-24.

- Morey, L. C. (2003). Measuring personality and psychopathology. In J. A. Schinka & W. F. Velicer (Eds.) *Handbook of psychology (vol. 2.): Research methods in psychology* (pp. 377-406). Hocoken, NJ: John Wiley & Sons, Inc.
- Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation*, 7(3), Retrieved March 20, 2016 from <http://PAREonline.net/getvn.asp?v=7&n=3>.
- Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10), Retrieved January 31, 2016 from <http://PAREonline.net/getvn.asp?v=7&n=10>.
- Mozaffari, H. (2013). An analytical rubric for assessing creativity in creative writing. *Theory and Practice in Language Studies*, 3(12), 2214-2219.
- Myers, M. (1980). *A procedure for writing assessment and holistic scoring*. Urbana, IL: National Council of Teachers of English.
- National Council of Teachers of English. (February 28, 2016). *Position statement on professional knowledge for the teaching of writing*. Retrieved from <http://www2.ncte.org/statement/teaching-writing/>
- National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards for English language arts*. Washington, DC: National Governors Association Center for Best Practices & Council of Chief State School Officers.

- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press. <https://doi.org/10.17226.10019>.
- Ostenson, J. W. (2013). Connecting assessment and instruction to help students become more critical producers of multimedia. *Journal of Media Literacy Education*, 4(2), 167-178.
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129-144.
- Pantaleo, S. (2012). Meaning-making with colour in multimodal texts: an 11-year-old student's purposeful 'doing.' *Literacy*, 46, 147-155.
- Petticrew, M. & Roberts, H. (2006). *Systematic reviews in social sciences*. Malden, MA: Blackwell Publishing.
- Phillips, N., & Smith, B. (2012). Multimodality and aurality: sound spaces in student digital book trailers. In P. J. Dunston, S. K. Fullerton, C. C. Bates, K. Headley, & P. M. Stecker (Eds.), *61st Yearbook of the Literacy Research Association*. Oak Creek, WI: Literacy Research Association.
- Pufpaff, L., Clarke, L., & Jones, R. E. (2015). The effects of rater training on inter-rater agreement. *Mid-Western Educational Researcher*, 27(2), 117-141.
- Ranker, J. (2012). Young students' uses of visual composing resources in literacy classroom contexts: a cross-case analysis. *Visual Communication*, 11(4), 461-483.

- Reinking, D., & Watkins, J. (2000). A formative experiment investigating the use of multimedia book reviews to increase elementary students' independent reading. *Reading Research Quarterly, 35*(3), 382-419.
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*, 18-39.
- Selfe, R. J., & Selfe, C. L. (2008). "Convince me!" Valuing Multimodal Literacies and Composing Public Service Announcements. *Theory into Practice, 47*(2), 83-92.
- Shanahan, L. (2012). Use of sound with digital text: Moving beyond sound as an add-on or decoration. *Contemporary Issues in Technology and Teacher Education, 12*(3). Retrieved from <http://www.citejournal.org/vol12/iss3/languagearts/article1.cfm>
- Shanahan, L. (2013a). Multimodal representations: a fifth-grade teacher influences students' design and production, *Pedagogies: An International Journal, 8*(2), 85-102.
- Shanahan, L. (2013b). Composing "kid-friendly" multimodal text: When conversations, instruction, and signs come together. *Written Communication, 30*(1), 194-227.
- Shweta, Bajpai, R. C., & Chaturvedi, H. K. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology, 41*(3), 20-27.
- Silseth, K., & Gilje, Ø. (2017). Multimodal composition and assessment: A sociocultural perspective. *Assessment in Education: Principles, Policy & Practice, 1-17*.

- Smith, B. (2014). Beyond words: A review of research on adolescents and multimodal composition. In R. E. Ferdig & K. E. Pytash (Eds.), *Exploring multimodal composition and digital writing* (pp. 1–19). Hershey, PA: IGI Global.
- Sorapure, M. (2006). Between modes: Assessing student new media compositions. *Kairos: A Journal of Rhetoric, Technology, and Pedagogy*, 10(2). Retrieved from <http://kairos.technorhetoric.net/10.2/coverweb/sorapure/>
- Spandel, V. (2009). *Creating writers: 6 Traits, process, workshop, and literature* (5th ed.). Boston, MA: Pearson.
- Stecher, B. (2014). Looking back: Performance assessment in an era of standards-based educational accountability. In L. Darling-Hammond & F. Adamson (Eds.) *Beyond the bubble test: How performance assessments support 21st century learning*. (pp. 17-52). San Francisco, CA: Jossey-Bass.
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4), 1-19.
- Stevens, D. D., & Levi, A. J. (2013). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning* (2nd ed.). Sterling, VA: Stylus Publishing, LLC.
- Stoddart, T., Abrams, R., Gasper, E., & Canaday, D. (2000). Concept maps as assessment in science inquiry learning-a report of methodology. *International Journal of Science Education*, 22(12), 1221-1246.

- Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Procedures and techniques for developing grounded theory*. Thousand Oaks, CA: Sage Publication.
- Taras, M. (2005). Assessment—summative and formative—some theoretical reflections. *British Journal of Educational Studies*, 53(4), 466-478.
- Towndrow, P. A., Nelson, M. E., & Yusuf, W. F. B. M. (2013). Squaring literacy assessment with multimodal design: An analytic case for semiotic awareness. *Journal of Literacy Research*, 45, 327–355.
- Unsworth, L. (2006). Towards a metalanguage for multiliteracies education: Describing the meaning-making resources of language-image interaction. *English Teaching: Practice and Critique*, 5(1), 55-76.
- U.S. Department of Education, Office of Educational Technology. (2017). *Reimagining the role of technology in education: 2017 national education technology plan update*, Washington, D.C.
- Vassilikopoulou, M., Retalis, S., Nezi, M., & Boloudakis, M. (2011). Pilot use of digital educational comics in language teaching. *Educational Media International*, 48, 115–126.
- Wierszewski, E. (2013). “*Something old, something new*”: *Evaluative criteria in teacher responses to student multimodal texts*. In H. A. McKee & D. N. DeVoss (Eds.) *Digital writing assessment and evaluation*. Logan, UT: Computers and Composition Digital Press/Utah State University Press. Retrieved from http://ccdigitalpress.org/dwae/05_wierszewski.html

- Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37, 3-14.
- Wise, J. B. (2016). *The effects of explicit instruction utilizing traditionally-printed and digital mentor texts to teach evidence-based written arguments to fifth grade students*. Unpublished doctoral dissertation. University of Delaware, DE.
- Wolcott, W., & Legg, S. M. (1998). *An Overview of Writing Assessment: Theory, Research, and Practice*. Urbana, IL: National Council of Teachers of English.
- Wren, D. G. (March 4, 2009). Performance assessment: A Key component of a balanced assessment system. *Research Brief: Report from the Department of Research, Evaluation, and Assessment*, 2. 1-12.
- Yancey, K. B. (2004). Looking for sources of coherence in a fragmented world: Notes toward a new assessment design. *Computers and Composition*, 21(1), 89-102.
- Yu, E. (2014). Let developmental students shine: Digital writing. *Research & Teaching in Developmental Education*, 30(2), 99-110.
- Zammit, K. (2009). Popular culture in the classroom: Interpreting and creating multimodal texts. In M. O'Donnell, R. Whittaker, & A. McCabe (Eds.), *Advances in Language and Education* (pp. 60-76). London: Continuum Publishers.
- Zammit, K. (2014). Creating multimodal text in the classroom: Shifting teaching practices, influencing student outcomes. In R. E. Ferdig & K. E. Pytash (Eds.), *Exploring multimodal composition and digital writing* (pp. 20-35). Hershey, PA: IGI Global.

Appendix A

PRESENTATION SLIDES USED FOR THE OFFLINE SCORING TRAINING

Scoring Training

Sohee Park
University of Delaware
December 19, 2016

Introduction of The Study

- Purposes
 - To identify key domains and criteria representing the construct of digital multimodal composition (DMC)
 - To develop a rubric for reliable and valid interpretation of scores for upper-elementary students' digital multimodal book reviews
- What is digital multimodal composition (DMC)?
 - Composing practices related to digital texts that incorporate multiple modes of representation such as **written and oral language and visual, audio, tactile, gestural, and spatial** representations.

Orientation to the assessment task

- The data were collected from a 4th grade classroom in 2013 as part of a previous research study.
- Task: Create a multimodal book review on a book *Frindle* using the iMovie app on the iPad
- No time limit
- Six components in the graphic organizer:
Introduction / Brief summary / 1 thing I liked / 1 think I did not like / Recommendation / Conclusion

Domains and Criteria of DMC assessment

- Domains for DMC assessment (Multimodal Assessment Project Group, 2013)

```
graph TD; C[Context] --> D[Domains for DMC assessment]; A[Audience] --> D; P[Process, management, and technique] --> D; M[Medium] --> D; S[Status of artifact] --> D;
```

Domains and Criteria of DMC assessment

- **Artifact** "is the finished product. Audiences expect artifacts to convey a coherent message with a clear focus created through an appropriate use of structure, medium, and technique."

Domains and Criteria of DMC assessment

- **Context** "is the world around the artifact, around the creation of the artifact, and how the artifact enters, circulates, and fits into the world. Authors attend to the context of a multimodal artifact when they make design decisions related to genre or to an artifact's intended uses. Given their purposes, authors consider the affordances, constraints and opportunities, given purpose, audience, composing environment, and delivery mode."

Domains and Criteria of DMC assessment

- **Substance** "refers to the content and overall quality and significance of the ideas presented.... Considering the substance of a piece encourages authors to think about elements such as quality of ideas, quality of performance, credibility, accuracy, and significance."

Clarification of the scoring criteria

Please refer to the scoring rubric

Criterion 1: Technical Aspects of Visual Mode

- Resolution
- Color
- Camera shots: long shot, medium shot, close-up
- Camera angles: the bird's eye view, high angle, eye level, low angle, oblique angle
- Lighting
- Ken Burns: Zoom and pan in iMovie / Default setting = random
- Transitions: Default setting = cross-dissolve

Criterion 2: Technical Aspects of Audio-Voice

- Volume
- Speed: overall flow of voice (check pause or stuttering with this aspect)
- Tone (monotonous vs. excited or lively)
- Pronunciation (check mumbling with this aspect)

Criterion 3 & 4: Conventions of Oral & Written language

- Conventions of Standard English for 4th grade students presented in CCSS – Please refer to the scoring protocol

Criterion 5: Relationship Between Visual and Oral Language

- **Concurrence**: equivalence of meaning between visuals and oral language
- **Complementary**: what is represented in visuals and what is represented in oral language may be different but complementary and joint contributors to an overall meaning that is more than the meanings conveyed by the separate modes.
- **Connection**: the quoting or reporting of speech or thoughts + conjunctive relations of time, place, and cause.

Criterion 8: Organization of Multimodal Content

- Six required sections
 - Introduction
 - Brief summary
 - One thing I liked
 - One thing I did not like
 - Recommendation
 - Conclusion
- Relative importance of sections

Criterion 9: Audience Awareness

- Awareness + Engagement
- Example of point 4:
 - Introduction: "Have you ever read the book Frindle? Well, if you haven't, it's a great book. (...) Do you know what that idea was?"
 - One thing I did not like: "... What do you think about that? Well I think that is disrespectful."
 - Conclusion: "Well, I told you my opinions, so you need to tell me yours."

Criterion 9: Audience Awareness

- Example of point 3:
 - Introduction: "Hi, I'm Art and I am doing a book review on Frindle..."
 - Conclusion: "I hope after watching this book review, you will read the book. I hope you like this book. Thank you for watching."
- Example of point 2:
 - Conclusion: "(...) everybody should read this book because it's a really good book."

Criterion 9: Audience Awareness

- Example of point 1: Students who get point 1 never use the word, "you" in the artifact. Instead...
 - Conclusion: "This is Amy signing up! Goodbye!"
OR
 - Recommendation: "I would recommend the book because it will be entertaining for kids and learn more about the 5th grade."

Criterion 10: Quality of Summary

- Setting: a 5th grade English class in the Lincoln Elementary school / or ELA class in an elementary school
- Main characters: Nick & Mrs. Granger
- List of major events
 - 1) Mrs. Granger was a very strict ELA teacher and made the class constantly check the dictionary.
 - 2) Nick made up a new word, Frindle, for pens.
 - 3) Nick spread this word to other classmates.
 - 4) Mrs. Granger was opposed to students who were using the word and kept them in detention.
 - 5) Judy Morgan, a news reporter, heard about the word Frindle and visited Lincoln elementary school to interview Nick and the related people. The news about the Frindle appeared in the first page of the local newspaper.
 - 6) The news was spread to other middle and high school students in the town, a local businessman and in the end to people across the country.

Criterion 10: Quality of Summary

- List of major events
 - 7) Bud Lawrence, a local businessman, sold pens named Frindle and earned lots of money.
 - 8) Bud Lawrence provided loyalty to Nick and Nick's father set up the Frindle trust fund for Nick.
 - 9) The fund got bigger and bigger, so Nick got rich.
 - 10) When Nick became 21-years old, he received a package from Mrs. Granger.
 - 11) The package included a new dictionary including the word Frindle and her letter explaining why she was against his idea 10 years ago...
 - 12) On a Christmas day morning, Mrs. Granger was notified the establishment of a permanent trust fund named after her name for college scholarships with a donation of one million dollars from Nick and received a letter and gift from Nick.

Guided Practice of Scoring

- Sohee's demonstration of scoring Amy's digital multimodal book review
- Guided practice with Aaron's digital multimodal book review
 - Please follow the procedures written on the scoring protocol
- Grading two products following the directions in the scoring protocol

Appendix B
INITIAL SCORING PROTOCOL
Scoring Protocol

Please read the followings before you start scoring.

<General Directions>

- You should be in a quiet place to evaluate some criteria related to oral language and sounds in the book reviews.
- Many of the book reviews do not include any voice and sound for 1~2 seconds. This is not a fault of the students, but the technical problem of the iMovie. So, please ignore the short soundless moments.
- Please do not grade more than two criteria in one day.
- Throughout the scoring procedures, please use the memo spaces on the scoring sheet to let you remember the reasons of your decision on the score. The memos will be used to recall your thoughts on the rubric during the interview.

<Scoring Procedures>

Please follow the procedures described below while you grade students' digital multimodal book reviews.

1. Connect to the VPN and School of Education OET server in order to access Frindle folder.
2. Prepare the printed rubric and the sheet for scores and memos.

<Scoring Criteria 1 and 2>

3. Open Dede's book review and count the number of frames, titles, and sounds (sound effects = SE & theme music = TM) while watching the review.
4. Read the descriptors for criteria 1 and 2 carefully.
5. Re-watch the review and evaluate technical aspects of visual and voice in each frame. If necessary, please refer to the page 3 of this scoring protocol.
6. Open next student's book review and count the number of frames, titles, and sounds (sound effects = SE and theme music = TM) while watching the review.
7. Re-watch the review and evaluate technical aspects of visual and voice in each frame. If necessary, please refer to the page 3 of this scoring protocol.

8. Repeat 6 and 7 until you complete scoring criteria 1 and 2 of all students' book reviews.

<Scoring Criteria 3 and 4>

9. Open Dede's book review and read the descriptors for criteria 3 and 4 carefully.
10. Watch the review and evaluate conventions of oral and written language in each frame. If necessary, please refer to pages 3-5 of this scoring protocol.
11. Open next student's book review and repeat 10 until you complete scoring criteria 3 and 4 of all students' book reviews.

<Scoring Criteria 5~7>

12. Open Dede's book review and read the descriptors for criterion 5 carefully.
13. Watch the review and evaluate the relationship between oral language and visual in each frame. If necessary, please refer to page 5 of this scoring protocol.
14. Open next student's book review and repeat 13 until you complete scoring criterion 5 of all students' book reviews.
15. Repeat 12 through 14 to grade criterion 6 and 7 separately.

<Scoring Criterion 8>

16. Open Dede's book review and read the descriptors for criterion 8 carefully.
17. Watch the review and count the number of required sections included in the review and evaluate if the student weighed sections differently by using more than one image/video in some sections that he/she wants to emphasize.
18. Open next student's book review and repeat 17 until you complete scoring criterion 8 of all students' book reviews.

<Scoring Criterion 9>

19. Open Dede's book review and read the descriptors for criterion 9 carefully.
20. Read page 5 of this scoring protocol for further understanding on different levels of audience awareness
21. Watch the review and evaluate the student's audience awareness and engagement.
22. Open next student's book review and repeat 21 until you complete scoring criterion 9 of all students' book reviews.

<Scoring Criteria 10 and 11>

23. Open Dede's book review and read the descriptors for criteria 10 and 11 carefully.
24. Read page 6 of this scoring protocol for further understanding on major events of Frindle.
25. Watch the review and evaluate the quality of summary and opinion.
26. Open next student's book review and repeat 25 until you complete scoring criteria 10 and 11 of all students' book reviews.

Supplemental Materials for Scoring

1. Criterion 1: Technical Aspects of Visual Mode

- Resolution: the degree of detail visible in a photographic or television image.
- Color
- Camera shots: long shot, medium shot, close-up
(<http://www.mediaknowall.com/camangles.html>)
- Camera angles: the bird's eye view, high angle, eye level, low angle, oblique angle (<http://www.mediaknowall.com/camangles.html>)
- Lighting
- Ken Burns: Zoom and pan in iMovie / Default setting = random
(https://support.apple.com/kb/PH14582?locale=en_US)
- Transitions: Default setting = cross-dissolve
(https://support.apple.com/kb/PH22902?viewlocale=ar_AE&locale=en_US)

2. Criterion 2: Technical Aspects of Audio Mode

- Volume
- Speed: overall flow of voice (check pause or stuttering with this aspect)
- Tone (monotonous vs. excited or lively)
- Pronunciation (check mumbling with this aspect)

3. Criterion 3 and 4: Conventions of Oral and Written Language

<Conventions of Standard English: Grade 4>

CCSS.ELA-LITERACY.L.4.1: Demonstrate command of the conventions of standard English grammar and usage when writing or speaking.

- CCSS.ELA-LITERACY.L.4.1.A: Use relative pronouns (who, whose, whom, which, that) and relative adverbs (where, when, why).
- CCSS.ELA-LITERACY.L.4.1.B: Form and use the progressive (e.g., I was walking; I am walking; I will be walking) verb tenses.
- CCSS.ELA-LITERACY.L.4.1.C: Use modal auxiliaries (e.g., can, may, must) to convey various conditions.
- CCSS.ELA-LITERACY.L.4.1.D: Order adjectives within sentences according to conventional patterns (e.g., a small red bag rather than a red small bag).
- CCSS.ELA-LITERACY.L.4.1.E: Form and use prepositional phrases.
- CCSS.ELA-LITERACY.L.4.1.F: Produce complete sentences, recognizing and correcting inappropriate fragments and run-ons.*
- CCSS.ELA-LITERACY.L.4.1.G: Correctly use frequently confused words (e.g., to, too, two; there, their).*

CCSS.ELA-LITERACY.L.4.2: Demonstrate command of the conventions of standard English capitalization, punctuation, and spelling when writing.

- CCSS.ELA-LITERACY.L.4.2.A: Use correct capitalization.
- CCSS.ELA-LITERACY.L.4.2.B: Use commas and quotation marks to mark direct speech and quotations from a text.
- CCSS.ELA-LITERACY.L.4.2.C: Use a comma before a coordinating conjunction in a compound sentence.
- CCSS.ELA-LITERACY.L.4.2.D: Spell grade-appropriate words correctly, consulting references as needed.

<Conventions of Standard English: Grade 3>

CCSS.ELA-LITERACY.L.3.1: Demonstrate command of the conventions of standard English grammar and usage when writing or speaking.

- CCSS.ELA-LITERACY.L.3.1.A: Explain the function of nouns, pronouns, verbs, adjectives, and adverbs in general and their functions in particular sentences.
- CCSS.ELA-LITERACY.L.3.1.B: Form and use regular and irregular plural nouns.
- CCSS.ELA-LITERACY.L.3.1.C: Use abstract nouns (e.g., childhood).
- CCSS.ELA-LITERACY.L.3.1.D: Form and use regular and irregular verbs.
- CCSS.ELA-LITERACY.L.3.1.E: Form and use the simple (e.g., I walked; I walk; I will walk) verb tenses.
- CCSS.ELA-LITERACY.L.3.1.F: Ensure subject-verb and pronoun-antecedent agreement.*
- CCSS.ELA-LITERACY.L.3.1.G: Form and use comparative and superlative adjectives and adverbs, and choose between them depending on what is to be modified.
- CCSS.ELA-LITERACY.L.3.1.H: Use coordinating and subordinating conjunctions.
- CCSS.ELA-LITERACY.L.3.1.I: Produce simple, compound, and complex sentences.

CCSS.ELA-LITERACY.L.3.2: Demonstrate command of the conventions of standard English capitalization, punctuation, and spelling when writing.

- CCSS.ELA-LITERACY.L.3.2.A: Capitalize appropriate words in titles.
- CCSS.ELA-LITERACY.L.3.2.B: Use commas in addresses.
- CCSS.ELA-LITERACY.L.3.2.C: Use commas and quotation marks in dialogue.
- CCSS.ELA-LITERACY.L.3.2.D: Form and use possessives.

- CCSS.ELA-LITERACY.L.3.2.E: Use conventional spelling for high-frequency and other studied words and for adding suffixes to base words (e.g., sitting, smiled, cries, happiness).
- CCSS.ELA-LITERACY.L.3.2.F: Use spelling patterns and generalizations (e.g., word families, position-based spellings, syllable patterns, ending rules, meaningful word parts) in writing words.
- CCSS.ELA-LITERACY.L.3.2.G: Consult reference materials, including beginning dictionaries, as needed to check and correct spellings.

4. Criterion 5: Relationship between Visual and Oral Language

- Concurrence: equivalence of meaning between visuals and oral language
- Complementary: what is represented in visuals and what is represented in oral language may be different but complementary and joint contributors to an overall meaning that is more than the meanings conveyed by the separate modes.
- Connection: the quoting or reporting of speech or thoughts + conjunctive relations of time, place, and cause.

5. Criterion 9: Audience Awareness

- Example of point 4:
 - Introduction: “Have you ever read the book Frindle? Well, if you haven’t, it’s a great book. (...) Do you know what that idea was?”
 - One think I did not like: “... What do you think about that? Well I think that is disrespectful.”
 - Conclusion: “Well, I told you my opinions, so you need to tell me yours.”
- Example of point 3:
 - Introduction: “Hi, I’m Art and I am doing a book review on Frindle...”
 - Conclusion: “I hope after watching this book review, you will read the book. I hope you like this book. Thank you for watching.”
- Example of point 2:
 - Conclusion: “(...) everybody should read this book because it’s a really good book.”
- Example of point 1: Students who get point 1 never use the word, “you” in the artifact. Instead...
 - Conclusion: “This is Amy signing up! Goodbye!” OR
 - Recommendation: “I would recommend the book because it will be entertaining for kids and learn more about the 5th grade.”

6. Criterion 10: Quality of Summary

- Setting: a 5th grade English class in the Lincoln Elementary school / or ELA class in an elementary school
- Main characters: Nick & Mrs. Granger
- List of major events
 - 1) Mrs. Granger was a very strict ELA teacher and made the class constantly check the dictionary.
 - 2) Nick made up a new word, Frindle, for pen.
 - 3) Nick spread this word to other classmates.
 - 4) Mrs. Granger was opposed to students who were using the word and kept them in detention.
 - 5) Judy Morgan, a news reporter, heard about the word Frindle and visited Lincoln elementary school to interview Nick and the related people. The news about the Frindle appeared in the first page of the local newspaper.
 - 6) The news was spread to other middle and high school students in the town, a local businessman and in the end to people across the country.
 - 7) Bud Lawrence, a local businessman, sold pens named Frindle and earned lots of money.
 - 8) Bud Lawrence provided loyalty to Nick and Nick's father set up the Frindle trust fund for Nick.
 - 9) The fund got bigger and bigger, so Nick got rich.
 - 10) When Nick became 21-years old, he received a package from Mrs. Granger.
 - 11) The package included a new dictionary including the word Frindle and her letter explaining why she was against his idea 10 years ago...
 - 12) On a Christmas day morning, Mrs. Granger was notified the establishment of a permanent trust fund named after her name for college scholarships with a donation of one million dollars from Nick and received a letter and gift from Nick.

Appendix C

LIST OF ONLINE DIGITAL MULTIMODAL BOOK REVIEWS ABOUT A BOOK, FRINDLE

- Online 1: <https://www.youtube.com/watch?v=Q8hCrZRT1K8>
- Online 2: <https://www.youtube.com/watch?v=Im-zITOM3wI>
- Online 3: <https://www.youtube.com/watch?v=DDZtWVIirQA>
- Online 4: http://www.showme.com/sh?h=IFTWIN2&jw_version=7
- Online 5: <https://www.youtube.com/watch?v=Ytx8TQGvk2o>
- Online 6: <https://www.youtube.com/watch?v=gIVT6OaRPGE>
- Online 7: <https://www.youtube.com/watch?v=ucLKsds3ha8>
- Online 8:
<https://www.schooltube.com/video/9df0121c5f0344d7892d/Frindle%20Book%20Review>

Appendix D

REVISED SCORING PROTOCOL

Scoring Protocol

(Revised on January 9th, 2017)

Please read the followings before you start scoring.

<General Directions>

- You should be in a quiet place to evaluate some criteria related to oral language and sound in the book reviews.
- Many of the book reviews do not include any voice and sound for 1~2 seconds. This is not a fault of the students, but a technical problem of the iMovie. Please ignore the short soundless moments.
- You should grade three or fewer criteria in a day.
- Throughout the scoring procedures, please use the memo spaces on the scoring sheet to allow you to remember your scoring decision. The memos will be used to recall your thoughts on the rubric during the interview.
- Whenever you have difficulties in assigning scores, please refer to the scores of the book reviews used during the scoring trainings and the supplemental information for scoring section (pp. 3-7 of this document).

<Scoring Procedures and Suggested Scoring Schedule>

Please follow the procedures described below while grading students' digital multimodal book reviews.

1. Connect to the VPN and School of Education OET server in order to access Frindle folder. If your Internet connection is weak, you may experience buffering while playing the book reviews. In that case, please download students' book reviews on your personal computer. You should use the downloaded videos for the grading purpose only and should delete them from your computer after the grading is done.
2. Prepare the printed rubric and the sheet for scores and memos.

Counting the Number of Frames, Titles, and Sound & Scoring Criteria 1 and 2

3. Open Dede's book review and count the number of frames, titles, and sound (sound effects = SE & theme music = TM) while watching the review.
4. Read the descriptors for criteria 1 and 2 carefully.
5. Re-watch the review and evaluate technical aspects of visual and voice in each frame. If necessary, please refer to the page 3 of this scoring protocol.

6. Open next student's book review and count the number of frames, titles, and sounds (sound effects = SE and theme music = TM) while watching the review.
7. Re-watch the review and evaluate technical aspects of visual and voice in each frame. If necessary, please refer to the page 3 of this scoring protocol.
8. Repeat 6 and 7 until you complete scoring criteria 1 and 2 for all students' book reviews.

Scoring Criteria 3 and 4

9. Open Dede's book review and read the descriptors for criteria 3 and 4 carefully.
10. Watch the review and evaluate "conventions of oral and written language" in each frame. If necessary, please refer to pages 3-5 of this scoring protocol.
11. Open next student's book review and repeat 10 until you complete scoring criteria 3 and 4 for all students' book reviews.

Scoring Criteria 5~7

12. Open Dede's book review and read the descriptors for criterion 5 carefully.
13. Watch the review and evaluate the relationship between oral language and visual in each frame. If necessary, please refer to page 5 of this scoring protocol.
14. Open next student's book review and repeat 13 until you complete scoring criterion 5 for all students' book reviews.
15. Repeat 12 through 14 to grade criterion 6 and 7 separately.

Scoring Criterion 8

16. Open Dede's book review and read the descriptors for criterion 8 carefully.
17. Watch the review and count the number of required sections included in the review and evaluate if the student weighs sections differently by using more than one image/video in some sections that he/she wants to emphasize.
18. Open next student's book review and repeat 17 until you complete scoring criterion 8 for all students' book reviews.

Scoring Criterion 9

19. Open Dede's book review and read the descriptors for criterion 9 carefully.
20. Read pages 5-6 of this scoring protocol for further understanding on different levels of audience awareness
21. Watch the review and evaluate the student's audience awareness and engagement.
22. Open next student's book review and repeat 21 until you complete scoring criterion 9 for all students' book reviews.

Scoring Criteria 10 and 11

23. Open Dede's book review and read the descriptors for criteria 10 and 11 carefully.
24. Read pages 6-7 of this scoring protocol for further understanding on major events of Frindle.
25. Watch the review and evaluate the quality of summary and opinion.

26. Open next student’s book review and repeat 25 until you complete scoring criteria 10 and 11 for all students’ book reviews.

<Suggested Schedule>

Date	Day 1	Day 2	Day 3	Day 4	Day 5
Task	* Count the # of frames, titles, & sound * Score Criteria 1 & 2	* Score Criteria 3 & 4	* Score Criteria 5, 6, 7	* Score Criteria 8 & 9	* Score Criteria 10 & 11
Expected hours	Approx. 3 hours	Approx. 3 hours	Approx. 3 hours	Approx. 3 hours	Approx. 3 hours

**Supplemental Information for Scoring
(Revised on January 9th, 2017)**

1. Criterion 1: Technical Aspects of Visual Mode

- Color
- Camera shots: long shot, medium shot, close-up
(<http://www.mediaknowall.com/camangles.html>)
- Camera angles: the bird’s eye view, high angle, eye level, low angle, oblique angle (<http://www.mediaknowall.com/camangles.html>)
- Lighting
- Transitions: Default setting = cross-dissolve
(https://support.apple.com/kb/PH22902?viewlocale=ar_AE&locale=en_US)
- Resolution: the degree of detail visible in a photographic or television image.
 - If an image is too blurry and you cannot exactly read the words or identify features (e.g., faces or objects) in the image, this is one of the cases that the resolution of an image does not help clearly convey meaning.
- Ken Burns: Zoom and pan in iMovie / Default setting = random
(https://support.apple.com/kb/PH14582?locale=en_US)
 - On each frame, the Ken Burns effect should be applied to present some words (e.g., “dictionary,” “Frindle,” and “recommendations”) or features (e.g., a person and thumbs) conveying important meaning of the image. The words or features should not be cut at an inappropriate point. For example, if a student wants to display a book cover image, the Ken Burns effect should fully include the author’s name and the book title in the book cover image.

2. Criterion 2: Technical Aspects of Audio Mode

- Volume
- Speed: overall flow of voice (check pause or stuttering with this aspect)
- Tone (monotonous vs. excited or lively)
- Pronunciation (check mumbling with this aspect)

3. Criterion 3 and 4: Conventions of Oral and Written Language

- When you find a number of errors in written language, count the same type of errors only once. The followings are examples of the types of errors you may find...
 - No capitalization of the book title (e.g., Frindle)
 - No capitalization of the first letter in each sentence
 - No capitalization of a person's title and name (e.g., andrew clements; mrs. granger; nick)
 - No use of the period at the end of a sentence.
 - No use of the apostrophe
 - A word misspelled multiple times in the same way
 - Missing one comma after a parentheses

<Conventions of Standard English: Grade 4>

CCSS.ELA-LITERACY.L.4.1: Demonstrate command of the conventions of standard English grammar and usage when writing or speaking.

- CCSS.ELA-LITERACY.L.4.1.A: Use relative pronouns (who, whose, whom, which, that) and relative adverbs (where, when, why).
- CCSS.ELA-LITERACY.L.4.1.B: Form and use the progressive (e.g., I was walking; I am walking; I will be walking) verb tenses.
- CCSS.ELA-LITERACY.L.4.1.C: Use modal auxiliaries (e.g., can, may, must) to convey various conditions.
- CCSS.ELA-LITERACY.L.4.1.D: Order adjectives within sentences according to conventional patterns (e.g., a small red bag rather than a red small bag).
- CCSS.ELA-LITERACY.L.4.1.E: Form and use prepositional phrases.
- CCSS.ELA-LITERACY.L.4.1.F: Produce complete sentences, recognizing and correcting inappropriate fragments and run-ons.*
- CCSS.ELA-LITERACY.L.4.1.G: Correctly use frequently confused words (e.g., to, too, two; there, their).*

CCSS.ELA-LITERACY.L.4.2: Demonstrate command of the conventions of standard English capitalization, punctuation, and spelling when writing.

- CCSS.ELA-LITERACY.L.4.2.A: Use correct capitalization.

- CCSS.ELA-LITERACY.L.4.2.B: Use commas and quotation marks to mark direct speech and quotations from a text.
- CCSS.ELA-LITERACY.L.4.2.C: Use a comma before a coordinating conjunction in a compound sentence.
- CCSS.ELA-LITERACY.L.4.2.D: Spell grade-appropriate words correctly, consulting references as needed.

<Conventions of Standard English: Grade 3>

CCSS.ELA-LITERACY.L.3.1: Demonstrate command of the conventions of standard English grammar and usage when writing or speaking.

- CCSS.ELA-LITERACY.L.3.1.A: Explain the function of nouns, pronouns, verbs, adjectives, and adverbs in general and their functions in particular sentences.
- CCSS.ELA-LITERACY.L.3.1.B: Form and use regular and irregular plural nouns.
- CCSS.ELA-LITERACY.L.3.1.C: Use abstract nouns (e.g., childhood).
- CCSS.ELA-LITERACY.L.3.1.D: Form and use regular and irregular verbs.
- CCSS.ELA-LITERACY.L.3.1.E: Form and use the simple (e.g., I walked; I walk; I will walk) verb tenses.
- CCSS.ELA-LITERACY.L.3.1.F: Ensure subject-verb and pronoun-antecedent agreement.*
- CCSS.ELA-LITERACY.L.3.1.G: Form and use comparative and superlative adjectives and adverbs, and choose between them depending on what is to be modified.
- CCSS.ELA-LITERACY.L.3.1.H: Use coordinating and subordinating conjunctions.
- CCSS.ELA-LITERACY.L.3.1.I: Produce simple, compound, and complex sentences.

CCSS.ELA-LITERACY.L.3.2: Demonstrate command of the conventions of standard English capitalization, punctuation, and spelling when writing.

- CCSS.ELA-LITERACY.L.3.2.A: Capitalize appropriate words in titles.
- CCSS.ELA-LITERACY.L.3.2.B: Use commas in addresses.
- CCSS.ELA-LITERACY.L.3.2.C: Use commas and quotation marks in dialogue.
- CCSS.ELA-LITERACY.L.3.2.D: Form and use possessives.
- CCSS.ELA-LITERACY.L.3.2.E: Use conventional spelling for high-frequency and other studied words and for adding suffixes to base words (e.g., sitting, smiled, cries, happiness).
- CCSS.ELA-LITERACY.L.3.2.F: Use spelling patterns and generalizations (e.g.,

word families, position-based spellings, syllable patterns, ending rules, meaningful word parts) in writing words.

- CCSS.ELA-LITERACY.L.3.2.G: Consult reference materials, including beginning dictionaries, as needed to check and correct spellings.

7. Criterion 5: Relationship between Visual and Oral Language

- Concurrence: equivalence of meaning between visuals and oral language
- Complementary: what is represented in visuals and what is represented in oral language may be different but complementary and joint contributors to an overall meaning that is more than the meanings conveyed by the separate modes.
- Connection: the quoting or reporting of speech or thoughts + conjunctive relations of time, place, and cause.

8. Criterion 6: Relationship between oral language and written language

- Among the descriptors of level 1-Needs Improvement, the statement, “written language in almost all frames is not different from the script of oral language,” refers to the case that a student copies and pastes what s/he writes on the script to the written language (titles).

9. Criterion 9: Audience Awareness

- Example of point 4:
 - Introduction: “Have you ever read the book Frindle? Well, if you haven’t, it’s a great book. (...) Do you know what that idea was?”
 - One thing I did not like: “... What do you think about that? Well I think that is disrespectful.”
 - Conclusion: “Well, I told you my opinions, so you need to tell me yours.”
- Example of point 3:
 - Introduction: “Hi, I’m Art and I am doing a book review on Frindle...”
 - Conclusion: “I hope after watching this book review, you will read the book. I hope you like this book. Thank you for watching.”
- Example of point 2:
 - Conclusion: “(...) everybody should read this book because it’s a really good book.”
- Example of point 1: Students who get point 1 never use the word, “you” in the artifact.
 - Conclusion: “This is Amy signing up! Goodbye!” OR
 - Recommendation: “I would recommend the book because it will be entertaining for kids and learn more about the 5th grade.”

10. Criterion 10: Quality of Summary

- Setting: a 5th grade English class in the Lincoln Elementary school / or ELA class in an elementary school
- Main characters: Nick & Mrs. Granger
- List of major events
 - 1) Mrs. Granger was a very strict ELA teacher and made the class constantly check the dictionary.
 - 2) Nick made up a new word, Frindle, instead of using the word pen.
 - 3) Nick spread this word to other classmates.
 - 4) Mrs. Granger was opposed to students who were using the new word and kept them in detention.
 - 5) Judy Morgan, a news reporter, heard about the word Frindle and visited Lincoln elementary school to interview Nick and the related people. The news about the Frindle appeared in the first page of the local newspaper.
 - 6) The news was spread to other middle and high school students in the town, a local businessman and in the end to people across the country.
 - 7) Bud Lawrence, a local businessman, sold pens named Frindle and earned lots of money.
 - 8) Bud Lawrence provided loyalty to Nick and Nick's father set up the Frindle trust fund for Nick.
 - 9) The fund got bigger and bigger, so Nick got rich.
 - 10) When Nick became 21-years old, he received a package from Mrs. Granger.
 - 11) The package included a new dictionary including the word Frindle and her letter explaining why she was against his idea 10 years ago...
 - 12) On a Christmas day morning, Mrs. Granger was notified the establishment of a permanent trust fund named after her name for college scholarships with a donation of one million dollars from Nick and received a letter and gift from Nick.
- Although the “list of major events” does not include the details of the story you found from a student’s book review, count those details as long as you think that they are correct and closely related to one of the listed major events.

11. Criterion 11: Quality of Opinion

- In order to grade this criterion, you need to count the number of details or reasons supporting the opinions and not the number of opinions.
- Example of point 4
 - Opinion: I would recommend this book to kids and adults
 - Supporting detail/reason 1: Because it is really funny book

- Supporting detail/reason 2: and it has a really good ending
 - Opinion: I really think you and your friends should read this book
 - Supporting detail/reason 3: because when my teacher read it to my class, I laughed and really enjoyed it very much.
- Example of point 3 (Aaron, Amy, Antonio, Art): The author presented three four opinions but provided only two supporting reasons.
 - Opinion 1: One thing I liked, it said in other grades he had a lot of other crazy ideas ...
 - Opinion 2: There was nothing I didn't like about the story.
 - Supporting detail/reason 1: It was a really good story If I were writing the story, I wouldn't change it.
 - Opinion 3: If I were to rate this book, I would give two thumbs up
 - Supporting detail/reason 2: because it is very good book
 - Opinion 4: I think everyone should try it and see if they like it.
- Example of point 2: The author provided only one detail or reason supporting the opinion(s).
 - Opinion: I recommend this book to all kids
 - Supporting detail/reason 1: so they don't give up in their dreams
- Example of point 1: The author presented one or some opinion(s) about the book, but s/he did not provide any details or reasons supporting the opinion(s).
- Example of point 0 (Caeden, Online 1, Online 2): No opinion about the book was presented.

Appendix E

A SAMPLE OF THE SCORING SHEET INCLUDING THE SPACE FOR SCORING LOG

Student Name	# of frames	# of titles	# of sound	1. Artifact: Technical aspect of non-linguistic modes		2. Artifact: Conventions of linguistic modes		3. Artifact: Coherence of multimodal product			4. Artifact: Organization	5. Content: Audience awareness	6. Substance: Quality of content		Total
				1) Technical aspects of visual mode	2) Technical aspects of audio: voice	3) Conventions of oral language (narration)	4) Conventions of written language (caption)	5) Relationship between visual and oral language	6) Relationship between oral and written language	7) Relationship between oral language and sound	8) Organization of multimodal product	9) Audience awareness	10) Quality of summary	11) Quality of opinion	
Dede															
Dede Memo															
Deirdre															
Deirdre Memo															
Doug															
Doug Memo															
Erin															
Erin Memo															
Ethan															
Ethan Memo															

Appendix F

A SEMI-STRUCTURED INTERVIEW QUESTIONNAIRE

Utility of the Rubric

1. How much time did you spend to grade a criterion of all students' digital multimodal book reviews on the average?
2. While grading digital multimodal book reviews, what were the advantages of using the rubric?
3. While grading digital multimodal book reviews, what were the challenges of using the rubric?
4. Which content of the rubric was clear and easy to understand to you? Why did you think that way?
5. Which content of the rubric was unclear or difficult to understand to you? Why did you think that way?
6. Do you think that the supplemental information for scoring provided with the scoring protocol was helpful?
7. If you evaluate the utility of the rubric from 1 to 4, 1 is the most difficult to use and 4 is very easy to use, which score will you give to the rubric? Why?
8. In the future, if you assign digital multimodal book reviews to your students, would you use the developed rubric to assess your students' performances?
 - a. If you said yes, why would you use the rubric?
 - b. If you said no, why would you not use the rubric?

Appropriateness of the Rubric

9. Do you think that the developed rubric helps you assess the different qualities of students' digital multimodal book reviews adequately?
10. Do you think that the 11 criteria included in the rubric are appropriate to evaluate the digital multimodal book review?
 - a. If it isn't, which criteria or characteristics that should be deleted or changed from the rubric?
 - b. If it isn't, which criteria or characteristics that should be added to the rubric?
11. Do you think that the descriptors of the rubric criteria represent developmental characteristics of upper-elementary grade (grades 3-5) students properly?
 - a. If it isn't, which developmental characteristics should be considered? How should the descriptors be revised?

Appendix G

CODEBOOK (35 CODES AND 11 CATEGORIES)

Category & Code		Definition	Example or Related Criteria
1	Benefits of grading per criterion	Rater mentions the benefits of grading by criterion of all 25 students at once instead of grading all criteria of one student.	See each code
1-1	Keeping focus on one criterion	Rater states that grading by criterion helps the rater focus on one criterion.	[Lindsey, p. 1] “I thought that I liked the idea of grading one criterion at a tie mainly because I think that it was easier to focus on the criteria.” [Kristen, p. 3] “I felt more confident in looking at each of the criterion individually and thus focusing on one at a time rather than the whole product as one complete piece.”
1-2	Crystalizing understanding of the rubric	Rater states that grading by criterion helps the rater understand the rubric better.	[Lindsey, p. 1] “I thought that I liked the idea of grading one criterion at a tie mainly because (...) it helped me kind of crystalize my understanding of the rubric.”
2	Usefulness of the scoring protocol	Rater provides her opinion on the usefulness of the scoring protocol for grading.	See each code
2-1	Useful for grading certain criteria	Rater talks that the scoring protocol is especially useful for certain criteria in the rubric.	[Lindsey, p. 2] “I found the protocol especially useful for the audience awareness criterion. (...) I like the examples that you had provided in there. It helped me be able to recognize, because those were some of the things that we struggled within

			the training, what is really audience awareness and what really shows that they know that they are presenting this for someone.” “I liked it for the quality of the summary, too.”
2-2	Burden: looking at between the rubric and the protocol	Rater states that the scoring protocol is not necessary for grading some criteria and it includes too much information.	[Lindsey, p. 2] “I thought it was a lot to look at though between the protocol and the rubric. I didn’t find that I referred to it all of the time. I would glance at it and go, “Okay, all right, I’ve got the idea”, until it came to audience awareness.
2-3	Very helpful for consistent grading	Rater states that the scoring protocol helps raters grade students’ products more consistently.	[Kristen, pp. 1-2] Sohee: “So, when you grade the quality of the summary part, was the supplementary information that I provided you within the scoring protocol helpful?” Kristen: “That was very helpful, yes.” ... “I think just from a consistency standpoint, knowing that I wasn’t the only grader, it helped me to know I’m using the same information that other scorers would be using or other teachers
3	Advantages/importance of using the rubric	Rater talks about the benefits or importance of using the rubric to grade students’ products. Sometimes, the rater refers to the names and numbers of the criteria that she thinks specifically important in the rubric and the reasons.	See each code
3-1	Evaluating multimodality (C1~7)	Rater reports that using the rubric is important and beneficial in order to evaluate multimodal aspects of the students’ products.	[Lindsey, p. 9] “I think that this, aspect 1, the technical aspects, is important because if you’re going multimodal, I mean really, this is mattering. This is important. How are they going to convey themselves and the information and some of the technical aspects of it.”
3-2	Consistent and objective grading	Rater states that the rubric helps raters grade students’ products more consistently and objectively.	[Kristen, p. 5] “As a teacher if I’m grading half of the class, then that particular grading session I may be harsher in one particular element, then I come back to it and I grade the second half of the class, maybe I get a little more lacks in

			terms of “well that’s okay.” But if I have a rubric, then there is no room for personal opinion.”
3-3	Consistent way to provide feedback to students	Rater talks that the rubric may help teachers provide feedback on students’ products more consistently.	[Kristen, p. 4] “I think it definitely gives them more consistent approach to providing students with their score and feedback to be specific as a teacher instead of saying that they needed to improve, you can say specifically where they needed to improve.”
3-4	A medium for conversation between the teacher and students	Rater reports that the rubric may help teachers and students have productive conversation and have the same understanding.	[Kristen, p. 5] I think it helps students and teachers both to be on the same page to understand. A student can understand a teacher’s expectation.”
4	Challenges of using the rubric	Rater states the challenges she faced while using the rubric for grading. Sometimes, rater refers to the names and numbers of the criteria that challenges her and the reasons.	See each code
4-1	Too many criteria	Rater describes that the rubric includes too many criteria than other traditional writing rubrics.	[Lindsey, p. 22] I think that they (other teachers) would be overwhelmed to start with, how long it was.
4-2	Inappropriate page-break	Rater points out the difficulty caused by the inappropriately divided rubric, which does not present all descriptors of a certain criterion in one page.	[Lindsey, p. 6] And really, with it being on a page break like this, I had to keep paying attention to ... I couldn’t just go five, five, three to four, two, because then I thought, “Oh, I have to turn the page. Oh no.” And then I had to fix a couple of them, because like, “Okay, three to four, two, whatever, that’s easy.” But then when I turn the page, I think understanding the difference between the theme music versus a sound effect.
4-3	Difficulty evaluating	Rater states her struggles to focus on and evaluate one mode	[Lindsey, p. 16] “Now, I’ll tell you also, though. When I did this, the audio voice, and I was really trying to focus on the voice. I closed my eyes, or I didn’t look at the screen, because

	different modes separately	among the multimodal ensemble.	I knew I was supposed to focus on and not get distracted by some other beautiful images. So I knew I was supposed to ... Or at least that's what I thought, I hope I did it right. I was trying to focus on what they were saying or the audio, what the sound was."
4-4	Mismatch (C3, C10)	Rater reports that her expected level of students' performances and described level of performance in the criterion do not match.	[Lindsey, p. 3] Even if I didn't agree as a teacher that I thought it was as good, but because of the tool that I was using, yes, they did what they were supposed to do, maybe ... "When we looked at the summary, it wasn't a very good summary. However, it met all of the pieces that they needed for this.
4-5	Tough language (C4)	Rater states that some words in the criterion descriptors are vague and difficult to interpret	[Lindsey, p. 3] Some of the language is still tough, because when you have "some" and you have "almost all" and you have "almost all" here, and you have "some." Those things are difficult." ...
4-6	Inappropriate label (C8)	Rater thinks that some labels of certain criteria do not represent the attributes of the criteria adequately.	[Lindsey, p. 11] "You want them to start really feeling that they have. That's why you want them to be aware of the audience because you want them to realize that they have a message to share, and that they can be kind of creative and take some risks in how they do it. So, I don't know if organization is the right label for that, then. That might end up being... It could be emphasis of required elements, you know, or something like that."
4-7	Interconnection of several criteria (C9, C10)	Rater states that some criteria evaluate aspects that are evaluated in different criteria.	[Lindsey, p. 16] This "quality of opinion", goes kind of into your audience awareness, too. Like how you're presenting your opinion, your thoughts. I found that sometimes the opinion was peppered throughout also. [Kristen, p. 1] Cause for certain students it seemed to be split in some ways than it always keeps all the summary elements together. Sometimes they wrapped up certain pieces in the conclusion where they had all the elements, but they didn't keep all the summary together, so I just want to make sure it

			wasn't a support for their opinion or their likes and dislikes if it was connected in some ways. So that threw me off a little bit, so I would re watch those videos several times to make sure the elements of the summary, that I was getting all of them clearly.
5	Unnecessary criteria and the reasons	Rater refers to names or numbers of the criteria that she thinks unnecessary for the rubric and the reasons.	See each code
5-1	Nothing	Rater reports that there is no unnecessary criterion in the rubric.	[Kristen, p. 3] Well for me, I think one of the questions I had was because it's a multimodal product, it's difficult to remove because you're assessing content you know which would be the summary, the opinion, all those things, organization. But then you're assessing the mode to which they're delivering that which is the iMovie in this case or the digital story telling basically. So if you're only going to score based on their use of technology, that would be its own rubric, if you're only looking for content, that would be its own rubric but since you're looking at both of them together, honestly, I don't know how you really would eliminate too many of these.
5-2	C9: Under certain condition	Rater presents her thoughts on how to cut out the audience awareness criteria under certain conditions.	[Lindsey, p. 14] I was trying to think what you would cut out, because you have really good information in here. So it's hard. (...) Most of these kids have watched YouTube by now, and they know those YouTubers and they like Dan DTM or whatever, so they like these characters, and they would like to emulate them. So they'll say, "Hey ..." Maybe just making sure that they're aware, so that might be like just something that the teacher tells them, "You will have an audience for this. It's not the book reports you're used to writing. You will have an audience." Then maybe you don't have to measure that. Maybe you don't need ...

6	Previous experience with training for rubric use and development	Rater reports their previous experience with systematic rubric development or training.	See each code
6-1	No	Rater states that she does not have any previous experience with training for rubric use and development.	[Kristen, p. 2] Never. I've never had this type of training for use of rubrics. I've generated my own rubrics or looked at curriculum alignment in terms of maybe within my own school, teachers that we would talk about similar projects and we would say "okay, what are we assessing or how are we going to score or grade this particular tasks" but I've never gone to a formal training for either rubric development or use of rubrics before this.
6-2	Yes	Rater shares her previous experience with training for rubric use and development.	[Lindsey, p. 6] "Yes, I used to work on a state committee, and so I helped develop the refine ... They kind of presented us with the idea of it, but we refined the writing rubric that the state used for a long time. Trying to really define the differences between, "So what makes it a three, what makes it a two?" And we did a lot of score validations, I did a lot of the validations for the ... It was writing was the main focus, but that was really where we had rubrics. What pushes it to a three, or what keeps it at a two, and trying to do a lot of those things. So we'd do a lot of these, like what you're doing, the inter-rater reliabilities, and things like that."
7	Feedback on scoring training	Rater provides feedback on both face-to-face and online scoring trainings and suggests some improvements for future studies.	See each code

7-1	Preference: face-to-face training	Rater states her preference on and benefits of face-to-face training.	[Lindsey, p. 18] I think that myself, I like face-to-face aspects, or I liked to be appeared that we did.
7-2	Potential of tutorial video for more guided training	Rater describes how tutorial video for scoring my help future users to be familiar with the rubric.	[Lindsey, p. 18] I think for some teachers, they might really appreciate it if you did like an online learning kind of video, where they could pause it. And if you did it kind of like a lesson, where you say, "Okay, let's watch ..." Who was one of the kids? Like Erin. "Let's watch Erin together, and let's count all of the frames first." And so you watch it, and say, "Let's count all the frames." And so they count all the frames, and then they can input their score. And they have those different things, right? Blackboard, or whatever.
7-3	Importance of benchmarks	Rater emphasizes the need of appropriate and enough number of benchmarks during scoring training.	[Lindsey [pp. 19-20] "I would choose one or two that were fairly different, so that then you're showing how you look at those different kinds of kids. Or even three, one that was a high one, one that was a medium, and one that a low. So that you walk people through, these are our benchmarks. That's what we use to do, is do benchmarks."
8	Suggestions for rubric revision	Rater suggests some possible ways of rubric revision.	See each code
8-1	Making the rubric as a separable one	Rater suggests several ways that some parts of the rubric can be used separately.	[Kristen, p. 3] what you're assessing if you look at them as individuals there, it really works out to about five parts for each. Five for the technology, five for the content.
8-2	Adoption of plus/minus system	Rater states that she thinks plus/minus grading system or more levels of performance may be required for the rubric.	[Lindsey, p. 5] "So then for some of this, especially here in the middle, I felt like I was pushed one way or another. I thought every once in a while on here, because if it was 51.2% or something like that, then that pushed it to a good, and "Well, I guess so." And it's usually in the middle aspect that it might feel pushed. But I guess in reality, you have to be pushed one way or another anyway. So this helped quantify it, at least in justified, well. It was good. It's a low good, but it was good. I

			don't know if you really want to get into a three minus, or three pluses, because those can be really messy by the time you really think, "Well I give it a three minus, because it's a low good." Or whatever."
8-3	Bigger memo space	Rater recommends making the blank space for memo bigger than the rubric template used during the grading.	[Kristen, p. 7] I love the way you gave the spreadsheet from excel, I would probably just make the memo boxes a little bigger in I were grading, obviously I would only have one student per page.
8-4	Combining the rubric and the supplementary information	The researcher suggests inclusion of the supplementary information of each criterion written in the scoring protocol under the rubric of the criterion and the rater agrees to the idea.	[Focus group; Kristen, p. 9] "Yeah, I think... I like the rubric the way it was that you gave us in terms of here's 1~11, but then I would have this as supplementary information for a teacher to review, like a teacher manual, when you're going through this how you would score this section."
8-5	Suggestions for revision of C3	Rater provides opinions on some possible ways that the descriptors of criterion 3 can be improved.	[Lindsey, pp. 8-9] "Well, is it partly sunny or mostly cloudy?" It's almost the same thing, right? Trying to find words like "most of the oral language has short and simple ...", "some of the oral language has short simple structured ..." (...) And you might, if you're going to bring this in, and a couple of longer sentences. This one has "mostly longer sentences that vary." Because if they're varying in sufficient length and structure, you're going to have some short ones. So even if it was a run-on sentence, like it was longer, so they gave me some different kinds of variety of the length and structure, but then the problem was it didn't necessarily follow this. So it kind of competed with each other for me to try to figure out what score it was.
8-6	Suggestions for revision of C4	Rater provides opinions on some possible ways that the descriptors of criterion 4 can be	[Focus group; Lindsey, pp. 17-18] "If you're looking at deleting one, I agree four would be the one to delete. (...) Because it's not a totally written product and you don't really want them presenting a script of what they're saying. A

		improved or deletion of the criterion.	PowerPoint where you read all the bullets or something like that. Really, what you're trying to see is how they highlight what they're saying, their oral language, with the written things."
8-7	Suggestions for revision of C6	Rater provides opinions on some possible ways that the descriptors of criterion 6 can be improved.	[Focus group; Kristen, pp. 15-16] "To me personally, I like the language more of the percentages better than a number of frames, because student work product is going to vary widely. (...) I think the percentage is going to get you a stronger student work product than saying a number of frames"
8-8	Suggestion for revision of C11	Rater provides opinions on some possible ways that the descriptors of criterion 11 can be improved.	[Focus group; Kristen, p. 8] "I look that as a clarifier of the opinion sections would be the like, dislike, and recommendation. That was an oversight on my part, but because it didn't appear in three, or reasons and the opinion section, i.e., I would remove that, unless you're going to have it here in all of the other. When you use the expression "opinion section," I wouldn't use the i.e. as a clarifier."
9	Intention of using the rubric in the future	Rater states whether she wants to use the rubric for her student teaching and assessing in the future or not.	See each code
9-1	Intend to use	Rater states that she wants to use the rubric in her class later.	[Kristen, p. 7] I honestly ... With the product itself, I think it's perfectly suitable. Yeah I would love to use it again in the future.
10	Appropriateness of the rubric language:	Rater provides opinion on the appropriateness of the rubric language to students and teachers.	See each code
10-1	Appropriate for teachers, not for students	Rater states that the rubric language is only appropriate for teachers, not for students.	[Kristen, p. 4] I think the rubric is clear from a teacher's standpoint. I think if I were to give this to my students, it would be overwhelming. It would be difficult for them to process what they're being asked to do. So, I don't know if

			there would be a way to modify for kid friendly language with this.
11	Other important aspects to be considered for the use of rubric	Rater provides more opinions on the conditions that should be considered for successful application of the developed rubric.	See each code
11-1	Teacher's level of knowledge on technology	Rater emphasizes that the teacher's level of knowledge may influence on their different understanding of the rubric and scoring training.	[Kristen, p. 6] "I think with a teacher who was trying to implement more of the digital text, digital age into their classroom, they're not going to be as confident and comfortable with the technical aspects. So I think that could be a challenge because in that case what a teacher review as you know poor quality may or, you know as be even done well, may not be exactly what the rubric is discussing because resolution, pixelation, angle, they may not understand. They won't even know how to change the Ken Burns effect so to them, they might feel like it's adequate. So I just think knowing ... The teacher would have to know something. They would have to have a foundation with technology."
11-2	Instructional context	Rater states that knowing what was taught to students is important to use the rubric adequately.	[Kristen, p. 4] Especially that was the one question I think I had to ask you when we first met was how much technical training did the students have because if they just know iMovie in terms of just the very basics or just the very standards cause sometimes I think some of the technical errors or loss of points for students who didn't perform well in those sections could have been a lack of instruction for them.

Appendix H

RELATIONSHIPS AMONG DOMAINS, NEW CRITERIA AND ORIGINAL CRITERIA IN BOTH THE RUBRIC AND NON-RUBRIC LITERATURE

Domain 1: Artifact		
New Criteria	Original Criteria in Rubric Literature	Original Criteria in Non-Rubric Literature
Coherence of multimodal product	<ul style="list-style-type: none"> • Design for medium (Burnett et al., 2014) • Organization (Ostenson, 2013) • Organization & Coherence (Borton & Hout, 2007) • Physical design: Timing (Morain & Swarts, 2012) 	<ul style="list-style-type: none"> • Coherence (Yancey, 2004) • Cohesion (Levy & Kimber, 2009) • Design for a print PSA (Selfe & Selfe, 2008) • Metaphor (Sorapure, 2006) • Metonymy (Sorapure, 2006) • Multimodality (Wierszewski, 2013) • Structure, organization, arrangement (Selfe & Selfe, 2008) • Use of modalities, media, and genre (Yu, 2014)
Organization of content	<ul style="list-style-type: none"> • Organization (Burnett et al., 2014) • Cognitive design: completeness (Morain & Swarts, 2012) 	<ul style="list-style-type: none"> • Organization (Wierszewski, 2013)
Linguistic mode: Conventions	<ul style="list-style-type: none"> • Convention (Burnett et al., 2014) • Grammar (Towndrow et al., 2013) • Source materials (Borton & Hout, 2007) 	<ul style="list-style-type: none"> • Convention (Yu, 2014) • Genre (Selfe & Selfe, 2008) • Grammar (Wierszewski, 2013) • Grammar/Mechanics (Yu, 2014)
Linguistic mode: Relational relevance	<ul style="list-style-type: none"> • Linguistic design (Hung et al., 2013) 	

Audio mode: Technical aspects	<ul style="list-style-type: none"> • Audio (Ostenson, 2013) • Auditory design (Hung et al., 2013) • Voice (Towndrow et al., 2013) • Fluency (Brown, 2013) • Viewability (Morain & Swarts, 2012) 	<ul style="list-style-type: none"> • Design for an audio PSA (Selfe & Selfe, 2008)
Audio mode: Relational relevance	<ul style="list-style-type: none"> • Audio (Ostenson, 2013) • Auditory design (Hung et al., 2013) • Voice (Towndrow et al., 2013) • Fluency (Brown, 2013) • Viewability (Morain & Swarts, 2012) 	<ul style="list-style-type: none"> • Design for an audio PSA (Selfe & Selfe, 2008)
Visual mode: Technical aspects	<ul style="list-style-type: none"> • Image (Ostenson, 2013) • Gestural design (Hung et al., 2013) • Physical design: Viewability (Morain & Swarts, 2012) 	<ul style="list-style-type: none"> • Aesthetic/visual appeal (Yu, 2014) • Design (Levy & Kimber, 2009) • Design for a video PSA (Selfe & Selfe, 2008)
Visual mode: Relational relevance	None	None
Spatial mode: Technical aspects	<ul style="list-style-type: none"> • Spatial design (Hung et al., 2013) 	<ul style="list-style-type: none"> • Formal arrangement (Wierszewski, 2013)
Spatial mode: Relational relevance	<ul style="list-style-type: none"> • Spatial design (Hung et al., 2013) 	<ul style="list-style-type: none"> • Formal arrangement (Wierszewski, 2013)
Domain 2: Context		
New Criteria	Original Criteria in Rubric Literature	Original Criteria in Non-Rubric Literature
Rhetorical awareness: Task	<ul style="list-style-type: none"> • Mode of presentation (Borton & Hout, 2007) • Physical design: Accessibility (Morain & Swarts, 2012) 	<ul style="list-style-type: none"> • Following the assignment; purpose (Wierszewski, 2013) • Rhetorical context (Selfe & Selfe, 2008) • Rhetorical knowledge (Yu, 2014)

Rhetorical awareness: Audience	<ul style="list-style-type: none"> • Rhetorical awareness (Burnett et al., 2014) • Voice (Howell et al., 2013) 	<ul style="list-style-type: none"> • Audience (Wierszewski, 2013) • Rhetorical context (Selfe & Selfe, 2008) • Rhetorical knowledge (Yu, 2014)
Domain 3: Substance		
New Criteria	Original Criteria in Rubric Literature	Original Criteria in Non-Rubric Literature
Quality of ideas	<ul style="list-style-type: none"> • Character analysis (Husbye & Rust, 2014) • Content (Towndrow et al., 2013) • Economy (Towndrow et al., 2013) • Interpretation (Husbye & Rust, 2014) • Themes (Husbye & Rust, 2014) • Theme/point of view (Towndrow et al., 2013) • Cognitive design: accuracy, pertinence (Morain & Swarts, 2012) 	<ul style="list-style-type: none"> • Content (Levy & Kimber, 2009) • Content (Yu, 2014) • Movement (Wierszewski, 2013)
Quality of opinions/arguments	<ul style="list-style-type: none"> • Critical thinking skills (Borton & Hout, 2007) • Idea & Organization (Howell et al., 2015) • Stance and support: Argument, evidence, and analysis (Burnett et al., 2014) • Affective design: confidence (Morain & Swarts, 2012) 	
Domain 4: Process Management and Technique		
New Criteria	Original Criteria in Rubric Literature	Original Criteria in Non-Rubric Literature
Collaboration	<ul style="list-style-type: none"> • Collaboration (Howell et al., 2015) 	<ul style="list-style-type: none"> • Collaboration (Yu, 2014)
Technical skills	<ul style="list-style-type: none"> • Development of new literacies (Brown, 2013) 	<ul style="list-style-type: none"> • Technical execution (Wierszewski, 2013)

	<ul style="list-style-type: none"> • ICT usage (Towndrow et al., 2013) 	
Writing processes and strategies	<ul style="list-style-type: none"> • Publication (Howell et al., 2015) • Writing process (Brown, 2013) 	
Domain 5: Habits of Mind		
New Criteria	Original Criteria in Rubric Literature	Original Criteria in Non-Rubric Literature
Creativity		<ul style="list-style-type: none"> • Creativity (Wierszewski, 2013) • Creativity/Originality (Yu, 2014)
Self-efficacy	<ul style="list-style-type: none"> • Confidence, Self-efficacy (Morain & Swarts, 2012) 	
Uncategorized Criteria		
New Criteria	Original Criteria in Rubric Literature	Original Criteria in Non-Rubric Literature
	<ul style="list-style-type: none"> • Mood (Husbye & Rust, 2014) • Reading Comprehension (Brown, 2013) 	<ul style="list-style-type: none"> • Overall (Wierszewski, 2013) • Style/tone (Yu, 2014)

Appendix I

LISTED CHARACTERISTICS OF EACH STUDENT'S PERFORMANCE RELATED TO THE RELATIONAL RELEVANCE OF WRITTEN LANGUAGE MODE

Student Pseudonym	Characteristics related to the relational relevance of written language mode
Aaron	<ul style="list-style-type: none"> • No use of title
Amy	<ul style="list-style-type: none"> • Titles are used minimally. • Only one title is used at the introduction section in order to inform the book title.
Antonio	<ul style="list-style-type: none"> • Titles are used minimally. • Only one title is used at the introduction section in order to inform the book title, author name, and the reviewer's name.
Art	<ul style="list-style-type: none"> • Titles are used minimally. • Two titles are used: one is used to inform the book title and the author's name and another one is used as a title of the "I think you did not like" section.
Caeden	<ul style="list-style-type: none"> • Titles are used extensively. • All frames included titles and it delivers content of the book review video instead of oral language. Since the student included too many words in one title, font size is too small, so it is less readable.
Dede	<ul style="list-style-type: none"> • Titles are used extensively. • Almost all frames included titles and they summarize content in each frame well.
Deirdre	<ul style="list-style-type: none"> • No use of title
Doug	<ul style="list-style-type: none"> • No use of title
Erin	<ul style="list-style-type: none"> • No use of title
Ethan	<ul style="list-style-type: none"> • No use of title
Eve	<ul style="list-style-type: none"> • No use of title
Gillian	<ul style="list-style-type: none"> • Titles are used moderately. • (e.g., introduction, summary, and conclusion section)
Ida	<ul style="list-style-type: none"> • No use of title
Jaci	<ul style="list-style-type: none"> • Titles are used minimally.

	<ul style="list-style-type: none"> • (e.g., only at a recommendation section)
Jenny	<ul style="list-style-type: none"> • Titles are used moderately. • The first title informs the book title, author name and the student's name as a reviewer; Second title is only representing a part of the summary; second to fourth titles were not used effectively.
Joey	<ul style="list-style-type: none"> • Titles are used minimally. • Only one title is used at the introduction section in order to inform the book title and the author.
John	<ul style="list-style-type: none"> • Titles are used extensively. • Written language replaces the role of oral language.
Jorge	<ul style="list-style-type: none"> • Titles are used minimally. • Only one title is used at the introduction section in order to inform the genre of the video and the reviewer's name.
Karen	<ul style="list-style-type: none"> • Titles are used minimally. • (e.g., She used titles for introduction and conclusion: the book title and the author's name; thank you and her name)
Kari	<ul style="list-style-type: none"> • Titles are used extensively. • Most frames included titles; Written languages complement or summarize the content of oral language
Mallory	<ul style="list-style-type: none"> • Titles are used extensively. • Very relevant use of titles - when she wants to clarify the role of images in the frame or the role of the section, she used titles
Maria	<ul style="list-style-type: none"> • Titles are used moderately. • Written language presents her opinion about the story, but each title is not relevant to the content in each frame.
Marla	<ul style="list-style-type: none"> • No use of title
Meghan	<ul style="list-style-type: none"> • Titles are used moderately. • Written language repeats some details in oral language or present different information with oral language... titles were used confusedly.
Michael	<ul style="list-style-type: none"> • No use of title
Phillip	<ul style="list-style-type: none"> • Titles are used minimally. • He used titles at the introduction and conclusion; In the summary, he used title in order to explain the meaning of the image he used.
Sam	<ul style="list-style-type: none"> • Titles are used minimally. • Only two titles were used. One informs the book title and the author and the other informs the reviewer's name.
Sarah	<ul style="list-style-type: none"> • Titles are used minimally. • Only one title is used at the introduction section in order to inform the book title and the genre of the video.

Sylvester	<ul style="list-style-type: none">• Titles are used moderately.• Very relevant use of titles - Written language complements the content of oral language in the same frame.
Veronica	<ul style="list-style-type: none">• Titles are used minimally.• Only one title is used at the introduction section in order to inform the book title, the author, and the reviewer's name.

Appendix J

GROUPED CHARACTERISTICS ON EACH CRITERION

Criterion	Grouped Characteristics
1. Technical aspects of visual mode	<ol style="list-style-type: none">1) The student used only one image per frame2) The student used more than one image per frame3) The student did not take video of him/herself.4) The student's self-video was shoot clearly under the best conditions such as a good angle, appropriate lighting, no noise other than his/her narration etc.5) The student used images
2. Technical aspects of audio-voice	<ol style="list-style-type: none">1) The student did not include oral language2) The student's volume of voice was appropriate.3) The student's volume of voice was too loud or too quiet.4) The student's speed of narration was appropriate.5) The student's speed or narration was too fast or too slow.6) The student's pronunciation was clear.7) The student's pronunciation was unclear.
3. Conventions of oral language	<ol style="list-style-type: none">1) The student did not use oral language.2) The student used vague words or pronouns.3) Most sentences in the student's oral language did not follow standard English conventions.4) Most sentences in the student's oral language were short and structures were simple. However, they followed standard English conventions.5) Sentences in the student's oral language were generally complete with sufficient variety in length and sentence.6) Sentences in the student's oral language were skillfully constructed with appropriate variety in length and structure.
4. Conventions of written language	<ol style="list-style-type: none">1) The student did not include titles (written language).2) Titles included many different types of errors (three or more types) in in punctuation, capitalization, grammar and/or spelling.

	<p>3) Titles included some errors (3~4) in punctuation, capitalization, grammar and/or spelling.</p> <p>4) Titles included a few errors (1~2) in punctuation, capitalization, grammar and/or spelling.</p> <p>5) Titles did not include any errors in punctuation, capitalization, grammar and/or spelling.</p>
5. Relationship between oral language and other modes	<p>1) An image was relevant to a keyword or short phrases of oral language in each frame.</p> <p>2) Images divided one section into several frames in order to represent a keyword of oral language in a frame</p> <p>3) An image added new information to the content of oral language in a frame</p> <p>4) Most images included in the video are not relevant to the content of oral language in each frame</p> <p>5) Sound effects added an appropriate mood onto the oral language and image.</p> <p>6) Sound effects were too loud and hindered listening of oral language.</p>
6. Relationship between written language and other modes	<p>1) The student did not include titles (written language).</p> <p>2) Titles were used minimally (e.g., only one title was included in the entire video).</p> <p>3) Titles were used moderately (e.g., titles were included in three frames of the entire video).</p> <p>4) Titles were used extensively (e.g., titles were included in five or more frames of the entire video.)</p> <p>5) Written language clarified the role of oral language.</p> <p>6) Written language substituted oral language.</p> <p>7) Written language in frames played a role as a title or short summary.</p>
7. Coherence of multimodal product	<p>1) Sound effects/theme music added an appropriate mood onto other modes.</p> <p>2) Sound effects/theme music were not harmonized with other modes.</p> <p>3) The student included oral language, written language, visual, and audio modes in the book reviews, but these modes convey information redundantly.</p> <p>4) All modes conveyed meaning harmoniously.</p>

8. Rhetorical awareness: Audience	1) There was evidence of the student's audience awareness in the artifact. 2) There was no evidence of the student's audience awareness in the artifact.
9. Quality of summary	1) The student did not include any summary. 2) The student included too short and vague summary. Main characters and events were not introduced. 3) The student included too long and detailed summary. It provides information about the ending. 4) The student included wrong information about the book in the summary.
10. Quality of opinion	1) The student provided clear opinion and enough supporting reasons at the three opinion sections (1 thing I liked, 1 thing I did not like, and recommendations). 2) The student provided clear opinion at the three opinion sections, but s/he did not provide supporting reasons thoroughly. 3) The student provided his/her opinion without any supporting reasons.

Appendix K

FIRST DRAFT OF THE RUBRIC (RUBRIC #1)

Domain	Category	Criteria	Points	
			Per Criterion	Per Category
Artifact	1. Technical aspect of non-linguistic modes	1) Technical aspect of visual mode (images and videos)	0~4	0~8
		2) Technical aspect of audio-voice	0~4	
	2. Conventions of linguistic modes	3) Conventions of oral language (narration)	0~4	0~8
		4) Conventions of written language (Caption)	0~4	
	3. Coherence of multimodal product	5) Relationship between visual and oral language	0~4	0~12
		6) Relationship between oral language and written language	0~4	
		7) Relationship between oral language and sound	0~4	
	4. Organization of multimodal product	8) Organization of multimodal product	0~4	0~4
Context	5. Audience awareness	9) Audience awareness	0~4	0~4
Substance	6. Quality of content	10) Quality of summary	0~4	0~8
		11) Quality of opinion	0~4	
Total			0~44	

Domain & Category	Criteria	Descriptors of Levels				
		4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
1. Artifact: Technical aspect of modes ----- Points of Category 1: Sum of Criteria 1) & 2) = ()	1) Technical aspects of visual mode (images and videos)	<ul style="list-style-type: none"> The artifact includes <u>both</u> static images and videos. <u>All</u> technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are in <u>good</u> quality such that visuals convey meaning clearly. 	<ul style="list-style-type: none"> The artifact includes <u>both</u> static images and videos. Three or four technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are in <u>good</u> quality such that visuals convey meaning clearly. 	<ul style="list-style-type: none"> The artifact includes <u>only</u> either static images or videos. One or two technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are in <u>good</u> quality, so they hinder visuals conveying meaning clearly. 	<ul style="list-style-type: none"> The artifact includes <u>only</u> either static images or videos. No technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are in <u>good</u> quality, so they hinder visuals conveying meaning clearly. 	The artifact does not include any visual mode.
	2) Technical aspect of audio-voice	<ul style="list-style-type: none"> <u>All</u> characteristics of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) are in <u>good</u> quality. The artifact includes voice that is pronounced clearly with appropriate volume, speed, and tone. 	<ul style="list-style-type: none"> <u>Two or three</u> characteristics of voice among <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u> are in <u>good</u> quality. 	<ul style="list-style-type: none"> <u>Only one</u> characteristic of voice among <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u> is in <u>good</u> quality. 	<ul style="list-style-type: none"> <u>All</u> characteristics of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) are in <u>poor</u> quality. Voice is too loud or quiet, too fast or slow, monotonous, and/or pronounced unclearly. 	The artifact does not include any voice.

Domain & Category	Criteria	Descriptors of Levels				
		4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
2. Artifact: Conventions of linguistic modes <hr/> Points of Category 2: Sum of Criteria 3) & 4) = ()	3) Conventions of oral language (narration)	<ul style="list-style-type: none"> Oral language in the artifact has sentences that are generally complete with sufficient variety in length and structure. Most sentences follow standard English conventions. 	<ul style="list-style-type: none"> Oral language in the artifact has sentences that are generally complete with sufficient variety in length and structure. Most sentences <u>do not</u> follow standard English conventions. 	<ul style="list-style-type: none"> Oral language in the artifact has short and simply structured sentences. Most sentences follow standard English conventions. 	<ul style="list-style-type: none"> Oral language in the artifact has short and simply structured sentences. Most sentences <u>do not</u> follow standard English conventions. 	The artifact does not include any oral language.
	4) Conventions of written language (Caption)	<ul style="list-style-type: none"> More than 50% of frames in the artifact include captions. Captions include no or less than three errors in punctuation, capitalization, grammar and/or spelling, so the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> More than 50% of frames in the artifact include captions. Captions include three or more errors in punctuation, capitalization, grammar and/or spelling, so the errors hinder conveying meaning clearly. 	<ul style="list-style-type: none"> Less than 50% of frames in the artifact include captions. Captions include no or less than three errors in punctuation, capitalization, grammar and/or spelling, so the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> Less than 50% of frames in the artifact include captions. Captions include three or more errors in punctuation, capitalization, grammar and/or spelling, so the errors hinder conveying meaning clearly. 	The artifact does not include any written language (caption).

Domain	Criteria	Descriptors of Levels				
		4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
3. Artifact: Coherence of multimodal product ----- Points of Category 3: Sum of Criteria 5) ~ 7) = ()	5) Relationship between visual and oral language	<ul style="list-style-type: none"> All visuals are relevant to the content of oral language in each frame, so the combination of visual and oral language makes sense. Visuals complement or augment meaning of oral language and there are no or little redundant visuals. 	<ul style="list-style-type: none"> Most (More than half of) visuals are relevant to the content of oral language in each frame, so the combination of visual and oral language makes sense. Some relevant visuals are redundant. 	<ul style="list-style-type: none"> Most (More than half of) visuals are irrelevant to the content of oral language in each frame, so the combination of visual and oral language makes little sense. Even a few relevant visuals are redundant. 	<ul style="list-style-type: none"> All visuals are irrelevant to the content of oral language in each frame, so the combination of visual and oral language makes no sense. 	There is no relationship between visual and oral language.
	6) Relationship between oral language and written language	<ul style="list-style-type: none"> All captions are relevant to the content of oral language in each frame, so the combination of written and oral languages makes sense. Written languages complement or augment meaning of oral language and there are no or little redundant written languages. 	<ul style="list-style-type: none"> Most (More than half of) captions are relevant to the content of oral language in each frame, so the combination of written and oral languages makes sense. Some written languages are redundant. 	<ul style="list-style-type: none"> Most (More than half of) captions are irrelevant to the content of oral language in each frame, so the combination of written and oral languages makes little sense. Even a few relevant written languages are redundant. 	<ul style="list-style-type: none"> All captions are irrelevant to the content of oral language in each frame, so the combination of written and oral languages makes no sense. 	There is no relationship between oral language and written language.

	<p>7) Relationship between oral language and sound</p>	<ul style="list-style-type: none"> • Volume of oral language and other sounds is appropriate, so it does not hinder listening of each mode. • Tone and/or mood of most sound effects and/or music complement or augment the message of the oral language, it is very smooth to engage in the artifact. 	<ul style="list-style-type: none"> • Volume of oral language and other sounds is too quiet or loud, so it hinders listening of each mode. • Tone and/or mood of most sound effects and/or music are somewhat relevant to the message of the oral language, so it is okay to engage in the artifact. 	<ul style="list-style-type: none"> • Volume of oral language and other sounds is okay, so it does not hinder listening of each mode. • Tone and/or mood of some sound effects and/or music are not relevant to the message of the oral language, so it is difficult to engage in the artifact. 	<ul style="list-style-type: none"> • Volume of oral language and other sounds is too quiet or loud, so it hinders listening of each mode. • Tone and/or mood of most sound effects and/or music are not relevant to the message of the oral language, so it is difficult to engage in the artifact. 	<p>There is no relationship between oral language and sound.</p>
<p>4. Artifact: Organization ----- Points of Category 4 ()</p>	<p>8) Organization of multimodal product</p>	<ul style="list-style-type: none"> • The artifact organizes content into six required sections of book review (i.e., introduction, brief summary, one thing you liked, one thing you did not like, recommendation, and conclusion). • It presents the content of each section appropriately. • The student weighs sections differently by using more than one image/video in some sections (e.g., brief summary) that they want to emphasize. 	<ul style="list-style-type: none"> • The artifact organizes content into six required sections of book review. • It presents the content in one or two sections inappropriately. • The student weighs all sections similarly by using only one image/video per section or emphasizes only one section too much. 	<ul style="list-style-type: none"> • The artifact misses one or two required sections of book review • It presents the content in one or two sections inappropriately. • The student weighs all sections similarly by using only one image/video per section or emphasizes only one section too much. 	<ul style="list-style-type: none"> • The artifact misses more than three required sections of book review • It presents the content in all included sections inappropriately. • The student weighs all sections similarly by using only one image/video per section or emphasizes only one section too much. 	<p>The artifact is not completed.</p>

Domain	Criteria	Descriptors of Levels				
		4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
5. Context: Audience Awareness ----- Points of Category 5 ()	9) Audience awareness	<ul style="list-style-type: none"> The author is aware of audiences clearly and let them engage in throughout the artifact. 	<ul style="list-style-type: none"> The author is aware of audiences and sometimes let them engage in the artifact. 	<ul style="list-style-type: none"> The author is aware of audience, but s/he does not let them engage in the artifact. 	<ul style="list-style-type: none"> The author is aware of audiences a little bit, but the audience is still vague. 	The author is not aware of audiences.
6. Substance: Quality of Content ----- Points of Category 6: Sum of Criteria 10) & 11) = ()	10) Quality of summary	<ul style="list-style-type: none"> The summary length is appropriate (no more than 50% of the entire video). It introduces main characters and major events effectively. It does not reveal the story's important ending. Information about the book is accurate. 	<ul style="list-style-type: none"> The summary length is moderate or a bit long (between 50~75% of the entire video). It introduces main characters and major events. It implies the story's ending to audiences. It includes some inaccuracies, but they do not hinder audience's understanding of the summary. 	<ul style="list-style-type: none"> The summary length is too long or short (more than 75% of the entire video). It introduces one of main characters and major events. It implies the story's ending to audiences. It includes wrong information about the book. 	<ul style="list-style-type: none"> The summary is too long or short (more than 75% of the entire video). It does not introduce main characters and major events. It let audiences know the story's important ending. It includes wrong information about the book. 	The artifact does not include any summary of the book.
	11) Quality of opinion	<ul style="list-style-type: none"> The author provides clear opinion and two or more supporting reasons at three opinion sections (i.e., "1 thing I liked," "1 thing I did not like," and "recommendation") 	<ul style="list-style-type: none"> The author provides clear opinion at the three opinion sections with only one or no supporting reason. 	<ul style="list-style-type: none"> The author provides opinion at two of three opinion sections with only one or no supporting reason. 	<ul style="list-style-type: none"> The author provides opinion at only one of three opinion sections with only one or no supporting reason. 	The author does not provide any opinion about the book.
Six Categories Total = ()						

Appendix L

A METARUBRIC FOR REVISION OF THE RUBRIC (ARTER & CHAPPUIS, 2006)

Criterion 1: COVERAGE/ORGANIZATION		
A. Covers the Right Content		
<i>5—Strong</i>	<i>3—Medium</i>	<i>1—Weak</i>
<ol style="list-style-type: none"> 1. The content of the rubric represents the best thinking in the field about what it means to perform well on the skill or product under consideration. 2. The content of the rubric aligns directly with the content standards/ learning targets it is intended to assess. 3. The content has the "ring of truth"—your experience as a teacher confirms that the content is truly what you do look for when you evaluate the quality of a student performance or product. In fact, the rubric is insightful; it helps you organize your own thinking about what it means to perform well. 	<ol style="list-style-type: none"> 1. Much of the content represents the best thinking in the field, but there are a few places that are questionable. 2. Some features don't align well with the content standards/learning targets it is intended to assess. 3. Much of the content is relevant, but you can easily think of some important things that have been left out or that have been given short shrift, or it contains an irrelevant criterion or descriptor that might lead to an incorrect conclusion about the quality of student performance. 	<ol style="list-style-type: none"> 1. You can't tell what learning target(s) the rubric is intended to assess, or you can guess at the learning targets, but they don't seem important, or content is far removed from current best thinking in the field about what it means to perform well on the skill or product under consideration. 2. The rubric doesn't seem to align with the content standards/learning targets it is intended to assess. 3. You can think of many important dimensions of a quality performance or product that are not in the rubric, or content focuses on irrelevant features. You find yourself asking, "Why assess this?" or "Why should this count?" or "Why should students have to do it this way?"

Criterion 1: COVERAGE/ORGANIZATION (Continued)		
B. Criteria are Well Organized		
5—Strong	3—Medium	1—Weak
<p>1. The rubric is divided into easily understandable criteria as needed. The number of criteria reflects the complexity of the learning target. If a holistic rubric is used, it's because a single criterion adequately describes performance.</p> <p>2. The details that are used to describe a criterion go together; you can see how they are facets of the same criterion.</p> <p>3. The relative emphasis on various features of performance is right—things that are more important are stressed more; things that are less important are stressed less.</p> <p>4. The criteria are independent. Each important feature that contributes to quality work appears in only one place in the rubric.</p>	<p>1. The number of criteria needs to be adjusted a little: either a single criterion should be made into two criteria, or two criteria should be combined.</p> <p>2. Some details that are used to describe a criterion are in the wrong criterion, but most are placed correctly.</p> <p>3. The emphasis on some criteria or descriptors is either too small or too great; others are all right.</p> <p>4. Although there are instances when the same feature is included in more than one criterion, the criteria structure holds up pretty well.</p>	<p>1. The rubric is holistic when an analytic one is better suited to the intended use or learning targets to be assessed; or the rubric is an endless list of everything; there is no organization; the rubric looks like a brainstormed list.</p> <p>2. The rubric seems "mixed up"—descriptors that go together don't seem to be placed together. Things that are different are put together.</p> <p>3. The rubric is out of balance—features of more importance are emphasized the same as features of less importance.</p> <p>4. Descriptors of quality work are represented redundantly in more than one criterion to the extent that the criteria are really not covering different things.</p>
C. Number of Levels Fits Targets and Uses		
5—Strong	3—Medium	1—Weak
<p>1. The number of levels of quality used in the rating scale makes sense. There are enough levels to be able to show student progress, but not so many levels that it is impossible to distinguish among them.</p>	<p>1. Teachers might find it useful to create more levels to make finer distinctions in student progress, or to merge levels to suit the rubric's intended use. The number of levels could be adjusted easily.</p>	<p>1. The number of levels is not appropriate for the learning target being assessed or intended use. There are so many levels it is impossible to reliably distinguish between them, or too few to make important distinctions. It would take major work to fix the problem.</p>

Criterion 2: CLARITY		
A. Levels Defined Well		
5—Strong	3—Medium	1—Weak
<p>1. Each score point (level) is defined with indicators and/or descriptors. <i>A plus:</i> There are examples of student work that illustrate each level of each trait.</p> <p>2. There is enough descriptive detail in the form of concrete indicators, adjectives, and descriptive phrases that allow you to match a student performance to the "right" score. <i>A plus:</i> If students are to use the rubric, there are student-friendly versions, and/or versions in foreign languages for ELL students.</p> <p>3. Two independent users, with training and practice, assign the same rating most of the time. <i>A plus:</i> There is information on rater agreement rates that shows that raters can exactly agree on a score 65% of the time, and within one point 98% of the time.</p> <p>4. If counting the number or frequency of something is included as an indicator, changes in such counts really are indicators of changes in quality.</p> <p>5. Wording is descriptive, not evaluative.</p>	<p>1. Only the top level is defined. The other levels are not defined.</p> <p>2. There is some attempt to define terms and include descriptors, but some key ideas are fuzzy in meaning.</p> <p>3. You have a question whether independent raters, even with practice, could assign the same rating most of the time.</p> <p>4. There is some descriptive detail in the form of words, adjectives, and descriptive phrases, but counting the frequency of something or vague quantitative words are also present.</p> <p>5. Wording is mostly descriptive of the work, but there are a few instances of evaluative labels.</p>	<p>1. No levels are defined; the rubric is little more than a list of categories to rate followed by a rating scale.</p> <p>2. Wording of the levels, if present, is vague or confusing. You find yourself saying such things as, "I'm confused," or "I don't have any idea what this means." Or, the only way to distinguish levels is with words such as <i>extremely, very, some, little, and none</i>; or <i>completely, substantially, fairly well, little, and not at all</i>.</p> <p>3. It is unlikely that independent raters could consistently rate work the same, even with practice.</p> <p>4. Rating is almost totally based on counting the number or frequency of something, even though quality is more important than quantity.</p> <p>5. Wording tends to be evaluative rather than descriptive of the work; e.g., work is "mediocre," "above average," or "clever."</p>
B. Levels Parallel		
5—Strong	3—Medium	1—Weak
<p>1. The levels of the rubric are parallel in content—if an indicator of quality is discussed in one level, it is discussed in all levels. If the levels are not parallel, there is a good explanation why.</p>	<p>1. The levels are mostly parallel in content, but there are some places where there is an indicator at one level that is not present at the other levels.</p>	<p>1. Levels are not parallel in content and there is no explanation of why, or the explanation doesn't make sense.</p>

Appendix M

RUBRIC ANALYSIS FORM (ARTER & CHAPPUIS, 2006)

Criterion	Indicator	Rating	Rationale (use words and phrases from the Rubric for Rubrics)
<i>Coverage/ Organization</i>	1A: Covers the Right Content		
	1B: Criteria Are Well Organized		
	1C: Number of Levels Fits Targets and Uses		
<i>Clarity</i>	2A: Levels Defined Well		
	2B: Levels Parallel		

Appendix N

FINAL VERSION OF THE RUBRIC USED FOR INDEPENDENT SCORING (RUBRIC #7)

1. Domains and Criteria Included in the Draft Rubric and Points

Domain	Category	Criteria	Points	
			Per Criterion	Per Category
Artifact	1. Technical aspects of non-linguistic modes	1) Technical aspects of visual mode (images and videos): The effects of camera shots and angles, lighting, color, size, movement, and sequencing on the quality of visual mode such as static or moving images	0~4	0~8
		2) Technical aspects of audio-voice: The effects of fluency, articulation, intonation, volume, pitch, and length on the quality of voice	0~4	
	2. Conventions of linguistic modes	3) Conventions of oral language: The effects of (grade-appropriate) conventions of Standard English on the quality of oral language	0~4	0~8
		4) Conventions of written language (titles): The effects of (grade-appropriate) grammar, mechanics, style, citation, and genre on the quality of written language	0~4	
	3. Coherence of multimodal product	5) Relationship between visual and oral language	0~4	0~12
		6) Relationship between oral language and written language	0~4	
		7) Relationship between oral language and sound	0~4	
	4. Organization of multimodal content	8) Organization of multimodal content: Logical structure of content or messages within and among frames or sections	0~4	0~4
Context	5. Audience awareness	9) Audience awareness: Consideration of explicit or implicit audiences and their engagement in the artifact	0~4	0~4
Substance	6. Quality of content	10) Quality of summary: Clarity, credibility, significance (depth and length), and interest of the content and the pace of content progress or development	0~4	0~8
		11) Quality of opinion: Clarity and persuasiveness of arguments and the use of analysis and evidence to support the argument	0~4	
Total			0~44	

2. Rubric

Domain & Category	Criteria	Descriptions of Levels				
		4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
1. Artifact: Technical aspects of non-linguistic modes <hr/> Points of Category 1: Sum of Criteria 1) & 2) = ()	1) Technical aspects of visual mode (images and videos)	<ul style="list-style-type: none"> Technical aspects of visuals (e.g., resolution, color, <u>camera shots and angles</u>, lighting, <u>Ken Burns</u>, and <u>transitions</u>) used in more than 75% of the frames convey meaning clearly. 	<ul style="list-style-type: none"> Technical aspects of visuals (e.g., resolution, color, <u>camera shots and angles</u>, lighting, <u>Ken Burns</u>, and <u>transitions</u>) used in 75-51% of the frames convey meaning clearly. 	<ul style="list-style-type: none"> Technical aspects of visuals (e.g., resolution, color, <u>camera shots and angles</u>, lighting, <u>Ken Burns</u>, and <u>transitions</u>) used in 50-26% of the frames convey meaning clearly. 	<ul style="list-style-type: none"> Technical aspects of visuals (e.g., resolution, color, <u>camera shots and angles</u>, lighting, <u>Ken Burns</u>, and <u>transitions</u>) used in 25% or fewer of the frames convey meaning clearly. OR The artifact includes visual mode in only one frame. 	<ul style="list-style-type: none"> The artifact does not include any visual mode.
	2) Technical aspects of audio-voice	<ul style="list-style-type: none"> All technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) convey meaning clearly for the duration of the artifact. For example, voice is pronounced clearly with appropriate volume, speed, and tone. 	<ul style="list-style-type: none"> Two or three technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) convey meaning clearly for the duration of the artifact. 	<ul style="list-style-type: none"> Only one technical aspect of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) conveys meaning clearly for the duration of the artifact. 	<ul style="list-style-type: none"> None of the technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) conveys meaning clearly for the duration of the artifact. For example, the voice is too loud or quiet, too fast or slow, monotonous, and/or pronunciation is unclear. 	<ul style="list-style-type: none"> The artifact does not include any voice.

Domain	Criteria	Descriptions of Levels				
		4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
2. Artifact: Conventions of linguistic modes <hr/> Points of Category 2: Sum of Criteria 3) & 4) = ()	3) Conventions of oral language (narration)	<ul style="list-style-type: none"> Oral language in the artifact has sentences that are generally complete with sufficient variety in length and structure. Almost all sentences follow the conventions of Standard English suggested in the Common Core State Standards (CCSS). 	<ul style="list-style-type: none"> Oral language in the artifact has sentences that are generally complete with sufficient variety in length and structure. Some sentences do not follow the conventions of Standard English suggested in the CCSS. 	<ul style="list-style-type: none"> Oral language in the artifact has short and simply structured sentences. Almost all sentences follow the conventions of Standard English suggested in the CCSS. 	<ul style="list-style-type: none"> Oral language in the artifact has short and simply structured sentences. Some sentences do not follow the conventions of Standard English suggested in the CCSS. 	<ul style="list-style-type: none"> The artifact does not include any oral language.
	4) Conventions of written language (Titles)	<ul style="list-style-type: none"> More than 50% of frames in the artifact include written language (titles). Written language (titles) include(s) three or fewer errors in punctuation, capitalization, grammar and/or spelling, but the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> 50% or fewer of frames in the artifact include written language (titles). Written language (titles) include(s) three or fewer errors in punctuation, capitalization, grammar and/or spelling, but the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> More than 50% of frames in the artifact include written language (titles). Written language (titles) include(s) more than three errors in punctuation, capitalization, grammar and/or spelling, and the errors may hinder conveying meaning clearly. 	<ul style="list-style-type: none"> 50% or fewer of frames in the artifact include written language (titles). Written language (titles) include(s) more than three errors in punctuation, capitalization, grammar and/or spelling, and the errors may hinder conveying meaning clearly. OR The artifact includes written language (titles) in only one frame. 	<ul style="list-style-type: none"> The artifact doesn't include any written language (titles).

Domain	Criteria	Descriptions of Levels				
		4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
3. Artifact: Coherence of multimodal product <hr/> Points of Category 3: Sum of Criteria 5) ~ 7) = ()	5) Relationship between visual and oral language	<ul style="list-style-type: none"> More than 75% of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to the audience without the author's explanation. Some visuals are used more than once to represent different content of the oral language purposefully. 	<ul style="list-style-type: none"> More than 75% of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to the audience without the author's explanation. Some visuals are used more than once to represent similar content of the oral language redundantly. 	<ul style="list-style-type: none"> 51~75% of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to the audience without the author's explanation. Some visuals may be used more than once to represent content of the oral language either purposefully or redundantly. 	<ul style="list-style-type: none"> Fewer than 50% of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to the audience without the author's explanation. Some visuals may be used more than once to represent content of the oral language either purposefully or redundantly. 	<ul style="list-style-type: none"> There is no relationship between visual and oral language.
	6) Relationship between oral language and written language (Titles)	<ul style="list-style-type: none"> Written language in 5 or more frames plays a role as a title or short summary of oral language, or presents (a part of) a few sentences in the oral language. Some of the written language may present bibliographic information of the book, student's name, and/or greetings mentioned in the oral language. 	<ul style="list-style-type: none"> Written language in 3~4 frames plays a role as a title or short summary of oral language, or presents (a part of) a few sentences in the oral language. Some of the written language may present bibliographic information of the book, student's name, and/or greetings mentioned in the oral language. 	<ul style="list-style-type: none"> Written language in 2 frames plays a role as a title or short summary of oral language, or presents only (a part of) a sentence in the oral language. All or most of the written language may present bibliographic information of the book, student's name, and/or greetings mentioned in the oral language. 	<ul style="list-style-type: none"> Written language in 1 frame presents only bibliographic information of the book reviewed or the student's name mentioned in the oral language. OR Written language in almost all frames is not different from the script of oral language. OR Written language in the artifact has little relationship with oral language. 	<ul style="list-style-type: none"> There is no relationship between oral language and written language.

	<p style="text-align: center;">7) Relationship between oral language and sound</p>	<ul style="list-style-type: none"> • 5 or more different kinds of sound (i.e. sound effects and theme music) in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is appropriate, so it does not hinder the listening of each mode. 	<ul style="list-style-type: none"> • 5 or more different kinds of sound (i.e. sound effects and theme music) in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. <p style="text-align: center;">----- OR -----</p> <ul style="list-style-type: none"> • 3~4 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is appropriate, so it does not hinder the listening of each mode. 	<ul style="list-style-type: none"> • 3~4 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. <p style="text-align: center;">----- OR -----</p> <ul style="list-style-type: none"> • 2 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is appropriate, so it does not hinder the listening of each mode. <p style="text-align: center;">----- OR -----</p> <ul style="list-style-type: none"> • Only 1 theme music set the relevant tone and/or mood on the message of the oral language. 	<ul style="list-style-type: none"> • 2 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. <p style="text-align: center;">----- OR -----</p> <ul style="list-style-type: none"> • Only 1 sound effect set the relevant tone and/or mood on the message of the oral language. 	<ul style="list-style-type: none"> • There is no relationship between oral language and sound.
--	---	--	--	---	--	---

Domain	Criteria	Descriptions of Levels				
		4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
4. Artifact: Organization <hr/> Points of Category 4 ()	8) Organization of multimodal content	<ul style="list-style-type: none"> The artifact organizes content into six required sections of the book review (i.e., introduction, brief summary, one thing you liked, one thing you did not like, recommendation, and conclusion). The student weighs sections differently by using more than one image/video in some sections (e.g., brief summary) that he/she wants to emphasize. 	<ul style="list-style-type: none"> The artifact organizes content into six required sections of the book review. ----- AND ----- The student weighs all sections similarly by using one or two image/video per section. OR The student emphasizes only one section too much (e.g., summary). 	<ul style="list-style-type: none"> The artifact misses one or two required sections of the book review ----- AND ----- The student weighs all sections similarly by using one or two images/videos per section. OR The student emphasizes only one section too much (e.g., summary). 	<ul style="list-style-type: none"> The artifact misses three or more required sections of the book review ----- AND ----- The student weighs all sections similarly by using only one image/video per section. OR The student emphasizes only one section too much (e.g., summary). 	<ul style="list-style-type: none"> The artifact is not completed.
5. Context: Audience Awareness <hr/> Points of Category 5 ()	9) Audience awareness	<ul style="list-style-type: none"> There is evidence that the author is aware of audiences clearly and asks them to engage throughout the artifact. For example, the author mentions “you” or specific audience(s) from the beginning to the end and recommends them to read the book explicitly. 	<ul style="list-style-type: none"> There is evidence that the author is aware of audiences clearly and sometimes asks them to engage with the artifact. For example, the author mentions “you” or specific audience(s) at only one or two sections of the artifact and recommends them to read the book explicitly. 	<ul style="list-style-type: none"> There is evidence that the author is aware of audiences, but does not ask them to engage with the artifact. OR There is evidence that the author seems to be aware of audiences very vaguely, but asks them to engage with the artifact. 	<ul style="list-style-type: none"> There is evidence that the author seems to be aware of audiences very vaguely and does not let them engage with the artifact. 	<ul style="list-style-type: none"> There is no evidence that the author is aware of audiences.

Domain	Criteria	Descriptions of Levels				
		4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
6. Substance: Quality of Content <hr/> Points of Category 6: Sum of Criteria 10) & 11) = ()	10) Quality of summary	<ul style="list-style-type: none"> The summary presents 6 or more major events coherently, so the audience can understand the story. The summary includes all of the story elements (i.e., setting, main characters, and major events) or includes all except the setting. 	<ul style="list-style-type: none"> The summary presents 6 or more major events less coherently, so the audience is a bit confused about the exact orders of events in the story. The summary does not introduce the setting and/or one of the main characters of the story. 	<ul style="list-style-type: none"> The summary presents 3-5 major events, so the audience still needs more details to understand the story. Details may be organized less coherently, so the audience is confused about the exact orders of events in the story. The summary does not introduce the setting and/or one of the main characters of the story. 	<ul style="list-style-type: none"> The summary presents only 1-3 major events and is organized illogically, so the audience cannot understand the point of the story. The summary does not introduce the setting and/or one of the main characters of the story. 	<ul style="list-style-type: none"> The artifact does not include any summary of the book.
	11) Quality of opinion	<ul style="list-style-type: none"> The author provides a total of three or more supporting details or reasons in the opinion sections (i.e., "1 thing I liked," "1 thing I did not like," and "recommendation") 	<ul style="list-style-type: none"> The author provides a total of two supporting details or reasons in the opinion sections. 	<ul style="list-style-type: none"> The author provides only one supporting detail or reason in the opinion sections. 	<ul style="list-style-type: none"> The author does not provide any supporting details or reasons in the opinion sections. 	<ul style="list-style-type: none"> The author does not provide any opinion about the book.
Six Categories Total = ()						

Appendix O

MAJOR CHANGES ON THE DESCRIPTORS OF EACH CRITERION IN DIFFERENT VERSIONS OF THE RUBRIC

Criterion 1: Technical aspects of visual mode (images and videos)

	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1	<ul style="list-style-type: none"> The artifact includes <u>both</u> static images and videos. <u>All</u> technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are in <u>good</u> quality such that visuals convey meaning clearly. 	<ul style="list-style-type: none"> The artifact includes <u>both</u> static images and videos. Three or four technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are in <u>good</u> quality such that visuals convey meaning clearly. 	<ul style="list-style-type: none"> The artifact includes <u>only</u> either static images or videos. One or two technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are in <u>good</u> quality, so they hinder visuals conveying meaning clearly. 	<ul style="list-style-type: none"> The artifact includes <u>only</u> either static images or videos. No technical aspects of images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) are in <u>good</u> quality, so they hinder visuals conveying meaning clearly. 	The artifact does not include any visual mode.
Rubric #2, 3	<ul style="list-style-type: none"> The artifact includes <u>both</u> static images and videos. <u>All</u> technical aspects of images (i.e., 	<ul style="list-style-type: none"> The artifact includes <u>both</u> static images and videos. <u>Three or less</u> technical aspects of 	<ul style="list-style-type: none"> The artifact includes <u>either</u> static images <u>or</u> videos. <u>All</u> technical aspects of images (i.e., 	<ul style="list-style-type: none"> The artifact includes <u>either</u> static images or videos. <u>One or no</u> technical aspect of images (i.e., 	The artifact does not include any visual mode.

	resolution and color) and videos (i.e., camera shots, angles, and lighting) convey meaning clearly.	images (i.e., resolution and color) and videos (i.e., camera shots, angles, and lighting) convey meaning clearly.	resolution and color) or videos (i.e., camera shots, angles, and lighting) convey meaning clearly.	resolution and color) or videos (i.e., camera shots, angles, and lighting) conveys meaning clearly.	
Rubric #4, 5	<ul style="list-style-type: none"> The artifact includes <u>both</u> static and moving visuals. <u>Almost all</u> technical aspects of visuals used in the artifact (e.g., resolution, color, <u>camera shots and angles</u>, lighting, and <u>visual effects</u>) convey meaning clearly. 	<ul style="list-style-type: none"> The artifact includes <u>both</u> static and moving visuals. <u>50% or less</u> technical aspects of visuals used in the artifact (e.g., resolution, color, camera shots and angles, lighting, and visual effects) convey meaning clearly. 	<ul style="list-style-type: none"> The artifact includes <u>either</u> static <u>or</u> moving visuals. <u>Almost all</u> technical aspects of visuals used in the artifact (e.g., resolution, color, camera shots and angles, lighting, and visual effects) convey meaning clearly. 	<ul style="list-style-type: none"> The artifact includes <u>either</u> static <u>or</u> moving visuals. <u>50% or less</u> technical aspects of visuals used in the artifact (e.g., resolution, color, camera shots and angles, lighting, and visual effects) convey meaning clearly. 	The artifact does not include any visual mode.
Rubric #6, 7	<ul style="list-style-type: none"> Technical aspects of visuals (e.g., resolution, color, <u>camera shots and angles</u>, lighting, <u>Ken Burns</u>, and <u>transitions</u>) used in <u>more than 75%</u> of frames convey meaning clearly. 	<ul style="list-style-type: none"> <u>Technical aspects of visuals</u> (e.g., resolution, color, <u>camera shots and angles</u>, lighting, <u>Ken Burns</u>, and <u>transitions</u>) used in <u>75-51%</u> of frames convey meaning clearly. 	<ul style="list-style-type: none"> <u>Technical aspects of visuals</u> (e.g., resolution, color, <u>camera shots and angles</u>, lighting, <u>Ken Burns</u>, and <u>transitions</u>) used in <u>50-25%</u> of frames convey meaning clearly. 	<ul style="list-style-type: none"> <u>Technical aspects of visuals</u> (e.g., resolution, color, <u>camera shots and angles</u>, lighting, <u>Ken Burns</u>, and <u>transitions</u>) used in <u>25% or less</u> frames convey meaning clearly. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> The artifact includes visual mode in only one frame. 	The artifact does not include any visual mode.

Criterion 2: Technical aspects of audio-voice

	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1-3	<ul style="list-style-type: none"> All characteristics of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) are in <u>good</u> quality. The artifact includes voice that is pronounced clearly with appropriate volume, speed, and tone. 	<ul style="list-style-type: none"> Two or three characteristics of voice among <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u> are in <u>good</u> quality. 	<ul style="list-style-type: none"> Only one characteristic of voice among <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u> is in <u>good</u> quality. 	<ul style="list-style-type: none"> All characteristics of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) are in <u>poor</u> quality. Voice is too loud or quiet, too fast or slow, monotonous, and/or pronounced unclearly. 	The artifact does not include any voice.
Rubric #4	<ul style="list-style-type: none"> All technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) convey meaning clearly. . For example, voice is pronounced clearly with appropriate volume, speed, and tone. 	<ul style="list-style-type: none"> Two or three technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) convey meaning clearly. 	<ul style="list-style-type: none"> Only one technical aspect of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) conveys meaning clearly. 	<ul style="list-style-type: none"> None of the technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) conveys meaning clearly. For example, voice is too loud or quiet, too fast or slow, monotonous, and/or pronunciation is unclear. 	The artifact does not include any voice.
Rubric #5, 6	<ul style="list-style-type: none"> All technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) convey meaning clearly for the whole 	<ul style="list-style-type: none"> Two or three technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) convey meaning 	<ul style="list-style-type: none"> Only one technical aspect of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) conveys meaning 	<ul style="list-style-type: none"> None of the technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) conveys meaning clearly for the whole 	The artifact does not include any voice.

	time. For example, voice is pronounced clearly with appropriate volume, speed, and tone.	clearly for the whole time.	clearly for the whole time.	time. For example, voice is too loud or quiet, too fast or slow, monotonous, and/or pronunciation is unclear.	
Rubric #7	<ul style="list-style-type: none"> • <u>All</u> technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) convey meaning clearly for the duration of the artifact. For example, the voice is pronounced clearly with appropriate volume, speed, and tone. 	<ul style="list-style-type: none"> • <u>Two or three</u> technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) convey meaning clearly for the duration of the artifact. 	<ul style="list-style-type: none"> • <u>Only one</u> technical aspect of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) conveys meaning clearly for the duration of the artifact. 	<ul style="list-style-type: none"> • <u>None of the</u> technical aspects of voice (i.e., <u>volume</u>, <u>speed</u>, <u>tone</u>, and <u>pronunciation</u>) conveys meaning clearly for the duration of the artifact. For example, the voice is too loud or quiet, too fast or slow, monotonous, and/or pronunciation is unclear. 	The artifact does not include any voice.

Criterion 3: Conventions of oral language (narration)

	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1-3	<ul style="list-style-type: none"> Oral language in the artifact has sentences that are generally complete with sufficient variety in length and structure. Most sentences follow Standard English conventions. 	<ul style="list-style-type: none"> Oral language in the artifact has sentences that are generally complete with sufficient variety in length and structure. Most sentences <u>do not</u> follow Standard English conventions. 	<ul style="list-style-type: none"> Oral language in the artifact has short and simply structured sentences. Most sentences follow Standard English conventions. 	<ul style="list-style-type: none"> Oral language in the artifact has short and simply structured sentences. Most sentences <u>do not</u> follow Standard English conventions. 	The artifact does not include any oral language.
Rubric #4-7	<ul style="list-style-type: none"> Oral language in the artifact has sentences that are generally complete with sufficient variety in length and structure. <u>Almost all</u> sentences follow the conventions of Standard English suggested in the Common Core State Standards (CCSS). 	<ul style="list-style-type: none"> Oral language in the artifact has sentences that are generally complete with sufficient variety in length and structure. <u>Some</u> sentences <u>do not</u> follow the conventions of Standard English suggested in the CCSS. 	<ul style="list-style-type: none"> Oral language in the artifact has short and simply structured sentences. <u>Almost all</u> sentences follow the conventions of Standard English suggested in the CCSS. 	<ul style="list-style-type: none"> Oral language in the artifact has short and simply structured sentences. <u>Some</u> sentences <u>do not</u> follow the conventions of Standard English suggested in the CCSS. 	The artifact does not include any oral language.

Criterion 4: Conventions of written language (captions)

Criterion 4	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1-2	<ul style="list-style-type: none"> More than 50% of frames in the artifact include captions. Captions include no or less than three errors in punctuation, capitalization, grammar and/or spelling, so the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> More than 50% of frames in the artifact include captions. Captions include three or more errors in punctuation, capitalization, grammar and/or spelling, so the errors hinder conveying meaning clearly. 	<ul style="list-style-type: none"> Less than 50% of frames in the artifact include captions. Captions include no or less than three errors in punctuation, capitalization, grammar and/or spelling, so the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> Less than 50% of frames in the artifact include captions. Captions include three or more errors in punctuation, capitalization, grammar and/or spelling, so the errors hinder conveying meaning clearly. 	The artifact does not include any written language (caption).
Rubric #3-4	<ul style="list-style-type: none"> More than 50% of frames in the artifact include written language (captions). Written language (captions) include(s) less than three errors in punctuation, capitalization, grammar and/or spelling, so the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> Less than 50% of frames in the artifact include written language (captions). Written language (captions) include(s) less than three errors in punctuation, capitalization, grammar and/or spelling, so the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> More than 50% of frames in the artifact include written language (captions). Written language (captions) include(s) three or more errors in punctuation, capitalization, grammar and/or spelling, so the errors hinder conveying meaning clearly. 	<ul style="list-style-type: none"> Less than 50% of frames in the artifact include written language (captions). Written language (captions) include(s) three or more errors in punctuation, capitalization, grammar and/or spelling, so the errors hinder 	The artifact does not include any written language (captions).

				<p>conveying meaning clearly.</p> <p>OR</p> <ul style="list-style-type: none"> The artifact includes written language (captions) in only one or two frames. 	
Rubric #5	<ul style="list-style-type: none"> 50% or more of frames in the artifact include written language (captions). Written language (captions) include(s) three or less errors in punctuation, capitalization, grammar and/or spelling, and the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> Less than 50% of frames in the artifact include written language (captions). Written language (captions) include(s) three or less errors in punctuation, capitalization, grammar and/or spelling, and the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> 50% or more of frames in the artifact include written language (captions). Written language (captions) include(s) more than three errors in punctuation, capitalization, grammar and/or spelling, and the errors may hinder conveying meaning clearly. 	<ul style="list-style-type: none"> Less than 50% of frames in the artifact include written language (captions). Written language (captions) include(s) more than three errors in punctuation, capitalization, grammar and/or spelling, and the errors may hinder conveying meaning clearly. <p>OR</p> <ul style="list-style-type: none"> The artifact includes written language (captions) in only one frame. 	The artifact does not include any written language (captions).
Rubric #6-7	<ul style="list-style-type: none"> <u>More than 50%</u> of frames in the artifact include written language (titles). 	<ul style="list-style-type: none"> <u>50% or fewer</u> of frames in the artifact include written language (titles). 	<ul style="list-style-type: none"> <u>More than 50%</u> of frames in the artifact include written language (titles). 	<ul style="list-style-type: none"> <u>50% or fewer</u> of frames in the artifact include 	The artifact does not include any written

	<ul style="list-style-type: none"> • Written language (titles) include(s) three or fewer errors in punctuation, capitalization, grammar and/or spelling, but the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> • Written language (titles) include(s) three or fewer errors in punctuation, capitalization, grammar and/or spelling, but the errors do not hinder conveying meaning clearly. 	<ul style="list-style-type: none"> • Written language (titles) include(s) more than three errors in punctuation, capitalization, grammar and/or spelling, and the errors may hinder conveying meaning clearly. 	<p>written language (titles).</p> <ul style="list-style-type: none"> • Written language (titles) include(s) more than three errors in punctuation, capitalization, grammar and/or spelling, and the errors may hinder conveying meaning clearly. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> • The artifact includes written language (titles) in only one frame. 	<p>language (titles).</p>
--	---	---	---	--	---------------------------

Criterion 5: Relationship between visual and oral language

Criterion 5	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1	<ul style="list-style-type: none"> All visuals are relevant to the content of oral language in each frame, so the combination of visual and oral language makes sense. Visuals complement or augment meaning of oral language and there are no or little redundant visuals. 	<ul style="list-style-type: none"> Most (More than half of) visuals are relevant to the content of oral language in each frame, so the combination of visual and oral language makes sense. Some relevant visuals are redundant. 	<ul style="list-style-type: none"> Most (More than half of) visuals are irrelevant to the content of oral language in each frame, so the combination of visual and oral language makes little sense. Even a few relevant visuals are redundant. 	<ul style="list-style-type: none"> All visuals are irrelevant to the content of oral language in each frame, so the combination of visual and oral language makes no sense. 	There is no relationship between visual and oral language.
Rubric #2	<ul style="list-style-type: none"> More than 75% of visuals are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of visual and oral language makes sense. 	<ul style="list-style-type: none"> More than 50% of visuals are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of visual and oral language generally makes sense. 	<ul style="list-style-type: none"> Less than 50% of visuals are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of visual and oral language makes little sense. 	<ul style="list-style-type: none"> Less than 25% of visuals are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of visual and oral language makes little or no sense. 	There is no relationship between visual and oral language.

	<ul style="list-style-type: none"> • Almost all relevant visuals are necessary. 	<ul style="list-style-type: none"> • Most relevant visuals are necessary. 	<ul style="list-style-type: none"> • Most relevant visuals are redundant. 	<ul style="list-style-type: none"> • Almost all relevant visuals are redundant. 	
Rubric #3-4	<ul style="list-style-type: none"> • More than 50% of visuals are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of visual and oral language generally makes sense. • Most relevant visuals are necessary. 	<ul style="list-style-type: none"> • More than 50% of visuals are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of visual and oral language generally makes sense. • Most relevant visuals are redundant. 	<ul style="list-style-type: none"> • Less than 50% of visuals are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of visual and oral language makes little sense. • Most relevant visuals are necessary. 	<ul style="list-style-type: none"> • Less than 50% of visuals are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of visual and oral language makes little or no sense. • Most relevant visuals are redundant. 	There is no relationship between visual and oral language.
Rubric #5	<ul style="list-style-type: none"> • More than 80% of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to audiences without the author's explanation. • Some visuals are used more than once to represent different content of oral 	<ul style="list-style-type: none"> • More than 80% of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to audiences without the author's explanation. • Some visuals are used more than once to represent similar content of oral 	<ul style="list-style-type: none"> • 50~80% of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to audiences without the author's explanation. • Some visuals may be used more than once to represent content of oral language either 	<ul style="list-style-type: none"> • Less than 50% of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to audiences without the author's explanation. • Some visuals may be used more than once to represent content of oral 	There is no relationship between visual and oral language.

	language purposefully.	language redundantly.	purposefully or redundantly.	language either purposefully or redundantly.	
Rubric #6-7	<ul style="list-style-type: none"> • <u>More than 75%</u> of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to the audience without the author's explanation. • Some visuals are used more than once to represent different content of the oral language purposefully. 	<ul style="list-style-type: none"> • <u>More than 75%</u> of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to the audience without the author's explanation. • Some visuals are used more than once to represent similar content of the oral language redundantly. 	<ul style="list-style-type: none"> • <u>51~75%</u> of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to the audience without the author's explanation. • Some visuals may be used more than once to represent content of the oral language either purposefully or redundantly. 	<ul style="list-style-type: none"> • <u>Fewer than 50%</u> of frames relate visual and oral language relevantly (e.g., concurrence, complementarity, or connection), which makes sense to the audience without the author's explanation. • Some visuals may be used more than once to represent content of the oral language either purposefully or redundantly. 	There is no relationship between visual and oral language.

Criterion 6: Relationship between oral and written language

Criterion 6	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1	<ul style="list-style-type: none"> All captions are relevant to the content of oral language in each frame, so the combination of written and oral languages makes sense. Written languages complement or augment meaning of oral language and there are no or little redundant written languages. 	<ul style="list-style-type: none"> Most (More than half of) captions are relevant to the content of oral language in each frame, so the combination of written and oral languages makes sense. Some written languages are redundant. 	<ul style="list-style-type: none"> Most (More than half of) captions are irrelevant to the content of oral language in each frame, so the combination of written and oral languages makes little sense. Even a few relevant written languages are redundant. 	<ul style="list-style-type: none"> All captions are irrelevant to the content of oral language in each frame, so the combination of written and oral languages makes no sense. 	There is no relationship between oral language and written language.
Rubric #2	<ul style="list-style-type: none"> More than 75% of captions are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of written and oral 	<ul style="list-style-type: none"> More than 50% of captions are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of written and oral 	<ul style="list-style-type: none"> Less than 50% of captions are irrelevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of 	<ul style="list-style-type: none"> Less than 25% of captions are irrelevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of 	There is no relationship between oral language and written language.

	<p>languages makes sense.</p> <ul style="list-style-type: none"> • Almost all relevant captions are necessary. 	<p>languages makes sense.</p> <ul style="list-style-type: none"> • Most relevant captions are necessary. 	<p>written and oral languages makes little sense.</p> <ul style="list-style-type: none"> • Most relevant captions are redundant. 	<p>written and oral languages makes no sense.</p> <ul style="list-style-type: none"> • Almost all relevant captions are redundant. 	
Rubric #3	<ul style="list-style-type: none"> • More than 50% of captions are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of written and oral language makes sense. • Most relevant captions are necessary. 	<ul style="list-style-type: none"> • More than 50% of captions are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of written and oral language makes sense. • Most relevant captions are redundant. 	<ul style="list-style-type: none"> • Less than 50% of captions are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of written and oral language makes little sense. • Most relevant captions are necessary. 	<ul style="list-style-type: none"> • Less than 50% of captions are relevant to the content of oral language in each frame (e.g., either complement or augment meaning of oral language), so the combination of written and oral language makes little sense. • Most relevant captions are redundant. 	There is no relationship between oral language and written language.
Rubric #4	<ul style="list-style-type: none"> • 5 or more captions in the artifact play a role as a title or short summary of oral language, or present (a part of) a few sentences in the oral language. Some of the captions may present bibliographic information of the book, student's 	<ul style="list-style-type: none"> • 3~4 captions in the artifact play a role as a title or short summary of oral language, or present (a part of) a few sentences in the oral language. Some of the captions may present bibliographic information of the book, student's 	<ul style="list-style-type: none"> • 2 captions in the artifact play a role as a title or short summary of oral language, or present only (a part of) a sentence in the oral language. All or most of the captions may present bibliographic information of the 	<ul style="list-style-type: none"> • 1 caption in the artifact presents only bibliographic information of the book reviewed or the student's name mentioned in the oral language. OR • Captions in the artifact have little 	There is no relationship between oral language and written language.

	name, and/or greetings mentioned in the oral language.	name, and/or greetings mentioned in the oral language.	book, student's name, and/or greetings mentioned in the oral language.	relationship with oral language.	
Rubric #5-7	<ul style="list-style-type: none"> Written language in 5 or more frames plays a role as a title or short summary of oral language, or it presents (a part of) a few sentences in the oral language. Some of the written language may present bibliographic information of the book, student's name, and/or greetings mentioned in the oral language. 	<ul style="list-style-type: none"> Written language in 3~4 frames play a role as a title or short summary of oral language, or it presents (a part of) a few sentences in the oral language. Some of the written language may present bibliographic information of the book, student's name, and/or greetings mentioned in the oral language. 	<ul style="list-style-type: none"> Written language in 2 frames plays a role as a title or short summary of oral language, or it presents only (a part of) a sentence in the oral language. All or most of the written language may present bibliographic information of the book, student's name, and/or greetings mentioned in the oral language. 	<ul style="list-style-type: none"> Written language in 1 frame presents only bibliographic information of the book reviewed or the student's name mentioned in the oral language. OR Written language in almost all frames is not different from the script of oral language. OR Written language in the artifact has little relationship with oral language. 	<ul style="list-style-type: none"> There is no relationship between oral language and written language.

Criterion 7: Relationship between oral language and sound

242

Criterion 7	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1	<ul style="list-style-type: none"> Volume of oral language and other sounds is appropriate, so it does not hinder listening of each mode. Tone and/or mood of most sound effects and/or music complement or augment the message of the oral language, it is very smooth to engage in the artifact. 	<ul style="list-style-type: none"> Volume of oral language and other sounds is too quiet or loud, so it hinders listening of each mode. Tone and/or mood of most sound effects and/or music are somewhat relevant to the message of the oral language, so it is okay to engage in the artifact. 	<ul style="list-style-type: none"> Volume of oral language and other sounds is okay, so it does not hinder listening of each mode. Tone and/or mood of some sound effects and/or music are not relevant to the message of the oral language, so it is difficult to engage in the artifact. 	<ul style="list-style-type: none"> Volume of oral language and other sounds is too quiet or loud, so it hinders listening of each mode. Tone and/or mood of most sound effects and/or music are not relevant to the message of the oral language, so it is difficult to engage in the artifact. 	There is no relationship between oral language and sound.
Rubric #2-4	<ul style="list-style-type: none"> Tone and/or mood of most sound effects and/or music complement or augment the message of the oral language, so it is very easy for the viewer to engage with the artifact. Volume of oral language and other sounds is appropriate, so it does not hinder the listening of each mode. 	<ul style="list-style-type: none"> Tone and/or mood of most sound effects and/or music are somewhat relevant to the message of the oral language, so the viewer is able to engage with the artifact. Volume of oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. 	<ul style="list-style-type: none"> Tone and/or mood of most sound effects and/or music are not relevant to the message of the oral language, so it is difficult for the viewer to engage with the artifact. Volume of oral language and other sounds is appropriate, so it does not hinder the listening of each mode. 	<ul style="list-style-type: none"> Tone and/or mood of most sound effects and/or music are not relevant to the message of the oral language, so it is difficult for the viewer to engage with the artifact. Volume of oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. 	There is no relationship between oral language and sound.

<p>Rubric #5-7</p>	<ul style="list-style-type: none"> • 5 or more different kinds of sound (i.e. sound effects and theme music) in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is appropriate, so it does not hinder the listening of each mode. 	<ul style="list-style-type: none"> • 5 or more different kinds of sound (i.e. sound effects and theme music) in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. <p>----- OR -----</p> <ul style="list-style-type: none"> • 3~4 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is appropriate, so it does not hinder the listening of each mode. 	<ul style="list-style-type: none"> • 3~4 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. <p>----- OR -----</p> <ul style="list-style-type: none"> • 2 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is appropriate, so it does not hinder the listening of each mode. <p>----- OR -----</p> <ul style="list-style-type: none"> • Only 1 theme music set the relevant tone and/or mood on the message of the oral language. 	<ul style="list-style-type: none"> • 2 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. <p>----- OR -----</p> <ul style="list-style-type: none"> • Only 1 sound effect set the relevant tone and/or mood on the message of the oral language. 	<p>There is no relationship between oral language and sound.</p>
-------------------------------	--	--	---	--	--

Criterion 8: Organization of multimodal product

244

Criterion 8	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1-2	<ul style="list-style-type: none"> The artifact organizes content into six required sections of book review (i.e., introduction, brief summary, one thing you liked, one thing you did not like, recommendation, and conclusion). It presents the content of each section appropriately. The student weighs sections differently by using more than one image/video in some sections (e.g., brief summary) that they want to emphasize. 	<ul style="list-style-type: none"> The artifact organizes content into six required sections of book review. It presents the content in one or two sections inappropriately. The student weighs all sections similarly by using only one image/video per section or emphasizes only one section too much. 	<ul style="list-style-type: none"> The artifact misses one or two required sections of book review It presents the content in one or two sections inappropriately. The student weighs all sections similarly by using only one image/video per section or emphasizes only one section too much. 	<ul style="list-style-type: none"> The artifact misses more than three required sections of book review It presents the content in all included sections inappropriately. The student weighs all sections similarly by using only one image/video per section or emphasizes only one section too much. 	The artifact is not completed.
Rubric #3-4	<ul style="list-style-type: none"> The artifact organizes content into six required sections of the book review (i.e., introduction, brief 	<ul style="list-style-type: none"> The artifact organizes content into six required sections of the book review. 	<ul style="list-style-type: none"> The artifact misses one or two required sections of the book review 	<ul style="list-style-type: none"> The artifact misses more than three required sections of the book review 	The artifact is not completed.

	<p>summary, one thing you liked, one thing you did not like, recommendation, and conclusion).</p> <ul style="list-style-type: none"> The student weighs sections differently by using more than one image/video in some sections (e.g., brief summary) that they want to emphasize. 	<ul style="list-style-type: none"> The student weighs all sections similarly by using one or two image/video per section. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> The student emphasizes only one section too much (e.g., summary). 	<ul style="list-style-type: none"> The student weighs all sections similarly by using one or two images/videos per section. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> The student emphasizes only one section too much (e.g., summary). 	<ul style="list-style-type: none"> The student weighs all sections similarly by using only one image/video per section. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> The student emphasizes only one section too much (e.g., summary). 	
Rubric #5-7	<ul style="list-style-type: none"> The artifact organizes content into six required sections of the book review (i.e., introduction, brief summary, one thing you liked, one thing you did not like, recommendation, and conclusion). The student weighs sections differently by using more than one image/video in some sections (e.g., brief summary) that he/she wants to emphasize. 	<ul style="list-style-type: none"> The artifact organizes content into six required sections of the book review. <p style="text-align: center;">----- AND -----</p> <ul style="list-style-type: none"> The student weighs all sections similarly by using one or two image/video per section. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> The student emphasizes only one section too much (e.g., summary). 	<ul style="list-style-type: none"> The artifact misses one or two required sections of the book review <p style="text-align: center;">----- AND -----</p> <ul style="list-style-type: none"> The student weighs all sections similarly by using one or two images/videos per section. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> The student emphasizes only one section too much (e.g., summary). 	<ul style="list-style-type: none"> The artifact misses three or more required sections of the book review <p style="text-align: center;">----- AND -----</p> <ul style="list-style-type: none"> The student weighs all sections similarly by using only one image/video per section. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> The student emphasizes only one section too much (e.g., summary). 	<ul style="list-style-type: none"> The artifact is not completed

Criterion 9: Audience Awareness

246

Criterion 9	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1-2	<ul style="list-style-type: none"> The author is aware of audiences clearly and let them engage in throughout the artifact. 	<ul style="list-style-type: none"> The author is aware of audiences and sometimes let them engage in the artifact. 	<ul style="list-style-type: none"> The author is aware of audiences, but s/he does not let them engage in the artifact. 	<ul style="list-style-type: none"> The author is aware of audiences a little bit, but the audience is still vague. 	The author is not aware of audiences.
Rubric #3	<ul style="list-style-type: none"> The author is aware of audiences clearly and lets them engage with throughout the artifact. 	<ul style="list-style-type: none"> The author is aware of audiences and sometimes lets them engage with the artifact. 	<ul style="list-style-type: none"> The author is aware of audiences, but s/he does not let them engage with the artifact. 	<ul style="list-style-type: none"> The author seems to be aware of audiences a little bit, but the audience is still vague. 	The author is not aware of audiences.
Rubric #4-5	<ul style="list-style-type: none"> The author is aware of audiences clearly and asks them to engage with throughout the artifact. For example, the author mentions “you” or specific audience(s) from the beginning to the end and recommends them to read the book explicitly. 	<ul style="list-style-type: none"> The author is aware of audiences clearly and sometimes asks them to engage with the artifact. For example, the author mentions “you” or specific audience(s) at only one or two sections of the artifact and recommends them to read the book explicitly. 	<ul style="list-style-type: none"> The author is aware of audiences, but s/he does not ask them to engage with the artifact. OR The author seems to be aware of audiences very vaguely, but s/he asks them to engage with the artifact. 	<ul style="list-style-type: none"> The author seems to be aware of audiences very vaguely and does not let them engage with the artifact. 	The author is not aware of audiences.

<p>Rubric #6-7</p>	<ul style="list-style-type: none"> • There is evidence that the author is aware of audiences clearly and asks them to engage throughout the artifact. For example, the author mentions “you” or specific audience(s) from the beginning to the end and recommends them to read the book explicitly. 	<ul style="list-style-type: none"> • There is evidence that the author is aware of audiences clearly and sometimes asks them to engage with the artifact. For example, the author mentions “you” or specific audience(s) at only one or two sections of the artifact and recommends them to read the book explicitly. 	<ul style="list-style-type: none"> • There is evidence that the author is aware of audiences, but s/he does not ask them to engage with the artifact. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> • There is evidence that the author seems to be aware of audiences very vaguely, but s/he asks them to engage with the artifact. 	<ul style="list-style-type: none"> • There is evidence that the author seems to be aware of audiences very vaguely and does not let them engage with the artifact. 	<p>There is no evidence that the author is aware of audiences.</p>
-------------------------------	--	--	---	---	--

Criterion 10: Quality of summary

248

Criterion 10	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1	<ul style="list-style-type: none"> The summary length is appropriate (no more than 50% of the entire video). It introduces main characters and major events effectively. It does not reveal the story's important ending. Information about the book is accurate. 	<ul style="list-style-type: none"> The summary length is moderate or a bit long (between 50~75% of the entire video). It introduces main characters and major events. It implies the story's ending to audiences. It includes some inaccuracies, but they do not hinder audience's understanding of the summary. 	<ul style="list-style-type: none"> The summary length is too long or short (more than 75% of the entire video). It introduces one of main characters and major events. It implies the story's ending to audiences. It includes wrong information about the book. 	<ul style="list-style-type: none"> The summary is too long or short (more than 75% of the entire video). It does not introduce main characters and major events. It let audiences know the story's important ending. It includes wrong information about the book. 	The artifact does not include any summary of the book.
Rubric #2	<ul style="list-style-type: none"> It introduces main characters and major events effectively. It does not reveal the story's important ending. Details about the book are accurate and they help audience's understanding of the summary. 	<ul style="list-style-type: none"> It introduces main characters and major events. It implies the story's ending to audiences. Some details about the book are inaccurate, but they do not hinder audience's 	<ul style="list-style-type: none"> It introduces one of main characters and major events. It implies the story's ending to audiences. Some details about the book are inaccurate and they hinder audience's understanding of the summary. 	<ul style="list-style-type: none"> It does not introduce main characters and major events. It let audiences know the story's important ending. Some details about the book are inaccurate and they hinder audience's 	The artifact does not include any summary of the book.

		understanding of the summary.		understanding of the summary.	
Rubric #3	<ul style="list-style-type: none"> The artifact introduces the main characters and major events. Details about the book are accurate and they help the audience's understanding of the summary. The artifact does not reveal the story's ending. 	<ul style="list-style-type: none"> The artifact introduces the main characters and major events. Some details about the book are inaccurate, but they do not hinder the audience's understanding of the summary. The artifact may or may not reveal the story's ending. 	<ul style="list-style-type: none"> The artifact introduces the main characters and major events very limitedly. Details about the book are accurate but very limited, so they are not helpful to help audience's understanding of the summary. The artifact may or may not reveal the story's ending. 	<ul style="list-style-type: none"> The artifact introduces the main characters and major events very limitedly. Most details about the book are inaccurate, so they hinder the audience's understanding of the summary. The artifact may or may not reveal the story's ending. 	The artifact does not include any summary of the book.
Rubric #4	<ul style="list-style-type: none"> The summary introduces a setting, main characters and major events in the story. Details about the book are organized coherently, so they help the audiences' understanding of the summary. The artifact does not reveal the story's ending. 	<ul style="list-style-type: none"> The summary introduces a setting, main characters and major events in the story. Details about the book are organized less coherently, but they do not hinder the audiences' understanding of the summary. The artifact reveals the story's ending. 	<ul style="list-style-type: none"> The summary introduces a setting, main characters and major events in the story limitedly. Details about the book are organized less coherently and they hinder the audiences' understanding of the summary. The artifact may or may not reveal the story's ending. 	<ul style="list-style-type: none"> The summary introduces a setting, main characters and major events in the story very limitedly Details about the book are too insufficient or organized illogically, so the audiences cannot understand the point of the story. The artifact may or may not reveal the story's ending. 	The artifact does not include any summary of the book.
Rubric #5-6	<ul style="list-style-type: none"> The summary presents 6 or more details about the 	<ul style="list-style-type: none"> The summary presents 6 or more details about the 	<ul style="list-style-type: none"> The summary presents 3-5 details about the major events, so the 	<ul style="list-style-type: none"> The summary presents only 1-3 details about the 	The artifact does not include any

	<p>major events coherently, so the audience can understand the story.</p> <ul style="list-style-type: none"> • The summary includes all of the story elements (i.e., setting, main characters, and major events) or does not introduce a setting. 	<p>major events less coherently, so the audience is a bit confused about the exact orders of events in the story.</p> <ul style="list-style-type: none"> • The summary does not introduce a setting and/or one of the main characters of the story. 	<p>audience still needs more details to understand the story. Details may be organized less coherently, so the audience is confused about the exact orders of events in the story.</p> <ul style="list-style-type: none"> • The summary does not introduce a setting and/or one of the main characters of the story. 	<p>major events and is organized illogically, so the audience cannot understand the point of the story.</p> <ul style="list-style-type: none"> • The summary does not introduce a setting and/or one of the main characters of the story. 	<p>summary of the book.</p>
<p>Rubric #7</p>	<ul style="list-style-type: none"> • The summary presents 6 or more major events coherently, so the audience can understand the story. • The summary includes all of the story elements (i.e., setting, main characters, and major events) or includes all except the setting. 	<ul style="list-style-type: none"> • The summary presents 6 or more major events less coherently, so the audience is a bit confused about the exact orders of events in the story. • The summary does not introduce the setting and/or one of the main characters of the story. 	<ul style="list-style-type: none"> • The summary presents 3-5 major events, so the audience still needs more details to understand the story. Details may be organized less coherently, so the audience is confused about the exact orders of events in the story. • The summary does not introduce the setting and/or one of the main characters of the story. 	<ul style="list-style-type: none"> • The summary presents only 1-3 major events and is organized illogically, so the audience cannot understand the point of the story. • The summary does not introduce the setting and/or one of the main characters of the story. 	<p>The artifact does not include any summary of the book.</p>

Criterion 11: Quality of opinion

Criterion 11	Descriptors of Levels				
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable
Rubric #1-2	<ul style="list-style-type: none"> The author provides clear opinion and two or more supporting reasons at three opinion sections (i.e., “1 thing I liked,” “1 thing I did not like,” and “recommendation”). 	<ul style="list-style-type: none"> The author provides clear opinion at the three opinion sections with only one or no supporting reason. 	<ul style="list-style-type: none"> The author provides opinion at two of three opinion sections with only one or no supporting reason. 	<ul style="list-style-type: none"> The author provides opinion at only one of three opinion sections with only one or no supporting reason. 	The author does not provide any opinion about the book.
Rubric #3-4	<ul style="list-style-type: none"> The author provides clear opinions at all three opinion sections (i.e., “1 thing I liked,” “1 thing I did not like,” and “recommendation”). The opinions in all sections are followed by one or more supporting reasons. 	<ul style="list-style-type: none"> The author provides clear opinions at all three opinion sections. The opinions in two sections are followed by one or more supporting reasons. 	<ul style="list-style-type: none"> The author provides opinions at two of the three opinion sections The opinions in two or one sections are followed by one or more supporting reasons. 	<ul style="list-style-type: none"> The author provides an opinion at only one of three opinion sections. The opinion in the section is followed by one or no supporting reasons. 	The author does not provide an opinion about the book.
Rubric #5-7	<ul style="list-style-type: none"> The author provides a total of three or more supporting details or reasons in the opinion sections (i.e., “1 thing I liked,” “1 thing I did not like,” and “recommendation”). 	<ul style="list-style-type: none"> The author provides a total of two supporting details or reasons in the opinion sections. 	<ul style="list-style-type: none"> The author provides only one supporting detail or reason in the opinion sections. 	<ul style="list-style-type: none"> The author does not provide any supporting details or reasons in the opinion sections. 	The author does not provide any opinion about the book.

Appendix P
REVISED SCORING RUBRIC

Domain	Category	Criteria	Points Per Criterion	Points Per Category
Artifact	1. Technical aspect of modes	1) Technical aspect of visual mode (images and videos): The effects of camera shots and angles, lighting, color, size, movement, and sequencing on the quality of visual mode as static or moving images	0~4	0~8
		2) Technical aspect of audio-voice: The effects of fluency, articulation, intonation, volume, pitch, and length on the quality of voice	0~4	
	2. Conventions of linguistic modes	3) Conventions of oral language (narration): The effect of (grade-appropriate) conventions of Standard English on the quality of oral language	0~4	0~8
		4) Conventions of written language (titles): The effects of (grade-appropriate) grammar, mechanics, style, citation, and genre on the quality of written language	0~4	
	3. Coherence of multimodal product	5) Relationship between visual and oral language	0~4	0~12
		6) Relationship between oral language and written language	0~4	
		7) Relationship between oral language and sound	0~4	
	4. Organization of multimodal product	8) Organization of multimodal product: Logical structure of content or messages within and among frames or sections	0~4	0~4
Context	5. Audience awareness	9) Audience awareness: Consideration of explicit or implicit audiences and their engagement in the artifact	0~4	0~4
Substance	6. Quality of content	10) Quality of summary: Clarity, credibility, significance (depth and length), and interest of the content and the pace of content progress or development	0~4	0~8
		11) Quality of opinion: Clarity and persuasiveness of arguments and the use of analysis and evidence to support the argument	0~4	
Total			0~44	

Domain 1: Artifact						
Category 1: Technical aspects of modes						
Criteria	Descriptions of Levels					
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable	Score
1) Technical aspects of visual mode (i.e., resolution, color, camera shots and angles , lighting, Ken Burns , and transitions)	Technical aspects of images and/or videos used in <u>more than 75%</u> of the frames convey meaning clearly.	Technical aspects of images and/or videos used in <u>75-51%</u> of the frames convey meaning clearly.	Technical aspects of images and/or videos used in <u>50-26%</u> of the frames convey meaning clearly.	Technical aspects of images and/or videos used in <u>25% or fewer</u> of the frames convey meaning clearly. OR The artifact includes visual mode in only one frame.	The artifact does not include any visual mode.	
2) Technical aspects of audio-voice (i.e., volume, speed, tone, and pronunciation)	<u>All</u> technical aspects of voice convey meaning clearly for the duration of the artifact. For example, voice is pronounced clearly with appropriate volume, speed, and tone.	<u>Two or three</u> technical aspects of voice convey meaning clearly for the duration of the artifact.	<u>Only one</u> technical aspect of voice conveys meaning clearly for the duration of the artifact.	<u>None of the</u> technical aspects of voice conveys meaning clearly for the duration of the artifact. For example, the voice is too loud or quiet, too fast or slow, monotonous, and/or pronunciation is unclear.	The artifact does not include any voice.	
Sum of Criteria 1 and 2						

Domain 1: Artifact						
Category 2: Conventions of linguistic modes						
Criteria	Descriptions of Levels					
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable	Score
3) Conventions of oral language (narration)	Almost all sentences in the oral language follow the conventions of Standard English suggested in the Common Core State Standards (CCSS) and convey meaning clearly.	More than half of the sentences in the oral language follow the conventions of Standard English suggested in the CCSS. Sentences with incorrect conventions do not hinder conveying meaning clearly.	Some sentences in the oral language follow the conventions of Standard English suggested in the CCSS. Sentences with incorrect conventions may hinder conveying meaning clearly.	Only a few or none of the sentences in the oral language follows the conventions of Standard English suggested in the CCSS. Sentences with incorrect conventions hinder conveying meaning clearly.	The artifact does not include any oral language.	
4) Conventions of written language (Title)	Written language (title) does not include any errors in punctuation, capitalization, grammar and/or spelling.	Written language (title) includes one type of errors in punctuation, capitalization, grammar and/or spelling.	Written language (title) includes two types of errors in punctuation, capitalization, grammar and/or spelling.	Written language (title) includes three or more types of errors in punctuation, capitalization, grammar and/or spelling.	The artifact does not include any written language (title).	
Sum of Criteria 3 and 4						

Domain 1: Artifact						
Category 3: Coherence of multimodal product						
Criteria	Descriptions of Levels					
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable	Score
5) Relationship between visual and oral language	<ul style="list-style-type: none"> More than 75% of frames relate visual and oral language relevantly, which makes sense to the audience without the author's explanation. Some visuals are used more than once to represent different content of the oral language purposefully. 	<ul style="list-style-type: none"> More than 75% of frames relate visual and oral language, which makes sense to the audience without the author's explanation. Some visuals are used more than once to represent similar content of the oral language redundantly. 	<ul style="list-style-type: none"> 51~75% of frames relate visual and oral language relevantly, which makes sense to the audience without the author's explanation. Some visuals may be used more than once to represent content of the oral language either purposefully or redundantly. 	<ul style="list-style-type: none"> Fewer than 50% of frames relate visual and oral language relevantly, which makes sense to the audience without the author's explanation. Some visuals may be used more than once to represent content of the oral language either purposefully or redundantly. 	There is no relationship between visual and oral language.	
6) Relationship between oral language and written language	<ul style="list-style-type: none"> Written language in 5 or more frames plays a role as a title or short summary of oral language, or it presents (a part of) a few sentences in the oral language. Some of the written language may present bibliographic information of the book, student's name, and/or greetings mentioned in the oral language. 	<ul style="list-style-type: none"> Written language in 3~4 frames plays a role as a title or short summary of oral language, or it presents (a part of) a few sentences in the oral language. Some of the written language may present bibliographic information of the book, student's name, and/or greetings mentioned in the oral language. 	<ul style="list-style-type: none"> Written language in 2 frames plays a role as a title or short summary of oral language, or it presents only (a part of) a sentence in the oral language. All or most of the written language may present bibliographic information of the book, student's name, and/or greetings mentioned in the oral language. 	<ul style="list-style-type: none"> Written language in 1 frame presents only bibliographic information of the book reviewed or the student's name mentioned in the oral language. OR Written language in almost all frames is not different from the script of oral language. OR Written language in the artifact has little relationship with oral language. 	There is no relationship between oral language and written language.	

<p style="text-align: center;">7) Relationship between oral language and sound</p>	<ul style="list-style-type: none"> • 5 or more different kinds of sound (i.e. sound effects and theme music) in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is appropriate, so it does not hinder the listening of each mode. 	<ul style="list-style-type: none"> • 5 or more different kinds of sound (i.e. sound effects and theme music) in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. <p style="text-align: center;">----- OR -----</p> <ul style="list-style-type: none"> • 3~4 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is appropriate, so it does not hinder the listening of each mode. 	<ul style="list-style-type: none"> • 3~4 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. <p style="text-align: center;">----- OR -----</p> <ul style="list-style-type: none"> • 2 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is appropriate, so it does not hinder the listening of each mode. <p style="text-align: center;">----- OR -----</p> <ul style="list-style-type: none"> • Only 1 theme music set the relevant tone and/or mood on the message of the oral language. 	<ul style="list-style-type: none"> • 2 different kinds of sound in the artifact set the relevant tone and/or mood on the message of the oral language. • Volume of the oral language and other sounds is too quiet or loud, so it hinders the listening of each mode. <p style="text-align: center;">----- OR -----</p> <ul style="list-style-type: none"> • Only 1 sound effect set the relevant tone and/or mood on the message of the oral language. 	<p>There is no relationship between oral language and sound.</p>	
Sum of Criteria 5, 6, and 7						

Domain 1: Artifact Category 4: Organization						
Criteria	Descriptions of Levels					
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable	Score
8) Organization of multimodal content	<ul style="list-style-type: none"> The artifact organizes content into six required sections of the book review (i.e., introduction, brief summary, one thing you liked, one thing you did not like, recommendation, and conclusion). The student weighs sections differently by using more than one image/video in some sections (e.g., brief summary) that he/she wants to emphasize. 	<ul style="list-style-type: none"> The artifact organizes content into six required sections of the book review. <p style="text-align: center;">----- AND -----</p> <ul style="list-style-type: none"> The student weighs all sections similarly by using one or two image/video per section. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> The student emphasizes only one section too much (e.g., summary). 	<ul style="list-style-type: none"> The artifact misses one or two required sections of the book review <p style="text-align: center;">----- AND -----</p> <ul style="list-style-type: none"> The student weighs all sections similarly by using one or two images/videos per section. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> The student emphasizes only one section too much (e.g., summary). 	<ul style="list-style-type: none"> The artifact misses three or more required sections of the book review <p style="text-align: center;">----- AND -----</p> <ul style="list-style-type: none"> The student weighs all sections similarly by using only one image/video per section. <p style="text-align: center;">OR</p> <ul style="list-style-type: none"> The student emphasizes only one section too much (e.g., summary). 	The artifact is not completed.	

Domain 1: Context						
Category 5: Audience Awareness						
Criteria	Descriptions of Levels					
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable	Score
9) Audience awareness	<ul style="list-style-type: none"> There is evidence that the author is aware of audiences clearly and asks them to engage throughout the artifact. For example, the author mentions “you” or specific audience(s) from the beginning to the end and recommends them to read the book explicitly. 	<ul style="list-style-type: none"> There is evidence that the author is aware of audiences clearly and sometimes asks them to engage with the artifact. For example, the author mentions “you” or specific audience(s) at only one or two sections of the artifact and recommends them to read the book explicitly. 	<ul style="list-style-type: none"> There is evidence that the author is aware of audiences, but s/he does not ask them to engage with the artifact. OR There is evidence that the author seems to be aware of audiences very vaguely, but s/he asks them to engage with the artifact. 	<ul style="list-style-type: none"> There is evidence that the author seems to be aware of audiences very vaguely and does not let them engage with the artifact. 	There is no evidence that the author is aware of audiences.	

Domain 3: Substance						
Category 6: Quality of Content						
Criteria	Descriptions of Levels					
	4-Excellent	3-Good	2-Fair	1-Needs Improvement	0-Not Assessable	Score
10) Quality of summary	<ul style="list-style-type: none"> The brief summary section presents 6 or more major events coherently, so the audience can understand the story. The summary includes all of the story elements (i.e., setting, main characters, and major events) or includes all except the setting. 	<ul style="list-style-type: none"> The brief summary section presents 6 or more major events less coherently, so the audience is a bit confused about the exact orders of events in the story. The summary does not introduce the setting and/or one of the main characters of the story. 	<ul style="list-style-type: none"> The brief summary section presents 3-5 major events, so the audience still needs more details to understand the story. Details may be organized less coherently, so the audience is confused about the exact orders of events in the story. The summary does not introduce the setting and/or one of the main characters of the story. 	<ul style="list-style-type: none"> The brief summary section presents only 1-3 major events and is organized illogically, so the audience cannot understand the point of the story. The summary does not introduce the setting and/or one of the main characters of the story. 	The artifact does not include any summary of the book.	
11) Quality of opinion	The author provides a total of three or more supporting details or reasons in the “1 thing I liked,” “1 thing I did not like,” and “recommendations” sections.	The author provides a total of two supporting details or reasons in the “1 thing I liked,” “1 thing I did not like,” and “recommendations” sections.	The author provides only one supporting detail or reason in the “1 thing I liked,” “1 thing I did not like,” and “recommendations” sections. .	The author does not provide any supporting details or reasons in the “1 thing I liked,” “1 thing I did not like,” and “recommendations” sections.	The author does not provide any opinion about the book.	
Sum of Criteria 10 and 11						
Total Score: ()						

Appendix Q

INSTITUTIONAL REVIEW BOARD APPROVAL LETTER



RESEARCH OFFICE

210 Halliher Hall
University of Delaware
Newark, Delaware 19716-1551
Ph: 302/831-2136
Fax: 302/831-2828

DATE: October 5, 2016

TO: Sohee Park
FROM: University of Delaware IRB

STUDY TITLE: [944930-1] Toward More Valid and Reliable Classroom Assessments of Digital Multimodal Composition Performances: Developing a Scoring Rubric for the Assessment of Fourth-Grade Students' Digital Multimodal Book Reviews

SUBMISSION TYPE: New Project

ACTION: APPROVED

APPROVAL DATE: October 4, 2016

EXPIRATION DATE: October 3, 2017

REVIEW TYPE: Expedited Review

REVIEW CATEGORY: Expedited review category # (7)

Thank you for your submission of New Project materials for this research study. The University of Delaware IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a study design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

This submission has received Expedited Review based on the applicable federal regulation.

Please remember that informed consent is a process beginning with a description of the study and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the study via a dialogue between the researcher and research participant. Federal regulations require each participant receive a copy of the signed consent document.

Please note that any revision to previously approved materials must be approved by this office prior to initiation. Please use the appropriate revision forms for this procedure.

All SERIOUS and UNEXPECTED adverse events must be reported to this office. Please use the appropriate adverse event forms for this procedure. All sponsor reporting requirements should also be followed.

Please report all NON-COMPLIANCE issues or COMPLAINTS regarding this study to this office.

Please note that all research records must be retained for a minimum of three years.

Based on the risks, this project requires Continuing Review by this office on an annual basis. Please use the appropriate renewal forms for this procedure.

If you have any questions, please contact Nicole Farnese-McFarlane at (302) 831-1119 or nicolefm@udel.edu. Please include your study title and reference number in all correspondence with this office.



RESEARCH OFFICE

210 Hulihan Hall
University of Delaware
Newark, Delaware 19716-1551
Ph: 302/831-2136
Fax: 302/831-2828

DATE: September 26, 2017

TO: Sohee Park
FROM: University of Delaware IRB

STUDY TITLE: [944930-2] Toward More Valid and Reliable Classroom Assessments of Digital Multimodal Composition Performances: Developing a Scoring Rubric for the Assessment of Fourth-Grade Students' Digital Multimodal Book Reviews

SUBMISSION TYPE: Continuing Review/Progress Report

ACTION: APPROVED
APPROVAL DATE: September 26, 2017
EXPIRATION DATE: October 3, 2018
REVIEW TYPE: Expedited Review

REVIEW CATEGORY: Expedited review category # (7)

Thank you for your submission of Continuing Review/Progress Report materials for this research study. The University of Delaware IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a study design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

This submission has received Expedited Review based on the applicable federal regulation.

Please remember that informed consent is a process beginning with a description of the study and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the study via a dialogue between the researcher and research participant. Federal regulations require each participant receive a copy of the signed consent document.

Please note that any revision to previously approved materials must be approved by this office prior to initiation. Please use the appropriate revision forms for this procedure.

All SERIOUS and UNEXPECTED adverse events must be reported to this office. Please use the appropriate adverse event forms for this procedure. All sponsor reporting requirements should also be followed.

Please report all NON-COMPLIANCE issues or COMPLAINTS regarding this study to this office.

Please note that all research records must be retained for a minimum of three years.

Based on the risks, this project requires Continuing Review by this office on an annual basis. Please use the appropriate renewal forms for this procedure.

If you have any questions, please contact Nicole Farnese-McFarlane at (302) 831-1119 or nicolefm@udel.edu. Please include your study title and reference number in all correspondence with this office.